H. FUJISAKI, EDITOR

RECENT RESEARCH TOWARDS ADVANCED MAN-MACHINE INTERFACE THROUGH SPOKEN LANGUAGE

ELSEVIER

RECENT RESEARCH TOWARDS ADVANCED MAN-MACHINE INTERFACE THROUGH SPOKEN LANGUAGE

This publication is made possible by
the Grant-in-Aid for the Publication of Scientific Research Results
from the Japanese Ministry of Education, Science and Culture

RECENT RESEARCH TOWARDS ADVANCED MAN-MACHINE INTERFACE THROUGH SPOKEN LANGUAGE

Edited by

H. FUJISAKI
Science University of Tokyo
Tokyo, Japan

1996
ELSEVIER
AMSTERDAM - LAUSANNE - NEW YORK - OXFORD - SHANNON - TOKYO

ELSEVIER SCIENCE B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

Library of Congress Cataloging-in-Publication Data

Recent research towards advanced man-machine interface through spoken language / edited by H. Fujisaki.
p. cm.
Includes bibliographical references.
ISBN 0-444-81607-0
1. Speech processing systems. 2. Human-machine systems. 3. Speech synthesis. 4. Natural language processing (Computer science) I. Fujisaki, H. (Hiroya)
TK7882.S65R423 1996
006.4'54--dc20 96-34509
CIP

The Japanese version of this title is published by Hokusen-sha

ISBN: 0 444 81607 0

This book is printed on acid-free paper.

Printed in The Netherlands.

PREFACE

Needless to say, the spoken language is the most important means of human information transmission. Hence man-machine interface through the spoken language becomes increasingly important as we enter the age of the so-called Information Society. Because of the depth and the width of the problems involved, however, full realization of such an interface calls for coordination of research efforts beyond the scope of a single group or institution. It is in this spirit that a nationwide research project was conceived and started in 1987 as one of the first Priority Research Areas supported by the Grant-in-Aid for Scientific Research from the Ministry of Education, Science and Culture of Japan. The project has been carried out by the collaboration of over 190 researchers in Japan, involving 90 expert members and over 100 collaborating members.

In order to encourage mutual exchange of ideas and results through open discussion and publication, both among the members of the project and among researchers engaged in similar research projects in other countries, annual symposia have been held on specific topics. The present volume starts with an overview of the project, followed by 41 papers presented at these symposia, revised by the authors whenever necessary. It includes not only the reports of our members, but also 14 contributions from foreign experts who were invited to participate in these symposia to present their own work and to give valuable advices. The present volume is expected to serve as an important source of information on each of the nine topics adopted for intensive study under the project. As leader of the project, I am grateful to the Ministry of Education, Science and Culture for a special grant that made the publication possible. I would also like to thank all the contributors, both domestic and abroad, for their valuable efforts toward the advancement of science and technology for spoken language processing, and also the members of the editorial committee, Drs. K. Shirai, S. Kiritani and S. Sekimoto for their enduring efforts.

Although the project was successfully terminated in 1990, research efforts will obviously be continued and expanded in the future. It is my sincere hope that this volume will serve as a milestone and as a guideline for further work in this important scientific and technological field of spoken language processing.

Hiroya Fujisaki

This Page Intentionally Left Blank

List of Contributors

Noriyuki Aoki
Department of Electrical Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169 Japan

Yasuharu Asano
Faculty of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113 Japan

Mats Blomberg
Department of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH)
Stockholm, Sweden

Rolf Carlson
Department of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH)
Stockholm, Sweden

Jiauwu Dang
Electronics Department, Faculty of Engineering, Shizuoka University
3-5-1 Jouhoku, Hamamatsu-shi, 432 Japan

Masatake Dantsuji
Faculty of Letters, Kansai University
3-3-35 Yamate-cho, Suita-shi, Osaka, Japan

Hikaru Date
Department of Information Engineering, Yamagata University
Yonezawa, 992 Japan

Renato De Mori
School of Computer Science, McGill University
3480 University Street, Montreal, Quebec, Canada, H3A 2A7

Kjell Elenius
Department of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH)
Stockholm, Sweden

Mitsuru Endo
Research Center for Applied Information Sciences, Tohoku University
Sendai, 980 Japan

Frank Fallside
Cambridge University Engineering Department
Trumpingdon Street
Cambridge CB2 1PZ, UK

Gunnar Fant
Dept. of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH)
Stockholm, Sweden

Hiroya Fujisaki
Dept. of Electronic Engineering, Faculty of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

Björn Granström
Department of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH)
Stockholm, Sweden

Jean-Paul Haton
CRIN/INRIA
B.P. 239, 54506 Vandoeuvre les Nancy Cedex, France

Shizuo Hiki
School of Human Sciences, Waseda University
Tokorozawa, Saitama, 359 Japan

Makoto Hirai
The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, 567 Japan

Yoshimitsu Hirata
Department of Information and Computer Sciences, Toyohashi University of Technology
Tenpakucho, Toyohashi, 441 Japan

Keikichi Hirose
Faculty of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113 Japan

Hsiao-Wuen Hon
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213, USA

Naoki Hosaka
Department of Electrical Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169 Japan

Xuedong Huang
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213, USA

Sheri Hunnicutt
Department of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH)
Stockholm, Sweden

Mei-Yuh Hwang
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213, USA

Satoshi Imaizumi
Research Institute of Logopedics and Phoniatrics, Faculty of Medicine, University of Tokyo
Bunkyo-ku, Tokyo, 113 Japan

Akira Ishida
Faculty of Technology, Tokyo University of Agriculture & Technology
2-4-16 Nakamachi, Koganei, Tokyo 184 Japan

Shuichi Itahashi
Institute of Information Sciences and Electronics, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305 Japan

Fumitada Itakura
Department of Electrical Engineering, Nagoya University
Chikusa-ku, Nagoya, 464-01 Japan

Akinori Ito
Research Center for Applied Information Sciences, Tohoku University
Sendai, 980 Japan

Mervyn Jack
Centre for Speech Technology Research, University of Edinburgh
Edinburgh, UK

Hideki Kashioka
The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, 567 Japan

Hideki Kasuya
Faculty of Engineering, Utsunomiya University
2753 Ishii-machi, Utsunomiya, 321 Japan

Hisashi Kawai
Faculty of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113 Japan

Ken'iti Kido
Research Center for Applied Information Sciences, Tohoku University
Sendai, 980 Japan

Shigeru Kiritani
Research Institute of Logopedics and Phoniatrics, Faculty of Medicine, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

Tadahiro Kitahashi
The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, 567 Japan

Shigeyoshi Kitazawa
Faculty of Engineering, Shizuoka University
3-5-1 Johoku, Hamamatsu-shi, Shizuoka, Japan

Hidefumi Kobatake
Faculty of Technology, Tokyo University of Agriculture & Technology
2-4-16 Nakamachi, Koganei, Tokyo, 184 Japan

Tetsunori Kobayashi
Department of Electrical Engineering, Hosei University
3-7-2 Kajino-cho, Koganei, Tokyo, 184 Japan

Yutaka Kobayashi
Department of Electronics and Information Science, Kyoto Institute of Technology
Matsugasaki, Sakyo-ku, Kyoto, Japan

Anita Kruckenberg
Dept. of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH)
Stockholm, Sweden

Roland Kuhn
School of Computer Science, McGill University
3480 University Street, Montreal
Quebec, Canada, H3A 2A7

John Laver
Centre for Speech Technology Research, University of Edinburgh
Edinburgh, UK

Kai-Fu Lee
School of Computer Science, Carnegie Mellon University, Pittsburgh
PA 15213, USA

Mats Ljungqvist
Faculty of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, Japan

Chengxiang Lu
Electronics Department, Faculty of Engineering, Shizuoka University
3-5-1 Jouhoku, Hamamatsu-shi, 432 Japan

Shozo Makino
Research Center for Applied Information Sciences, Tohoku University
Sendai, 980 Japan

Joseph-Jean Mariani
LIMSI-CNRS, BP 30, 91406 Orsay Cedex, France

Hiroshi Matsumoto
Deparment of Electrical and Electronic Engineering, Shinshu University
500 Wakasato, Nagano-shi, 380 Japan

J. McAllister
Centre for Speech Technology Research, University of Edinburgh
Edinburgh, UK

M. McAllister
Centre for Speech Technology Research, University of Edinburgh
Edinburgh, UK

Nobuhiro Miki
Research Institute for Electronic Science, Hokkaido University
Sapporo, Japan

Jouji Miwa
Faculty of Engineering, Iwate University
4-3-5 Ueda, Morioka, 020 Japan

Riichiro Mizoguchi
The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, 567 Japan

Kunitoshi Motoki
Faculty of Engineering, Hokkai-Gakuen University
Sapporo, Japan

Isao Murase
Department of Information and Computer Sciences, Toyohashi University of Technology
Tenpakucho, Toyohashi, 441 Japan

Seiichi Nakagawa
Department of Information and Computer Sciences, Toyohashi University of Technology
Tenpakucho, Toyohashi, 441 Japan

Takayoshi Nakai
Electronics Department, Faculty of Engineering, Shizuoka University
3-5-1 Jouhoku, Hamamatsu-shi, 432 Japan

Takayuki Nakajima
Integrated Media Laboratories Sharp Corporation
1-9-2 Nakase, Mihama-ku, Chiba, 261 Japan

Kazuo Nakata
Faculty of Technology, Tokyo University of Agriculture and Technology
2-24-16 Nakamachi, Koganei, Tokyo, 184 Japan

Yasuhisa Niimi
Department of Electronics and Information Science, Kyoto Institute of Technology
Matsugasaki, Sakyo-ku, Kyoto, Japan

Lennart Nord
Dept. of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH)
Stockholm, Sweden

Louis C.W. Pols
Institute of Phonetic Sciences, University of Amsterdam
Herengracht 338, 1016 CG
Amsterdam, The Netherlands

P. V. S. Rao
Tata Institute of Fundamental Research, Bombay, India

Shuzo Saito
Department of Electronics Engineering, Kogakuin University
2665-1 Nakanocho, Hachioji, Tokyo Japan

Yoshihiro Sekiguchi
Faculty of Engineering, Yamanashi University, 4 Takeda, Kofu, 400 Japan

Stephanie Seneff
Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Inst. of Technology, Cambridge, MA 02139, USA

Minoru Shigenaga
Faculty of Engineering, Yamanashi University, 4 Takeda, Kofu, 400 Japan

Katsuhiko Shirai
Department of Electrical Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169 Japan

Harald Singer
Department of Electrical Engineering, Nagoya University
Chikusa-ku, Nagoya, 464-01 Japan

Kenneth N. Stevens
Research Laboratory of Electronics, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Akihiko Sugiura
Faculty of Technology, Tokyo University of Agriculture and Technology
2-24-16 Nakamachi, Koganei, Tokyo, 184 Japan

Hisayoshi Suzuki
Electronics Department, Faculty of Engineering, Shizuoka University
3-5-1 Jouhoku, Hamamatsu-shi, 432 Japan

Torazo Suzuki
Machine Understanding Division, Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Japan

Atsuko Takano
The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, 567 Japan

Kazuhiro Tamaribuchi
Department of Electronics Engineering, Kogakuin University
2665-1 Nakanocho, Hachioji, Tokyo, Japan

Ryunen Teranishi
Kyushu Institute of Design
Shiobaru, Minami-ku, Fukuoka, Japan

Jun'ichi Toyoda
The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, 567 Japan

Kuniaki Uehara
The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, 567 Japan

Taizo Umezaki
Department of Electrical Engineering, Nagoya University
Chikusa-ku, Nagoya, 464-01 Japan

Wayne H. Ward
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Tomio Watanabe
Department of Information Engineering, Yamagata University
Yonezawa, 992 Japan

Tetsuya Yamamoto
Faculty of Engineering, Kansai University
3-3 Yamate-cho, Suita, 564 Japan

Yasuki Yamashita
Dept. of Electrical and Electronic Engineering, Shinshu University
500 Wakasato, Nagano-shi, 380 Japan

Yoichi Yamashita
The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, 567 Japan

Sheryl R. Young
Computer Science Department, Carnegie
Mellon University
Pittsburgh, PA 15213, USA

Victor W. Zue
Spoken Language Systems Group,
Laboratory for Computer Science
Massachusetts Inst. of Technology,
Cambridge, MA 02139, USA

CONTENTS

Chapter 3. Feature Extraction

Chapter 4. Speech Recognition

Chapter 5. Speech Understanding

Chapter 6. Speech Synthesis

Chapter 7. Dialogue Systems

Chapter 8. Speech Enhancement

Chapter 9. Evaluation

Chapter 10. Speech Database

Chapter 1

OVERVIEW

This Page Intentionally Left Blank

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Overview of Japanese Efforts Toward an Advanced Man-Machine Interface Through Spoken Language

Hiroya Fujisaki

Dept. of Electronic Engineering, Faculty of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

1. INTRODUCTION

Research on speech science and technology has been quite active in Japan for almost 30 years, and has brought about significant contributions to the development of the field. One of the important factors that contributed to these high activities was the role played by the Technical Committee on Speech Research of the Acoustical Society of Japan, which, under the leadership of the present author, has been holding regular monthly meetings since 1973, where at least several reports of on-going research projects are presented and discussed. For the first several years, the activity of the committee was supported in part by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science and Culture awarded to the present author for a project on the integration of various research activities in the field of speech. The annual funding was on the order of a few million yen (10^4 U.S. dollars). It was quite helpful for sustaining the activity of the committee (meeting and publication), and was instrumental in establishing a forum for information exchange among various research groups in the academic, governmental, and industrial circles. The committee, currently sponsored jointly by the Acoustical Society of Japan and the Institute of Electronics, Information and Communication Engineers of Japan, continues to be an important tie among various research groups in Japan.

In spite of these activities, however, no nationwide project has been carried out in the field of speech processing to combine and coordinate research efforts by various groups toward a common goal. Although the Fifth Generation Computer Project, started in 1981 under the sponsorship of the Ministry of International Trade and Industry, had in its early planning stage the computer input/output through speech and natural language as one of its important features, later dropped it completely and concentrated on hardware/software technologies for the realization of fast and efficient deductive inference. As a MITI project, it also emphasized industrial funding and participation but did not encourage academic involvement.

When the author was consulted in 1984 by the Ministry of Posts and Telecommunications for proposals of long-range large-scale projects, one of his suggestions was to start a very long-range basic study toward the realization of "the interpreting telephony," to combine speech processing and natural language translation in a whole system. To

the author's pleasure, the proposal was quickly adopted and a study group was formed to investigate the various possibilities, and an institute (Advanced Telecommunications Research International) was founded in April 1986. As a private enterprise, however, it consists solely of researchers drawn from industries but none from the academic circles.

On the other hand, various national and international projects have been initiated in Europe as well as in the United States, such as the Alvey Programme, the GRECO project, the ESPRIT project, the DARPA Strategic Computing Project, etc. These projects, more or less stimulated by the initial plan of the Japanese FGCS Project, have been pushing into foreground speech and natural language interface as one of their main features. Unlike the FGCS project and the Interpreting Telephony Project of Japan, these projects have been carried out not by a single institution, but by a number of research groups/institutions working toward the realization of a common goal or at least nationally approved individual goals. In order to pursue a long-range goal by a nationwide cooperation of researchers, it seemed to be essential to ask for active participation of academic people — not only faculty members but also graduate students as potential researchers in every circle of the speech community —, as well as researchers from other governmental institutions and private industries.

It was with this realization that a proposal was drafted by the present author toward a national project on the realization of "Advanced Man-Machine Interface Through Spoken Language," to be supported by the Ministry of Education, Science and Culture. It was adopted in September 1986 as one of the Priority Areas under a newly created category of the Grant-in-Aid for Scientific Research, and was started from 1987. Like most of the research projects funded by the Ministry, the project was run for a period of three years, with the fourth year dedicated to summarize the outcome.

2. AIM OF THE PROJECT

As soon as the Ministry of Education, Science and Culture announced its plan to create a new category of Grant-in-Aid for Scientific Research in October 1985 to support certain large-scale research projects of foremost importance as Priority Areas, an informal discussion was held by several leading scientists in the field of speech technology in Japan to explore the possibility of starting a large-scale project under the new category. Through months of discussion it was agreed that a project involving essentially all the active research groups, rather than a few selected groups, would be essential to combine and coordinate individual research efforts at the national level and thus to boost the country's speech technology. It was also agreed that the project should have strong ties with related technologies such as natural language processing, artificial intelligence and knowledge engineering, and the new generation computer technology. Rather than selecting a few specific technical targets to be fulfilled within a span of several years, therefore, it was decided that the main goal of the project should be to establish a system of cooperation among individual research groups scattered all over the country, in which they are assigned different tasks that supplement each other and thus can avoid duplication of research efforts, by sharing common speech data, adopting standardized facilities and analysis techniques as much as possible, and exchanging their experiences and results whenever it is appropriate.

These preparation led to a proposal, drafted and submitted by the present author with the collaboration of eleven other experts, to the Ministry in March 1986. Through stages of screening and hearing, the proposal was adopted in September 1986 as one of the Priority Areas to be started from April 1987. At the same time, a grant was given to start further planning and preparations for the project.

Although several core projects have already been conceived and listed in the original plan, the formal procedure was to announce the Priority Area to accept individual proposals that would fit in the framework of "Advanced Man-Machine Interface Through Spoken Language," in order to provide an opportunity to include new proposals that would complement the core projects and thus would strengthen the whole plan. Following the Ministry's announcement in October 1986, applications were received by December, and notification of accepted proposals were made in May 1987, with certain modifications of the budget to suit the financial requirements and limitations of the Ministry. The current annual budget is approximately 10^6 U.S. dollars.

3. ORGANIZATION

The following eight areas were adopted for intensive research by the project.

1. Advanced Techniques for Speech Analysis
2. Advanced Techniques for Feature Extraction of Speech
3. Advanced Techniques for Speech Recognition
4. Advanced Techniques for Speech Understanding
5. Advanced Techniques for Speech Synthesis
6. Knowledge Processing and Conversational Techniques for Speech Interface
7. Advanced Techniques for Speech Processing in the Presence of Noise and/or Interference
8. Evaluation of Techniques and Systems for Man-Machine Interface

For each of these areas, a research group was organized from several members who were experts already working on related topics, and a larger number of collaborating members who were also actively involved in research but were in general less experienced. These members were mostly from academic institutions, but some were also from governmental research laboratories or private industries. These eight groups were the "core groups" and their research projects form the core of the entire project. In addition to these core projects, however, a total of seven research proposals were adopted. Each of these seven additional projects belonged to either one of the above-mentioned area, and was carried out keeping close contact with the core group of the respective area. In addition, a steering group was organized for the purpose of coordinating the individual projects and encouraging cooperation among various groups. The steering group, headed by the present

author, consisted of the leaders of the eight core groups and eleven other advisory members who were experts either in speech technology or in related fields such as natural language processing and computer science. The number of researchers involved in the project varied slightly from year to year. As of November 1988, the entire project was carried out by a total of 185 researchers, consisting of 87 (expert) members and 98 collaborating members. The organization of the entire project at the time of its start is summarized in Table 1.

Table 1. Organization of the Japanese National Project on Advanced Man-Machine Interface Through Spoken Language

Project Leader and Chairman of the Steering Group		
H. Fujisaki	University of Tokyo	
Core Group	Group Leader	Affiliation
1. Speech Analysis	S. Saito	Kogakuin University
2. Feature Extraction	K. Shirai	Waseda University
3. Speech Recognition	K. Kido	Tohoku University
4. Speech Understanding	M. Shigenaga	Yamanashi University
5. Speech Synthesis	S. Kiritani	University of Tokyo
6. Conversational Systems	O. Kakusho	Osaka University
7. Speech Enhancement	K. Nakata	Tokyo Univ. Agr. Tech.
8. Evaluation and Assessment	S. Hiki	Waseda University
Advisory Members of the Steering Group		
J. Oizumi (Chiba Inst. Tech.)	T. Sakai (Kyoto University)	
M. Nagao (Kyoto University)	S. Itahashi (Univ. of Tsukuba)	
J. Suzuki (Comm. Res. Lab.)	Y. Kato (NEC)	
K. Fuchi (ICOT)	A. Kurematsu (ATR International)	
K. Nakajima (Electrotechnical Lab.)	A. Ichikawa (Hitachi Limited)	
S. Furui (NTT)	M. Nakatsui (Comm. Res. Lab.)	

4. OUTLINES OF CORE PROJECTS

4.1. Advanced Techniques for Speech Analysis

The core group in this research area consists of four expert members and six collaborating members from three universities. The aim of the group is to explore and develop new techniques for the acoustic analysis of speech, especially for the analysis of time-varying characteristics, of individual variations, as well as for accurate analysis of both source and vocal tract characteristics. The group also collects and examines speech analysis software developed by other groups and distributes those that are suitable for the common use. The names, affiliations, and main tasks of the expert members are as follows (* indicates group leader):

*	S. Saito	Kogakuin University	Analysis of time-varying characteristics
	F. Itakura	Nagoya University	Use of group delay spectrum
	N. Miki	Hokkaido University	Source and vocal tract estimation
	K. Tamaribuchi	Kogakuin University	Acoustic and articulatory features

4.2. Advanced Techniques for Feature Extraction of Speech

The core group in this research area consists of four expert members and four collaborating members from two universities and one governmental institution. The aim of the group is to develop novel techniques for deriving both articulatory and auditory parameters that are linguistically important.

*	K. Shirai	Waseda University	Articulatory modeling and vector quantization
	T. Nakajima	Electrotech. Lab.	New methods for spectral analysis
	T. Kobayashi	Hosei University	Perceptually important parameters
	H. Kasahara	Waseda University	Hardware environments

4.3. Advanced Techniques for Speech Recognition

The core group in this research area consists of seven expert members and 18 collaborating members from three universities and two other institutions. The aim of the group is to develop new technique for phoneme recognition, speaker adaptation, and phrase spotting. These techniques will eventually be combined to build a dictation system for carefully pronounced, grammatically correct sentences of Japanese, without resorting to specific information concerning the task.

*	K. Kido	Tohoku University	Sound-phoneme-grapheme conversion
	M. Kato	Tohoku University	Analysis of dialectal variations
	H. Matsumoto	Shinshu University	Techniques for speaker adaptation
	J. Miwa	Iwate University	Speaker-independent phoneme recognition
	S. Makino	Tohoku University	Techniques for phrase spotting
	J. Ujihara	NHK	Use of auditory parameters
	M. Nakatsui	Comm. Res. Lab.	Use of articulatory parameters

4.4. Advanced Techniques for Speech Understanding

The core group of in this research area consists of five expert members and five collaborating members from three universities. The aim of the group is to develop new techniques for speech understanding of carefully spoken sentences from a vocabulary size of 200 to 1000 words. Although the three universities have already been working on their individual speech understanding systems, they will cooperate in adopting a common task and common speech data, so that a comparative evaluation will be possible of the three methods of approach.

*	M. Shigenaga	Yamanashi University	Controls for speech understanding
	I. Sekiguchi	Yamanashi University	Use of prosodic information
	Y. Niimi	Kyoto Inst. Tech.	Use of high-level knowledge
	Y. Kobayashi	Kyoto Inst. Tech.	Use of high-level knowledge
	S. Nakagawa	Toyohashi Inst. Tech.	System architecture

4.5. Advanced Techniques for Speech Synthesis

The core group in this research area consists of seven expert members and 10 collaborating members from one university and two other institutes. The aim of the group is to develop new techniques for both speech synthesis by rule and speech synthesis from concepts, with special emphasis on natural language processing and prosody.

*	S. Kiritani	University of Tokyo	Synthesis of voice quality
	S. Imaizumi	University of Tokyo	Synthesis of segmental features
	H. Fujisaki	University of Tokyo	Natural language processing
	H. Morikawa	University of Tokyo	Text synthesis system
	Y. Sato	Fujitsu Limited	Text synthesis system
	Y. Sagisaka	ATR International	Synthesis prosodic features
	N. Higuchi	KDD	Synthesis prosodic features

4.6. Knowledge Processing and Conversational Techniques for Speech Interface

The core group in this research area consists of eight expert members and 17 collaborating members from four universities and two other institutions. The aim of the group is to develop advanced techniques for knowledge processing and for conversational modeling.

*	O. Kakusho	Osaka University	Supervision of group project
	J. Toyoda	Osaka University	Natural language processing
	T. Kitahashi	Osaka University	Conversational modeling
	M. Yanagida	Comm. Res. Lab.	Knowledge base
	R. Mizoguchi	Osaka University	Knowledge processing
	K. Uehara	Osaka University	Knowledge base
	Y. Miyoshi	Himeji Inst. Tech.	Natural language processing
	Y. Kato	NEC	Conversational system

4.7. Advanced Techniques for Speech Processing in the Presence of Noise and/or Interference

The core group in this research area consists of six expert members from five universities and another institution. The aim of the group is to develop advanced techniques for speech enhancement against noise, interference, and distortion.

*	K. Nakata	Tokyo Univ. Agr. Tech.	Supervision of group project
	H. Kobatake	Tokyo Univ. Agr. Tech.	Speech enhancement
	T. Ifukube	Hokkaido University	Speech signal restoration
	M. Ebata	Kumamoto University	Knowledge-based processing
	A. Ichikawa	Hitachi Limited	Speech through telephone systems

4.8. Evaluation of Techniques and Systems for Man-Machine Interface

The core group in this research area consists of four expert members and nine collaborating members from four universities and three other institutions. The aim of the group is to establish methods for evaluating both the objective performances of various speech processing techniques and the subjective acceptability of such techniques for man-machine interface in various environments.

*	S. Hiki	Waseda University	Assessment of techniques
	H. Kasuya	Utsunomiya Univ.	Psychological scaling
	K. Kakehi	NTT	Evaluation of transmission/processing
	H. Yamamoto	KDD	Evaluation of recognition/synthesis

5. PRINCIPLES OF PROJECT MANAGEMENT

For the purpose of efficient utilization of the grant and available resources for the fulfillment of the goal, the steering committee has set up the following basic principles for the management of the projects.

1. Common use of speech data
2. Common use of software tools
3. Standardization of workstations and other hardware facilities
4. Establishment of a network among major participating groups
5. Encourage information exchange and discussion within the project
6. Encourage information exchange with large-scale projects of other countries
7. Open publication of results

5.1. Common Use of Speech Data

Using common speech data is obviously indispensable for the objective comparison/evaluation of various speech processing techniques and systems. Although previous efforts by Japanese electronics industries to collect speech materials for their common use have resulted in a database of isolated spoken words, and several other attempts are being made by individual organizations, no systematic effort has been made so far to construct a database of connected speech for the common use of research organizations at the

national level. In the current project, isolated utterances (109 syllables, 216 phonetically balanced words) and sentence/discourse material of the common Japanese were collected from 20 speakers (10 male and 10 female, from 20 years to 60 years of age), and made available to all the participating groups.

5.2. Common Use of Software Tools

Common use of software resources is another important factor for the efficient management of the total project. A working group was set up to survey and collect information concerning both existing software tools and those that are yet to be developed, and to encourage their exchange for the mutual benefit among all the groups. As it was already mentioned, the Speech Analysis Group was responsible for the standardization of software tools for speech analysis.

5.3. Standardization of Workstations and Other Hardware Facilities

Because of diversity of participating research groups, existing research facilities cannot obviously be standardized. However, a firm principle was set up to standardize workstations and other equipments to be procured by the current grant, in order to maximize exchangeability of data and results.

5.4. Establishment of a Network among Major Participating Groups

It is hardly necessary to mention the advantage of having a high-speed data communication network among participating groups scatted all over the country. As a start, a network was established via commercial telephone lines among personal computers and workstations.

5.5. Information Exchange and Discussion among Participating Groups

In order to encourage information exchange and discussion among various participating groups, joint meetings of groups with common interests are held at intervals of three months. In addition, the steering committee holds several events and meetings a year for all the members of the project. In principle, a work presented at these meetings are printed as a research report.

5.6. Information Exchange with Large-scale Projects of Other Countries

We consider that an in-depth discussion of various approaches and their results are of vital importance also at the international level, especially among researchers working in

somewhat similar large-scale projects, national or international. In order to encourage exchange at this level, an international symposium/workshop was held every year, where foreign experts were invited to participate and exchange ideas and results. The first of such symposia was held in Tokyo on 12 and 13 of January 1988, the second one in Hawaii during 19 – 22 of November 1988, and the third one in Tokyo during 11 – 14 of December 1989. The major outcomes of the project have also been presented at international conferences including the First International Conference on Spoken Language Processing, held in Kobe during 18 – 22 of November 1990.

5.7. Open Publication of Results

Because of the academic nature of the project, an open policy was adopted for the publication of results.

6. OUTLINES OF CLOSELY RELATED PROJECTS AND ACTIVITIES

6.1. Advanced Natural Language Processing

A national project has been conducted from 1986 to 1989 by the support of MESC toward the realization of advanced natural language processing. The project is similar in the number of researchers and in the annual budget as the above-mentioned 'spoken language' project, but deals exclusively with the written language. The project was run by six subgroups, working on the following themes:

1. Studies of theoretical linguistics as a basis for machine processing of natural language
2. Comparative studies of linguistic structures of various languages for machine processing
3. Studies on contextual information processing
4. Studies on collection and processing of language data
5. Studies on languages for information documentation
6. Studies on natural language processing systems based on human processes of linguistic information.

The subgroup on the last theme involved researchers in the fields of cognitive psychology, psycholinguistics, as well as computer science, and was headed by the present author. In view of the importance of crossfertilization of the two areas, i.e., the processing of the written language and the processing of the spoken language, a symposium was held to discuss topics of common interest for both areas.

6.2. Investigation of Prosodic Features of Spoken Japanese

Another national project was started in 1989 also by the support of MESC for the investigation of prosodic features of the spoken Japanese, with applications to the teaching of the Japanese language. The project consists of 10 subgroups:

1. Collection and Analysis of Speech Data
   (1) Tokyo Dialect including Radio Announcers
   (2) Dialects of Eastern Japan
   (3) Dialects of Western Japan
   (4) Dialects of Ryukyu Islands
2. Construction and Utilization of Database of Collected Speech Materials
3. Analysis and Interpretation of Speech Data
   (1) Acoustic Analysis
   (2) Physiological Analysis
   (3) Linguistic Analysis
4. Guidelines for Teaching Japanese
   (1) Teaching Japanese to Foreigners
   (2) Teaching Japanese to Native Speakers

Although the emphasis of the project is on the study of prosody from linguistic, dialectal, and educational points of view, the database of the dialectal speech will be of great value also as a material for spoken language processing.

6.3. Industry-Academia Cooperative Committee on Intelligent Processing of Spoken and Written Languages

Speech processing for man-machine interface almost invitably involves natural language processing. Or more precisely, the processing of the spoken language and the processing of the written language are quite closely related and share many important elements. For instance, a high-quality speech synthesis from text requires a deep understanding of the text based on syntactic, semantic, and discourse analyses. Unfortunately, however, speech processing and natural language processing have been studied quite separately in the past.

In order to provide a forum for better and closer communication between researchers working in these two areas of study, as well as between researchers in the academia and those in the industries, an Industry-Academia Cooperative Committee was established in 1987 under the auspices of the Japan Society for the Promotion of Science. The committee consists of 28 members from the academia and 25 from the industries, and holds biomonthly meetings to discuss problems and topics that are of common interest to researchers in both fields.

7. CONCLUSION

An overview has been given on some Japanese efforts toward the realization of advanced man-machine interface through spoken language. Since most of the coordinated projects in Japan are rather new, much has to be learned from the experiences of national projects of other countries as well as of international projects such as the ESPRIT projects. I am pleased, however, that our emphasis on the integration of speech and natural language processing into one well-defined area of spoken language processing is being shared by other national and international projects. Although the project on "Advanced Man-Machine Interface Through Spoken Language" was terminated in March 1991, the ties of cooperation among various research groups, established both within and outside Japan, will surely be strengthend and even broadened through our continued efforts.

This Page Intentionally Left Blank

Chapter 2

SPEECH ANALYSIS

This Page Intentionally Left Blank

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Composite Cosine Wave Analysis and its Application to Speech Signal

Shuzo Saito and Kazuhiro Tamaribuchi

Department of Electronics Engineering, Kogakuin University
2665-1 Nakanocho, Hachioji, Tokyo, Japan

Abstract

A method for the analysis of the frequency components of an acoustic signal expressed in composite cosine waves is described. It is shown theoretically that three parameters of each of the frequency components of an input acoustics signal, that is the frequency, amplitude, and initial phase of each component, can be determined from $3m$ discrete sampled data points for an input signal composed of m frequency components, which not necessary have harmonic relations. This method of analysis is applied to speech signals and several results are described.

1. INTRODUCTION

To analyze the frequency components of acoustic signals we have several calculation procedures, like Fourier analysis and so on. The digital Fourier transformation procedure (DFT) is regarded as a useful means for acoustic signals represented by time-sampled sequential data of the signal waveforms. The results of a DFT analysis, however, are expressed as a set of frequency components with a harmonic structure, a lot of frequency components with a harmonic relation are required to represent the result of the analysis, even if the input signal is composed of only two kinds of sinusoidal waves with an inharmonic relation.

There is another approach to analyze the frequency components of an input signal for which the frequency components do not necessarily have a harmonic relation, i.e. by use of the autocorrelation functions of the signal. In this approach the results of the analysis include several errors caused by the calculating procedure, in which the integration over an infinite region is replaced by a summation over a finite region to cope with the discrete signal data.

This paper reports an analysis method for acoustic signals which is able to determine the characteristics of the components of the signal from the discrete sampled data of the signal within a restricted region of analysis. It is shown theoretically that the required number of discrete data points of the signal in this analysis is only three times the number of frequency components, provided that the signal is composed from a finite number

of frequency components within the region of analysis. The validity of this method of analysis is verified by use of an artificial signal composed of five frequency components. Then several results using this method for a speech signal are described.

2. THEORY OF FREQUENCY ANALYSIS

A composite cosine wave which is composed of m cosine waves, at the time rT can be represented as follows:

$$f(r) = \sum_{n=0}^{m-1} a_n \cos(r\omega_n T + \phi_n) \tag{1}$$

where a_n is the amplitude, ω_n the angular frequency, ϕ_n the initial phase and T the sampling interval.

2.1. Determination of the Frequency Components

To determine ω_n, the function $F_s(r-s)$ is defined as

$$F_s(r-s) \equiv \frac{1}{2}(1-\delta_{s0})\{F_{s-1}(r-s) + F_{s-1}(r-s+2)\} + \delta_{s0} f(r-s) \tag{2}$$

where δ_{s0} is Kronecker's delta , $0 < s < [(3m+1)/2] - 1, s < r < 3m - 2s - 1$,
and [ ]of $[(3m+1)/2]$ is Gauss's symbol.

F_s can be calculated from $3m$ sequentially sampled data points being in the ranges of s and r indicated above.

From eqs. (1) and (2) the following equation is derived :

$$F_s(r-s) = \sum_{n=0}^{m-1} a_n \cos(r\omega_n T + \phi_n) \cos^s(\omega_n T). \tag{3}$$

Since $F_s(r-s)$ is a function of $\cos^s(\omega_n T)$, the following equations are mth-order algebraic equations of $\cos(\omega_n T)$.

$$F_m(r-m) + \sum_{s=0}^{m-1} b_s F_s(r-s) = 0 \tag{4}$$

where $r = m, m+1, ..., 2m-1, b_s$ are constants, and $s = 0, 1, ..., m-1$.

From eqs. (4), the coefficients b_s are determined and then used to solve the following equation:

$$x^m + \sum_{s=0}^{m-1} b_s x^s = 0. \tag{5}$$

Since the roots of eq. (5) are equal to $\cos(\omega_n T)$, the frequencies can be determined from the following equations:

$$x_n = \cos(\omega_n T), \tag{6}$$

where $n = 0, 1, ..., m-1$.

2.2. Determination of the Initial Phase and Amplitude

Since the values of $F_s(r-s)$ and $\cos(\omega_n T)$ in eq. (3) can be determined, solutions for $a_n \cos(r\omega_n T + \phi_n)$ can be derived and are denoted as $A_n(r)$. Then the phase, ϕ_n, is determined by use of $A_n(r_1)$ and $A_n(r_2)$ for r_1 and r_2, where $r_1 \neq r_2$, $m-1 \leq r_1, r_2 \leq 2m$, as follows:

$$\phi_n = \arctan \frac{C_n(r_1, r_2)}{S_n(r_1, r_2)}, \tag{7}$$

where

$$C_n(r_1, r_2) = A_n(r_1)\cos(r_2\omega_n T) - A_n(r_2)\cos(r_1\omega_n T), \tag{8}$$

$$S_n(r_1, r_2) = A_n(r_1)\sin(r_2\omega_n T) - A_n(r_2)\sin(r_1\omega_n T). \tag{9}$$

Finally, the amplitude a_n is represented as

$$a_n = \frac{A_n(r_1)}{\cos(r_1\omega_n T + \phi_n)} \tag{10}$$

3. VERIFICATION OF THE METHOD OF ANALYSIS BY APPLYING IT TO AN ARTIFICIAL SIGNAL

The analysis procedure described above is applied to artificial signals composed of five sinusoidal frequency components with frequency lower than 5 kHz. The three parameters of the five frequency components are determined by use of a random variable with a uniform distribution. The artificial signal is fed to a 4.5 kHz low-pass filter, then sampled at 10 kHz, and its amplitude is quantized using 12 bits.

The frequency component analysis is performed using 15 sequentially sampled data points of the waveform, that is 1.5 ms long. The analysis of the frequencies, initial phases, and amplitudes of the five frequency components is done with 64 bit double precision using an Eclipse MV/7800 computer.

The results are shown in Table 1. It is seen that the results of the analysis of the components agree with the true values with an accuracy of 9 digits. In any result an error is observed at the 8th to 9th digit after the decimal point, which seems to be within the computation error of the computer.

4. RESULTS OF THE ANALYSIS FOR A SPEECH SIGNAL

4.1. Recursive Procedure of Component Analysis for a Speech Signal

In this analysis method, the number of sequentially sampled data points of a signal is taken as $3m$, that is three times of the number of frequency components m. Applying this method of analysis to a speech signal, it is necessary to determine the number of frequency components of the speech signal in advance. Although the number of frequency components for speech is essentially unknown, the necessary maximum number of frequency components can be estimated experimentally.

Table 1. Results of the analysis of a signal composed of five frequency components

	FREQUENCY[Hz]	PHASE[deg]	AMPLITUDE
TRUE	4110.5270386	65.7023621	0.0075531006
	2830.0285339	-6.4599609	0.5151557922
	2501.2969971	210.7960510	0.4859390259
	2253.0555725	5.6661987	0.3516044617
	318.3174133	108.5696411	0.1190757751
DOUBLE	4110.5270386	65.7023621	0.0075531006
	2830.0285339	-6.4599609	0.5151557922
	2501.2969971	210.7960510	0.4859390259
	2253.0555725	5.6661987	0.3516044617
	318.3174133	108.5696411	0.1190757751
12 bits	4110.8529017	65.6202150	0.0075929833
	2829.9996775	-6.4369952	0.5151029358
	2501.2877524	210.8099101	0.4857993445
	2253.0759978	5.6491724	0.3516911544
	318.2913003	108.5781091	0.1190958824

TRUE : *True values of the components of the input signal.*
DOUBLE : *Resules for 64 bit double-precision data.*
12 bits : *Resules for 12 bit quantized data.*

A speech signal is, however, substantially unstable, so it is necessary to use a kind of analysis-by-synthesis technique when applying this method of analysis to a speech signal. The first step is to analyze the speech signal with a maximum number of frequency components, then a speech-like signal is reconstructed using the result of the analysis and compared with the input speech signal, and then the error between the reconstructed and original signal is calculated. If the error value is less than the threshold value, then the analysis of the present sampled data is finished and one can proceed to the succeeding sampled data. But in the case that the error value is over the threshold value, the maximum number of frequency components is reduced by one and is the same analysis procedure repeated. This recursive procedure used for a speech signal is illustrated in Figure 1. The threshold value used mainly in this paper is 0.5 LSB.

4.2. Estimation of the Maximum Number of Frequency Components

The effects of the maximum number of frequency components on the results of the analysis are measured for a maximum number of 10, 15, 20, 25, and 30, respectively. The speech signal of the Japanese word /namae/ (name) is used in this experiment. The setting of the sampling frequency and the amplitude quantization is similar to that in

the preceding experiments. The results are shown in Figure 2. It seems that a maximum value of 25 or 20 may be sufficient for speech analysis.

4.3. Comparison of the Analyzed Spectrum with the LPC and FFT Procedures

Setting the maximum number of frequency components to 20, an analysis of the frequency components of several Japanese words is performed. The experimental conditions for the analysis are similar to the preceding ones. As an example the results of the analysis for Japanese vowel /i/ which is preceded by the consonant /s/, are shown in Figure 3. the frequency component analysis is done using 54 sequentially sampled data points, that is 5.4 ms long, of the /i/ sound. Similar analyses are executed by the LPC and FFT procedures for the same speech material and the results are shown in the same figure. The number of sampled data points used in the FFT analysis is 256 points, whereas that used in the LPC analysis is 54 points.

Comparing the results derived from the three kinds of procedures, the frequency-spectrum envelopes estimated are very similar in general, but a few interesting differences are observed in the fine structures of the speech spectrum, especially between the composite cosine method and LPC. The effects of such spectrum differences on the speech quality should be checked using some synthetic procedures in the near future.

4.4. Distributions of Segment Lengths Utilized for Frequency Component Analysis

It is expected that the maximum number of frequency components may be a kind of measure for the temporal stability of the input signal in the composite cosine wave analysis method. So speech material similar to that referred to in section 4.2. were analyzed and the distributions of the segment lengths which are utilized in the frequency component analysis, are calculated. Speech material from nine speakers in six repetitions is used for this experiment. The experimental conditions are similar to those mentioned in section 4.3..

The results are shown in Figures 4 (a), (b). The abscissa of each figure represents the number of frequency components in the analysis and the ordinate represents the percentage of the number of discrete sampled data points corresponding to the value of the abscissa when speaking the word. It should be noted that the percentage value used in the ordinate of this figures differs from that in Figure 2. The distributions shown in Figures 4 (a) and (b) are corresponding to those for different speakers and for one speaker, respectively. It is seen that fair differences are observed in the distributions for different speakers, but not in the distributions for one speakers. It is necessary to make further study of the individual differences contained in a speech signal using the composite cosine analysis method.

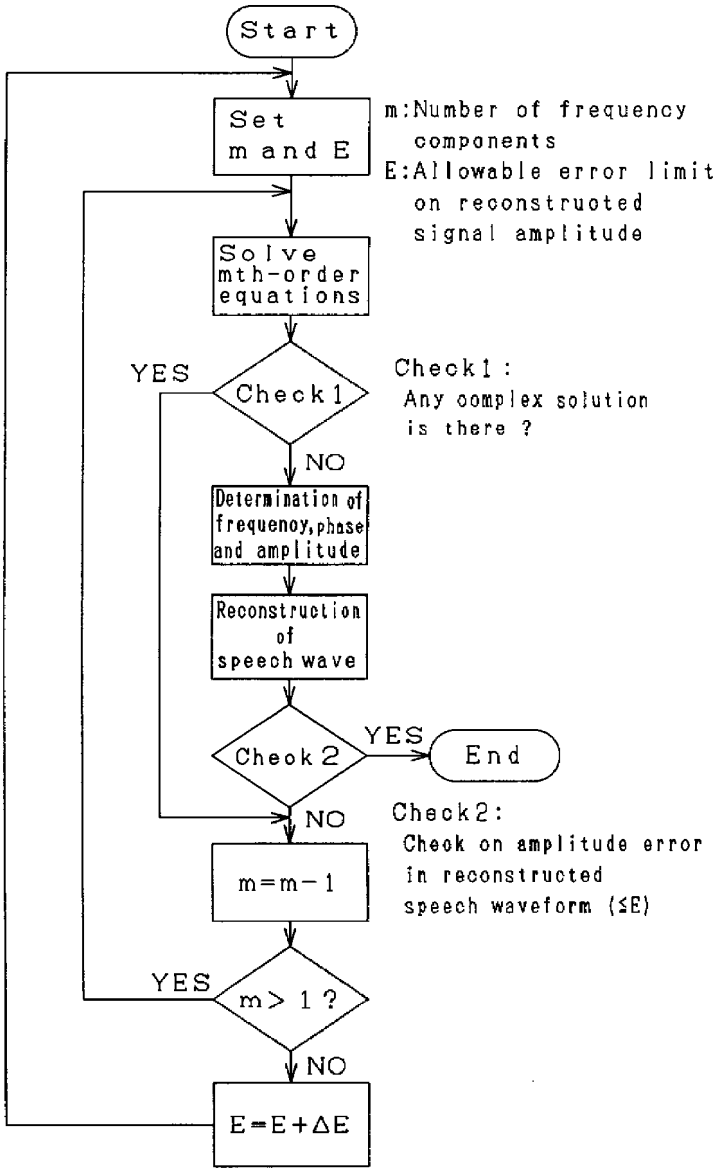


Figure 1. Analysis algorithm used for a speech signal.

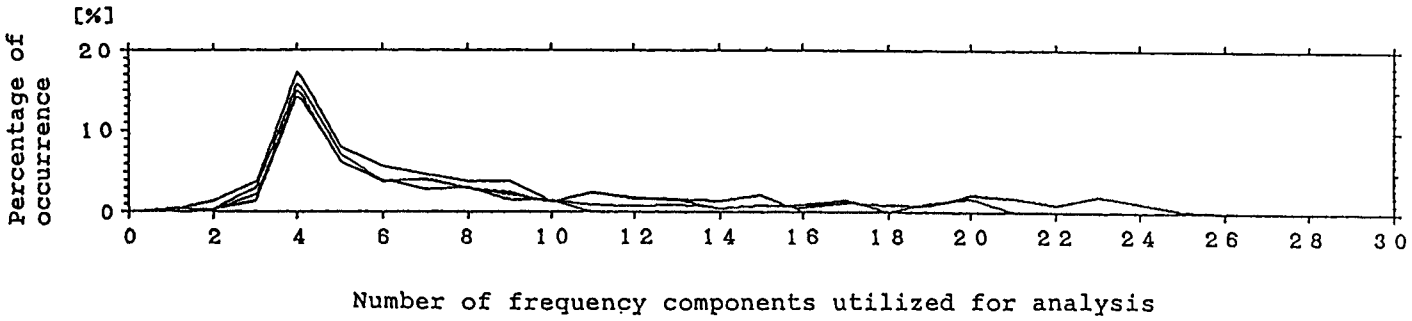


Figure 2. Effect of the maximum number of frequency components in the composite cosine wave analysis

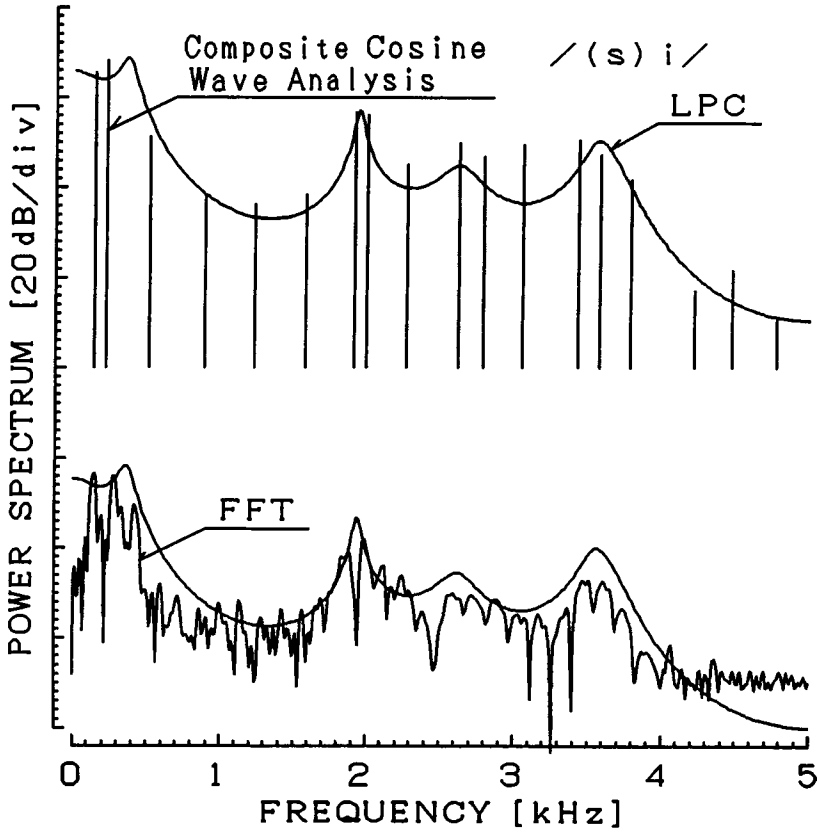


Figure 3. Results of the composite cosine wave analysis for the vowel /i/ and comparison with LPC and FFT analyses.

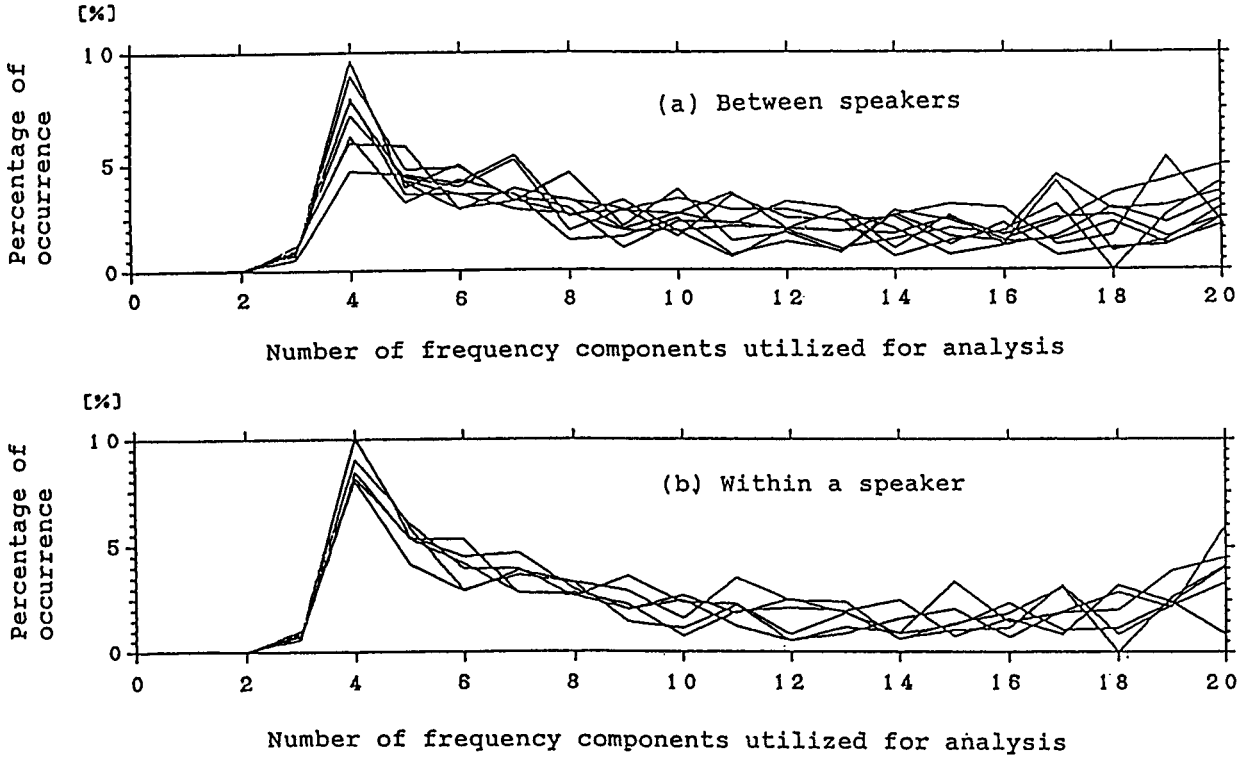


Figure 4. Distribution of the segment length utilized for analysis (a) for nine different male speakers, (b) for one speaker.

5. CONCLUSION

The theory of the composite cosine wave method of analysis and its application to speech analysis are described in this paper. The validity of this method is examined in the case of steady acoustic signals containing five frequency components. The results show that not only the frequency, but also the initial phase and amplitude can be determined with an accuracy of nine digits or more for analyzed data with double-precision representation.

This analysis method is also applied to a signal for which the number of frequency components is unknown and for which the frequency components are unstable temporally, like in speech. For this analysis, several modifications are added to the analyzing procedure and examined with Japanese speech signals. The results show that there are several uses of this method for fine analysis of the speech structure. It seems that the composite cosine wave method of analysis plays a role as a cooperative tool with other speech analysis procedures as LPC, FFT and so on, to achieve further advances in speech information processing.

References

1. S. Saito and K. Tamaribuchi: "A study on dynamic characteristics of speech signal," *Preprints of The First Symposium on Advanced Man-Machine Interface Through Spoken Language*, p.103, 1988.

2. S. Sagayama and F.Itakura: "Composite sinusoid modeling applied to spectrum analysis of speech," *Trans. Committee on Speech Res., Acous. Soc. Japan*, S79-06, 1979.

3. K. Tamaribuchi and S. Saito: "A new analysis method for acoustic signals composed of cosine waves," *Proc. ICASSP*, p.2440, 1988.

4. K. Tamaribuchi and S. Saito: "Spectrum estimation based on an analysis method of speech signal composed of cosine waves," *Trans. IEE Japan*, vol.108-C, No.10, p.781, 1988.

5. S. Saito and K. Tamaribuchi: "Spectral estimation of speech based on a composite cosine wave model," *J.A.S.A.*, Suppl. No.1, vol.84, p.S13, 1988.

6. K. Tamaribuchi and S. Saito: "An analysis method for acoustic waves composed of cosine waves," *Trans. IEICE Japan*, vol.J72-A, No.1, p.49, 1989.

Recent Research Towards Advanced Man-Machine Interface
Through Spoken Language, H. Fujisaki (Editor)

Smoothed Group Delay Analysis and its Applications to Isolated Word Recognition

Harald Singer, Taizo Umezaki and Fumitada Itakura

Department of Electrical Engineering, Nagoya University
Chikusa-ku, Nagoya, 464-01 Japan

Abstract

In previous work [1, 2] the effectiveness of the smoothed group delay distance measure was shown. The coefficients of the smoothed group delay spectrum (SGDS) were there calculated by multiplying the LPC Cepstrum coefficients with a smoothing weight function, i.e. the representation was in the time domain.

In this paper, we calculated the SGDS coefficients using a DTFT (Discrete Time Fourier Transform) of the linear-prediction coefficients, i.e. the representation is in the frequency domain. We report isolated-word recognition experiments with low bit quantization of these SGDS coefficients. We show that the recognition accuracy can be maintained using only 26 bits per frame, as compared to the conventional calculation with floating point accuracy. Using a bark scale representation the error rate can even be further reduced.

1. INTRODUCTION

For speech recognition, extraction of the relevant features of the speech signal, commonly called analysis, is of vital importance for the recognition process. Important speech information removed at this stage cannot easily be recovered later on.

Analysis and the subsequent similarity calculation have been intensively studied over the past years. One common standard is the LPC Cepstrum, which performs very well in a well-behaved environment. On the other hand, under adverse conditions, like variable frequency characteristics (different microphones, changing transmission lines, etc.), or additive noise, LPC based measures are strongly affected.

To overcome this problem, weighting of the Cepstral coefficients has been proposed, with the objective to emphasize the spectral peaks and to separate the influence of neighboring poles. One of these proposed analysis methods is the smoothed group delay spectrum [1]. It was shown that the Fourier coefficients $G(n)$ of the group delay spectrum are identical to the Cepstral coefficients weighted by a weighting function $W(n) = n$. To avoid overemphasis of the Cepstral components with large index n and problems with

truncation effects due to a finite number of coefficients, a generalized weighting function $W(n)$ having the form of a Gaussian window has been proposed:

$$W(n) = n^s \exp(-\frac{n^2}{2\tau^2}) \quad (1)$$

Multiplication with this Gaussian window in the quefrency domain is equivalent to smoothing with a Gaussian window in the logarithmic spectral domain. The necessity of smoothing the group delay spectrum can easily be seen from Figure 1. The peaks at the formant frequencies of the group delay spectrum, shown in Figure 1d, are very sharp, which may be appropriate for estimation of the formant frequencies. For speech recognition, on the other hand, a broadening of the peaks is indispensable, since just a slight deviation in formant frequencies between two patterns would result in a large distance value. We can see that this smoothing is achieved by using Gaussian weighting according to eq.(1) (see Figure 1e). Optimal recognition results for this smoothed group delay spectrum were obtained for τ =10.0 (corresponds to a window bandwidth of 300 Hz) and $s = 1.0$. This distance measure performs very well even under noisy conditions.

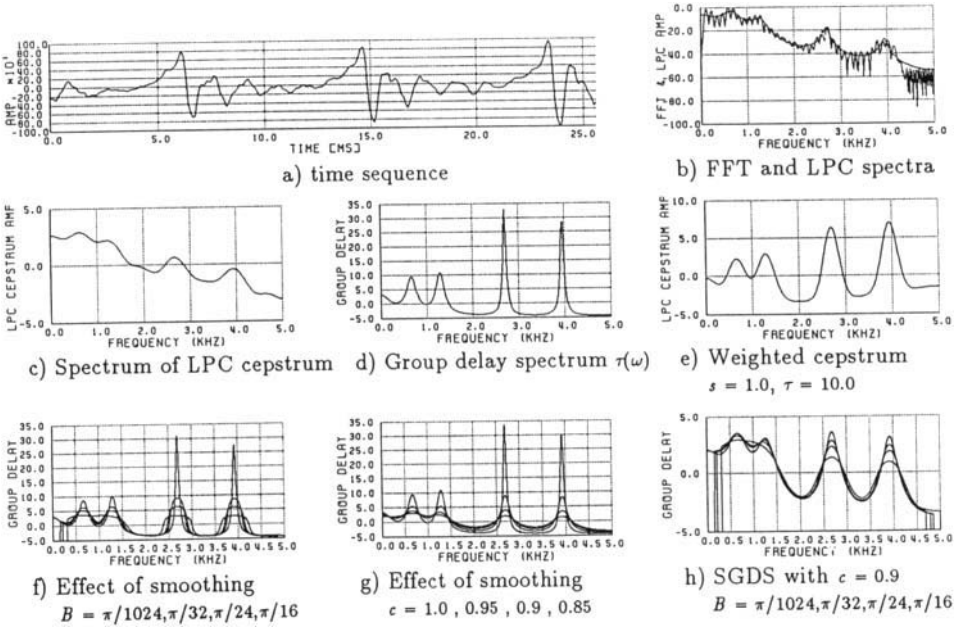


Figure 1. Second vowel /a/ in MIKASA.

In section 2 of this paper we show that a similar (not identical) smoothed group delay function can be directly calculated from the LPC coefficients. We introduce the basic algorithm, give details on the databases and the pattern matching algorithm (standard DTW)

and report recognition results, comparing the new method with the above-mentioned smoothed group delay weighted LPC Cepstrum and the standard LPC Cepstrum (CEP). For ease of explanation we will call the smoothed group delay spectrum calculated by the Gaussian weighted LPC Cepstrum from now on simply weighted Cepstrum (WCEP) to distinguish it from the directly calculated smoothed group delay spectrum (SGDS).

In section 3, we report the results of modifications of the basic algorithm. A rough quantization in 1 to 3 bits per parameter and the use of the perception based bark scale are investigated. Finally, we discuss our findings in section 4.

2. BASICS

2.1. Basic Algorithm

In the LPC analysis each frame of speech can be represented by a pth order all pole filter with transfer function

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{2}$$

The LPC coefficients a_k are calculated with the autocorrelation method.

With $H(w) = H(z)|_{z=e^{jw}}$ and $A(w) = A(z)|_{z=e^{jw}} = \sum_{k=0}^{p} a_k e^{-j\omega k}$, the phase $\Theta(\omega)$ of the transfer function H(z) can be written as

$$\Theta(\omega) = -\arctan\frac{\mathrm{Im}\{A(\omega)\}}{\mathrm{Re}\{A(\omega)\}} \tag{3}$$

For the calculation of $\Theta(\omega)$, we used a discrete time Fourier transform (DTFT) with the Goertzel algorithm[3].

The group delay spectrum is defined as the derivative of the phase

$$\tau(\omega_i) = -\left.\frac{\partial\Theta(\omega)}{\partial\omega}\right|_{\omega=\omega_i} \tag{4}$$

Instead of this derivative we used the finite difference

$$\hat{\tau}(\omega_i) = -\frac{1}{\omega_i - \omega_{i-1}}\int_{\omega_{i-1}}^{\omega_i}\tau(\omega)d\omega = -\frac{\Theta(\omega_i) - \Theta(\omega_{i-1})}{\omega_i - \omega_{i-1}}, \quad 1 \le i \le L, \tag{5}$$

where L is the number of channels of the SGDS and $\omega_i > \omega_{i-1}$ with $\omega_0 \ge 0, \omega_L \le \pi$.

The averaging operation in eq.(5) is equivalent to the multiplication of the Cepstral coefficients with a sinusoidal window. The logarithmic spectrum of the LPC Cepstrum and the group delay spectrum can be expressed as

$$\log|H(\omega)| = \sum_{k=0}^{\infty} c_k \cos k\omega, \quad \tau(\omega) = \sum_{k=0}^{\infty} k c_k cos(k\omega), \tag{6}$$

where c_k are the Cepstrum coefficients. The phase $\Theta(\omega)$ is therefore

$$\Theta(\omega) = -\int_0^{\omega} \tau(x)dx = -\int_0^{\omega} \sum_{k=0}^{\infty} kc_k \cos(kx)dx = -\sum_{k=0}^{\infty} kc_k \frac{\sin(k\omega)}{k} = -\sum_{k=0}^{\infty} c_k \sin(k\omega) \tag{7}$$

If we define $\hat{\tau}(\omega, B)$ as the difference equation smoothed group delay function, where the bandwidth is $2B$ we get

$$\begin{aligned} \hat{\tau}(\omega, B) &= \frac{\Theta(\omega + B) - \Theta(\omega - B)}{2B} = \frac{\sum_{k=0}^{\infty} c_k \sin(k(\omega + B)) - \sum_{k=0}^{\infty} c_k \sin(k(\omega - B))}{2B} \\ &= \frac{\sum_{k=0}^{\infty} c_k \cos(k\omega)\sin(kB)}{B} = \sum_{k=0}^{\infty} \frac{\sin(kB)}{B} c_k \cos(k\omega) \end{aligned} \tag{8}$$

If we compare eq.(6) with eq.(8) we see that we have in fact weighted the Cepstrum coefficients c_k with the sinusoidal function $\sin(kB)/B$. For 16 sampling points, B would be $\pi/32$. The use of a sinusoidal window for smoothing was also advocated in [4], although the window length is different from our approach (here the window length is not limited). The effect of different values of B can be seen in Figure 1f.

An additional smoothing effect is achieved by multiplying the coefficients a_k with an exponentially decaying window prior to the DTFT

$$\hat{a}_k = c^k a_k, \quad c < 1.0, \tag{9}$$

which moves the poles z_k of $H(z)$ away from the unit circle. The poles of $\hat{H}(z)$ are then located at $\hat{z}_l = z_l c$. The effect of this smoothing for different values of c on the GDS is shown in Figure 1g. In Figure 1h both types of smoothing were used with $c = 0.9$ and different values for B.

The complete basic algorithm is depicted as a block diagram in Figure 2. In the recognition phase, we define the distance metric of the SGDS between to speech frames as

$$d = \sum_{i=0}^{L} \{\hat{\tau}_R(\omega_i) - \hat{\tau}_T(\omega_i)\}^2 \tag{10}$$

where $\hat{\tau}_R(\omega_i)$ and $\hat{\tau}_T(\omega_i)$ are the values of the ith channel of the reference pattern and the test pattern, respectively.

2.2. Databases and Matching Algorithms

2.2.1. Databases

Database 1 consists of 550 Japanese citynames recorded twice and spoken by 5 Japanese male speakers. The second utterances were recorded 1 week after the first utterances.

68 easily confusable pairs (see Table 1) were chosen for the recognition experiments. In preliminary experiments we found that nearly all errors occurred as a confusion of these

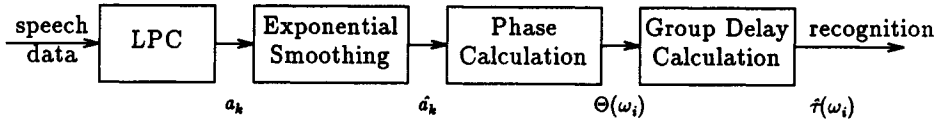


Figure 2. Block diagram of the basic algorithm (LPC parameters a_k, smoothed LPC parameters $\hat{a}_k$, phase $\Theta(\omega_i)$, and smoothed group delay $\hat{\tau}(\omega_i)$.

pairs. Therefore we only performed DP Matching between these pairs, to speed up our experiments. The analysis conditions for this database are shown in Table 2.

Database 2 consists of 15 digits (see Table 3) recorded once and spoken by 50 Japanese male speakers. The analysis conditions for database 2 are shown in Table 4.

The pattern matching algorithm is a standard endpoint constrained DP algorithm. Database 1 was used for intra-speaker recognition (speaker dependent) with reference patterns and test patterns from the same speaker. We also investigated inter-speaker recognition with reference and test patterns from different speakers.

Database 2 was only used for speaker independent recognition. 10 speakers were designated as reference speakers and their utterances were stored as reference templates, the other 40 speakers were test speakers. A variant of the popular K-mean algorithm was used: an "unknown" test pattern was DP-matched with all 10 15 reference templates and the calculated distances were saved. For each of the labels the 3 best results were averaged ($K = 3$). The label with the best averaged result was chosen as final result.

2.2.2. Addition of Noise

To investigate the robustness of the distance measure under noisy conditions, we added multiplicative signal-dependent white noise as follows:

$$\hat{S}_n = S_n(1 + AR_n) \quad with \quad A = \sqrt{3}\, 10^{-\frac{\mathrm{SNR[dB]}}{20}} \tag{11}$$

S_n is the "clean" speech signal, $\hat{S}_n$ the noisy speech signal, A the relative noise amplitude, SNR[dB] the desired signal to noise ratio, and R_n a uniform distributed random number between -1 and 1. We only investigated two cases, namely no noise and SNR = 20dB.

2.3. Comparison of Experimental Results

First, we tried to find an optimal value for the smoothing factor c. The number of SGD channels was fixed at 16. In accordance with ref.[5] an increase in the number of channels did not improve the recognition accuracy. Figure 3 shows the error rate for database 1 for different values of c. Both intraspeaker and interspeaker experiments were also performed with added segmental noise (see section 2.2.4). As a consequence, we choose $c = 0.9$ for all subsequent experiments (if not explicitly mentioned otherwise). This value was also confirmed in ref [5].

Tab. 1. Confusable pairs in database 1

wako:	ako:	oga	koga	toda	noda	kuji	huji
toba	tosa	tama	zama	sakai	kasai	nara	naha
uji	huji	kuji	uji	kitami	itami	kamaishi	takaishi
takahagi	takasaki	hamada	yamada	mikasa	mitaka	mutsu	huttsu
hirakata	hirata	hirata	hirara	hino	chino	chiba	chita
takikawa	tachikawa	kashiwara	kashihara	sunagawa	sukagawa	morioka	tomioka
hukagawa	sukagawa	yono	ono	yashio	yachiyo	o:da	onoda
toyosaka	toyonaka	mobara	obama	oyama	toyama	okaya	okayama
cyo:hu	ko:hu	o:tsu	go:tsu	o:tsu	o:bu	o:date	o:take
o:muta	o:mura	to:no	o:no	gobo:	gojo:	o:me	ko:be
kakuda	katsuta	matsuzaka	matsubara	matsuyama	matsubara	matsuzaka	matsuyama
takayama	takahama	takahama	nagahama	takayama	wakayama	handa	sanda
sagae	sabae	sanjo:	anjo:	hukui	tsukumi	kiryu	chiryu
hukushima	hukuchiyama	utsunomiya	ichinomiya	ichinomiya	nishinomiya	ito	mito
iwatsuki	iwakuni	kamogawa	kakogawa	ichikawa	ichihara	hamamatsu	takamatsu
matsudo	matsuto:	tsushima	kushima	atsugi	yasugi	izumi	izumo
nagaoka	takaoka	takaoka	kasaoka	enzan	sennan	kaga	saga

Tab. 2. Analysis conditions for database 1

sampling rate	10kHz
sampling accuracy	12bit
window type	Hamming
window length	25.6ms
window period	8ms
order of LPC	10

Tab. 3. Labels in database 2

0	zero, rei, maru	5	go
1	ichi	6	roku
2	ni	7	nana, shichi
3	san	8	hachi
4	yon,shi	9	kyuu, ku

Tab. 4. Analysis conditions for database 2

sampling frequency	8kHz
sampling accuracy	12bit
window type	Hamming
window length	32ms
window period	10ms
order of LPC	10

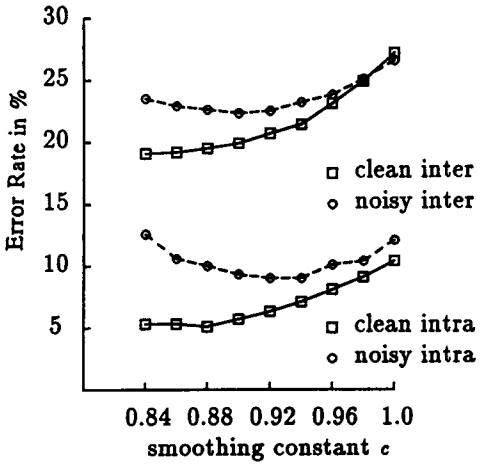


Figure 3. Dependency of the error rate on the smoothing constant c for interspeaker and intraspeaker tests.

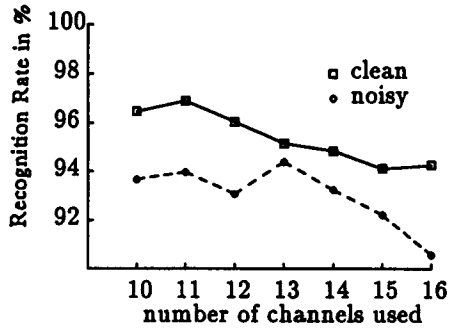


Figure 4. Effect of not using the SGD parameters of higher frequencies.

We also found that not using the higher channels improved the recognition rate considerably (see Figure 4). The best results were obtained using only channels 1 to 11. At 10 KHz sampling rate and evaluation of $\tau(\omega_i)$ at 16 equally spaced frequency points, the bandwidth of 1 channels corresponds to 5KHz/16 $\approx$ 300Hz. Not using channels 12 to 16 is thus equivalent to using a low-pass filter with cutoff frequency at 3.3KHz. The relevant information is apparently confined to the region up to around 3.3KHz. This result is interesting in view of the fact that telephonic speech uses the frequency range between 300Hz and 3.4KHz.

We then compared our results with standard LPC Cepstrum (CEP) and weighted LPC Cepstrum ($s = 1.0, \tau = 10.0$ from eq.(1)), both with 16 parameters (channels) per frame. Figure 5 shows the recognition results for database 1 (intraspeaker only) and database 2. Without noise the 3 methods have comparable results. With noise added, the score for CEP drops considerably. Especially database 1, with its many confusable consonants, is strongly affected by noise. Similar results have been reported in ref [5]. On the other hand, we could show that both WCEP and SGD are quite robust against white noise influences.

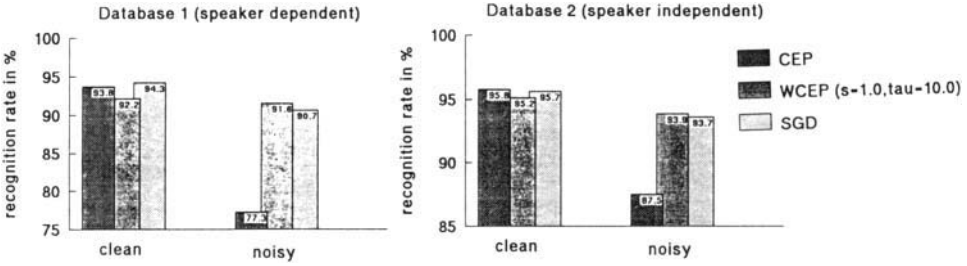


Figure 5. Comparison of Cepstrum; weighted Cepstrum, and SGD.

3. MODIFICATIONS

3.1. Low Bit Quantization

Under the assumption that the essential features for speech recognition are the frequencies of the formants, we can further reduce the information contents by low bit quantization of the SGD parameters. Three problems have to be addressed:

- How many bits per channel are necessary?
- Where should the quantization thresholds be set for optimal recognition results?
- Which value should be assigned to each quantization level?

We tried to solve these problems heuristically, since minimum distortion is not necessarily equivalent to a maximum recognition rate. Quantization also works as an additional

smoothing effect, which can even raise the recognition score. The optimal quantization thresholds for quantization in 1,2 and 3 bits are shown in Table 5. Note that the LPC order is $p = 10$. The theoretical minimum for the group delay $\tau(\omega_i)$ can be shown to be $-p/2$ (here $-p/2 = -5$). It is obvious that the threshold values depend also on the smoothing factor c (here $c = 0.9$).

Figure 6 demonstrates how the SGDS channels $\tau(\omega_i)$ were quantized according to Table 5. Figure 7 shows how the formant structure is preserved even for quantization with 1 bit.

The recognition results for databases 1 and 2 are shown in Figure 8. Obviously, quantization with 1 bit is too rough, whereas 2 bits gives comparable results to calculation with floating point precision.

Only using channels 1 to 11, as explained in section 2.3, gives better results than using all 16 channels also for the quantized SGDS. This signifies that one frame of speech can be represented by 22 bits. A hypothetical large-vocabulary recognition system with 5000 words, average word length 0.5s and 125 frames/s would thus necessitate $5000 \times 0.5 \times 125 \times 22/8$ Bytes $\approx$ 1MByte of memory space. SGDS is thus well suited for the implementation of an inexpensive large-vocabulary recognition system.

Tab. 5. Thresholds for quantization

Bits	Thresholds						
1	0						
2	0	2.5	5				
3	-2.5	0	1	2	3	4	5

3.2. Bark scale and Mel scale

Many researchers suggest a perception based approach, using filterbanks which model the human hearing. In this paper, we simply use a bark scale, which models the critical bandwidth in the cochlea. Bark scale conversion can be applied either in the time (or quefrency) domain or in the frequency domain. We choose a frequency domain, that is the frequencies f_i are now chosen equally spaced on the bark scale. Bark scale to frequency scale conversion is performed by

$$f_i = 600 \sinh(m_i/6), \qquad m_i = 6 \operatorname{arcsinh}(f_i/600) \tag{12}$$

with f_i hertz, m_i in bark.

Figure 9 shows the positive effect of "bark scale sampling". The first and second formant, which would have merged into one peak if sampled linearly on the frequency scale, are clearly distinct if bark scale sampling is applied.

We also found that not using the first, the 15th and the 16th channel gave equal or better results. Since the data has not been high-pass filtered, deleting this first parameter is equivalent to a high-pass filter with 100Hz cutoff frequency. (For linear frequency spacing this equivalent high-pass filter would have a cutoff frequency of about 300Hz, eliminating also information on the first formant.) Not using channels 15 and 16 is

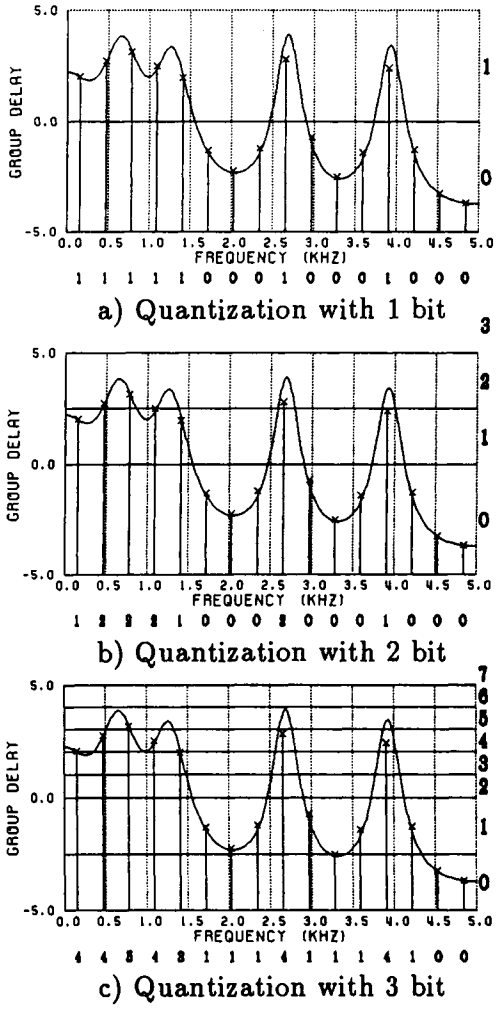


Figure 6. Quantization of $\tau(\omega_i)$. The continuous line represents the SGD spectrum of the vowel /a/ in KITAMI, using only the smoothing of eq.(9); $\tau(\omega)$ was calculated according to eq.(4). The crosses represent the SGDS parameters according to eq.(5). The numbers at the right margin stand for the output of each quantization range, the numbers at the bottom for the actual quantizer output.

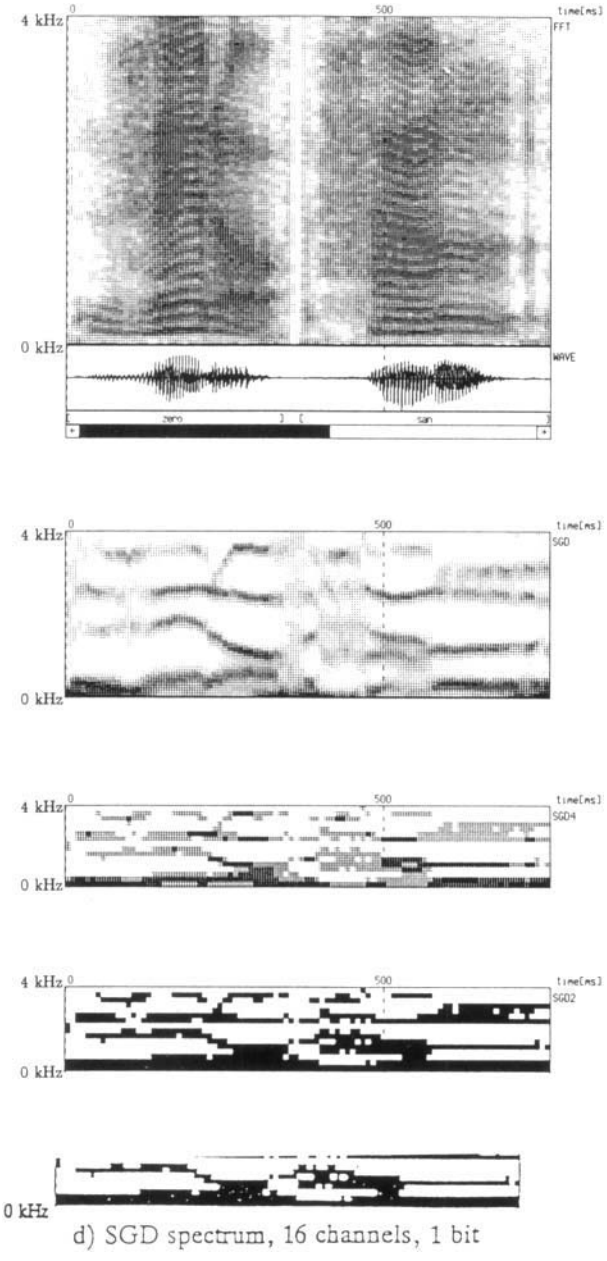


Figure 7. Time frequency pattern of FFT and SGD spectra.

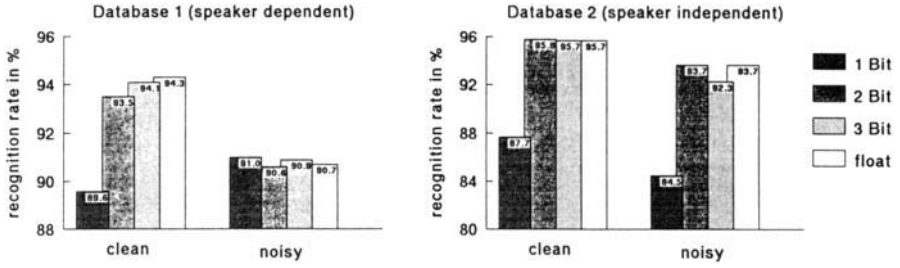


Figure 8. Recognition results for quantized SGDS.

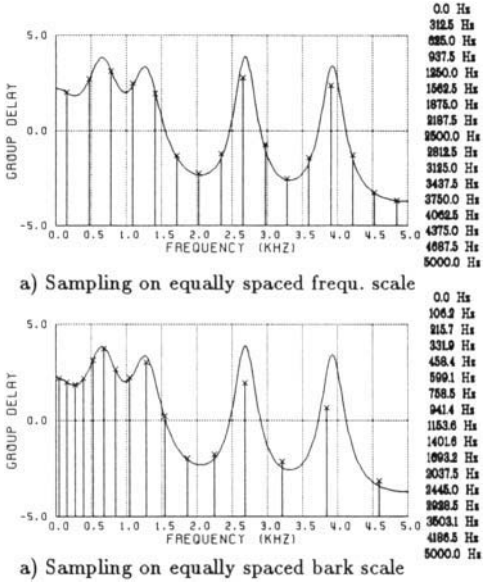


Figure 9. Linear frequency scale versus bark scale.

equivalent to a low-pass filter with cutoff frequency of about 3500Hz, a result that is consistent with section 2.3.

Table 6 compares the results for database 1. For "linear frequency sampling" (linear) only channels 1 to 11 were used, for "bark scale sampling" (bark) channels 2 to 14. The quantization thresholds were chosen according to table 5. The mel Cepstrum was calculated by warping the LPC Cepstrum coefficients according to an allpass filter with warping constant $a = 0.45$ [6].

Tab. 6. Recognition rate in % on database 1 (speaker dependent)

SNR		SGD				LPC-CEP
		float	1bit	2bit	3bit	float
No Noise	bark	97.2	91.9	95.6	96.9	96.0
	linear	96.9	90.3	94.8	95.7	93.8
20 dB	bark	95.9	90.3	93.7	95.4	86.2
	linear	94.0	90.0	91.5	90.9	77.3

4. CONCLUSIONS

Our results showed that (after carefully tuning various parameters, like quantization thresholds, etc.) the recognition rates are better for the bark scale than for the linear frequency scale. We can thus conclude, that frequency domain techniques (the bark scaling, frequency weighting, and low bit quantization) can be easily used with the SGDS. Contrary to the Cepstrum representation, the SGDS can also be easily visualized as a time frequency pattern and therefore easily interpreted.

The low bit representation allows high-speed computation in the pattern matching stage of the recognizer. Furthermore, considerable (1 order of magnitude) saving of memory space can be achieved, allowing an inexpensive implementation of a large vocabulary ASR.

References

1. F.Itakura and T.Umezaki: "Distance measure for speech recognition based on the smoothed group delay spectrum," *Proc.ICASSP*, pp.1257-1260, 1987.

2. T.Umezaki and F.Itakura: "Speech Analysis by Group Delay Spectrum of All-Pole Filters and Its Application to the Spectrum Distance Measure for Speech Recognition," *IECE Trans.*, vol.I72-D-II, No.8, 1989.

3. A. V. Oppenheim and R. W. Schafer: *Digital Signal Processing*, Prentice-Hall, 1975.

4. B.H.Juang: "On the Use of Bandpass Liftering in Speech Recognition," *Proc.ICASSP*, pp.765-768, 1986.

5. T. Umezaki, H. Singer, and F. Itakura: "Using low bit quantization of the smoothed group delay spectrum (SGDS) for speech recognition," *Research Report PASL No.1-1-1*, 1989.

6. T. Kobayashi, S. Kondo, and S. Imai: "Evaluation of Generalized Cepstral Distance Measures for Isolated Word Recognition," *IECE Technical Report*, SP87-18, 1987.

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

A New Method of Speech Analysis — PSE

Takayuki Nakajima*and Torazo Suzuki

Machine Understanding Division, Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba Science City, Japan

Abstract

Giving a new definition of the speech power-spectrum envelope (PSE) characteristics, a PSE analysis method is proposed according to this definition. The application of a 2048 point high-resolution FFT is studied and becomes the basis of the analysis, and is used in all the FFT and IFFT processing.

In the method, first, the fundamental frequency, f_0, is estimated by the Cepstrum technique. The PSE data series are obtained by sampling at intervals f_0 along a linear frequency scale on the short term speech power spectrum. The cosine-series PSE model is proposed for representing the pole-zero structure by choosing the log power spectrum in the linear frequency domain. Parameters in the model are estimated by minimizing the error between the model and the PSE data series. Experimental results on nasal murmur sounds are shown in the form of a comparison with those of the LPC and Cepstrum method. An outline of the PSE analysis-synthesis system is given. The wave-form generation part is based on the superposition of the zero-phase impulse responses, and is able to synthesize a high quality female speech wave-form with very high pitch.

1. INTRODUCTION

Superior speech analysis techniques and descriptions of phoneme characteristics based on such techniques are key elements for future development of speech recognition.

In the field of speech analysis, the following four subjects remain to be studied: (1) the problem of spectral pole-zero estimation (the zero is indispensable for describing the essential characteristics of consonants [1]. (2) the essential relation between the fundamental frequency (pitch) and short-time power spectra, (3) the kinds of physical characteristics that the human speech hearing sense receives, and (4) mechanisms that extract an optional voice from mixed voices or noise (are their any constraints other than semantic or linguistic constraints?).

In this paper, the authors, while remaining aware of the above four subjects, have returned to the basics of speech production, and have tried to develop a speech analysis

*present address: Integrated Media Laboratories Sharp Corporation, 1-9-2 Nakase, Mihama-ku, Chiba-shi, Chiba, 261 Japan

technique which is valid for both vowels and consonants. Here, they propose a power spectrum envelope (PSE) speech analysis technique which is based on a more exact and concrete definition of the power spectrum envelope [2].

With the PSE technique, the fundamental frequency plays a much more active role in speech analysis than it does with existing methods.

2. NEW DEFINITION OF THE SPEECH POWER SPECTRUM ENVELOPE (PSE)

Speech sound is produced by the speech organ (Figure 1(a)).Voiced sounds, seen in terms of their functions, are composed of three types of characteristics: vocal chord wave characteristics, vocal tract characteristics, and radiation characteristics. These frequency characteristics are shown in Figure 1(b).

Vocal chord waves have a periodic wave form (fundamental frequency f_0) which is called a sawtooth waveform. Their overall power spectrum has a smooth spectrum outline. But since their waveform is a periodic one, they have a line spectrum structure which starts with zero frequency and has energy only in a position which is a whole number multiple of the fundamental frequency.

Vocal tract frequency transfer characteristics, which carry the phoneme information, have a continuous spectrum, the same as that of radiation characteristics. When the different types of characteristics are expressed in the form of time wave forms (impulse responses), the speech waves are produced by the convolutions of each impulse response. Therefore, the short-time power spectrum of voiced sounds has a line spectrum structure which starts with zero frequency and has energy only in a position which is a whole number multiple of the pitch.

Until now, the power spectrum envelope has been simply defined as a curved line smoothly connecting the peaks in the fine structure of the power spectrum. But generally, since determining where the peaks are is a problem, we cannot call this definition an exact one. By bearing in mind that, in the case of steady state voiced sounds, reliable data in the frequency area only exist in a position which is a whole number multiple of the fundamental frequency, we can redefine the speech power spectrum envelope as follows: the speech power spectrum envelope is expressed in the form of the PSE model whose parameters are estimated using the PSE data series which is extracted at the whole number multiple on the short-term power spectrum along the power spectrum. The above definition means that only reliable samples should contribute to the results. The object of most of the existing speech analysis methods is to extract what is also called "the power spectrum envelope". But if these methods are compared according to the redefinition above, it is seen that they differ from each other in the results.

3. HIGH PRECISION FFT TECHNIQUE BY THE INTERPOLATION EFFECT

Figure (b) shows an FFT analysis of a standard logarithmic power spectrum, using as an example a male speaker's /e/. The figure shows the results of a 256 point FFT with a

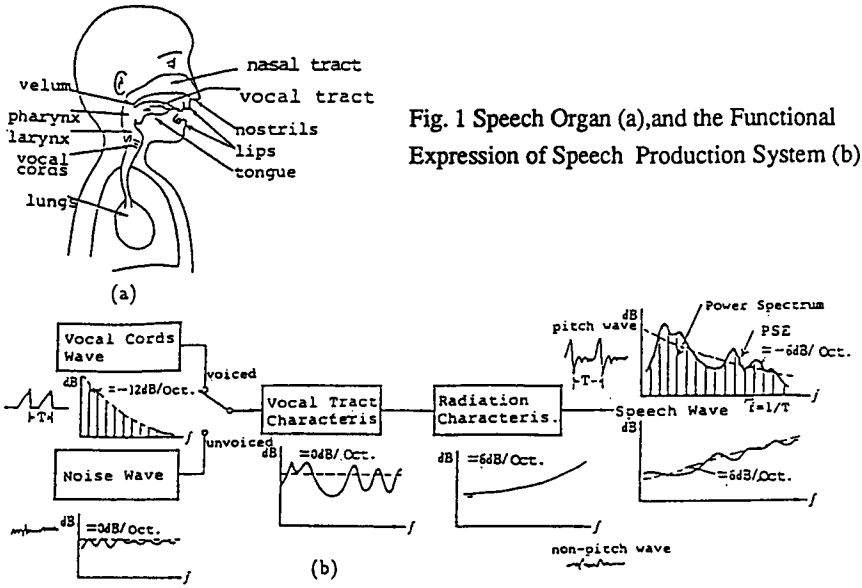


Fig. 1 Speech Organ (a),and the Functional Expression of Speech Production System (b).

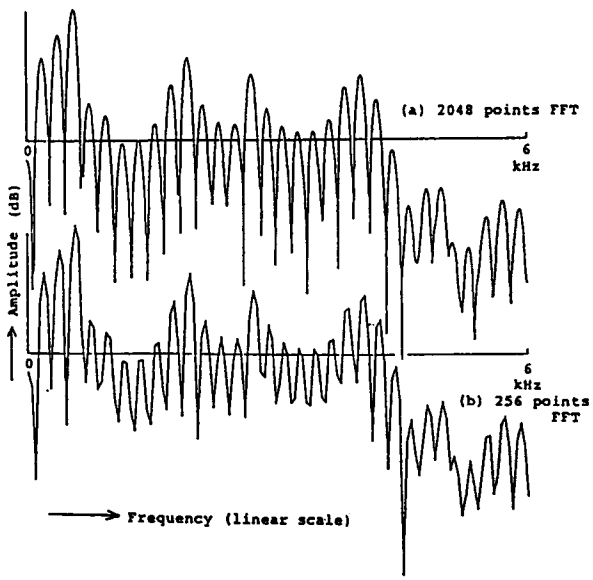


Fig. 2 Comparison of two different resolution Log Power Spectrum.
(male speaker's /e/, sampling: 12 kHz)
(a) Result of 2048 point FFT 256 point speech data cut out using Hanning Window from continuous speech , and 1792 point zero data are connected to the former data .
(Total : 171 msec)
(b) Result of 256 point FFT 256 point speech data cut out using Hanning Window from continuous speech.
(Total: 21.3 msec.)

waveform sampling frequency of 12 kHz and 256 point speech data cut out using a Hanning window (21.3 ms. length). The frequency resolution of the spectrum is (1/0.0213 = 43) Hz. When each frequency interval between two adjacent samples is interpolated, the frequency resolution can be expected to improve. This actually occurs very easily when the true waveform, being cut out using a Hanning window, is continued with zero data, and the total waveform time length is increased. Figure 2(a) shows the result of a 2048 point FFT. In the method, the same wave data as in Figure 2(b) are continued with zero data for increasing the wave length (total 2048 points in about 171 ms.). The frequency resolution of the logarithmic power spectrum is 5.86Hz.

By comparing the two figures, it can be seen that the interpolation effect enables the spectrum's fine structure to express itself extremely clearly. This tells us that it is now possible to give a concrete approach to the newly defined PSE analysis technique.

For this paper, without sacrificing generality, we handled 12 kHz sampled data. For this reason, a 2048 points FFT and IFFT are applied in the whole process, such as the power spectrum conversion from the wave and the Cepstrum conversion from logarithmic power spectrum.

4. SELECTING THE POWER SPECTRUM ENVELOPE SAMPLE SERIES BASED ON FUNDAMENTAL FREQUENCY INTERVAL SAMPLING

A flow diagram of the speech power spectrum envelope (PSE) analysis, proposed in this paper, is shown in the upper half of Figure 3. The upper part of the analysis flow diagram is divided into two parts. The first part starts with speech waveform processing and ends with finding the sampled spectral data series for fitting the following PSE model. The operations included are those numbered (1) to (5) in Figure 3. The second part starts with the sampled spectral data series and ends with the estimated PSE. This operation is numbered (6) in Figure 3.

This section is an explanation of the five operations that make up the first part. Operations (1) and (2) are the operations that extract the logarithmic power spectrum. Then, (3) and (4) are the operations that use the Cepstrum technique to estimate the fundamental frequency. Each operation is explained below in detail.

Operation (1). When the sampling frequency is 12 kHz, 384 speech waveform data points are cut out by getting the inner product with a Hanning window which is 32 ms. long. By padding zeros after three data, 2048 waveform data points $x(n)$ $(n = 1, 2, \ldots, 2048)$ are prepared.

Operation (2). The Fourier transform is applied to the wave form $x(n)$ and the logarithmic power spectrum $H(n)$ $(n = 1, 2, \ldots, 2048)$ is calculated. $H(n)$ is an even function. The frequency range of $H(n)$ is from -6 to 6 kHz.

Operation(3). The purpose of calculating the Cepstrum in the PSE analysis is to extract the fundamental frequency highly accurately. When the frequency is higher than 4kHz, it often happens that the fine structure that appears in the logarithmic

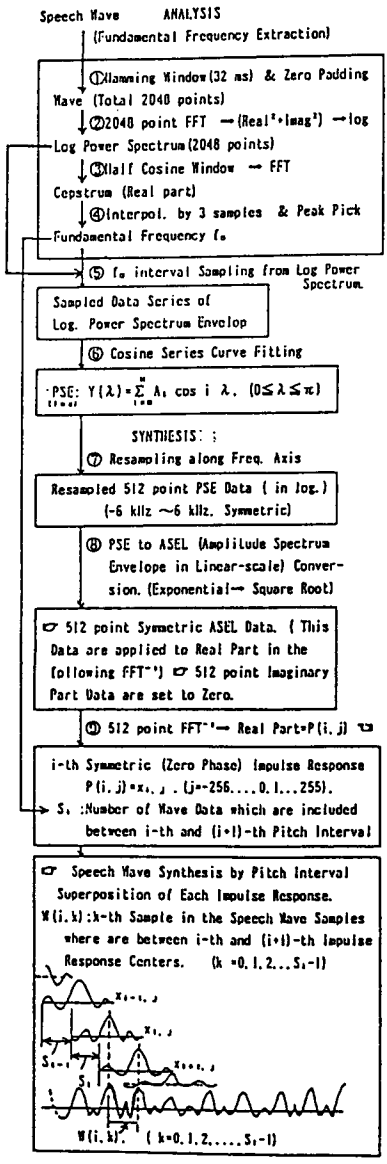


Fig. 3 PSE Analysis-synthesis System Flow Diagram.

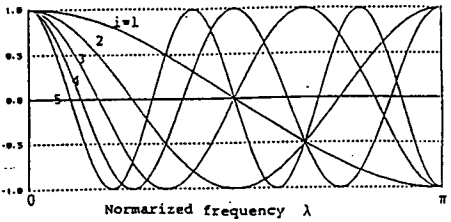


Fig. 4 PSE is modeled by Cosine Series in Log. Power Spectrum Domain.

Each curve shows lower five terms of the cosine series.

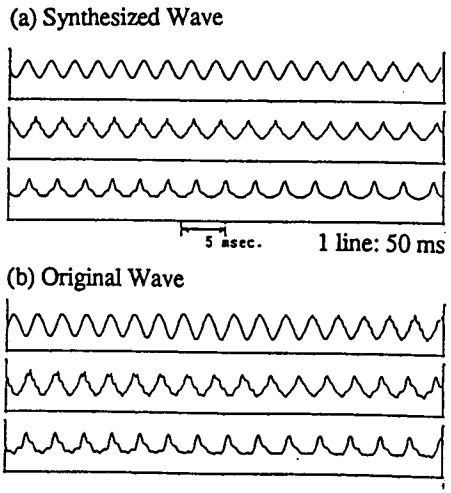


Fig. 5 Comparison between Synthesized Speech Wave-form and Original One.

Sample: Female Voice,Text:/...dono*youni*.../.

power spectrum becomes disturbed. For that reason, the data from this area is removed by introduction of the following window function.

$$w(f) = \begin{cases} \cos(f/4000) + 1 & \text{when } -4000 < f < 4000 \\ 0 & \text{in the rest of the frequency range} \end{cases}$$

The inverse Fourier transform is, therefore, applied to the inner product of $H(n)$ and $w(f)$ to obtain the Cepstrum.

Operation (4). The fundamental interval is given by the frequency value which corresponds to the peak of the Cepstrum [3]. Since the frequency time resolution is equal to the speech wave sampling interval, it is 83 micro second, when the sampling frequency is 12 kHz. The present level of accuracy for simple peak detection, when the fundamental frequency is tentatively 100 Hz, is +0.8% within the range of 99.17 to 100.83 Hz (between 1 / (0.01 +0.000083) and 1 / (0.01-0.000083)).

To improve this level of accuracy in pitch frequency, a total of three samples-the sample giving the maximum Cepstrum value and two neighboring samples in the upper and lower direction along the frequency axis – are used to obtain a more accurate quefrency value, which corresponds to the peak of the quadratic interpolation curve of the three samples. From this the fundamental frequency, f_0, is determined.

Operation (5). The logarithmic power spectrum envelope sample series, y_i ($i = 0, 1, 2, \ldots, N-1$), is obtained by sampling $H(n)$ at the positions of positive integral multiples of the fundamental frequency f_0 in the frequency range from 0Hz to the upper frequency limit F.

5. THE PSE MODEL AND THE ESTIMATION OF ITS PARAMETER

5.1. Making a model of the PSE

The power spectrum envelope characteristics reflect the vocal tract characteristics. Therefore, their poles and zeroes should be expressed with equal weight in the logarithmic domain.

ARMA is known as a pole and zero time domain model. The least-squares law is normally applied in the time domain for the parameter estimation. But some points about the criterion for evaluating the error between the data and the model are still unclear in the frequency domain. Moreover, generally, the problem of estimating ARMA parameters becomes a nonlinear problem requiring recurrent calculations, and causes it to be an unstable solution problem.

In the frequency domain, the speech power spectrum envelope characteristics on a linear frequency scale, $G(z)$, are theoretically expressed by a rational function consisting of a polynomial numerator, $P(z)$, which expresses the zero characteristics, and a polynomial denominator, $Q(z)$, which expresses the pole characteristics. For this reason, the amplitude parameter estimation of each term generally becomes a nonlinear problem. But if we think of the logarithmic power spectrum envelope characteristics, then the mixture

of terms with different error criteria in the numerator and denominator disappears and becomes a homogeneous structure concerning pole and zero, as shown in the following formula:

$$\ln G(z) = \ln P(z) - \ln Q(z) \tag{1}$$

In this formula, the poles and zeros work with the same error criteria. Here, since a logarithmic function is something which can be approximated by means of a finite-term polynomial, $\ln P(z)$ can be represented by a polynomial with a finite terms. Since $\ln Q(z)$ can be expressed in the same way, expressing $\ln G(z)$, the difference between the two, by means of a polynomial with a finite number of terms is effective to the same degree.

In this formula, is particular, we wish to point out that both the pole and zero error are evaluated in the same way when the logarithmic power spectrum envelope is represented by a finite-term polynomial.

Now our object is how to model the vocal tract frequency characteristics within the -4 to 4 kHz range, where the frequency characteristics of the speech wave appear most reliable. Concerning the logarithmic power spectrum envelope characteristics, we can make the general judgment that:

1. These characteristics are an even function;
2. The linear-scale frequency characteristics of the MA process, that express zero, can be expressed by means of a cosine series; and
3. The linear-scale frequency characteristics of the AR process, that express the poles, can be expressed by means of the inverse numbers of the cosine series.

For the above reason, a finite-term cosine series is supposed to be suitable for the logarithmic power spectrum envelope (PSE) model. Figure 4 shows the first five cosine functions in the frequency range 0 to 4kHz.

5.2. Estimating the parameters (Operation (6) in Figure 3)

Let F be the upper frequency limit in the PSE model. However, since F is the integer $N-1$ multiple of f_0, its minimum value is higher than 4000Hz. Now, using the logarithmic power spectrum envelope's sampled data series $y_i, (i = 0, 1, ..., N-1)$ of N (depend on f_0) with bandwidth 0 to F, the amplitude parameter $A_i, (i = 0, 1, ..., M)$ of the M-term cosine series:

$$Y(i) = \sum_{i=0}^{M} A_i \cos i\lambda \quad (0 \leq \lambda \leq \pi), \tag{2}$$

which represents the PSE model, is estimated by means of minimization of the squared error on the logarithmic power scale. Here, the y0 value is not reliable in real speech data, because HPF is normally used in the speech wave acquisition. Therefore, $0.99 \cdot y_1$ is substituted for it as an approximate value.

The sum of the squared error, J, between the data series y_i and the PSE model $Y(i)$ is:

$$J = \sum_{i=0}^{N-1} (Y(\delta i) - y_i)^2 \quad \text{where} \quad \delta = \pi/(N-1).$$

To minimize the sum of the squared error, J, the formulas are partially differentiated to $A_0, A_1, ..., A_M$, and each equation is set to zero. Then a set of simultaneous linear equations of the order $M + 1$ is obtained. By solving this set, $A_0, A_1, ..., A_M$ in eq. (2) are found.

6. OUTLINE OF THE PSE ANALYSIS-SYNTHESIS SYSTEM[2]

PSE analysis is a technique that can extract the pole and zero characteristics in a straightforward way. The speech, which is re-synthesized from the parameters obtained from it, is expected to be of very high quality. The synthesis part of this PSE analysis-synthesis system is shown in the lower half of Figure 3.

In the system, the impulse response is obtained by a 512 point complex IFFT under the assumption of zero phase shift of the impulse response. So that, the real part of the input of the IFFT is the 512 point symmetric re-sampled exponent data of the PSE (linear power scale PSE) in the frequency range from -6000 to 6000 Hz. The imaginary part of the input is set to 0 . The real part of the output of the IFFT is the 512 point symmetric impulse response.

Speech wave synthesis is done by pitch interval superposition of each impulse response. A high quality, clear female voice with high pitch (f_0 is even more than 500 Hz) can be synthesized by the system. A comparison between the synthesized speech waveform and the original one is shown in Figure 5 as an example. This kind of voice has been very hard to synthesize by the existing LPC and other methods, in which the impulse-response phase shift is not definitely controlled. In the older vocal tract filter source model, the wave-form is obtained under the assumption of steady state in spite of the fact that both the vocal tract filter and source pitch parameter values are renewed frame by frame, simultaneously.

7. EXAMPLES OF NATURAL SPEECH ANALYSIS AND COMPARISON WITH OTHER ANALYSIS TECHNIQUES

In order to realize automatic speech recognition of unspecified speakers, many speech analysis techniques used until now have been tested to distinguish sounds in pairs. These tests have shown that there are some phonemes or syllable pairs that are very difficult to tell apart. An example of such a pair are the two nasal syllables /mo/ and /no/. Both the nasal consonant components and the vowel components closely resemble each other. Figure 6 shows examples of analyses using three different techniques, of the nasal consonant component of a male speaker's /mo/ sound, while Figure 7 shows corresponding examples of the same male speaker's /no/ sound. Part (a) of each figure shows the results of analysis using the PSE technique (number of terms: m=26); part (b) shows the results using the Ceptsrum analysis technique (number of terms: 54); and part (c) shows the results using the linear prediction analysis (LPC) technique (number of terms: 18). To make the comparison easy, the analysis windows are arranged identically, and in both figures the displays overlap with the short-time power spectra on the same scale.

According to the acoustic knowledge of speech, the feature that distinguishes the nasal consonants /m/ and /n/ is the difference in the zero frequencies that correspond to the distance between the vellum (branch position to nasal tract) and the lips (in the case of /m/) and between the vellum and the tip of the tongue (in the case of /n/) [1]. PSE analysis, consistent with the theory, shows clearly that the zero frequency for nasal murmur /m/ in /mo/ is 480 Hz and the zero frequency for nasal murmur /n/ in /no/ is 720 Hz.

Also, the vertical lines with equal spacing seen in the results of the PSE analysis represent the fundamental frequency spaces and show that only the power spectrum data at these this positions is sam pled and used to estimate the power spectrum envelope. Although the zero is important, we know from observing the results of the other analysis techniques that it is hard to extract in the LPC. In the Cepstrum analysis shown in Figure 7 (b), the zero is observed, but its value is only half that obtained by PSE analysis, despite the large number of terms value. This is because all the power spectrum data have in the parameter estimation with equal weight.

With the LPC analysis, even though the number of terms used is a little more than normal, an estimation of the zero is not expected.

8. CONCLUSION

It has been pointed out that the fundamental frequency, f_0, of natural speech and the length of the vocal tract are interrelated, and that they work to normalize the difference of vocal tract shapes and contribute to the naturalness of vowels. But we can say that the participation of f_0 in this case is a rather passive factor, which helps us determine a general framework that separates categories such as children, women, and men.

In this paper, we claim, from the purely physical point of view of vocal chord wave periodicity under the steady-state assumption, that the only reliable PSE data is the very small amount of data sampled at the whole number multiple of f_0 on the spectrum data. It is indispensable that information on f_0, since it changes in every period, should be given before the PSE model introduction.

Sounds other than pure voiced sounds also contain periodic components. Fundamental periods are also observed in the plosive wave forms of voiced plosives, of voiced fricatives, of course, and even of unvoiced sounds. When these latter sounds are observed within the Cepstrum domain, the Cepstrum peaks, which correspond to the fundamental periodic intervals, are sometimes much less clear than those of pure voiced sounds, but plural periodic components can often be seen. When this is the case, the maximum peak should be respected. When there are no eminent fundamental periodic components, we propose that the sampled data series be obtained by very short interval sampling of the power spectrum domain. The results become asymptotic to the results by Cepstrum analysis in that case.

As an extension of the present paper, the authors have developed a non-steady-state based speech wave structure model. That led to a the synchronous pitch-pair PSE analysis valid for a non-steady-state speech wave with very rapid spectral transitions [4].

According to speech production theory, the zero of the spectrum is an important feature of consonants, because the vocal tract zero carries the information of the sound source

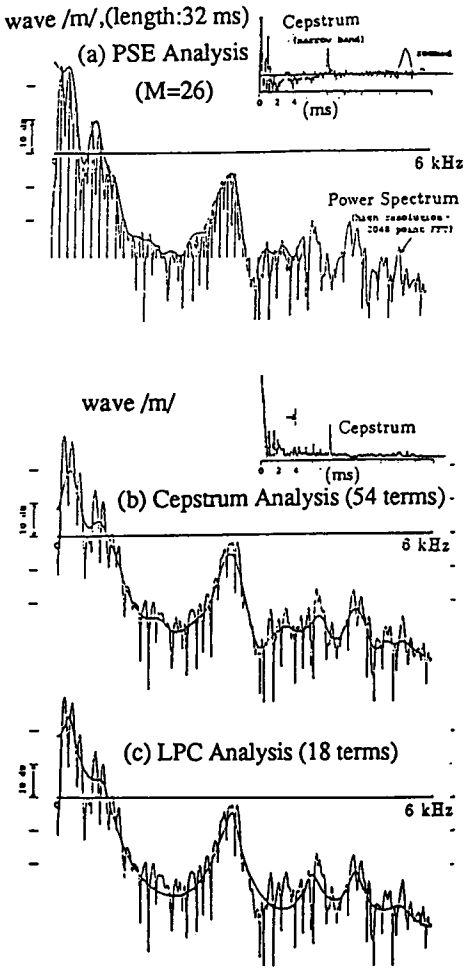


Fig. 6 Comparison of Results from Three Different Speech Analysis to Same Nasal Murmur */mo/*.

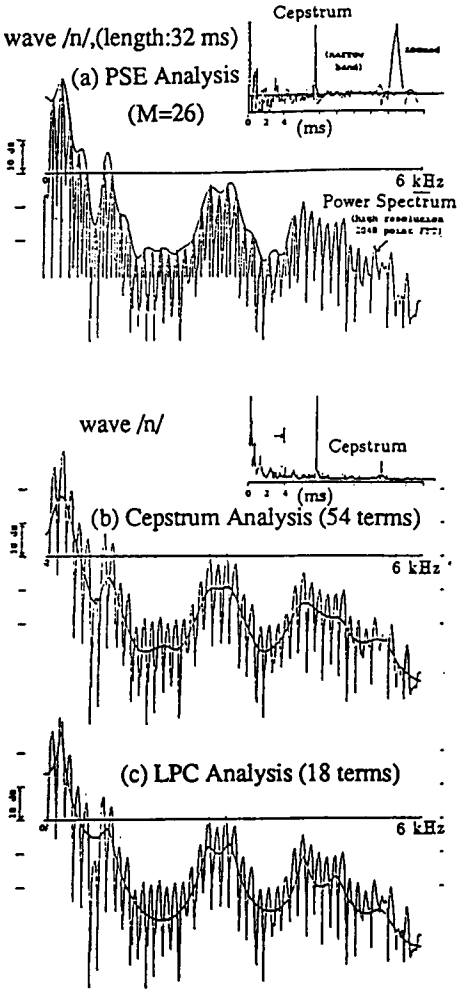


Fig. 7 Comparison of Results from Three Different Speech Analysis to Same Nasal Murmur */no/*.

position in the vocal tract. But it has been rated as relatively unimportant according to the results of speech perception experiments until now. The authors will also give a great deal of attention to the pole-zero structure of phoneme pairs which are difficult to tell apart using the developed analysis techniques, and will try hard to discover the invariant features of such pairs. It is an urgent necessity, they feel, to have phoneme feature descriptions that can deal more accurately and completely with the speech of unspecified speakers.

References

1. G. Fant: "Acoustic theory of speech production," p.191, p.139, 1970.

2. T. Nakajima and T. Suzuki: "Power spectrum envelope (PSE) speech analysis-synthesis system," J. Acous. Soc. Jpn., 44, pp.824-832, 1988.

3. A. V. Oppenheim: "Speech analysis-synthesis based homomorphic filtering," JASA, 45, No.2, pp.458-464, 1969.

4. T. Nakajima and T. Suzuki: "Pitch pair synchronous PSE analysis based on a non-steady state wave spectral model," J. Acous. Soc. Jpn., 44, pp.900-908, 1988.

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Estimation of Voice Source and Vocal Tract Parameters Based on ARMA Analysis and a Model for the Glottal Source Waveform

Hiroya Fujisaki and Mats Ljungqvist

Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo, Japan

Abstract

Conventional speech analysis methods based on linear prediction often fail to separate and estimate the source and vocal tract characteristics, especially in the case of voiced sounds, because of oversimplified assumptions regarding the voice source. We have already proposed a model that is capable of expressing a wide range of voice source characteristics, and demonstrated that source and vocal-tract parameters can be well separated and correctly estimated, for vowels and vowel-like sounds, by combining the proposed source model with linear predictive analysis. The present paper extends our approach to apply to a wider variety of speech sounds including nasal vowels and nasal consonants, by combining the proposed source model with ARMA analysis. The validity of the system was demonstrated by analysis of synthetic and natural speech.

1. INTRODUCTION

Nearly all speech analysis methods in practical use today are based on the linear source-filter model of speech production, which states that voice source, vocal tract, and radiation can be modeled linearly and non-interactively. Conventional speech analysis methods model the combined effects of these three factors in one filter, while assuming white noise excitation. This approach results in simple calculations and has, for the all-pole case, gained a wide acceptance in the form of linear predictive analysis [1, 2]. The approach has also been extended to apply to a wider variety of speech sounds by the introduction of pole-zero modeling. There are, however, inherent weaknesses in this approach. A major inconsistency is that, while white noise excitation is assumed in the analysis, voiced speech is periodic. Consequently, impulse or pulse-like excitation has to be used in the synthesis. This is often a source of error in the estimation of the formant frequencies and bandwidths, since the estimated spectral envelope will describe not only the vocal-tract transfer function, but will also contain voice source information. Furthermore, the voice-source function, which often has rather complicated spectral characteristics, cannot be separated easily from a combined source-tract-radiation filter. The influence of the voice source on spectral envelope estimation can be summarized as follows:

(1) the estimation is affected by the location of harmonic peaks of the source spectrum,

(2) the spectral characteristics of the glottal pulse is included in the estimated envelope,

(3) zeros pertaining to the voice source may be mistakenly interpreted as vocal tract-zeros (a consequence of (2)).

All these undesirable effects stem from the crude modeling of the voice source in conventional speech analysis methods. From this point of view, we have already proposed explicit modeling of the voice source in combination with conventional all-pole vocal-tract modeling, as a means of separating the voice source from the vocal-tract transfer function, we call this GAR (Glottal AR) modeling [3, 4]. In the present paper, a straight forward extension of the method is proposed to allow for pole-zero modeling of the vocal-tract transfer function (henceforth GARMA modeling).

2. ANALYSIS METHOD

Conventional methods for ARMA analysis of speech generally operate on an estimate of the vocal-tract impulse response, thus assuming impulse excitation of the vocal tract. As a consequence, the glottal source characteristics is incorporated in the "vocal-tract" impulse response. The poles and zeros of the combined impulse response are then estimated sequentially [5], or simultaneously [6] by iterative methods. In the present method, on the other hand, the glottal source and the vocal tract are represented by separate models. A parametric representation of the glottal-source signal is used in the ARMA estimation of the true vocal-tract characteristics.

2.1. Least-Squares Identification of a One-input, One-output Linear Model

A linear, time-invariant, discrete-time model with one input and one output can be represented, in the time domain by

$$s_n + \sum_{i=1}^{p} a_i s_{n-i} = \sum_{j=0}^{q} b_j g_{n-j}, \tag{1}$$

where g is the input (glottal source), s is the output (speech) and p and q are the pole and zero orders of the model [7]. Taking the z-transform of eq. (1) gives

$$A(z)S(z) = B(z)G(z), \tag{2}$$

Though it may look most natural to use the error criterion $V = \sum e_n^2$, where the z-transform of e_n is

$$E_1(z) = S(z) - \frac{B(z)}{A(z)}G(z), \tag{3}$$

this leads to a non-linear minimization problem. Instead we use a modified criterion:

$$E_2(z) = A(z)S(z) - B(z)G(z), \tag{4}$$

for which the error e_n is linear in the parameters a_i and b_j. The computation of this error measure is illustrated in Figure 1. Two special cases can be observed from eq.(4). By setting $G(z) = 0$ the equation reduces to conventional LPC, and by setting $B(z) = 1$ we obtain the previously proposed GAR model. The minimum error is found in the conventional way by setting grad$(V) = 0$, and by solving the following system of linear equations for the parameters a_i and b_j:

$$\begin{bmatrix} S_{i,j} & -X_{i,j} \\ -X_{j,i} & G_{i,j} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -S_{0,j} \\ X_{0,j} \end{bmatrix} \tag{5}$$

where,

$$\begin{aligned} S_{i,j} &= \sum_{n=0}^{N-1} s_{n-i}s_{n-j}, \quad 1 \le i,j \le p, \\ X_{i,j} &= \sum_{n=0}^{N-1} s_{n-i}g_{n-j}, \quad \begin{cases} 1 \le i \le p, \\ 0 \le j \le q, \end{cases} \\ G_{i,j} &= \sum_{n=0}^{N-1} g_{n-i}g_{n-j}, \quad 0 \le i,j \le q, \\ a &= [a_1, a_2, \ldots, a_p]^T, \\ b &= [b_0, b_1, \ldots, b_q]^T. \end{aligned}$$

The matrix is symmetric and can be efficiently inverted by standard methods such as the Cholesky decomposition method.

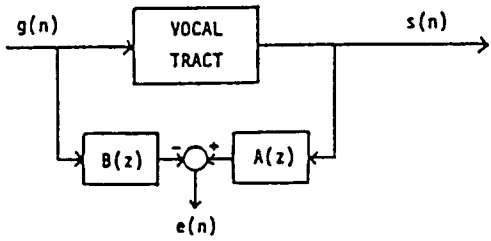
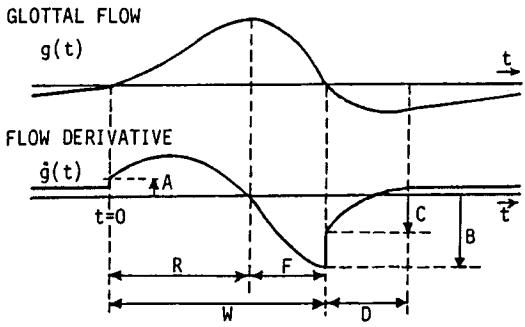


Figure 1. Computation of the error measure used in the analysis.

2.2. Modeling of the Voice Source

The glottal-voice source consists of quasi-periodic pulses which are created by the vibrating vocal cords as air is pressed up from the lungs. It is desirable, from both the theoretical and the practical point of view, to obtain a parametric representation, i.e. a model, of the glottal source. Models for the glottal waveform, mostly defined in the time domain, have been proposed by numerous authors. These models emphasize, in one way or another, certain aspects of the voice source. Our proposed model incorporates important features of almost all the previous models as well as additional features [4], and can thus be applied to various types of voices and situations. The model, illustrated



$$\dot{g}(t) = \begin{cases} A - \frac{2A + R\alpha}{R}\, t + \frac{A + R\alpha}{R^2}\, t^2, & 0 < t \le R, \\ \alpha(t-R) + \frac{3B - 2F\alpha}{F^2}(t-R)^2 - \frac{2B - F\alpha}{F^3}(t-R)^3, & R < t \le W, \\ C - \frac{2(C-\beta)}{D}(t-W) + \frac{C-\beta}{D^2}(t-W)^2, & W < t \le W + D, \\ \beta, & W + D < t \le T, \end{cases}$$

where $\alpha = \frac{4AR + 6FB}{2R^2 - F^2}$ *and* $\beta = \frac{CD}{D - 3(T - W)}$,

T : *Fundamental Period.*

Figure 2. Waveforms and formulas for the glottal model. The differentiated glottal flow $\dot{g}(t)$ includes the radiation factor.

by Figure 2, is composed of polynomial segments and can be fully described by the following six parameters: the open phase duration (W), the pulse skewness (S), the interval from glottal closure to maximum negative flow (D), the slope at glottal opening (A), the slope immediately before glottal closure (B), the slope immediately after closure (C). The notation $g(t)$ in Figure 2 indicates the glottal flow derivative, i.e., the voice source model combined with the radiation characteristics, which can be modeled with good accuracy by a +6dB/octave spectral slope.

2.3. Joint Estimation of the Voice-Source and Vocal-Tract Parameters

In Section 2.1 it was shown that, with a suitable choice of error criterion, the AR and MA parameters can be simultaneously estimated in a linear procedure, provided that both the input g and the output s of the model are known. The input signal is obviously not known in advance. Therefore we use the previously described parametric waveform model as input, and employ iterative search for the model parameters which give the best description of the input signal in terms of the minimum prediction error as given by eq. (4). The procedure for estimating the glottal and vocal-tract parameters is based on Analysis-by-Synthesis [8] and consists of two types of error minimization: (a) local

minimization of V based on a linear, least-squares estimation of the AR parameters p_i and the MA parameters b_j, under the assumption of a known input signal $g(t)$, and (b) global minimization of V based on an iterative search for the optimum parameters of $g(t)$. In each iteration of (b), the local minimization (a) is also carried out. The overall procedure, outlined in Figure 3, can be summarized as follows:

(1) generation of the voice source signal,

(2) estimation of the vocal tract transfer function using the generated source signal and the speech signal,

(3) evaluation of the prediction error. The sequence (1)-(3) is iterated as the glottal parameter space is searched for a parameter combination that gives the global minimum prediction error.

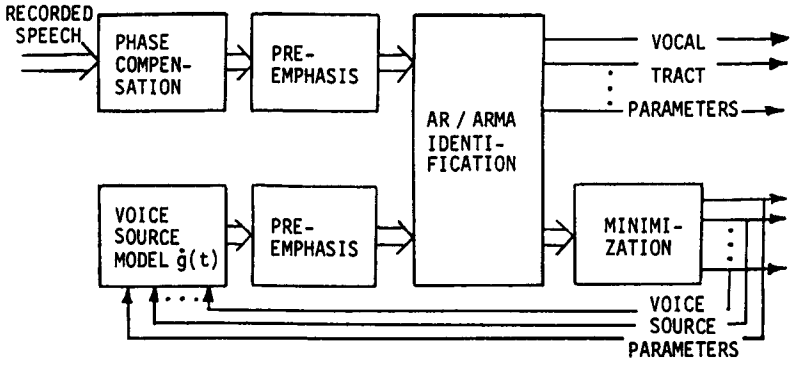


Figure 3. Block diagram of the parameter estimation process.

Among many possible methods for non-linear minimization, we use, in this study, a simple relaxation method, which essentially examines one dimension at a time with a successively decreasing step-size. The analysis is carried out pitch-synchronously with adaptable frame length and frame step. Since the minimization is sensitive to the positioning of the glottal model pulse it was found necessary to provide for pulse positioning with an accuracy of 1/10 of the sampling period (0.01 ms at 10kHz sampling frequency). To achieve pre-whitening of the inputs to the linear estimation procedure, both the speech signal and the voice source signal ($g(t)$) are preemphasized with the factor $(1-\mu z^{-1})$ where μ is close to unity.

In studies of the detailed waveform of the glottal source, it is important to assure freedom from the serious low frequency phase distortion that ordinary speech recordings often are subject to. We assume the phase distortion to be time invariant and carry out compensation, prior to the analysis, by filtering the time-reversed speech signal [9]. In the case of time-varying distortion it may instead be desirable to insert a simulation filter after the voice source model in the iteration loop. Such an arrangement does not require time-reversal of the signal.

3. ANALYSIS OF NASALIZED SPEECH SOUNDS

Earlier studies have shown that nasalized vowels often can be distinguished from their oral counterparts by the presence of pole-zero pairs in the spectrum. One pole-zero pair is usually prominent and located between the second and third formant, while a second pair appears in the low frequency region [10].

3.1. Synthetic Speech

Nasal vowels were synthesized using a terminal analog synthesizer with an excitation signal of the type shown in Figure 2, using five complex poles, one complex pole-zero pair for the nasal coupling, and a single differentiation $(1 - Z^{-1})$ to simulate the radiation factor. Four systems were compared:

(1) ARMA model with glottal waveform model (GARMA),

(2) ARMA model with impulse excitation (ARMA),

(3) AR model with glottal waveform model (GAR),

(4) AR model with impulse excitation (AR).

Table 1. Original and Estimated Values of Frequencies and Bandwidths of Poles (and Zeros) for a Synthetic Nasal Vowel /ẽ/.

	FREQUENCIES/BANDWIDTHS (Hz)							NORMALIZED
	POLES				ZEROS			ERROR
AR	662 / 121	1647 / 980	1709 / 132	2849 / 342	3551 / 194	4520 / 145		0.11
GAR	681 / 83	1695 / 123	1757 / 971	2851 / 378	3551 / 197	4520 / 144		0.07
ARMA	661 / 100	1579 / 287	1783 / 140	2753 / 219	3501 / 182	4501 / 209	1998 / 751	0.05
GARMA	689 / 72	1636 / 95	1942 / 96	2766 / 138	3502 / 164	4501 / 206	2256 / 260	0.001
ORIGINAL	690 / 70	1640 / 100	1940 / 110	2760 / 130	3500 / 160	4500 / 200	2260 / 250	

The systems (ARMA) and (AR) are evaluated using the same method as the (GARMA) and (GAR) systems but with the glottal model replaced by an impulse model. In the GARMA and ARMA systems the pole and zero orders were $p = l2$ and $q = 2$, while in the GAR and AR systems they were $p = 12$ and $q = 0$. Figure 4 shows the original and estimated spectral envelopes from analysis of the synthetic nasal vowel /ẽ/. It can be seen from the figure that the ARMA model without an explicit voice source model, while performing better than the all-pole systems, is not capable of accurately representing the

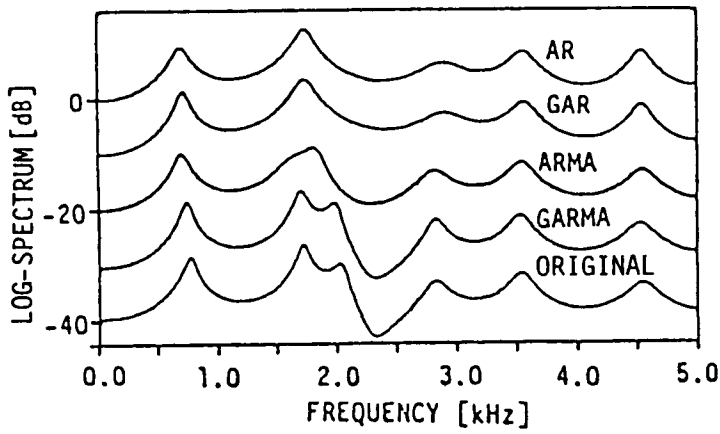


Figure 4. Original and estimated spectral envelopes for a synthetic nasal vowel /ɛ̃/ using several AR and ARMA analysis methods.

original vocal-tract transfer function. Table 1 lists the values of the frequencies and bandwidths of the poles (and zeros) for the synthetic speech signal as well as their estimated values. The normalized prediction errors are also given in the table.

3.2. Natural Speech

Experiments were also carried out on natural nasal vowels. Figure 5 shows spectral envelopes obtained by analysis, using several analysis methods, of the French nasal vowel /æ̃/ uttered by a male French speaker. For comparison, also the GARMA spectral envelope, including the spectral characteristics of the estimated voice source and radiation models, is shown. It can be seen that this spectral envelope, which represents the total system, agrees well with the FFT spectrum of the same speech segment. In Figure 6, waveforms obtained from the analysis of Figure 5 are shown. The figure shows that the methods that employ a model for the voice source (GAR and GARMA) give smaller prediction errors than the other methods.

4. CONCLUSIONS

The present study has shown the feasibility of combining voice source estimation with vocal-tract estimation in analysis of voiced speech. Especially, it was shown that our previously proposed GAR scheme for combining voice source modeling with linear predictive analysis can be generalized in a straightforward way to include pole-zero modeling of the vocal tract transfer function. The proposed system allows for simultaneous estimation of the voice-source and the vocal-tract ARMA parameters, based on iterative minimization of the mean-squared error. It is thus possible to separate the voice source from the vocal tract transfer function, which not only gives improved estimation of the vocal-tract

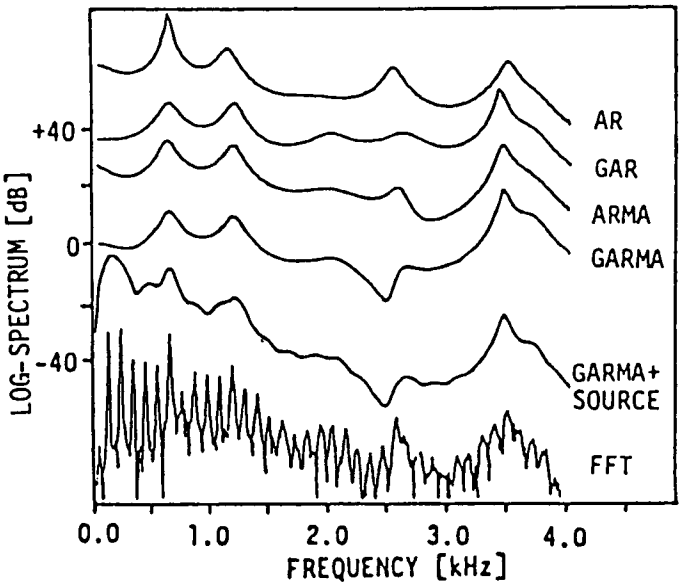


Figure 5. Estimated spectral envelopes for a natural nasal vowel using several AR and ARMA analysis methods (vowel /œ̃/, male voice).

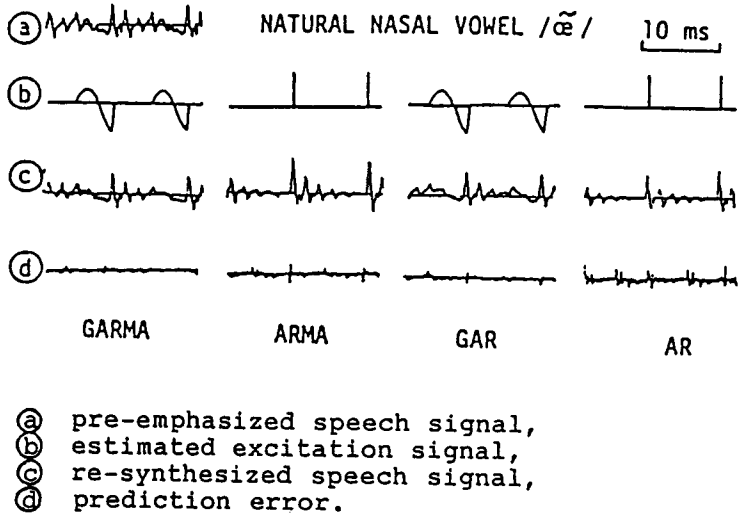


Figure 6. Results from analysis of the same vowel as in the previous figure using several AR and ARMA analysis methods.

transfer function, but also allows for a parametric representation of the glottal-source signal.

Experiments on synthetic speech confirmed the validity of the method, and indicated improved performance over both ARMA analysis with an impulse model for the source and AR analysis, as shown by the better spectral fit and the smaller prediction error. The results from analysis of natural speech further confirmed these observations.

References

1. B. S. Atal and M. R. Schroeder: "Predictive Coding of Speech Signals," *Reports of 6th Int. Congr. Acoust.*, C-5-4, 1968.

2. F. Itakura and S. Saito: "Analysis Synthesis Telephony Based Upon the Maximum Likelihood Method," *Reports of 6th Int. Congr. Acoust.*, C-5-5, 1968.

3. M. Ljungqvist and H. Fujisaki: "A Method for Simultaneous Estimation of Voice Source and Vocal Tract Parameters Based on Linear Predictive Analysis," *Trans. Committee on Speech Research, Acoust. Soc. Japan*, No. S85-21, 1985.

4. H. Fujisaki and M. Ljungqvist: "Proposal and Evaluation of Models for the Glottal Source Waveform," *Proc. ICASSP*, 31.2, 1986.

5. G. E. Kopec, A. V. Oppenheim and J. M. Tribolet: "Speech Analysis by Homomorphic Prediction," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-25, pp. 40-49, 1977.

6. H. Morikawa and H. Fujisaki: "Adaptive Analysis of Speech Based on a Pole-Zero Representation," *IEEE Trans. Acoust. Speech, and Signal Processing*, vol. ASSP-30, pp. 77-88, 1982.

7. K. J. Astrom and P. Eykhoff: "System Identification - A Survey," *Automatics*, vol. 7, pp. 123-162, 1971.

8. C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens and A. S. House: "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," *J.Acoust. Soc. Am.*, vol. 33, No. 12, 1961.

9. M. Ljungqvist and H. Fujisaki: "Correction of Low Frequency Distortion in Speech Recordings and its Effect on the Glottal Wave Shape," *Proc. Spring Meeting of Acoust. Soc. Japan*, pp. 161-162, 1985.

10. E. Bognar and H. Fujisaki: "Analysis, Synthesis and Perception of the French Nasal Vowels," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 31.1, 1986.

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Estimation of Sound Pressure Distribution Characteristics in the Vocal Tract

Nobuhiro Miki* and Kunitoshi Motoki**

*Research Institute for Electronic Science, Hokkaido University, Sapporo, Japan
**Faculty of Engineering, Hokkai-Gakuen University, Sapporo, Japan

Abstract

In this report the characteristics of the sound pressure distribution in the vocal tract are described. Since it is very difficult to directly measure the sound pressure distribution in the real vocal tract, we made plaster replicas which duplicate the actual shapes of the oral cavities, and measured the complex sound pressure distribution. Vectorial maps of active and reactive sound intensity are shown. From the vectorial maps we can see the equi-phase and the equi-amplitude lines in the replicas. From the experimental results, it is shown that the wave front at high frequencies forms a quite complicated curve like a circular pattern, and the assumption of plane wave propagation is not always valid even at low frequencies especially around the tongue tip.

1. INTRODUCTION

The characteristics of wave propagation in the vocal tract have been assumed to be those of plane waves because the wavelength of sound in the frequency range of speech is relatively long compared with the cross-section size of the vocal tract. Based on this assumption, a model of the vocal tract is constructed as a cascade of connected uniform tubes having cross-sectional areas corresponding to the predetermined vocal tract area function. The vocal tract area function has been estimated from the results of analysis of real speech or direct acoustic measurement. Plane-wave propagation is implicitly assumed in these estimation algorithms.

The distribution of sound pressure in the vocal tract model has been studied [1]. The real vocal tract, however, is regarded as a nonuniform acoustic circuit in which local reflection of sound may occur at any point. Thus we can imagine that the distribution of the sound pressure in the real vocal tract is different from that of a vocal tract model with plane-wave propagation.

In this report we show the experimental results of the measurement of the complex sound pressure distribution in plaster replicas of the oral cavities. Complex sound pressure distribution means the spatial distribution of the amplitude and phase of the sound pressure for a pure tone. The replicas were made using impression material to copy the

actual shapes of the oral cavities. Two-dimensional measurement was performed in the vertical and horizontal planes in the replicas.

By using the spatial distribution of the complex sound pressure, we can calculate the particle velocity and then obtain complex sound intensity vectors at each measuring point. The intensity vectors are composed of two parts. One is called the active intensity and the other is the reactive intensity. From the vectorial maps of the intensity we can see equi-phase lines, namely, wave fronts, and equi-amplitude lines. And if the rotation of the vectorial field of active intensity is zero, which implies that the particle trajectories are straight, the route of the acoustic power flow may be estimated as any continuous line tangential to the active intensity vectors. For these advantages and clear visualization, the measurement results are mainly shown in the form of vectorial maps of the complex intensity.

2. MEASUREMENT OF THE COMPLEX SOUND PRESSURE DISTRIBUTION

2.1. Calculation of the complex sound intensity from the measured complex sound pressure distribution

In this section we describe the definition of the complex sound intensity and its properties briefly. By omitting the time factor, the complex sound pressure $p(r)$ (r is a position vector) in the replica can be represented as

$$p(r) = P(r)e^{j\phi(r)}, \tag{1}$$

where $P(r)$ and $\phi(r)$ are the spatial distribution of the amplitude and the phase of the sound pressure, respectively, and should be measured at many points in the replica. The particle velocity, $\nu(r)$, is related to the complex sound pressure by Euler's equation:

$$\nabla p(r) + j\omega\rho\nu(r) = 0, \tag{2}$$

where ω is the angular frequency and ρ is the air density. Using the above relation, $\nu(r)$ is obtained as follows:

$$\nu(r) = V_x e^{j\theta_x}\mathbf{i}_x + V_y e^{j\theta_y}\mathbf{i}_y + V_z e^{j\theta_z}\mathbf{i}_z = \{j\nabla P(r) - P(r)\nabla\phi(r)\}e^{j\phi(r)}/\omega\rho, \tag{3}$$

where V_k, θ_k, i_k ($k = x, y, z$) are the amplitude, phase, and unit vectors for each direction, respectively. The complex sound intensity, $C(r)$, and active and reactive intensity, $I(r)$ and $Q(r)$ are defined by

$$C(r) = p(r)v^*(r)/2, \tag{4}$$

$$I(r) = \mathrm{Re}\{C(r)\}, \tag{5}$$

$$Q(r) = \mathrm{Im}\{C(r)\}. \tag{6}$$

* denotes taking the conjugate of the complex value. From eqs. (1), (3), and (4), $I(r)$ and $Q(r)$ are written as

$$I(r) = -P^2(r)\nabla\phi(r)/2\omega\rho, \tag{7}$$

$$Q(r) = -\nabla P^2(r)/4\omega\rho. \tag{8}$$

From eqs. (7) and (8), it is seen that the vectors $I(r)$ and $Q(r)$ are normal to the equi-phase and equi-amplitude lines, respectively. If the sound field can be assumed to be one-dimensional, $I(r)$ and $Q(r)$ should have the same tangential direction. The vectorial properties of $I(r)$ and $Q(r)$ can be examined by calculating their rotation and divergence [2, 3]. Particularly, the rotation of the active intensity is calculated as,

$$\begin{aligned} \nabla \times I(r) &= \{I(r) \times Q(r)\} 4\omega\rho / P^2(r) \\ &= \{V_y V_z \sin(\theta_y - \theta_z)\mathbf{i_x} + V_z V_x \sin(\theta_z - \theta_x)\mathbf{i_y} + V_x V_y \sin(\theta_x - \theta_y)\mathbf{i_z}\}\omega\rho. \quad (9) \end{aligned}$$

If the rotation of the active intensity is zero, the phases of the velocity component to each direction are the same, and thus the particle trajectories are straight. When this condition holds, the direction of power flow at each point can be regarded as tangential to the active intensity vector.

2.2. Experimental setup

Figure 1 shows a block diagram of the measurement system. The replica is placed in a plane baffle. A uniform acoustic tube (area 5.5 cm^2) is connected to the replica and a speaker is attached at the end of the uniform tube. The speaker is driven by a pure tone from an analyzer. The complex sound pressure in the replica is picked up by a condenser microphone, Mic. #2, which is attached to a pole on a movable XYZ-stage. A probe tube for Mic. #2 is made of glass, with a diameter of 3.0 mm and a length of about 40 cm. The front position of the end of the probe is measured on a vernier scale labeled on the XYZ-stage. A signal from Mic. #1, whose location is fixed at 100 mm distant from the replica-side end of the uniform tube, is used as a reference signal; the amplitude and phase in the replica are measured relative to the reference signal. Typical errors of the analyzer are 0.03 dB and 0.05°. The analyzer is controlled by a minicomputer. It might be possible to use the output signal from the analyzer as the reference signal. In that case, however, if the measurement circumstances, such as room temperature, slightly change between the start time and the end time of the measurement, the phase distribution is influenced by the small change in sound speed. This influence becomes greater as the length of the uniform tube becomes longer. Using the signal picked up near the replica, we can diminish this influence.

2.3. Replicas of the oral cavities

Since it is very difficult to directly measure the complex sound pressure distribution in a real vocal tract, we made replicas duplicating the actual shapes of the oral cavities. Alginate impression material was used to obtain molds of the oral cavities and the replicas were formed with plaster. The impression material has sufficient fluidity and is gelatinized within 2 min. Thus, it is not hard for subjects to keep the same articulatory position. The replicas were made for two male subjects (called replica (I) and replica (II)) with articulation of /a/. For the facilitation of measurement, the articulation is somewhat more enhanced than usual by lowering the mandible. The vertical distribution was measured for replica (I) and both the vertical and horizontal distributions were measured for replica

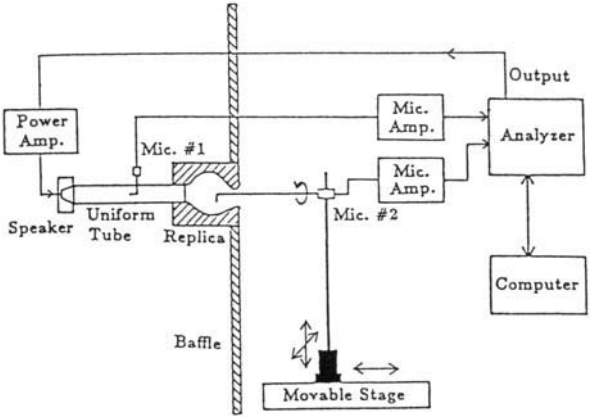


Figure 1. Block diagram of the measurement system.

(II). The anterior-posterior spacing of the measurement positions is 2.5 mm for replica (I), 3.0 mm for replica (II), and vertical and horizontal spacings are 5.0 mm.

3. EXPERIMENTAL RESULTS AND DISCUSSION

Figure 2(a) shows the equi-amplitude and equi-phase lines in the vertical plane for replica (I). The measurement frequencies are 1.5 and 3.5 kHz. The values of the amplitude and the phase are relative to those of the farthest point from the lips. The interval between the solid lines for the phase is 4° for 1.5 kHz and 20° for 3.5 kHz. The interval for the amplitude is 1.0 dB for both frequencies. Dashed lines represent half values of the interval. At the frequency of 1.5 kHz, although each equi-line is not exactly straight, we can consider with rough evaluation that both the equi-amplitude and equi-phase lines are aligned along a vocal tract axis which may be determined by sight from the replica shape. Therefore, the wave of this frequency may justifiably be assumed to be a plane wave. In the figure for 3.5 kHz, the equi-phase lines are deformed much more; but again with rough evaluation, the phase characteristic of the wave may be assumed to be that of a plane wave to some extent. The equi-amplitude lines, however, show that there exists a significant pressure gradient in the vertical direction at the region from the tongue tip to the maxilla. The vertical distance is about 4 cm and is less than half the wave length. In this region the amplitude distribution along the anterior-posterior direction includes local maxima, and the vertical difference of these maxima amounts to 3.9 dB. As the plane wave characteristics should be evaluated from both the amplitude and phase characteristics, it is difficult to recognize plane wave propagation in this region at 3.5 kHz.

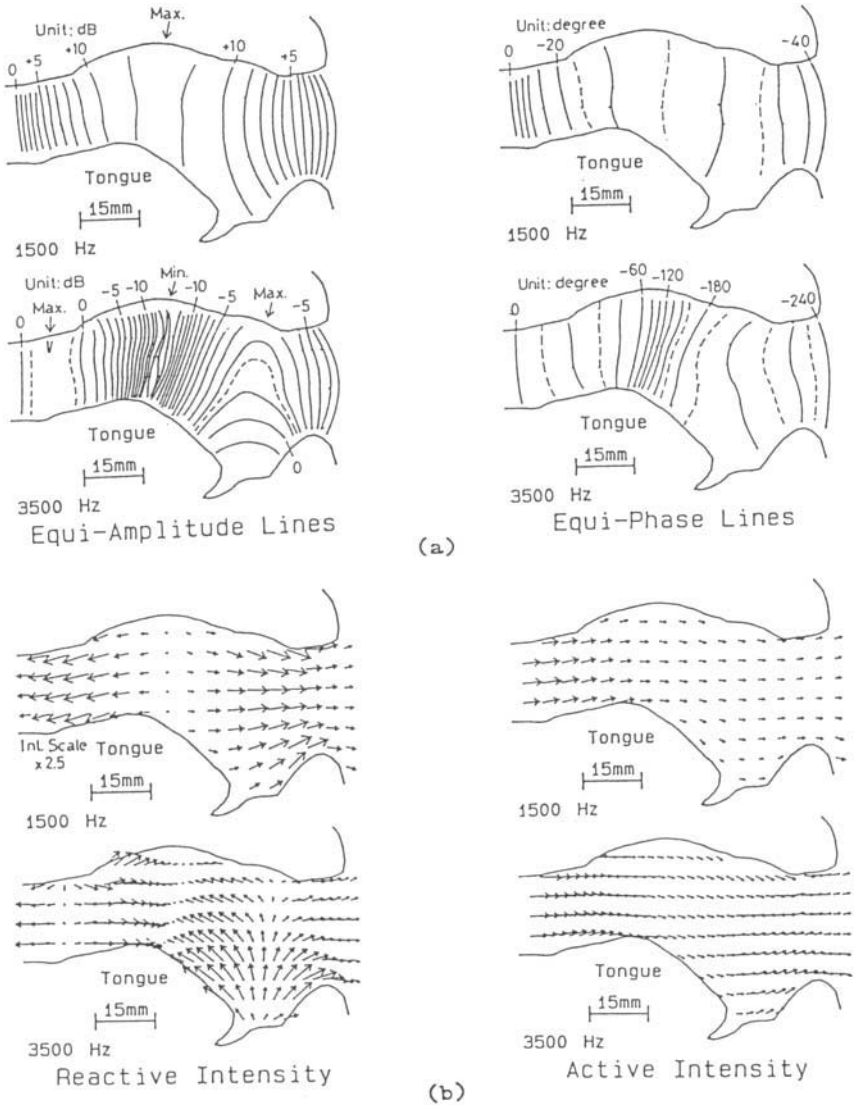


Figure 2. (a) Complex sound pressure distribution. (b) Vectorial intensity maps.

Figure 2(b) shows the maps of the active and reactive intensity vectors corresponding to Figure 2(a). The vectors for 1.5 kHz are plotted with an anterior-posterior spacing of 5.0 mm. The starting point of each arrow is the measurement position and the length of the arrow is proportional to the linear magnitude relative to the active intensity at the farthest point from the lips. The scaling factor of the reactive intensity relative to the active intensity is shown by the number preceded by "Int. Scale" if it is not 1. As mentioned in section 2.1, the active and reactive intensity vectors are normal to the equi-phase and equi-amplitude lines. Comparing Figure 2(b) with (a), we can clearly see the characteristics of the sound field above described, except that the amplitude and the phase at each point are not known directly. In the following discussion, the vectorial intensity maps are used to show the characteristics of the sound field. The amplitude is described in the text if needed.

Figure 3 shows the results for replica(I) with frequencies of 2.0, 4.0 and 6.0 kHz. In the maps for 4.0 kHz, the active intensity is quite uniform; but similar to the result at 3.5 kHz, a vertical gradient in the amplitude is seen in the region of the front of the tongue tip. The pressure gradient is also seen in the result for 2.0 kHz. At the frequency of 2.0 kHz, the vertical distance is less than a quarter of the wave length. It is usually assumed that the wave in a nonuniform acoustic tube is a plane wave if the diameter of the tube is shorter than half the wave length. This assumption is quite valid for the phase characteristics. As for the amplitude characteristics, however, this is not always valid, as seen in these experimental results. In the maps for 6.0 kHz, the active intensity vectors form a vortex above the tongue. It must be noted that the active intensity vectors under the condition of nonzero rotation do not represent the direction of the acoustic energy flow. This result implies that the equi-phase lines rotate clockwise around the center of the vortex. And the direction of the instantaneous particle velocity in the vortex region changes greatly with time. The phase characteristics are no longer like a plane wave. In the map for reactive intensity, we can see the complicated pressure gradients. The global minimum of the amplitude is at the center of the vortex formed by the active intensity. The decrease of the amplitude near the vortex center is very rapid. The vertical difference of the amplitude between this center and a point 5.0 mm upward from the center is about -23 dB. The measurement for replica(I) was performed by varying the frequency in 500 Hz steps. The first appearance of a strong vortex was at 5.0 kHz with counterclockwise rotation. Thus the phase rotation may be in either direction with respect to the frequency.

The measurement frequencies for replica(II) were determined in the following manner. The characteristics of the reflection coefficient at the junction between the replica and the uniform tube were measured first using 50 Hz steps. The method of this measurement is described in refs. [4] and [5]. The amplitude characteristics of the reflection coefficient were obtained as shown in Figure 4. Then the frequencies corresponding to the local minimum amplitude of the reflection coefficient and some other frequencies between the local minima, which are shown as arrows, were selected as the measurement frequencies.

The vertical and horizontal intensity maps for these frequencies are shown in Figure 5(a) and (b), respectively. In the vertical aspects, characteristics similar to those observed in replica(I) can be seen. The replica contour in the horizontal plane, which is located at mid-height in the vertical plane, is almost symmetrical. Corresponding to this physical figuration, up to 5 kHz, the active and reactive intensity vectors form quite symmetric

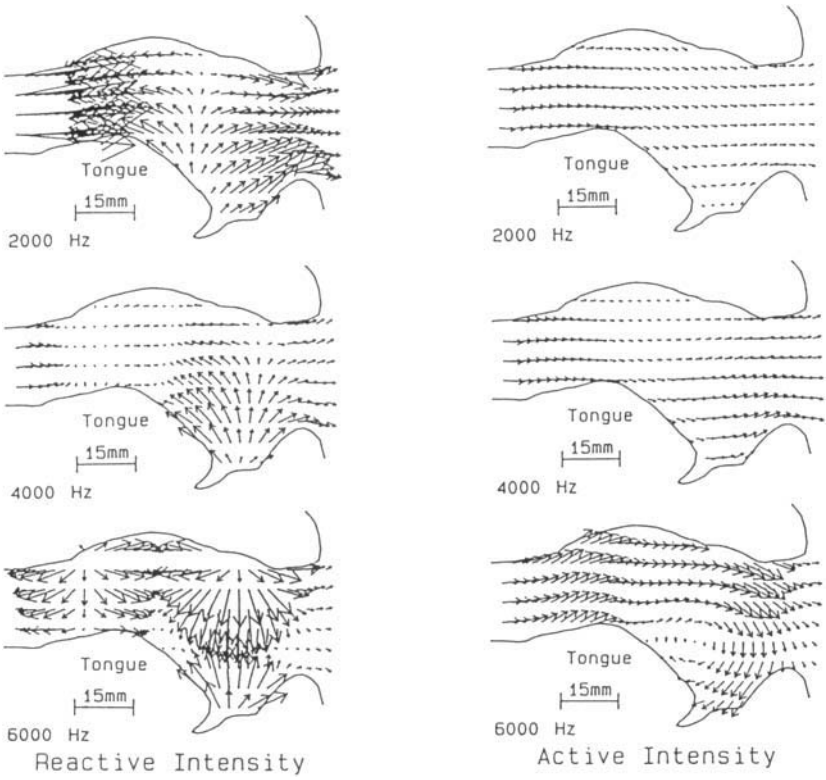


Figure 3. Intensity maps for replica(I) (vertical plane).

fields though the fields, contain transverse amplitude gradients and phase rotation in the horizontal plane at higher frequencies.

At 3.0 kHz, the horizontal differences of the local amplitude maxima near the lips region (the region near the anterior end of the replica contours) are 1.3 dB, and the tangential directions of the active and reactive intensity vectors at each point, except near the lips region, are approximately the same. The plane wave assumption seems to be valid in the horizontal plane. The vertical amplitude gradient, however, is relatively large and, as a whole, the entire sound field is not well explained by the progressive plane-wave assumption.

At 4.3 kHz, it can be seen from the horizontal reactive intensity map that the transverse amplitude distribution varies greatly. The amplitude near the wide spread wall is about +17 dB higher than the central amplitude. In the vertical active intensity map, a weak vortex with counterclockwise rotation is hardly observable. The sound field is quite different from the plane wave both in the vertical and in the horizontal plane. At frequencies higher than 4.3 kHz, the appearance of a vortex is always observed in the vertical and/or horizontal plane. At 5.6 kHz, the symmetry of the intensity maps begins to break up; and the sound field is extremely complicated.

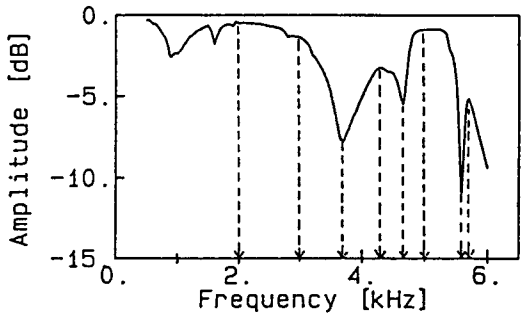


Figure 4. Amplitude characteristics of the reflection coefficient. Arrows indicate the measurement frequencies.

4. CONCLUSIONS

The characteristics of the sound field in the vocal tract are examined, based on acoustic measurement in replicas of the oral cavities. As result, the phase distribution can be approximated by plane waves up to a frequency of about 4 kHz. The amplitude distribution, however, varies significantly at the region above the tongue tip in the vertical plane even at low frequencies of about 2 kHz. And at high frequencies, above about 4 kHz, the phase distribution start to rotate in the vertical and the horizontal planes, associated with an extreme amplitude decrease at the center of the phase rotation.

The measurement was for two particular replicas and was in planes specified by sight. For more accurate evaluation of the sound field in the vocal tract, especially of the limit of the plane wave assumption, three-dimensional measurement in the whole region of the replicas with various articulatory shapes should be performed. We are now expanding the experimental device to enable this measurement.

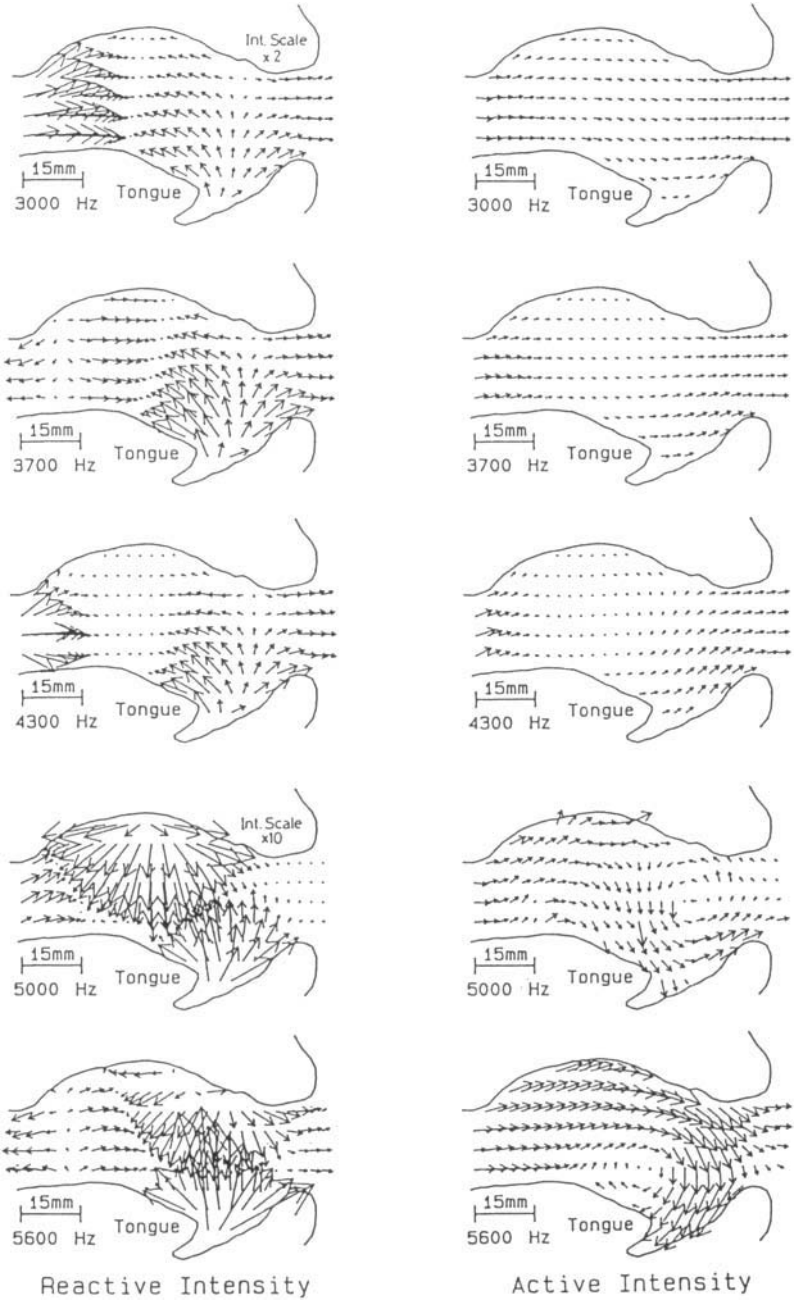


Figure 5. (a) Intensity maps for replica (II) (vertical plane).

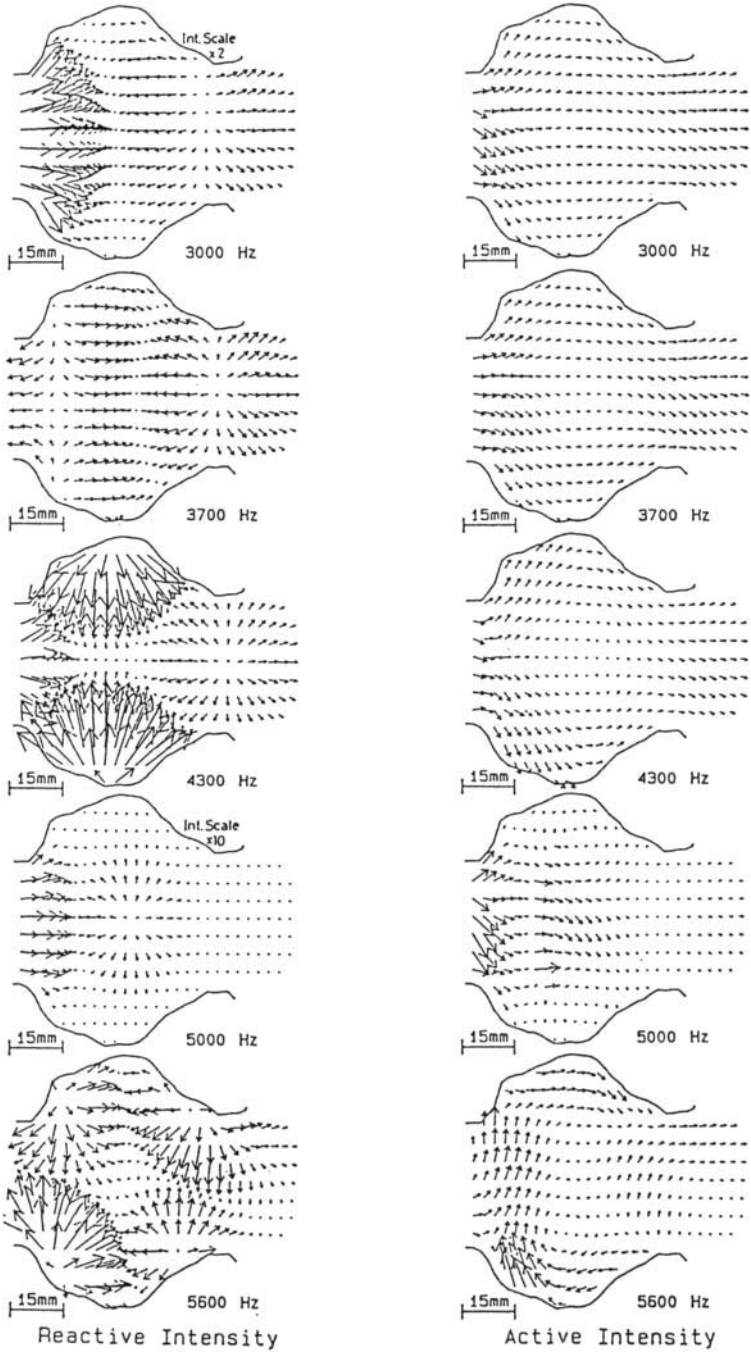


Figure 5. (b) Intensity maps for replica (II) (horizontal plane).

References

1. G. Fant, "coustic Theory of Speech Production," Mounton The Hague Paris (1970)

2. F. J. Fahy, "Sound Intensity," Elsevier Applied Science (1989)

3. J. Adin Mann, III, J. Tichy and A. J. Romano, "Instantaneous and Time-averaged Energy Transfer in Acoustic Fields," J. Acoust. Soc. Amer., 82, 1, pp.17-30 (1987)

4. J. Y. Chung and D. A. Blaser, "Transfer Function Method of Measuring In-duct Acoustic Properties. I. Theory," J. Acoust. Soc. Amer., 68, 3, pp.907-913 (1980)

5. K. Motoki, N. Miki and N. Nagai, "A Method to Measure the Area Function from Distribution of Complex Sound Pressure in the Nonuniform Acoustic Tube," Trans. IEICE Japan, J-72A, 8, pp.1222-1229 (1989)

6. K. Motoki, N. Miki and N. Nagai, "Measurement of Complex Sound Pressure Distribution in the Replica of the Oral Cavity," Proc. Autumn Meeting of Acoust. Soc. Japan, pp.225-226 (1989)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Speech Production Model Involving the Subglottal Structure and Oral-Nasal Coupling due to Wall Vibration

Hisayoshi Suzuki, Takayoshi Nakai, Jiauwu Dang and Chengxiang Lu

Electronics Department, Faculty of Engineering, Shizuoka University
3-5-1 Jouhoku, Hamamatsu-shi, 432 Japan

Abstract

This paper describes a speech production model considering factors such as the impedance of the velum, the inflation of the vocal tract volume, and the effects of the subglottal system. Based on our measurements of acoustic waveforms and mechanical vibrations at several points of the vocal organ, we propose a speech production model having oral-nasal coupling through the velum even when it closes. We assumed that the velum is composed of two vibrating plates connected to each other by a spring and a mechanical resistance. A syllable /bi/ is synthesized and examined, as an example, by the model, considering the velum leakage as well as a small inflation of the vocal tract volume by the inner air pressure just before the mouth opening. Concerning the speech production model, which has both a subglottal structure and a supraglottal one, it is shown that the waveform of glottal flow is changed, and synthesized vowels have zeros in their spectra because of an interaction between the sub- and supraglottal structures through the small opening area due to incomplete vocal cords closure.

1. INTRODUCTION

Human speech contains information regarding the physical individuality of the speaker as well as the linguistic information of speech. One of the possible ways to synthesize natural sounding speech is to develop a speech production model considering the acoustic properties of a human speaker, such as the influence on glottal vibration by the supra- and subglottal system, sound leakage from the oral cavity to the nasal cavity through the closed velum, inflation of the vocal tract volume by the inner pressure, and sound radiations from the nostrils, the face, and the neck as well as the mouth.

Observation of the radiation from other sources than the mouth was performed by the authors[1,2], and an attempt to make a speech production model considering part of these factors was reported previously[3]. In this paper, a modified model will be discussed. In this model the following factors are considered: sound leakage from the oral cavity to the

nasal cavity through the velum, inflation of the vocal tract volume when a stop consonant is produced, the subglottal structure and glottal leakage.

2. SPEECH PRODUCTION MODEL CONSIDERING THE VIBRATION OF THE VELUM

2.1. Observation of sound and vibration of speech

The sound and vibration at six points of the vocal organ was observed simultaneously in an anechoic room. These signals were separated from one another using a sound proof box and an isolation board, as shown in figure 1[2]. Japanese 100 /CV/-syllables were spoken by six male speakers. The database contains waveforms of sound radiated from (a)the mouth, (b)the nostrils, and (c)the skin near the pharynx, and those of vibrations of (d)the cheeks, (e)the nose, and (f)the skin near the pharynx.

Figure 2 shows a comparison of the sound pressure levels of radiations from the mouth, the nostrils, and the skin near the pharynx for various consonants and vowels. From figure 2, it can obviously be seen that the sounds radiated from the nostrils, especially for the vowel /i/ and a few voiced consonants, are not negligibly small.

Figure 3 shows the power spectra of the vowels /a/ and /i/. It should be noticed that the power in the higher frequency region above 1kHz of sound from the nostrils is much weaker than that form the mouth.

Figure 4 shows the waveforms of six signals accompanied with a voiced plosive /bi/. It is seen that the sounds radiated from other sources than the mouth and all vibrations begin at a time about 100ms before the onset of sound from the mouth. This causes the buzz bar of the voiced plosive /b/. The buzz bar is radiated from both the nostrils and the pharynx skin, but interestingly, the radiation from the nostrils is far larger than that from the pharynx skin.

To cope with these phenomena, we propose a speech production model that has sound leakage from the oral cavity to the nasal cavity even when the velum is closed, and that has some amount of air flow which produces a vocal cords vibration before the mouth is opened. The leakage is assumed to be produced by the vibration of the velum. The air flow before mouth opening is assumed, in this model, to be a result of inflation of the volume of the oral cavity.

2.2. A model of the velum

The velum has conventionally been treated as an on-off switch between nasal and non-nasal sound. However, the synthesized sounds by such a model are quite different from natural speech. For instance, the power spectra of synthesized sound from the nostrils of vowels are not restricted to below about 1kHz, as shown in figure 3. Therefore, the system of velum and nasal cavity should work as a low-pass filter rather than a simple on-off switch.

For vowels and non-nasal consonants, the velum works as a closed lid at the entrance of the nasal cavity. Let us assume that the velum is a vibrating plate, whose area, A_v, is

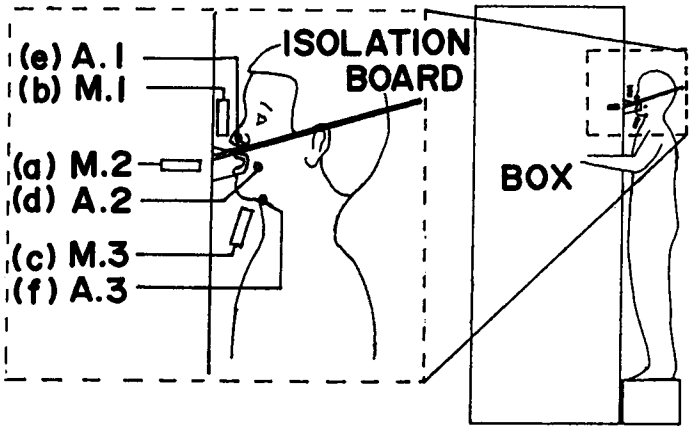


Figure 1. Illustration of the observation of sound and vibration using a sound proof box; M: microphone, A: acceleration pickup.

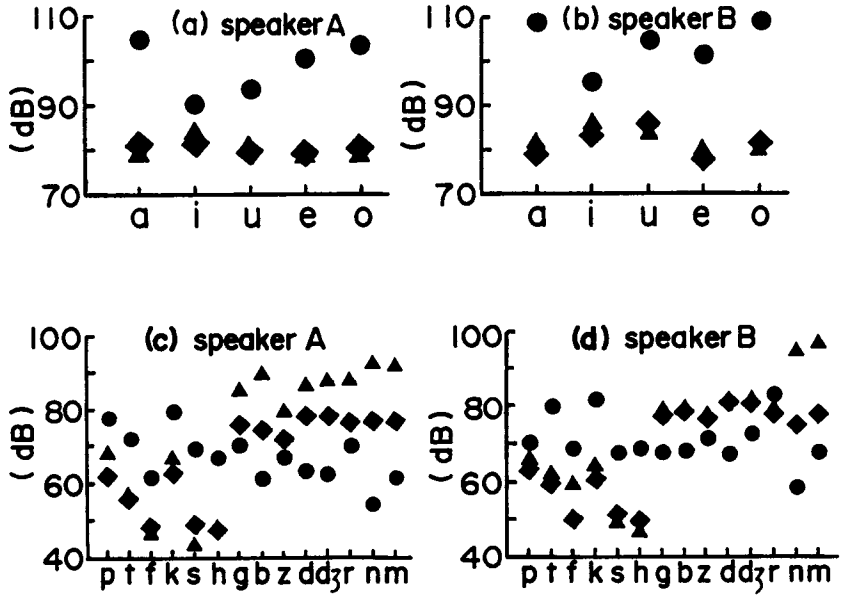


Figure 2. Sound pressure level for: (a) and (b): non-nasal vowels, (c) and (d): consonants; ● is from the mouth opening, ▲ is from the nostrils, and ◆ is from skin near the pharynx.

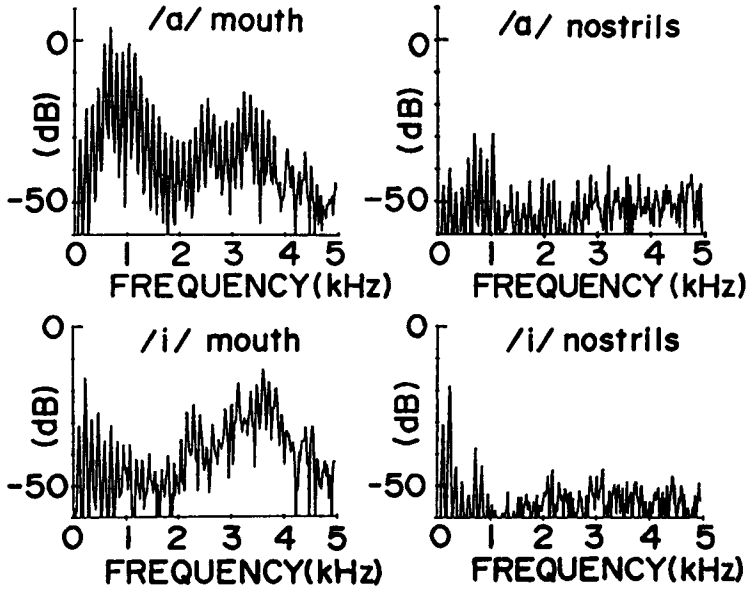


Figure 3. Spectra of sound from the mouth opening and nostrils for the vowels /a/ and /i/.

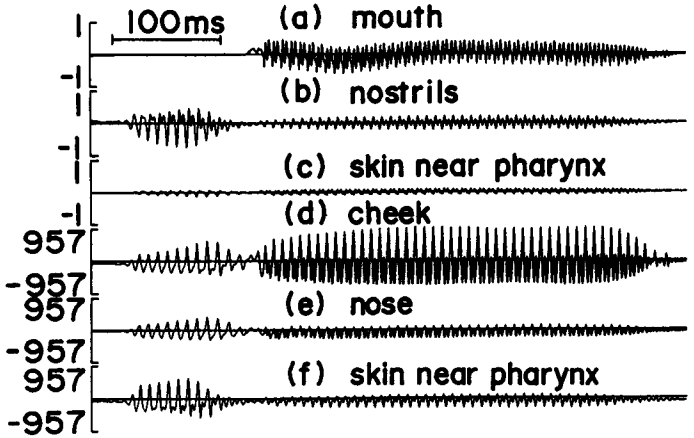


Figure 4. Waveforms of sound pressure and acceleration of /bi/; (a), (b), and (c): relative sound pressure, (d), (e), and (f): acceleration of wall vibration (cm/s^2).

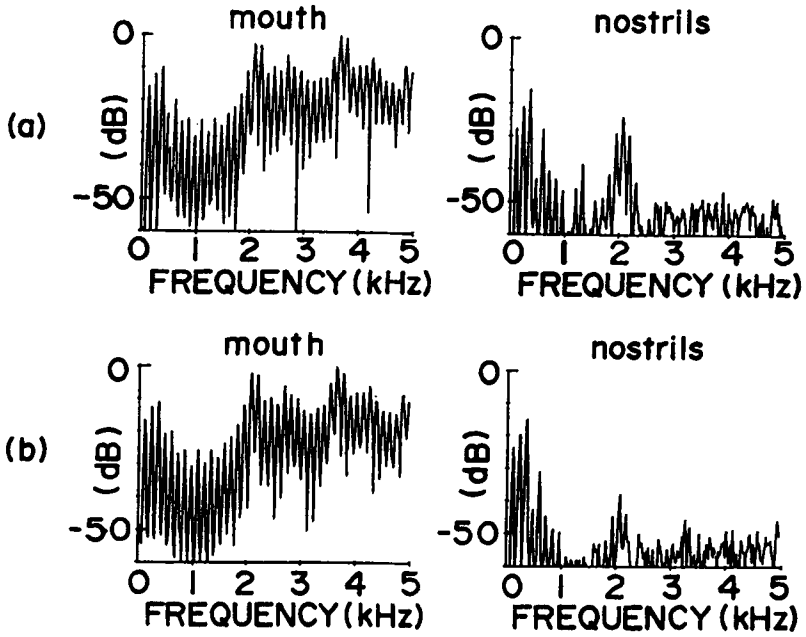


Figure 5. Spectra of radiations from the mouth and nostrils; (a) for the velum model with a single vibrating plate, and (b) for the velum model with a double vibrating plate shown in figure 6.

4cm^2, the stiffness per unit area, K_v, is 8.45×10^4dyn/cm^3, the same as that of the vocal tract wall, and the density is 1g/cm^3, which is approximately the same as the human average density. The effective thickness of the velum is assumed to be 0.15cm; i.e., its mass per unit area, M_v, is 0.15g/cm^2. To cause the relative levels of sound pressure of sounds from the mouth and nostrils for the synthesized vowel /i/ around 300Hz to be adjusted the natural ones, its mechanical resistance per unit area, R_v, is set to 50g/s/cm^2. By this setting, the sound pressure from the mouth and nostrils are matched to those of natural speech around 300Hz, as shown in figure 5(a). But, the sound pressure level around 2000Hz of synthesized sound from the nostrils is still larger than that of natural sound, as shown in figure 3. The difference in the level between synthesized and natural sound is about 10~15dB.

When the velum is assumed to be a simple vibrating plate, as mentioned above, the system of velum and nasal cavity is equivalent to a second-order LC type low-pass circuit, because the nasal cavity works approximately as a compliance in the high frequency region. Therefore, in order to simulate natural speech, the equivalent circuit of velum and nasal cavity should be a low-pass filter with a higher order than second order. From this consideration, we introduce a model of the velum which is composed of two vibrating plates connected by a spring and a mechanical resistance, as shown in figure 6(a), whose

equivalent circuit is shown in figure 6(b), where C_{tv} is chosen so that the synthesized vowels /a/ and /i/ are as similar as possible to the natural /a/ and /i/, respectively. Figure 5(b) shows the spectrum of sound from the nostrils synthesized by the velum model of figure 6. It is seen that the power spectra around both 300 Hz and 2000Hz have a reasonable level.

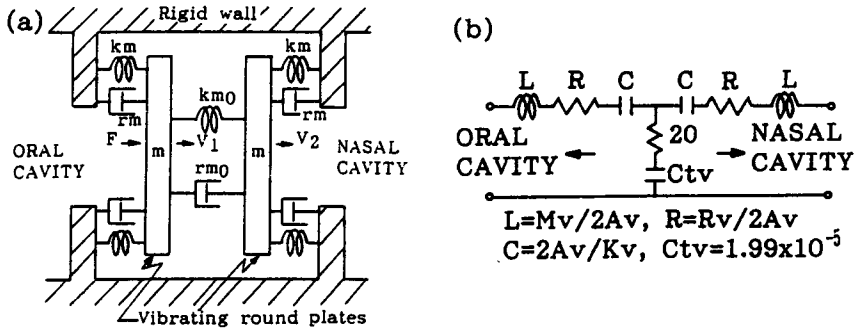


Figure 6. (a) Velum model with a double vibrating plate, and (b) equivalent circuit of the double vibrating plate model.

2.3. Inflation of the cavity volume

Next, we consider to produce the voiced plosives having a natural sounding buzz bar. To produce a buzz bar, for instance of /bi/, before mouth opening, there should be some amount of air flow through the glottis. This causes some inflation of the vocal tract volume as well as vocal cords vibration[4]. We assume here that the pharynx is enlarged by about 8.4cm^3 during buzz bar. Figure 7 shows the waveforms produced for synthesizing /bi/, where (a), (b) and (d) are the waveforms corresponding to figure 4, (c) is the waveform corresponding to figure 4(f), (e) is the air pressure at the velum, and (f) is the glottal volume flow. A listening test was performed to evaluate the quality of synthesized sound. The result shows that the synthesized /bi/ shown in figure 7 has a reasonable naturalness score.

3. THE SUBGLOTTAL STRUCTURE AND INCOMPLETE CLOSURE OF THE VOCAL CORDS

3.1. Model of the subglottal system

The vocal system is composed of the subglottal system, glottis, and supraglottal system, which includes the vocal tract, the nasal tract, the velum, the mouth, etc., as shown in figure 8. The subglottal system, which consists of the trachea, the bronchi, and the lungs, can be treated as a whole as a non-uniform acoustic tube[5]. Figures 9(a) and (b) show the equivalent circuit of the subglottal system and its input impedance seen from the glottis side, respectively. There is some leakage of air flow at the glottis even when

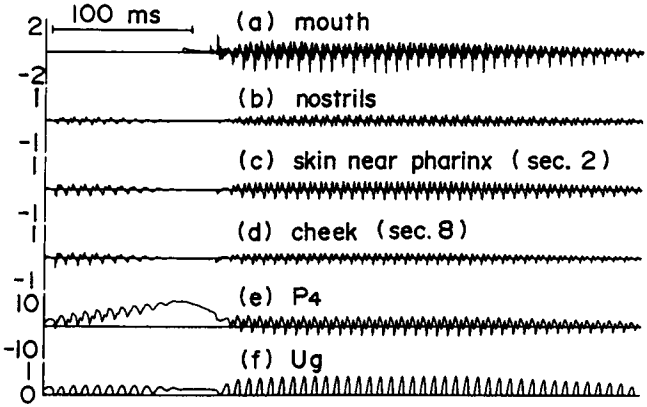


Figure 7. Waveforms of sound pressure and acceleration of synthesized /bi/; (a) and (b): sound pressure, (c) and (d): acceleration of the wall vibration ($10^4 cm/s^2$), (e): air pressure at the velum (cmH_2O), (f): volume velocity of glottal wave ($10^3 cm^3/s$).

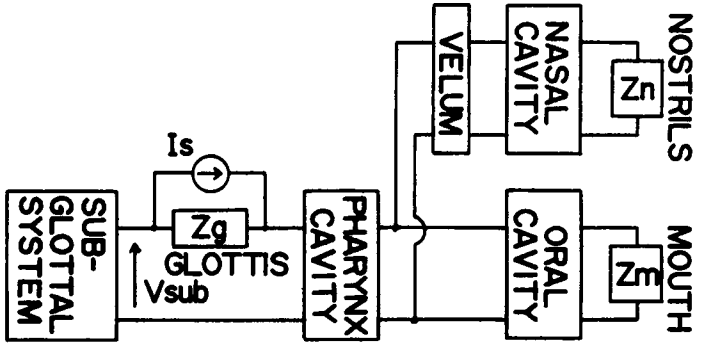


Figure 8. A total system for speech production with subglottal system and glottal leakage.

the vocal cords appear to be closed. Holmberg[6] reported that the ratio of leakage air flow, U_{gmin}, to the peak to peak value of the glottal volume velocity, U_{gmax} - U_{gmin}, is about 0.38- 1. The reason for this leakage is that there is a gap between the arytenoid cartilages of the glottis and that a part of vocal cords closes incompletely when a soft voice is uttered.

By the incomplete closure in the vocal cords, the air flow to the supraglottal system is changed, and the transfer function of the vocal tract is affected more or less by the subglottal system. Thus, the speech sound can be considerably affected by this phenomenon.

3.2. Transfer function

Figure 10 shows (a)the vocal tract transfer function, i.e., the ratio of the output volume flow I_o to the input flow I_s, I_o/I_s, for the vowel /a/, and (b)the output impedance of the subglottal system, i.e., the ratio V_{sub}/I_s, in the cases with and without the subglottal

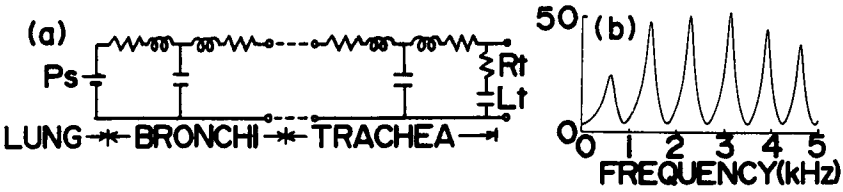


Figure 9. (a): Equivalent circuit of the subglottal system, and (b): its impedance seen from the glottis.

system, when the area of glottis, A_g, is varied from 0 to 0.25cm^2. It is seen that, when A_g is increased, the transfer function of the system with subglottal system has zeros at 1470, 2270, and 3080Hz, while the one without subglottal system has no zero. Figure 10(b) shows that V_{sub}/I_s has two zeros, which nearly coincide with the first and second formant frequencies of the vocal tract for /a/.

3.3. Leakage area

The leakage air flow is treated by introducing a leakage area A_{gs} in two-mass model of vocal cords[7]. We assume that the areas of the lower and upper vocal cords, A_{g1} and A_{g2}, are given as follows:

$$A_{g1} = \begin{cases} A_{go1} + 2\,l_g\,x_1 + A_{gs} & A_{go1} + 2\,l_g\,x_1 \geq 0 \\ A_{gs} & A_{go1} + 2\,l_g\,x_1 < 0 \end{cases}$$

$$A_{g2} = \begin{cases} A_{go2} + 2\,l_g\,x_2 + A_{gs} & A_{go2} + 2\,l_g\,x_2 \geq 0 \\ A_{gs} & A_{go2} + 2\,l_g\,x_2 < 0 \end{cases}$$

where l_g is the length of vocal cord, and x_1 and x_2 are the displacements in the opening direction of vocal cords.

We synthesized vowels by this model with A_{gs} values of 0cm^2 to 0.06cm^2. Figure 11(a) shows an example of the area wave of A_{g1} and A_{g2} when A_{gs}=0.06cm^2.

Figures 11(b) and (c) show the waveforms of the glottal volume velocity for the vowel /o/, where the A_{gs} values are 0cm^2 and 0.06cm^2, respectively.

By comparing the cases A_{gs}=0.06cm^2 and A_{gs}=0cm^2, we can see that the glottal volume velocity, U_g, is not constant even when the upper or lower vocal folds are closed. The reason why it occurs is that, as explained by Cranen et al.[8], there is an out-of-phase motion between the upper vocal cord and the lower vocal cord.

Figure 12 shows the power spectra of the mouth radiation of the synthesized /a/, in the cases without subglottal system (left) and with subglottal system (right). It is clearly seen that the spectra of the vowel with both the subglottal system and glottal leakage have some spectral dips due to zeros. In the region below 1kHz, when A_{gs} is 0.06cm^2, there are a few spectral dips not corresponding to the characteristics I_o/I_s and V_{sub}/I_s. They must be caused by changed vibration of the vocal cords under the influence of the leakage area. Thus, when the area of incomplete closure, A_{gs}, is increased, the spectra and quality of sounds are varied considerably.

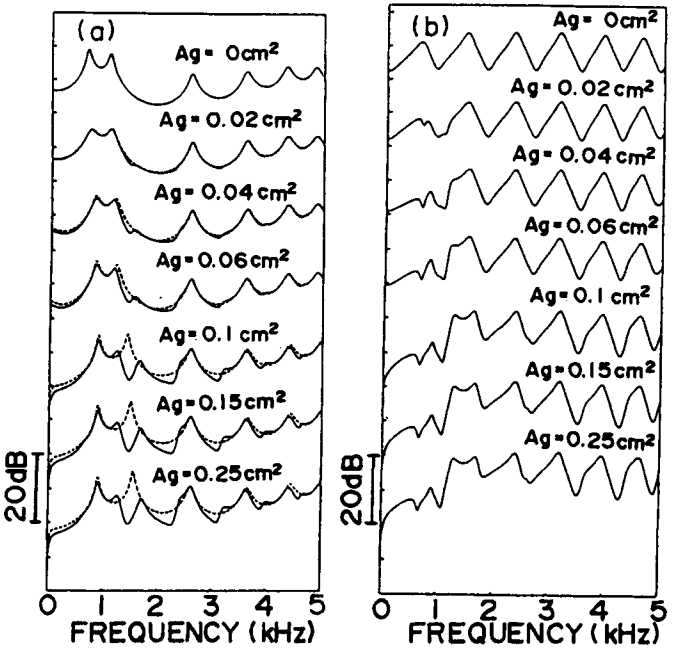


Figure 10. (a): Transfer function of a synthesis system with (solid curves) and without (broken curves) subglottal system for the vowel /a/, (b): frequency characteristics V_{sub}/I_s with subglottal system for the vowel /a/.

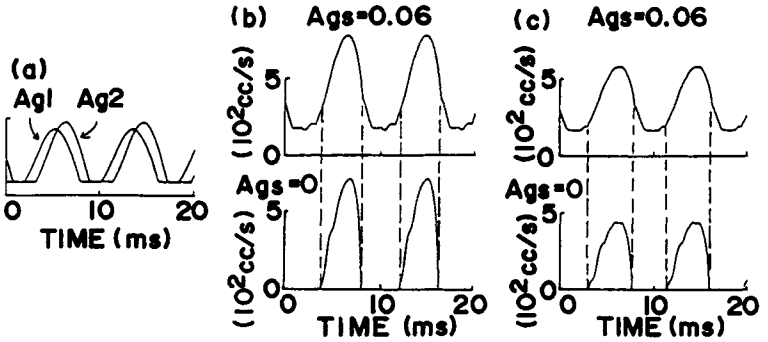


Figure 11. (a): Waveforms of synthesized A_{g1} and A_{g2} when $A_{gs} = 0.06\text{cm}^2$, (b): glottal volume velocity U_g for the vowel /a/ without subglottal system, (c) U_g with subglottal system.

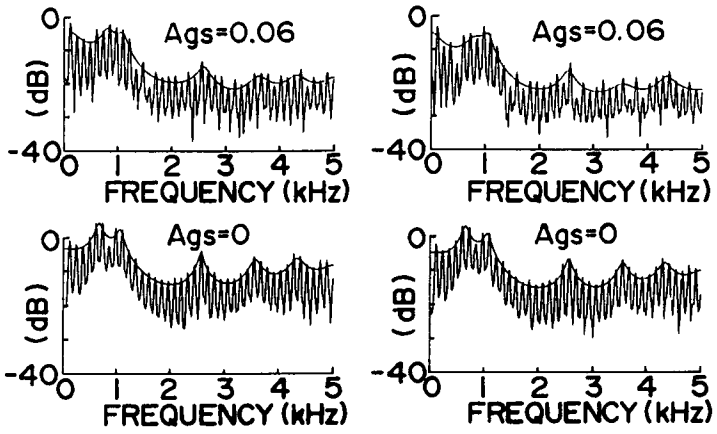


Figure 12. Spectra of sound from the mouth for the synthesized vowel /a/; left: without subglottal system, right: with subglottal system.

4. CONCLUSION

Observation has evidenced that the sounds radiated from the nostrils are unexpectedly large in most speech sounds. We proposed a speech production model to explain this fact by sound leak from the vocal tract to the nasal tract through the vibrating velum. The leakage sound is low-pass filtered by the closed and vibrating velum, then transferred to the nostrils and finally emitted from it. To produce a buzz bar of voiced plosives, we introduced an inflation of the volume of the pharynx cavity, which allows a limited amount of air flow and results in glottis vibration. This flow goes through the nasal tract and is emitted from the nostrils before the mouth is opened. Synthesized voices /a/, /i/ and /bi/, for example, were evaluated to be quite similar to natural speech. It is also shown that the subglottal system affects the sound quality whenever an incomplete vocal-cords closure occurs.

Acknowledgement

This study is partly supported by a Grant-in-Aid for Scientific Research (Priority Area No. 62608501) from the Japanese Ministry of Education, Science and Culture.

References

1. H. Suzuki, T. Nakai and K. Shimizu, "Measurement and Analysis of Speech Sound Radiated from Vocal Tract Wall," *ICASSP-86*, pp.1625-1628 (1986)

2. H. Suzuki, J. Dang and T. Nakai, "Measurement of Sound and Vibration at the Lips, Nostrils and Pharynx Wall in Speech Utterance and Simulation of Sound Leakage from the Oral Cavity to the Nasal Cavity in Non-nasal Sounds," IEICE Trans. A, Vol-J74-A, No.12, pp.1705-1714 (1991)

3. H. Suzuki and T. Nakai, "Speech production by a vocal cords - vocal tract - vocal tract wall vibration model," The 2nd symposium on Advanced Mann Machine Interface through Spoken Language, pp.7-1 - 7-8, at Hawaii (1988)

4. J. R. Westbury, "Enlargement of the supraglottal cavity and its relation to stop consonant voicing," J. Acoust. Soc. Am., 73(4), pp.1322-1366 (1983)

5. K. Ishizaka, "Input acoustic-impedance measure of the subglottal system," J. Acoust. Soc. Am., 60, pp.190-197 (1976)

6. E. Holmberg, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," J. Acoust. Soc. Am., 84, pp.511-529 (1988)

7. K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," Bell System Techn. J., 51, pp.1233-1268 (1972)

8. B. Cranen and L. Boves, "On subglottal formant analysis," J. Acoust. Soc. Am., 81, pp.734-746 (1987)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

On the Analysis of Predictive Data such as Speech by a Class of Single Layer Connectionist Models

Frank Fallside

Cambridge University Engineering Department
Trumpingdon Street, Cambridge CB2 1PZ, UK

Abstract

The class of single layer connectionist models analysed is

$$\sigma_k = \sum_{i=0}^{L} w_i y_{k-i}; \quad w_0 = 1$$

$$O_k = f(\sigma_k)$$

where the non-linearity includes the logistic function $\sum_N \sigma_k^2$ commonly used in error back propagation analysis. In this $\{y_k\}$ is the input data, the weights, the nett input to the non-linearity and the output of the connectionist model. It is shown that when the input data can be modelled by a linear predictive or autoregressive process, with

$$e_j = \sum_{i=0}^{p} a_i y_{j-i}; \quad a_0 = 1$$

a solution exists for the weights which minimises the cost function and hence an output error cost function. This establishes weight sets for linear predictive processes such as speech, leading in turn to sets of single layer connectionist models which provide a form of vector quantisation (VQ) analysis of speech. Examples are given of the analysis of speech and other data by the method and a comparison is made with the equivalent error back propagation analysis.

1. INTRODUCTION

Since the first application of linear predictive analysis to speech by Atal [1] its use has become widespread through the analysis of speech, for coding, recognition and synthesis; see for example successive Proceedings of the International Conference on Acoustics, Speech & Signal Processing (ICASSP).

There has recently been an upsurge in interest in the analysis of pattern data by connectionist models/artificial neutral networks, see for example, Rumelhart & MacClelland

[2]. This has included the analysis of speech, e.g. [3], using the error-back-propagation algorithm. Most such studies preprocess speech into the frequency domain and employ spectral pattern data.

It is of interest to study the direct analysis of speech, in the time domain, and this is the subject of the present paper. It is concerned mostly with the analysis of a class of single layer connectionist models, the relationship of the class to conventional linear predictive analysis and the relationship with the back propagation algorithm [2]. A few computational results are also given and some conclusions drawn on the application of the method to speech processing.

The paper starts with a brief review of conventional linear predictive analysis.

2. CONVENTIONAL LINEAR PREDICTIVE ANALYSIS

2.1. Linear Predictive (lp) Analysis

In the conventional lp analysis of a data sequence $\{y_i\}$, and the general sample y_i is estimated as $\hat{y}_i$ a linear combination of p past samples, the observations, see Figure 1,

$$\hat{y}_i = -\sum_{j=1}^{p} a_j y_{i-j} \tag{1}$$

Then the estimation error e_i is given by

$$\begin{aligned} e_i &= y_i - \hat{y}_i & (2)\\ &= \sum_{j=0}^{p} a_j y_{i-j} & (3) \end{aligned}$$

with $a_0 = 1$. To derive the lp coefficients, the sum of errors squared summed over some data length

$$E = \sum_{N_1}^{N_2} e_i^2 \tag{4}$$

is minimised with respect to the coefficients $a_1, \ldots, a_p$. This can be done by setting partial derivatives of E to zero or by using the orthogonality principle. Following the latter, via Parsons [4] we can write the sequence of estimation errors as

$$\begin{array}{lllllllll} e_r &= y_r &+ a_1 y_{r-1} &+ a_2 y_{r-2} &+ \cdots &+ a_p y_{r-p} \\ e_2 &= y_2 &+ a_1 y_1 &+ a_2 y_0 &+ \cdots &+ a_p y_{2-p} \\ e_1 &= y_1 &+ a_1 y_0 &+ a_2 y_{-1} &+ \cdots &+ a_p y_{1-p} \end{array} \tag{5}$$

or matrix in form

$$\boldsymbol{e} = \boldsymbol{y} + \boldsymbol{X}\boldsymbol{a} \tag{6}$$

Here we have reversed the normal order of the elements in the vectors, for convenience later, it has no effect on the results We now wish to find the conditions under which a minimises the errors squared summed over the data, E, with

$$E = \boldsymbol{e}^T \boldsymbol{e} \tag{7}$$

where $(^T)$ denoted transpose. Consider any other weight vector b with resulting error vector f. Now

$$\begin{aligned} f &= y + Xb && (8) \\ &= e + X(b-a) && (9) \end{aligned}$$

so

$$\begin{aligned} f^T f &= [e + X(b-a)]^T [e + X(b-a)] \\ &= e^T e + e^T X(b-a) + (b-a)^T X^T e + [X(b-a)]^T [X(b-a)] \end{aligned} \quad (10)$$

Now under the condition

$$X^T e = X^T (y + Xa) = 0 \quad (11)$$

eqn (10) yields

$$\begin{aligned} f^T f &= e^T e + |X(b-a)|^2 \\ &\geq e^T e \end{aligned} \quad (12)$$

with equality only for $b = a$.

Thus under the condition (11), that $e^T X = 0$, in the filter of eqn (3) each error e_i is required to be orthogonal to the observations $y(i-1), y(i-2), \ldots, y(i-p)$.

Eqn (11) gives the normal equations

$$X^T X a = -X^T y \quad (13)$$

with the minimum cost function

$$E_{\min} = e^T y \quad (14)$$

as p simultaneous linear equation in a_i. These can be solved in broadly one of two ways, depending on the limits N_1 and N_2. If $N_1 \to \infty, N_2 \to \infty$ and the data is windowed such that it is only non-zero over N points, there results the autocorrelation analysis, and if $N_1 = 0$ and $N_2 = N-1$ there results the covariance analysis [4].

2.2. Structure of the lp Filter

The lp filter of eqn (3)

$$\frac{y_i}{e_i} = \frac{1}{1 + a_1 z^{-1} + \cdots + \lambda a_p z^{-p}} \quad (15)$$

is essentially sequential or serial. Linear predictive analysis using the normal equations (13) is essentially a 'block' analysis using the entire data set N or can be sequential or serial in the PARCOR analysis [4] or variants, which take in one sample y_i at a time and update the analysis of a.

The structure of the analysis is shown in Figure 1(b) where a $(p+1)$ point filter window is moved incremently along the N point data set to establish the N errors $\{e_i\}$.

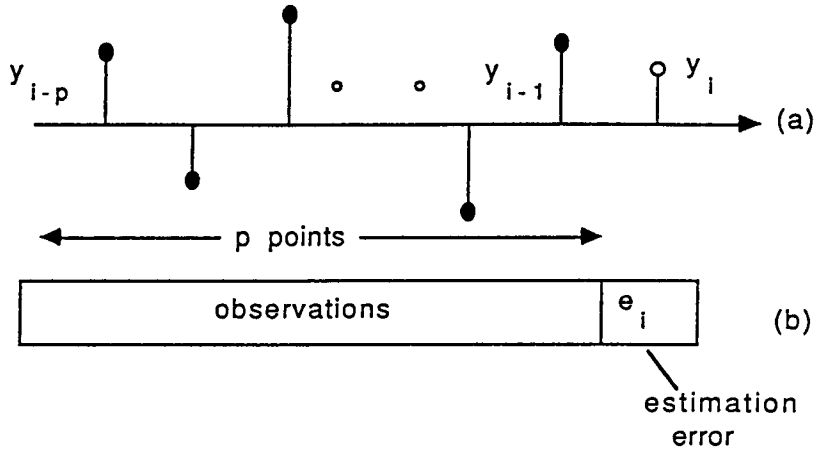


Figure 1. Linear prediction analysis (1) estimation (after Parsons(4)) (b) estimation error and observations.

3. ANALYSIS BY A SINGLE-LAYER CONNECTIONIST MODE

3.1. General

The model to be considered is shown in Figure 2. At this stage it is linear and has an $(L+1)$ column weight vector W, with $w_0 = 1$. For convenience we will also write

$$\begin{aligned} W &= [1\ w_1\ w_2 \ldots\ w_L]^T \\ &= [1\ \boldsymbol{w}]^T \end{aligned} \tag{16}$$

For an $(L+1)$ column input vector

$$y_k = [y_k\ y_{k-1}\ y_{k-2}\ \cdots\ y_{k-L}]^T \tag{17}$$

the output model for the k-th window of data is

$$\begin{aligned} \sigma_k &= y_k^T W \\ &= y_k + \boldsymbol{z}\boldsymbol{w} \end{aligned} \tag{18}$$

where $\boldsymbol{z}$ is an L row vector $\{y_{k-1}, \ldots, y_{k-L}\}$ of observations.

3.2. Linear Predictive Analysis of Data

Suppose we apply a data set $\{y_i\}$ of length N, to the network to form k and increment the data by 1 to form each new analysis frame, then

$$\begin{aligned} \min_{\boldsymbol{w}} \sigma^T \sigma &= \min_{\boldsymbol{w}} \sum_N \sigma_k^2 \\ &= (\boldsymbol{y} + \boldsymbol{Z}\boldsymbol{w})^T (\boldsymbol{y} + \boldsymbol{Z}\boldsymbol{w}) \end{aligned} \tag{19}$$

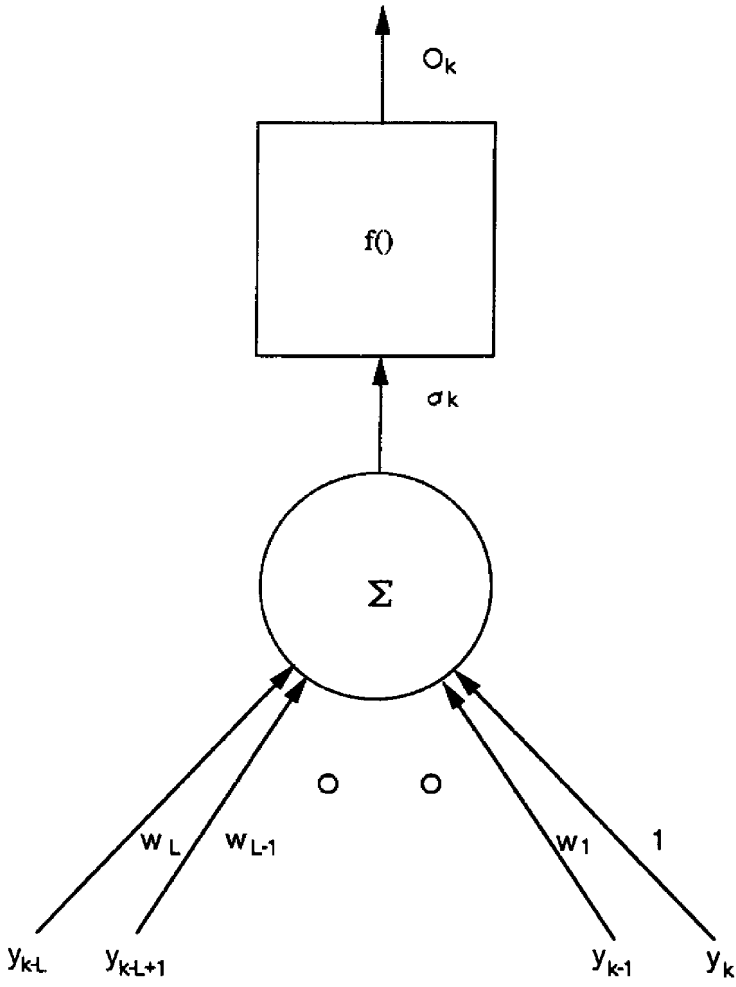


Figure 2. Single layer connectionist model.

carries out a linear predictive analysis of the data, as in Section 2, with $p = L$ lp coefficients. (We might note that if the complete data set y is windowed, the w would be those of the autocorrelation analysis and if it was not, and we made $N + L$ samples available they would be those of the covariance analysis).

The solution can also be obtained by a steepest descent or back propagation algorithm [2] with weight changes

$$\Delta w_i = -\eta \sigma_k y_{k-i} \tag{20}$$

Equally if the network has a non-linearity with

$$O_k = f(\sigma_k) \tag{21}$$

then if the non-linearity is monotonic increasing and differentiable the weight changes can again be calculated. For example for the logistic function

$$f(\sigma) = \frac{1}{1 + e^{-\sigma}} \tag{22}$$

The appropriate criterion is

$$I = \frac{1}{2} \sum_N (\frac{1}{2} - f(\sigma_k))^2$$

and

$$\Delta w_i = -\eta (\frac{1}{2} - O_k) O_k (1 - O_k) y_{k-i} \tag{23}$$

3.3. Analysis of Linear Predictive Data

We now address a more general problem, where the data $\{y_i\}$ is known to represent a linear predictive, or autoregressive process with $p < L$. In other words where the connectionist model or filter is longer than the linear predictive filter which models the data. Such an example is speech, where it is known for example that an lp filter with $p = 10$ accurately models the data, and which might apply to a connectionist model with 100 input nodes.

As before the connectionist model of Figure 2 has

$$\boldsymbol{\sigma_k = y_k + Z w} \tag{24}$$

Now the y_i obey a linear predictive process with

$$e_i = y_i + \boldsymbol{x a} \tag{25}$$

as in Section 2, or for the whole data set $\{y_i\}$

$$\boldsymbol{e = y + X a} \tag{26}$$

As a result k, the minimisation min T and w are constrained. Since the data y can be modelled by an lp filter, let us look for a solution to the minimisation problem with

$$\boldsymbol{w} = \boldsymbol{I}_L + \alpha \boldsymbol{a} \tag{27}$$

where

$$\boldsymbol{a} = [a_1 \dots a_p]^T \tag{28}$$

and where $\boldsymbol{I}_L$ is a unit L column vector and is a constant $L \times p$ matrix. As a result the output of the network is

$$\begin{aligned} \sigma_k &= y_k + \boldsymbol{z}\boldsymbol{I}_L + \boldsymbol{z}\alpha\boldsymbol{a} && (29) \\ &= y_k^T \boldsymbol{I}_{L+1} + \boldsymbol{z}\alpha\boldsymbol{a} && (30) \end{aligned}$$

where y_k is the $L+1$ column vector of inputs to the net and $\boldsymbol{I}_{L+1}$ is a unit $L+1$ column vector, and the vector of outputs for all the data $\{y_i\}$ is

$$\sigma = \boldsymbol{y}^T \boldsymbol{I}_{L+1} + \boldsymbol{Z}\alpha\boldsymbol{a} \tag{31}$$

We now set up a sum of squares cost function

$$\sigma^T\sigma = (\boldsymbol{y}^T \boldsymbol{I}_{L=1} + \boldsymbol{Z}\alpha\boldsymbol{a})^T(\boldsymbol{y}^T \boldsymbol{I}_{L+1} + \boldsymbol{Z}\alpha\boldsymbol{a}) \tag{32}$$

and seek the condition on that this is a minimum. Proceeding as in Section 2, let and suppose the correspondingly cost function is T . Then from (32),

$$\begin{aligned} \Theta^T\Theta &= (\boldsymbol{y}^T \boldsymbol{I}_{L=1} + \boldsymbol{Z}\beta\boldsymbol{a})^T(\boldsymbol{y}^T \boldsymbol{I}_{L+1} + \boldsymbol{Z}\beta\boldsymbol{a}) && (33) \\ &= (\sigma + \boldsymbol{Z}(\beta-\alpha)\boldsymbol{a})^T(\sigma + \boldsymbol{Z}(\beta-\alpha)\boldsymbol{a}) && (34) \\ &= \sigma^T\sigma + \sigma^T\boldsymbol{Z}(\beta-\alpha)\boldsymbol{a} + \boldsymbol{a}^T(\beta^T-\alpha^T)\boldsymbol{Z}^T\sigma && (35) \\ &\quad +|(\sigma + \boldsymbol{Z}(\beta-\alpha)\boldsymbol{a})|^2 \end{aligned}$$

Now under the condition

$$\sigma^T\boldsymbol{Z} = 0 \tag{36}$$

$$\begin{aligned} \Theta^T\Theta &= \sigma^T\sigma + |\sigma - \boldsymbol{z}(\beta-\alpha)\boldsymbol{a}|^2 && (37) \\ &\geq \sigma^T\sigma && (38) \end{aligned}$$

and equality holds only under . Thus the condition (36) is the condition for minimisation, that is orthogonal to $\boldsymbol{Z}$.

3.4. A Class of Connectionist Models

We now seek out a particular minimisation such that is dependent on e alone. Returning to the output of the network

$$\sigma_k = y_k{}^T I_{L+1} + Z\alpha a \tag{39}$$

or

$$\begin{aligned}\sigma_k &= [y_k y_{k-1} \cdots y_{k-L}] I_{L+1} \\ &\quad + [y_k y_{k-1} \cdots y_{k-L}] \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1p} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{L1} & \alpha_{L2} & \cdots & \alpha_{Lp} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}\end{aligned} \tag{40}$$

Now since the connectionist filter is $(L+1)$ wide and the lp filter has p observations, there can be $m = (L+1) - p$ estimates within the connectionist filter. Thus

$$\sigma_k = \begin{bmatrix} e_k \\ e_{k-1} \\ \vdots \\ e_{k-m+1} \\ \hline y_{k-m} \\ \vdots \\ y_{k-L} \end{bmatrix}^T I_{L+1} - \left[\left\{ \begin{array}{c} x \\ - \\ O \end{array} \right\} a \right]^T I_{L+1} + z\alpha a \tag{41}$$

We now seek out the conditions on α that σ_k is a function of the m vector e_m. Since the second and third terms in eqn (41) are functions of a then the last p elements of I_{L+1} must be zero to null the y_{k-j} elements in the first term and we will describe the result as $I_{L+1,p}$, viz.

$$I_{L+1} = I_{L+1,p} \tag{42}$$

Hence also

$$w = I_{L,p} + \alpha a \tag{43}$$

Thus eqn (41) becomes

$$\sigma_k = e_m^T I_m + z\alpha a - (xa)^T I_m \tag{44}$$

where I is a unit m column vector.
The condition sought is thus

$$z\alpha a - a^T x^T I_m = 0 \tag{45}$$

or

$$z\alpha - I_m^T x = 0 \tag{46}$$

Now

$$I_m^T x = z \begin{bmatrix} 1 & 1 & & 0 \\ 1 & 1 & & \\ \cdot & \cdot & & 1 \\ 1 & 1 & \cdot & 1 \\ 0 & & & \cdot \\ & & & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \overset{p}{\overbrace{}} \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ p\text{-}1 \\ \\ L \\ \\ m \end{matrix} \tag{47}$$

$$= zS_m \tag{48}$$

Hence the condition is $\alpha = S_m$ and so

$$w = I_{L,p} + S_m a \tag{49}$$

This gives a solution for w which makes k a function of e_m alone, in fact

$$\begin{aligned} \sigma_k &= e_m^T I_m \\ &= \sum_{j=0}^{m-1} e_{k-j} \end{aligned} \tag{50}$$

The estimation error of the net, σ^k, consists of the sum of the m lp estimation errors from the L observations, as shown in Figure 3.

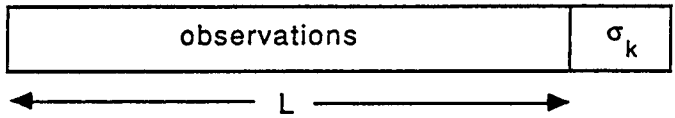


Figure 3. Estimation error and observations.

3.5. Minimisation of Cost Functions by the Class

The solution for w given in the previous section makes k a function of e_m alone. We now increment the data $\{y_i\}$ through the network and set up a cost function. The value of the increment affects the cost function. For example with $L = 6$, $p = 2$, there are $m = L + 1 - p = 5$ estimates in each window of data. Accordingly if the data is incremented by 1 we form sets of inputs, say

$$\begin{aligned} \sigma_N &= e_N + e_{N-1} + e_{N-2} + \ldots + e_{N-4} \\ \sigma_{N-1} &= e_{N-1} + e_{N-2} + \ldots + e_{N-4} + e_{N-5} \\ &\textit{etc.} \end{aligned} \tag{51}$$

Then on setting up he cost function $\sigma^T\sigma$, terms such as $d_j e_j$ appear with $d_j \neq 1$. Choosing m as the increment it is seen that $d_j = 1$ and

$$\sigma^T\sigma = (e_N + \ldots + e_{N-m+1})^2 + (e_{N-m} + \ldots + e_{N-2m+1})^2 + \ldots \tag{52}$$

since

$$\sigma = \begin{bmatrix} \boldsymbol{I}_m^T & & \boldsymbol{O} \\ & \boldsymbol{I}_m^T & \\ \boldsymbol{O} & & \boldsymbol{I}_m^T \end{bmatrix} \boldsymbol{e} \tag{53}$$

Without pursuing the minimisation completely here we notice that if the estimation errors are uncorrelated, if

$$< e_i e_j >= 0; \quad i \neq j \tag{54}$$

then the cost function

$$J = \min_a \sigma^T \sigma \tag{55}$$

$$\rightarrow \min_a \boldsymbol{e}^T \boldsymbol{e} \tag{56}$$

of Subsection 2.1. In other words if the estimation errors are uncorrelated, the minimisation or training of the connectionist model establishes the same coefficients a as does conventional lp analysis.

It is well known that for speech the estimation errors are not correlated but that the conventional lp analysis produces a useful parameterisation of speech and we might expect the same for the connectionist analysis.

3.6. Properties of the Class

Some of these are best seen by example. Take the case $L = 6$ as shown in Figure 4 (a) and assume $p = 2$. Therefore the number of estimates in a connectionist frame of data is $m = L + 1 - p = 5$. Hence

$$\sigma_k = e_k + e_{k-1} + e_{k-2} + e_{k-3} + e_{k-4} \tag{57}$$

$$\boldsymbol{I}_{L,p} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{S}_m = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \tag{58}$$

$$\boldsymbol{w} = \boldsymbol{I}_{L,p} + \boldsymbol{S}_m [a_1 a_2]^T \tag{59}$$

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} a_1 \\ a_1 + a_2 \\ a_1 + a_2 \\ a_1 + a_2 \\ a_1 + a_2 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 + a_1 \\ 1 + a_1 + a_2 \\ 1 + a_1 + a_2 \\ 1 + a_1 + a_2 \\ a_1 + a_2 \\ a_2 \end{bmatrix} \tag{60}$$

We note three features of the net

(i) If we redraw the example net of Figure 4 (a) as in Figure 4 (b) we can see how it is incorporating the constraints

$$e_i = y_i + a_i y_{i-1} + a_2 y_{i-2} \tag{61}$$

and forming the output as the sum of m estimation errors e as a result, in

$$\sigma_k = e_k + e_{k-1} + e_{k-2} + e_{k-3} + e_{k-4} \tag{62}$$

(ii) A feature of the weight values is that they are a function of the two variables and the constant in $[1 a1 a2]^T$ and that $L + 1 - 2p$ of the weights from w_p to w_{L-p} have the same value of

$$w_i = \sum_{j=0}^{p} a_j; \quad a_0 = 1, \; p \leq i \leq L - p \tag{63}$$

(iii) For a data set of say N points, the first network window is $(y_1, y_2, ..y_7)$, the second window, shifted along $m = 5$ points $(y_6, y_7, ..y_{12})$ and so on, thus

$$\begin{aligned} \boldsymbol{J} &= (e_3 + e_4 + e_5 + e_6 + e_7)^2 + (e_8 + e_9 + e_{10} + e_{11} + e_{12})^2 + \dots \\ &\quad + (e_{N-4} + \dots + e_{N-1} + e_N)^2 \end{aligned} \tag{64}$$

and if the e_i are uncorrelated then

$$\boldsymbol{J} \rightarrow \boldsymbol{E} = e_3^2 + e_4^2 + \dots + e_N^2 \tag{65}$$

3.7. Alternative Form of Network

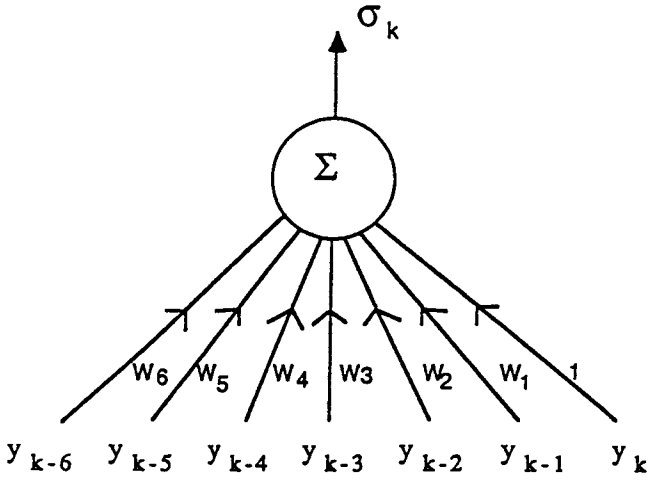
Because of the form of the solution for the weights

$$\boldsymbol{w} = \boldsymbol{I}_{L,p} + \boldsymbol{S}_m \boldsymbol{a} \tag{66}$$

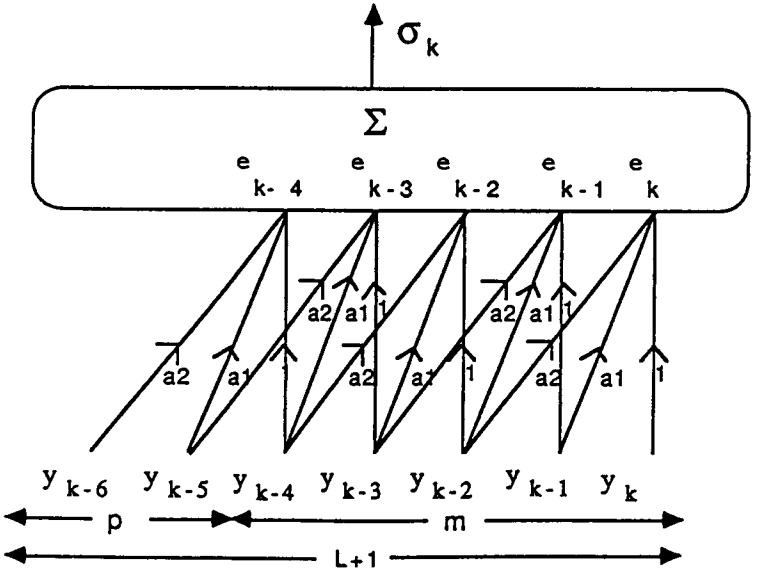
or viewed in another way, since

$$\begin{aligned} \sigma_k &= \boldsymbol{I}_m^T e_m \\ &= y_k + a_1 y_{k-1} + a_2 y_{k-2} + \cdots + a_p y_{k-p} \\ &+ y_{k-1} + a_1 y_{k-2} + a_2 y_{k-3} + \cdots + a_p y_{k-p-1} \\ &\vdots \\ &+ y_{k-L+p} + a_1 y_{k-L+p-1} + a_2 y_{k-L+p-2} + \cdots + a_p y_{k-L} \end{aligned} \tag{67}$$

$$\sigma_k = \boldsymbol{Y}_k + a_1 \boldsymbol{Y}_{k-1} + a_2 \boldsymbol{Y}_{k-2} + a_p \boldsymbol{Y}_{k-p} \tag{68}$$



(a)



(b)

Figure 4. Examples L=6, p=2 (a) single layer model (b) single layer model with expanded weight values.

where

$$\begin{aligned} \boldsymbol{Y}_k &= (y_k + y_{k-1} + \cdots + y_{k-L+p}) \\ &= \boldsymbol{I}_m^T \boldsymbol{y}_k \end{aligned} \tag{69}$$

and

$$\sigma_k = (e_k + e_{k-1} + \cdots + e_{k-m+1}) \tag{70}$$

This linear transformation of y_i defines a new filter and connectionist model such as shown in Figure 5 for the $L = 6$, $p = 2$ case with

$$\boldsymbol{Y}_k = (y_k + y_{k-1} + y_{k-2} + y_{k-3} + y_{k-4}) \tag{71}$$

and

$$\sigma_k = (e_k + e_{k-1} + e_{k-2} + e_{k-3} + e_{k-4}) \tag{72}$$

This form of filter, eqn (67), can be solved for min J directly by the linear predictive analysis of Section 2.

Again it can also be computed by steepest descent with

$$\Delta a_i = -\eta \sigma_k Y_{k-i} \tag{73}$$

We notice that if the e_k are uncorrelated, $< e_i e_j >= 0$; $i \neq j$ then

$$J \rightarrow \boldsymbol{e}^T \boldsymbol{e} \tag{74}$$

Then again the alternative form of the network solves the conventional lp problem, if the estimation errors are uncorrelated.

4. RESULTS

A few preliminary results are now given. These are based on the two speech waveforms shown in Figure 6 (a), for the fricative SH and (b) for the voiced XX.

4.1. Weights

Results for the weights $\boldsymbol{a}$ for the two single layer connectionist models, which minimise J, for the two sounds given in Table 1. These were evaluated by an lp analysis of the alternative linear form of network of Subsection 3.7 for $N = 255$, $p = 10$ and values $m = 1, 2, 10$ corresponding to $L = 9, 10, 19$. In each case the Y_k vectors of eqn (68) were Hamming windowed and the speech data was differenced.

It can be seen that the weight vectors for each sound, which would be constant if the e were correlated, are not particularly constant. Also we would expect them to be more constant in the case of the fricative SH, than the varied sound XX, since in the former case the excitation is less correlated.

However it is well known that the coefficients a are sensitive to the excitation even in the $m = 1$, lp case and a strict constancy cannot be expected as a result. Also a rather longer dataset would be needed for the lack of correlation of e to make itself felt.

Table 1. Weight vectors **a** for the two waveforms.

	m	1	2	10
SH	a[0]	1.0000e+00	1.0000e+00	1.0000e+00
	a[1]	1.9182e+00	1.0612e+00	9.5530e-01
	a[2]	3.1613e+00	2.2368e+00	1.2481e+00
	a[3]	3.6465e+00	1.7159e+00	6.5549e-01
	a[4]	3.3922e+00	1.8578e+00	-1.4257e-02
	a[5]	2.4098e+00	7.7389e-01	-6.2322e-01
	a[6]	1.4004e+00	6.3805e-01	-5.4530e-01
	a[7]	6.1027e-01	4.7522e-02	-1.6082e-01
	a[8]	2.6004e-01	1.7009e-01	1.5763e-01
	a[9]	8.8384e-02	8.7972e-03	2.0119e-01
	a[10]	8.1398e-02	7.3252e-02	3.4179e-01
XX	a[0]	1.0000e+00	1.0000e+00	1.0000e+00
	a[1]	-5.7751e-01	-1.3096e-01	-1.4638e+00
	a[2]	-2.7614e-02	9.9194e-01	4.7755e-01
	a[3]	1.3903e-01	-6.1721e-01	1.2726e-01
	a[4]	5.8045e-01	1.0194e+00	7.3267e-01
	a[5]	-4.8825e-01	-1.1269e+00	-1.0681e+00
	a[6]	1.2586e-01	7.7592e-01	1.2313e-01
	a[7]	1.7913e-01	-2.1720e-01	4.2442e-01
	a[8]	7.2381e-02	1.1211e-01	3.6772e-02
	a[9]	-1.7274e-01	-5.1237e-02	-1.4219e-01
	a[10]	4.9896e-01	2.6461e-01	7.2076e-02

Table 2. Normalised output cost functions J'.

m	1	2	10
SH net input SH	2.9233e-02	1.6996e-02	3.3537e-03
SH net input XX	3.1062e+01	1.7694e+01	7.1154e+00
XX net input XX	1.6661e-01	3.3856e-02	1.1707e-02
XX net input SH	1.0611e+00	6.2910e-01	2.7708e-01

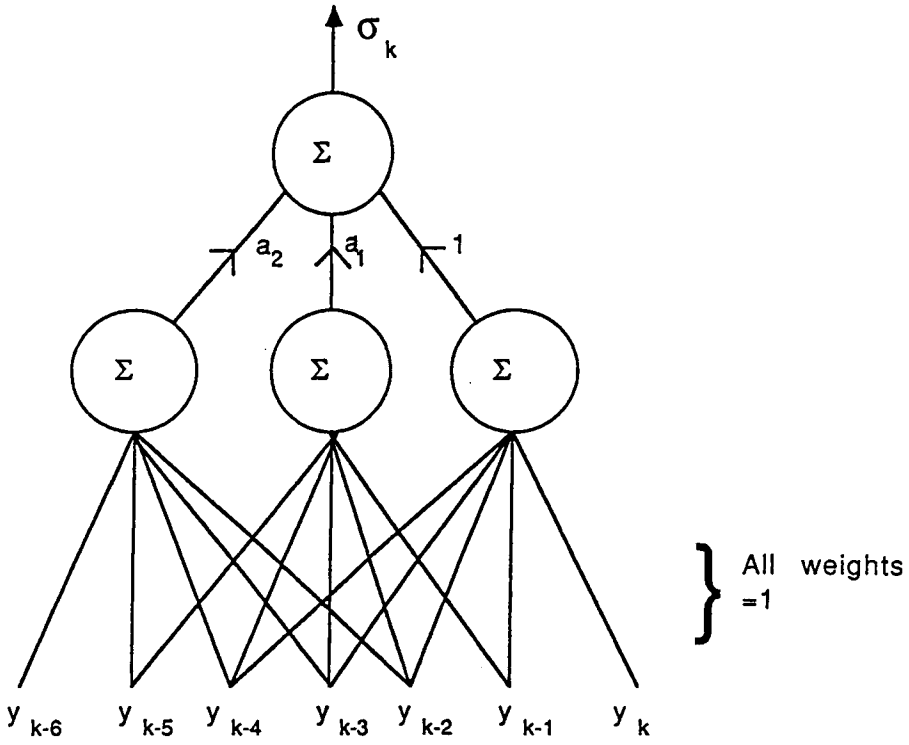


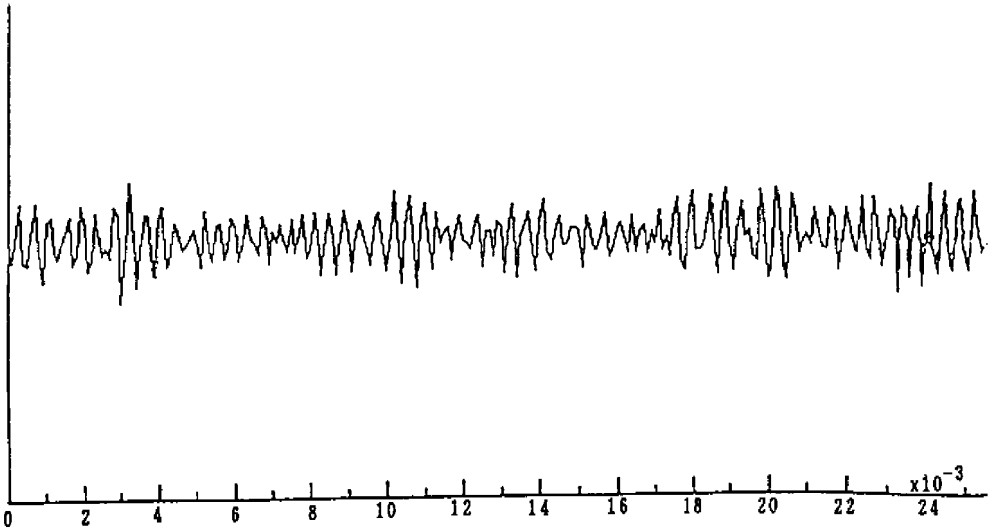
Figure 5. Alternative form of network.

4.2. Cost Functions

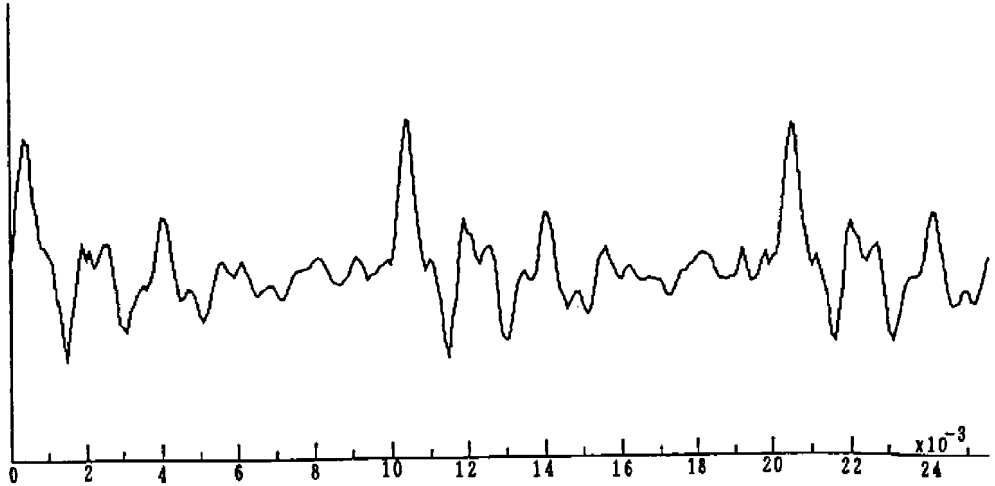
A more practical test of the value of the class is how well they classify different sounds. This is investigated here in a preliminary way by computing the output cost functions and the results are given in Table 2. Here, using the values of a shown in Table 1 above, the value of the normalised output cost function J for the SH networks with input SH and then input XX are given for $m = 1, 2, 10$ and then similarly for the XX networks. In each case the alternative form of networks was used, the Y_k vectors were Hamming windowed, the speech data was differenced and $J' = J/(\text{energy of } Y_k \text{ data})$.

We see that for this limited set of data the nets are performing a useful classification of the two sounds. Also that where a net has its 'own' sound as input J' is broadly constant as would be expected from Subsection 3.6. It would be precisely constant if the estimation errors were uncorrelated.

These results are perhaps the most practical attributes of the class, suggesting that a form of connectionist vector quantisation (CVQ) structure as shown in Figure 7 is



(a)



(b)

Figure 6. Speech data (a) SH (b) XX.

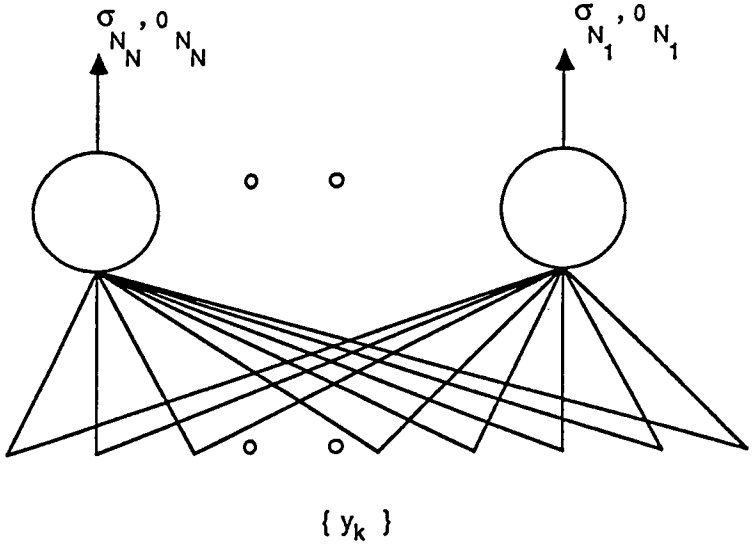


Figure 7. Sets of nets for connectionist vector quantisation (CVQ) classification.

possible, where a set of connectionist models of the type studied are trained to span an appropriate region of a space, analogous to conventional VQ structures [5]. The codebook entry for the current speech data is then specified by the network $O_{N,n}$, with the lowest output cost function. Results for this will be published later.

5. CONCLUSIONS

The paper has concentrated on the analysis of speech by connectionist models directly in the time domain, rather than in the frequency domain and including the constraint that the input speech can be modelled by a linear predictive process.

It has focussed on a class of networks where the order p of the linear predictive process is specified and produced an analytical solution for the weights of this class to minimise the network cost function. It has shown that when the estimation errors are uncorrelated this solution is the same as that of conventional lp analysis. The class chosen forces the analysis to be pth order, the effect of varying p for given data to establish global minima has not been explored in this paper.

The few preliminary results given show that the nature of speech, with its correlated excitation mitigates against the weight values of the class corresponding exactly to the conventional lp coefficients, as predicted by the theory. However as in the case of conventional lp analysis, good classification of sounds has been indicated by preliminary results and this in turn has suggested a form of connectionist vector quantisation (CVQ) structure.

Finally theweight values of the class exhibit an interesting uniformity, which speculatively might have a physiological analogue.

References

1. B. Atal and S. L. Hanauer: "Speech analysis & synthesis by linear prediction," J. Acoust. Soc. Amer., 50, pp.637-655, 1971.

2. D. E. Rumelhart and J. L. McClelland: "Parallel distributed processing: Explorations in the microstructure of cognition," Bradford/MIT Press, 1986.

3. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang: "Phoneme recognition using time-delay neutral networks," ATR Interpreting Telephony Research Laboratories, Tech. Rep. TR-1-006, 1987.

4. T. Parsons: "Voice & speech processing," Mc-Graw-Hill, 1986.

5. A. Buzzo, A. H. Gray, R. M. Gray and J. D. Markel: "Speech coding based on vector quantisation," IEEE Trans., vol. ASSP-28, 5, pp.562-574, 1980.

Chapter 3

FEATURE EXTRACTION

This Page Intentionally Left Blank

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Phoneme Recognition in Continuous Speech Using Feature Selection Based on Mutual Information

Katsuhiko Shirai, Noriyuki Aoki and Naoki Hosaka

Department of Electrical Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169 Japan

Abstract

An optimal statistical method to recognize phonemes in continuous speech is discussed. The novelty of this method is the evaluation of the effectiveness of acoustic features in each acoustic level using the criterion of mutual information between acoustic feature vectors and phoneme labels assigned to speech wave. In the proposed method for phoneme recognition, the power and its variational pattern, the LPC Mel-Cepstrum and its pattern of temporal change are adopted as the acoustic features. Multi-level clustering is suitable to discriminate phonemes by detecting the most reliable features in that context and using on effective combination of the various acoustic characteristics.

1. INTRODUCTION

In order to construct a large-vocabulary continuous speech recognition system, it is very important to develop a highly reliable phoneme recognizer.

Phoneme characteristics which are reliable enough for phoneme discrimination, do not necessarily correspond to the acoustic features obtained by short-time analysis of a frame. Therefore the temporal pattern of features over a suitable length for the features should be considered and the various kinds of contextual effects should be organized as compact as possible to classify phonemes.

The proposed method is completely statistical, relying on the traditional clustering method. However, in this direction of approach, recently many new trials have been performed to perform clustering of the time frequency pattern of speech waves considering a large time span or more complex context. the speciality of our method is to do clustering based on the mutual information criterion.

2. CLUSTERING ANALYSIS

2.1. Acoustic Analysis

The total system is shown in Figure 1. One part is the Multi-level clustering of acoustic features which is optimized to give the maximum information characterizing the phonetic nature in a frame unit. The second part is phoneme recognition by accumulating the probability obtained for each frame.

The speech database consists of 100 city names uttered twice by 12 male speakers, and the reference data which forms each codebook and phoneme dictionary is the first utterance of 6 speakers (group A). The second utterance of group A is used as the test data. And, further, the data uttered by the other 6 speakers (group B) is used in the speaker independent recognition test. The sampling frequency is 12.5kHz. The frame length is 20ms. and each frame is analyzed using 10ms intervals. Input speech is pre-emphasized by a differential filter and a Hamming window is applied.

Level 1 Coding of the power-change pattern that shows the power pattern of the speech waves including the neighboring frames. the power pattern is vector-quantized using the optimal code book. this code is effective in classifying each frame into several groups which reduce the entropy of phonemic labels and makes the classification in the following level easier.

Level 2 The LPC Mel-Cepstral coefficients are vector-quantized, from which the code book is made for each group classified by the power code of level 1.

Level 3 The LPC Mel-Cepstrum changes are extracted using the regression coefficient of each order from neighboring frames.

Level 4 The power code of each frame is concatenated to a group of power code appearance sequence in the neighboring frames. This means that new code is generated considering the dynamic features.

The Phoneme label of each frame can be determined by combination of these four codes. The phoneme dictionary contains the conditional probability of the phonemes after the above four codes are given, and the conditional entropy of each code. To determine the optimal phoneme, the most important point is that an effective combination of the feature set to discriminate phonemes is very critically dependent on the context and different for every utterance even if the context seems to be the same. Therefore, a complicated combination of the codes of neighboring frames should be carefully examined in this process.

2.2. Integration of Clusters Based on Mutual Information

The set of phonetic labels is denoted by X and the set of acoustic features by Y. If the code $y_l \in Y(l = 1, \ldots, n)$ is obtained and the label $x_k \in X(k = 1, \ldots, m)$ is observed, the

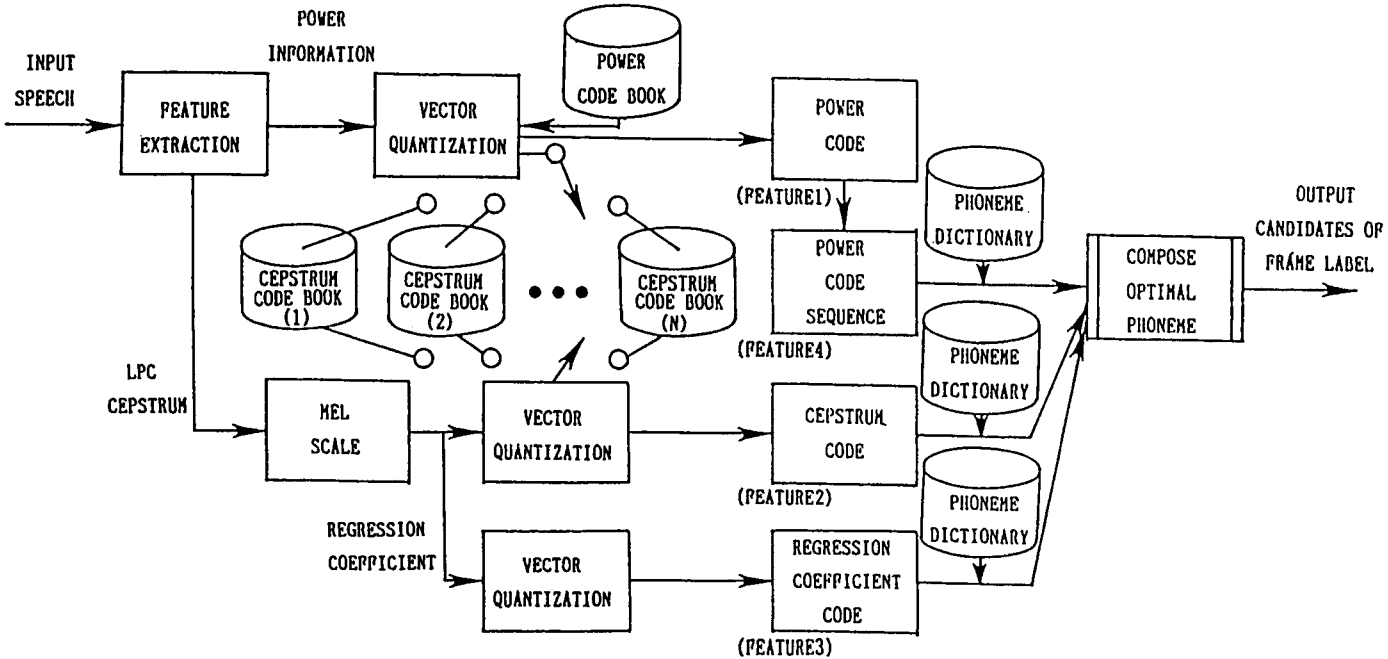


Figure 1. Block diagram of the system.

conditional probability $P(x_k|y_l)$ can be defined. The conditional entropy of X, $H(X|y_l)$ is defined as

$$H(X|y_l) = \sum_{k=1}^{m} -P(x_k|y_l) \cdot \log P(x_k|y_l). \quad (1)$$

and the mutual information $I(X;Y)$ between acoustic feature and the phoneme label is calculated from the entropies $H(X)$ and $H(X|Y)$.

There often occur several redundant clusters which represent the same categories of phonemes, and the separation of those clusters is useless for the phoneme recognition. Therefore, analyzing phonetic characteristics of each cluster in detail, several clusters can be integrated. The following merging algorithm decreases the number of centroids, maintaining the value of mutual information between the class of acoustic feature and the phonemic label.

step 1: for $i = 1$ to n
for $k = i + 1$ to n
$P(X|y_{ik}) = P(X|y_i) + P(X|y_k)$
$I_{ik}(X|Y) = H(X) - H_{ik}(X|Y)$

step 2: find i, k maximizing $I_{ik}(X|Y)$ $(i = k = 1, \ldots, n : i \neq k)$
$P(X|y_{ik}) = P(X|y_i) + P(X|y_k)$
$n = n - 1 \rightarrow$ step 1

n	:	the number of centroids
$P(X\|y_{ik})$	:	conditional probability
I_{ik}	:	mutual information
$H_{ik}(X\|Y)$	:	conditional entropy when clusters i and k are merged

2.3. Phoneme Information in the Power

Figure 2 shows the result of vector quantization using the level 1 feature. In the first level, the role of the clustering is to make a rough classification of the phoneme groups. Then, in this level, phoneme categories which have the same characteristics are gathered into a smaller member of groups. The more the number of centroids increases, the more the mutual information and recognition rates increase. the recognition rate saturates at $N = 128$ when the simple clustering technique is applied (solid line).

Starting from this situation the above merging algorithm is applied. The dashed line shows the result. The ability to classify phonemes is not very different for $N = 16$ and $N = 128$, and the performation reaches about 1.0 bit. In the following experiment, this set of code is adopted as level 1.

3. CLUSTERING USING POWER CODE SEQUENCE

The power code represents the pattern of power change over 7 frames. However, we can find further contextual information in sequence of power code, since there are strong correlations or very frequent patterns in the consecutive power codes.

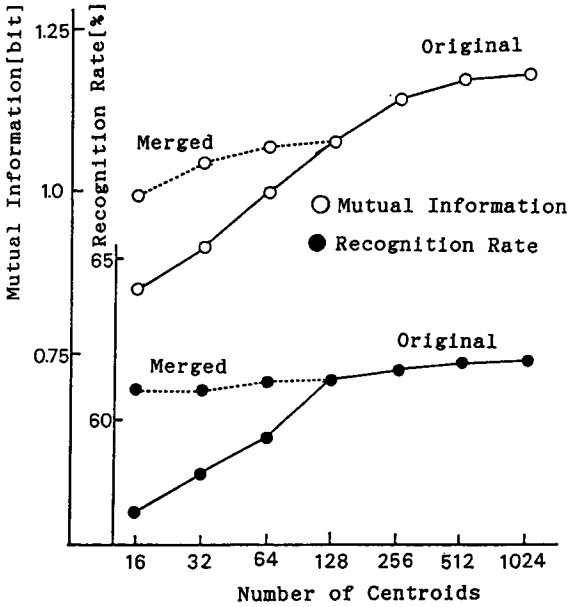


Figure 2. Characteristics of phonemes including in cluster using merged power code ($N = 16$).

To extract such contextual characteristics, we propose an algorithm which introduces new clusters to maximize the mutual information using the combined pattern of power codes (p-code). We define a block of power codes If the same code continues for several frames, those frames will be aggregated into one block code (b-code). An example is shown in Figure 3. The following sequence which consists of five blocks (both the previous and succeeding two blocks of the analyzed frame) is considered, where the block code is expressed as $C(h)$.

$$\text{Code sequence : } (C(h-2), C(h-1), C(h), C(h+1), C(h+2)).$$

If we select the position ($k = 2, 1, -1, -2$) and the block code of the position j, a new cluster can be generated. For example, (*,*,5,9,*) means that we use the cluster $C(h) = 5$ and consider the class where the block code which follows No.5 is No.9.

If j and k are prescribed, the probability of frame label x for the cluster i (code y_i) is supposed to be $P_{jk}(x|y_i)$, the difference of the probability is defined by;

$$\Delta P(x|y_i) = P(x|y_i) - P_{jk}(x|y_i).$$

i : the cluster which should be divided $(1, \ldots, n)$.
j : the position of block code (2, 1, -1, -2).
k : the block code prescribed at $k(1, \ldots, m)$.

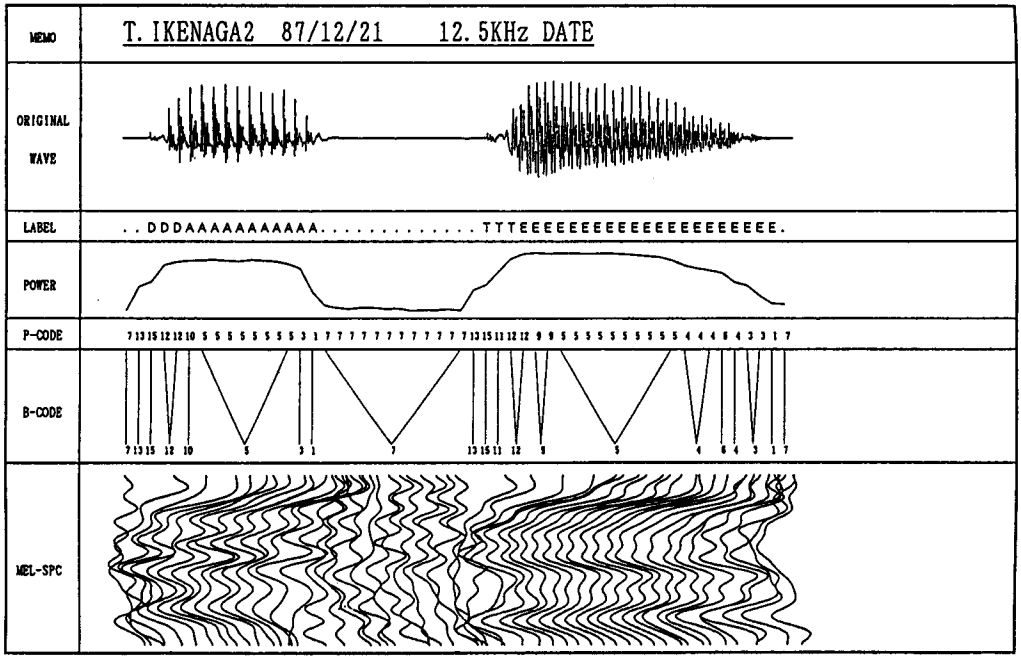


Figure 3. Acoustic features and block codes.

The mutual information $I_{ijk}(X|Y)$ given by the combination of i, j, k can be calculated using $\Delta P(X|Y_i)$. Therefore, if a set of i, j, k is found to maximize the mutual information, we select it as a new cluster.

Example : (*, *, C_5, *, *), $C(h) = 5$
with $k = 1, j = C_3$
separated new cluster (*, C_3, C_5, *, *)
residual cluster (*, $\tilde{C}_3$, C_5, *, *)

After the global classification explained in section 2, still large ambiguity remains and the entropy is about 3.0 bit. For example, the content of cluster No.5 is the central part of vowel sounds but 10% of them comes from semi vowels.

If we use the block codes is adopted as level 4. the number of clusters of level 4 is 256.

4. RECOGNITION EXPERIMENT

To extract a certain phoneme characteristics of one frame, it is necessary to use information from the adjacent frames. To make optimal phoneme decision depending on

neighboring frames, we consider the weighing the probability by the conditional entropy of each code. Then the probability based on the effective feature is emphasized by the lower entropy.

$$C(x_i) = \prod_{j=-M}^{M} \prod_{k=1}^{N} P(x_i|y_{jk})/H(X|Y_{jk}) \tag{2}$$

$2M+1$	:	the number of considered frames.
N	:	the number of features for recognition.
y_{ik}	:	the code of features j at the frame preceding or succeeding j.
$P(x_i\|y_{jk})$	:	the conditional probability.
$H(X\|Y_{jk})$	:	the conditional entropy.
$C(x_i)$	:	the score of frame label.

The recognition rate at the frame level was 92.1% and 81.6% for vowels and all phonemes, respectively. The case of $M = 7$ gave the highest score. If weighing by the conditional entropy was not applied, the recognition rate decreased about 8% and 10%, respectively. To obtain the final phoneme sequence, the phoneme segments should be determined. In the next experiment of phoneme recognition, we provide the segment marks manually. The most suitable phoneme sequences can be obtained by calculating the score of each phoneme in each segment from the results of the frame label output.

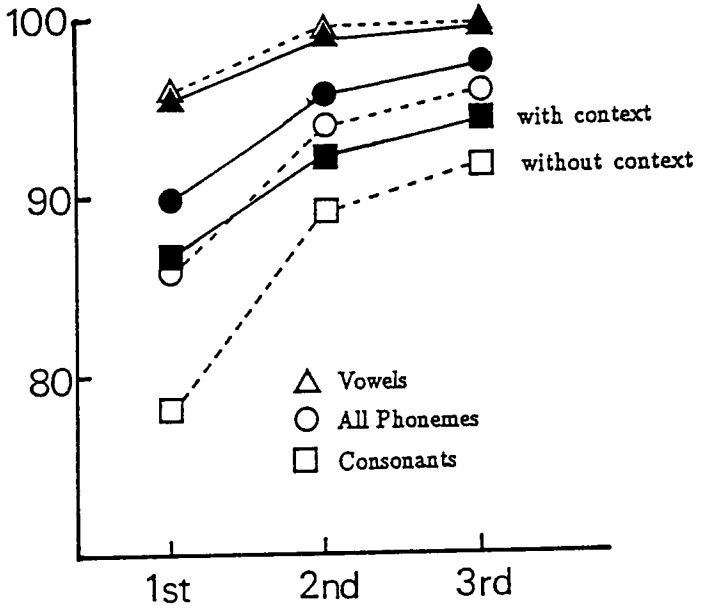
The phoneme recognition rate varies according to the level 1 codes and if the difference of the scores between the 1st and 2nd candidates is larger, the result is more reliable. Therefore, first we select the phoneme segments where the score is high enough and a correct phoneme decision can be done, and under the phoneme context assumption that those phoneme segments are actually existing, other segments are recognized.

Figure 4 shows a comparison of the recognition rate with and without considering the phoneme context. The recognition rate for consonants improves by considering the context, which is given usually by the preceding and following vowels. The recognition rates of vowels, all phonemes, and consonants were 94.8%, 90.0% and 86.9% for the second set of the first 6 speakers and in the speaker independent case 91.8%, 85.9% and 82.0%, respectively.

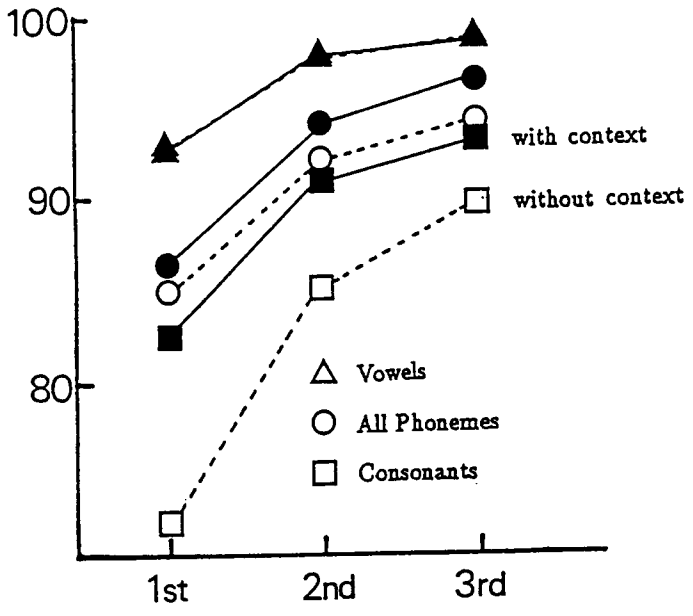
5. SPEAKER ADAPTATION

5.1. Expression of Speaker Individuality

The Phoneme recognition rate in the speaker independent case (group B) is considerably lower than that in the speaker dependent case (group A). Therefore, in order to improve the performance, a speaker adaptation method should be introduced there have been several researches on speaker adaptation in a vector quantization environment. In this paper, that one is considered with the same framework of the multi-level clustering scheme. The conditional probabilities stored in the phoneme dictionary must be speaker characteristics is a rather difficult task, modification of the centroid vectors is attempted.



(a) multi speaker (A group)



(b) speaker independent (B group)

Figure 4. Comparison of the recognition rates of phonemes with and without considering the context.

At first, the mean spectral differences ΔV_i for the average vectors of the 5 vowels V_i (i = 1, 2, 3, 4, 5) is calculated for a new speaker K.

A vector $S(K) = \{\Delta V_1, \ldots, \Delta V_5\}$ or its reduced expression by principal component analysis can represent the speaker individuality. For the calculation of ΔV_i, utterance of 10 words is used.

Figure 5 shows the separability of speakers by $S(K)$, that is the minimum ratio of two distances, one is with the talker K himself and the other is obtained between K and another talker nearest to K.

The recognition rate of vowels when the individual dictionary is changed to is shown if Figure 6. The recognition date is inversely proportional to the distance $|S(K) - S(J)|$.

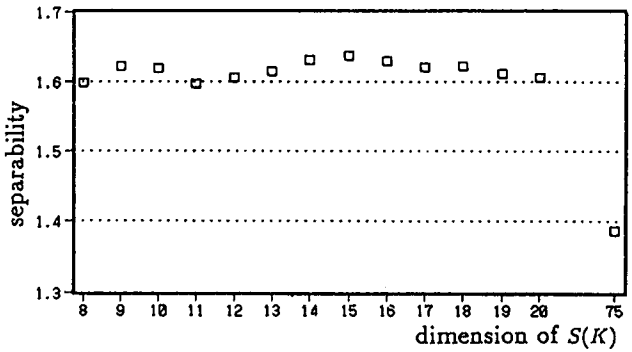


Figure 5. Separability of speaker by $S(K)$.

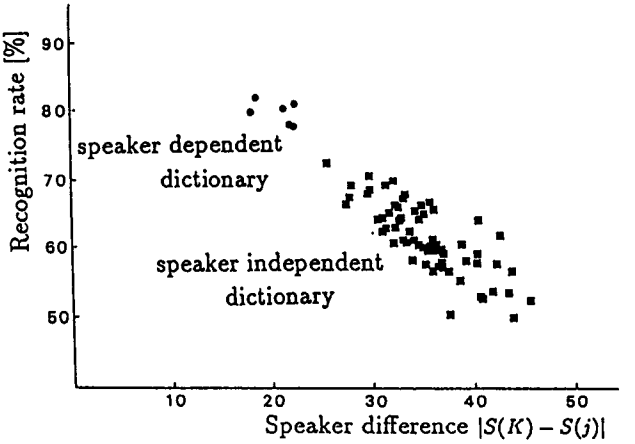


Figure 6. Relation between vowel recognition rate and speaker difference.

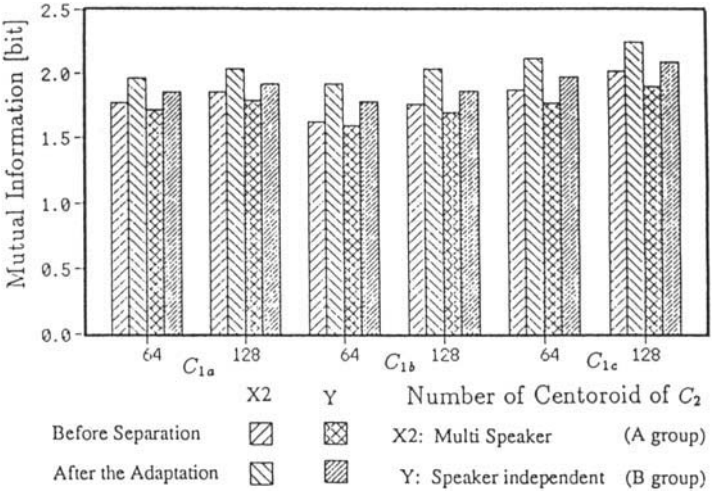


Figure 7. (a) Mutual information given by the level 2 feature.

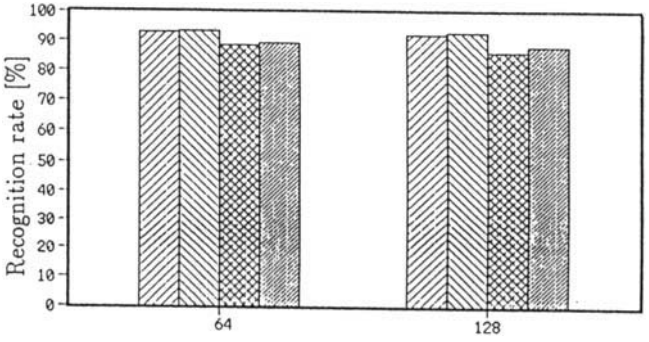


Figure 7. (b) Improvement of the recognition rate by the separation of dictionary.

5.2. Adaption Procedure

Each acoustic feature space is divided into N_s groups by the speaker individuality parameter. In the next experiment, three clusters $C_{1i}(i = a, b, c)$ in level 1 are subjected to speaker adaptation, in which the main part of the speech frames corresponding to the steady part, the ascending and the descending part of vowels are included. Therefore, for each cluster of C_{1i}, N_s groups of clusters of Level 2 are constructed.

The centroids of cluster C_{2j} which is in the central part of the vowels are modified by the average value of the data coded to C_{2j} in the learning phase.

Figure 7(a) shows that the mutual information given by the level 2 codes increases by the separation of the codebook into N_s groups ($N_s = 4$). And the vowel recognition rate by the level 2 feature is shown Figure 7(b). It is seen that the mutual information that decides the category of the vowel can increase by separating the dictionary of level 2 corresponding to the speaker. Therefore, it is shown that multilevel clustering can be also applied for speaker adaption by introducing another parameter which represents speaker individuality.

6. CONCLUSION

A new method to organize a hierarchical phoneme recognizer for continuous speech was presented. The high performance of the system is due to the effective hierarchical clustering of acoustic features based on the mutual information between the acoustic feature and the phoneme label. And also two considerations concerning the context, one is the power pattern and the other is the phonemic one, were verified to be very useful for phoneme recognition.

And, further, the possibility for speaker adaptation was shown in the same frame work.

References

1. S. Sagayama: "Phoneme Environment Clustering for Speech Recognition," *Proc. ICASSP-89*, pp.397-400, 1989.

2. S. Nakajima and H. Hamada: "Automatic Generation of Synthesis Units Based on Context Oriented Clustering," *Proc.ICASSP-88*, pp.659-662, 1988.

3. Y. Linde, A. Buzo and R. M. Gray: "An Algorithm for Vector Quantizer Design," *IEEE Trans. Comm.*, vol.COM-28, No.1, pp.84-95, 1980.

4. K. Mano, S. Ishige and K. Shirai: "Phoneme Recognition in Connected Speech Using Both Static and Dynamic Properties of Spectrum described by Vector Quantization," *Proc.ICASSP-86*, pp.2243-2246, 1986.

5. K. Shirai, N. Aoki and N. Hosaka: "Multi-Level Clustering of Acoustic Features for Phoneme Recognition Based on Mutual Information," *Proc ICASSP-89*, pp.604-607, 1989.

6. H. Matsumoto, Y. Yamashita and S. Nishizawa: "Unsupervised speaker adaptation for speech recognition based on a minimum fuzzy objective function criterion," *Research Report of PASL*, 1988.

7. Y. Niimi and Y. Kobayashi: "Speaker-Adaptation of a Code Book of Vector Quantization," *Proc. of ASJ Autumn Meeting*, 1988.

8. K. Shirai, N. Hosaka and E. Kitagawa: "Speaker Adaptive Phoneme Recognition by Multi-level Clustering Based on Mutual Information Criterion," *Proc.ICASSP-90*, pp.169-172, 1990.

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Dependency of Vowel Spectra on Phoneme Environment

Tetsunori Kobayashi

Department of Electrical Engineering, Hosei University
3-7-2 Kajino-cho, Koganei, Tokyo, 184 Japan

Abstract

Conventional quantification theory and a new theory are applied to investigate the dependence of vowel spectra on the phoneme environment. Conventional quantification theory assumes that the influence of multiple categories can be expressed as a linear combination of that of the single categories. In the new method, the theory is modified to deal with cross-category factors. Using this method, the multiple correlation coefficients are improved by 6.7 ~ 9.1% compared with the conventional linear method and some nonlinear categorical factors which affect the vowel distribution became apparent.

1. INTRODUCTION

It is well known that the spectrum of a phoneme is influenced by the manner and the place of articulation around it. However, the quantitative relation with these effects is not clear. It is very important to reveal it. If the relation between phoneme environment and spectra can be modeled effectively, it becomes possible to adjust reference according to the context and a high-performance speech recognition system may be realized. As for synthesis by rules, it becomes possible to adjust the target spectra according to the context and a high quality may be obtained.

In this paper, we apply quantification theory to this problem. However, conventional categorical factor analysis methods, such as quantification theory [1], assume that the influence of multiple categories can be expressed as a linear combination of that of the single categorise. This assumption makes it difficult to construct a strict model because there are so many nonlinear factors with contextual influence. For example, the influence of the next preceding sound highly depends on the preceding sound. These relations must be dealt with nonlinear model.

In this paper, we propose a nonlinear categorical factor analysis method and apply this method to vowel data in word speech. In the following section, conventional quantification theory is surveyed briefly, and then, a modified quantification theory which can deal with nonlinear factors is proposed. In section 4, an outline of the experiments using this theory is presented. In section 5, the experimental results are described.

2. CONVENTIONAL QUANTIFICATION THEORY

In conventional quantification theory, the criterion variable (variable to be modeled) $x^i(i = 1, 2, \ldots, N)$, is represented by summation of dummy variables, $d^i_{jk}(j = 1, 2, \ldots, J; k = 1, 2, \ldots, K_j)$, with coefficients called category weight, $w_{jk}(j = 1, 2, \ldots, N; k = 1, 2, \ldots, K_j)$, where d^i_{jk} depends on whether the data x^i belongs to the k-th category of the j-th item or not (if that is true, $d^i_{jk} = 1$, else, $d^i_{jk} = 0$), and N, J and K_J are the number of sample data, item and category of j-th item, respectively.

$$x^i = \overline{x} + \sum_j \sum_k w_{jk} \cdot d^i_{jk} + e^i. \tag{1}$$

Here, item means the view point from which data will be formulated. For example, if we attempt to formulate the variation of the vowel spectra with the kind of preceding phoneme and kind of of succeeding phoneme, then 'preceding phoneme' and 'succeeding phoneme' are items. Category means the kind of preceding phoneme and the kind of succeeding phoneme in this case. The category weights represent the intensity of the influence of the categories. They are calculated by the least-squares method (LSM). In this model, the influence of multiple categorical factors on the variation of a criterion variable is represented as a linear combination of the influences of all categorical factors.

3. NONLINEAR QUANTIFICATION THEORY

In case that the influence of some category highly depends on some of the other categories, it is impossible to express the categorical effect by a linear model, so it is necessary to consider nonlinear factors. To extend the model to involve cross-terms seems to be a simple solution of this problem,

$$x^i = \overline{x} + \sum_j \sum_k w_{jk} \cdot d^i_{jk} + \sum_j \sum_{l \neq j} \sum_k \sum_m w_{jklm} \cdot d^i_{jk} \cdot d^i_{lm} + e^i. \tag{2}$$

In this case, however, the number of parameter becomes too large compared with the number of data to be analyzed. The matrix which is used in the least-squares method tends to be singular, and, in most cases, the equation cannot be solved. Even if the equation can be solved fortunately, the result is not reliable.

It is required to select only effective parameters and reduce the number of parameters in order to perform a reliable factor analysis. This problem can be formalized in the framework of the combinatorial optimization problem. Here, we propose two parameter selection methods, on the basis of the best first search and on an apprxoximation of this algorithm.

Before illustrating the methods, we describe the definition of the terms, operational symbols, and formulation which are used in the k.algorithms.

N : Sample number.
K : Total dummy variable number.
$x = \{(x^1, x^2, x^3, \ldots, x^N)\}^T$: Data sample to be modeled.
$e = \{(e^1, e^2, e^3, \ldots, e^N)\}^T$: Error vector.
E : power of error vector. $E = e^T \cdot e$
c^i_j : j-th dummy variable of sample x^i.

Dummy variables are prepared for all categories and for all possible combinations of categories. They are numbered sequentially. For example, in the case that there are two items and there are two categories for each item, we prepare four dummy variables, $c_1 - c_4$, for the single terms $d_{11}, d_{12}, d_{21}, d_{22}$ and four dummy variables, $c_5 - c_8$, for the cross terms $d_{11} \cdot d_{21}, d_{11} \cdot d_{22}, d_{12} \cdot d_{21}, d_{12} \cdot d_{22}$.

Ω : Universal set of dummy variables.
ω_S : Selected dummy variable set.
ω_N : $\omega_N = \Omega - \omega_S$
$\lambda^k(\omega)$ or $\lambda^k(\omega; w)$: model

$$x^i = \overline{x} + \sum_{(j|c_j \in \omega)} w_j \cdot c_j^i + e^i.$$

The formulation $\lambda^k(\omega)$ is used in the case of weight estimation, $\lambda^k(\omega; w)$ is used in case of error calculation with estimated weights w. k indicates the number of parameters.

$w \Leftarrow \lambda^k(\omega) < x$:
Estimate the category weights w by LSM with a k-parameter model $\lambda^k(\omega)$ and data x.

$e = \lambda^k(\omega; w) < x$:
Calculate the error vector with a k-parameter model $\lambda^k(\omega; w)$ and data x.

Using the above symbolic notation, the algorithms of the nonlinear quantification mechanisms are expressed as follows.

Nonlinear method I

$\omega_S = \{\}$;
$\omega_N = \Omega$;
for $k = 1$ to K {
  for all $c_j \in \omega_N$ {
    $w \Leftarrow \lambda^k(\omega_S \cup \{c_j\}) < x$;
    $e = \lambda^k(\omega_S \cup \{c_j\}; w) < x$;
    $E_j = e^T \cdot e$;
  }
  $\hat{j} = argmin E_j$;
  add $c_{\hat{j}}$ to ω_S;
  remove $c_{\hat{j}}$ from ω_N;
}
ω_S is the final dummy variable set;
w is the final category weight vector;

Nonlinear method II

$\omega_S = \{\}$;
$\omega_N = \Omega$;
for $k = 1$ to K {
  $y = \lambda^{k-1}(\omega_S; w) < x$
  for all $c_j \in \omega_N$ {
    $w \Leftarrow \lambda^1(\{c_j\}) < y$;
    $e = \lambda^1(\{c_j\}; w) < y$;
    $E_j = e^T \cdot e$;
  }
  $\hat{j}$ = argmin E_j;
  add $c_{\hat{j}}$ to ω_S;
  remove $c_{\hat{j}}$ from ω_N;
}
ω_S is the final dummy variable set;
$w \Leftarrow \lambda^K(\omega_S) < x$;
w is the final category weight vector;

Method I is constructed on the basis of the strict best first search algorithm. In this method, many times ([size of ω_N] times) a k-parameter estimation problem must be solved using LSM to find the k-th dummy variable, therefore, this method is computationally expensive. The method II is an approximation of method I. Since this method require only 1-parameter estimation, the calculation speed is much higher than for method I.

4. EXPERIMENTS

Experiments are done to investigate the context effect on the vowel spectra in the stationary parts using the linear model and the nonlinear model. The criterion variables and dummy variables used in the experiments are as follows.

A. Criterion variables

Vowel spectra in the stationary parts are used. They are selected from 5240 word tokens (ATR word data base) which are spoken by one speaker (speaker id.: MAU).

Speech is sampled at 12kHz and quantized into 16bit. Then, 13th order selective LPC analysis is performed in the 0-3 kHz band. Then, 20th order Cepstral coefficients are analyzed. Then, the first three principal components are calculated for each vowel. (Namely, the different eigenvectors are obtained for each vowel.) The eigenvalues of principal components are shown in Table 1. Proportions and accumulated proportions of them are shown in Table 2. These three principal components are used as criterion variables.

Table 1. Eigenvalue of each principal compoment.

	Eigenvalue		
	1st	2nd	3rd
a	0.1270	0.0453	0.0280
i	0.1007	0.0618	0.0326
u	0.2029	0.0886	0.0708
e	0.1083	0.0418	0.0384
o	0.1860	0.0580	0.0472

Table 2. Proprtion and accumulated proportion of each principal component.

	Proportion (Accumulated proportion)		
	1st	2nd	3rd
a	46.8 (46.8)	16.7 (63.5)	10.3 (73.8)
i	36.5 (36.5)	22.4 (59.0)	11.8 (70.8)
u	41.4 (41.4)	18.1 (59.5)	14.4 (73.9)
e	44.5 (44.5)	17.1 (61.6)	10.4 (72.0)
o	46.1 (46.1)	14.4 (60.5)	11.7 (72.2)

B. Items for dummy variables

(1) Linear model (conventional quantification theory):

(1-1) cV model — items: place of articulation of the preceding consonant, manner of articulation of the preceding consonant.
total number of dummy variables : 7(i,u), 8(a,e,o).

(1-2) vcV model — items: (1-1) + kind of preceding vowel of the preceding consonant
total number of dummy variables : 13(i,u),14(a,e,o).

(1-3) cVc model — items: (1-1) + place of articulation of the succeeding consonant, manner of articulation of the succeeding consonant.
total number of dummy variables : 15(i,u),16(a,e,o)

(1-4) Vc model — items: place of articulation of the succeeding consonant, manner of articulation of the succeeding consonant.
total number of dummy variables : 8.

(1-5) Vcv model — items: (1-4) + kind of succeeding vowel of the succeeding consonant.
total number of dummy variables : 16.

(2) Nonlinear model (nonlinear quantification theory):

(2-1) NL1 model (method I) — items: (1-3) + kind of preceding vowel of the preceding consonant.
total number of dummy variables : 16

(2-2) NL2 model (method II) — same as (2-1).

C. Categories for dummy variables

The categories for each item are shown in Table 3.

5. EXPERIMENTAL RESULTS

Figure 1 shows the multiple correlation coefficients versus the different principal components for the case of NL1 model. As for /a/, /i/ and /e/, good values are obtained for the first principal components. As for /u/, good values are obtained for the first and second principal components. As for /o/, all values are not so good. These results show that the main parts of the distributions of the vowels /a/, /i/, /u/, and /e/ can be explained in terms of the phonemic context. As for /a/,/i/ and /e/, only a one-dimensional distribution can be modeled, and as for /u/, a two-dimensional distribution can be modeled effectively. As for /o/, some other factor must be more essential, so it is impossible to model the distribution using only the phonemic context.

Figure 2 shows the multiple correlation coefficients for the different models. The values for the Vc model and the Vcv model are very low. This fact shows that it is impossible to model the vowel distribution without considering preceding sounds. For most vowels, the the value for the cVc model is the best of all linear models (1-1)-(1-5), the vcV model is

Table 3. Relations between items, categories and phonemes.

Item	Category	Phoneme
Place of articulation	Labial	p, b, m, w, f
	Alveolar	t, d, n, r, ts, z, s
	Palatal	k, g, j, ch, sh, h(i)
	Glottal	h(a, e, o)
Manner of articulation	Fricative	s, sh, ts, ch, f, h, z
	Plosive	p, t, k, b, d, g
	Sonorant	j, w, r
	Nasal	n, m
Kind of vowel	/a/	a
	/i/	i
	/u/	u
	/e/	e
	/o/	o
	/N/	N
	Word initial	<
	Word final	>

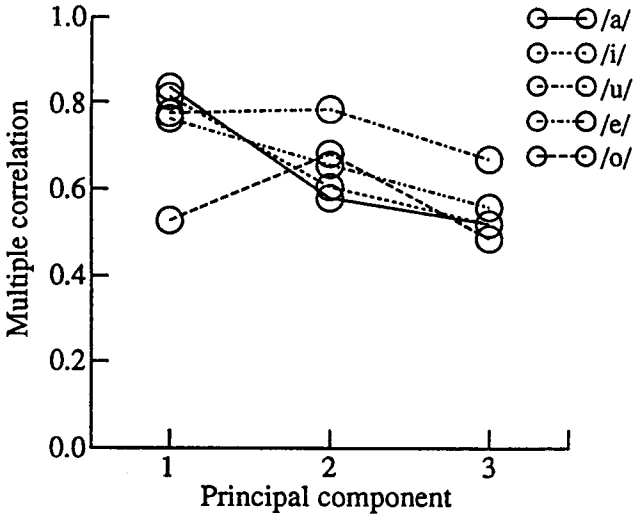


Figure 1. Multiple crrelation coeffients obtained using the NL1 model, versus the differnt principal components.

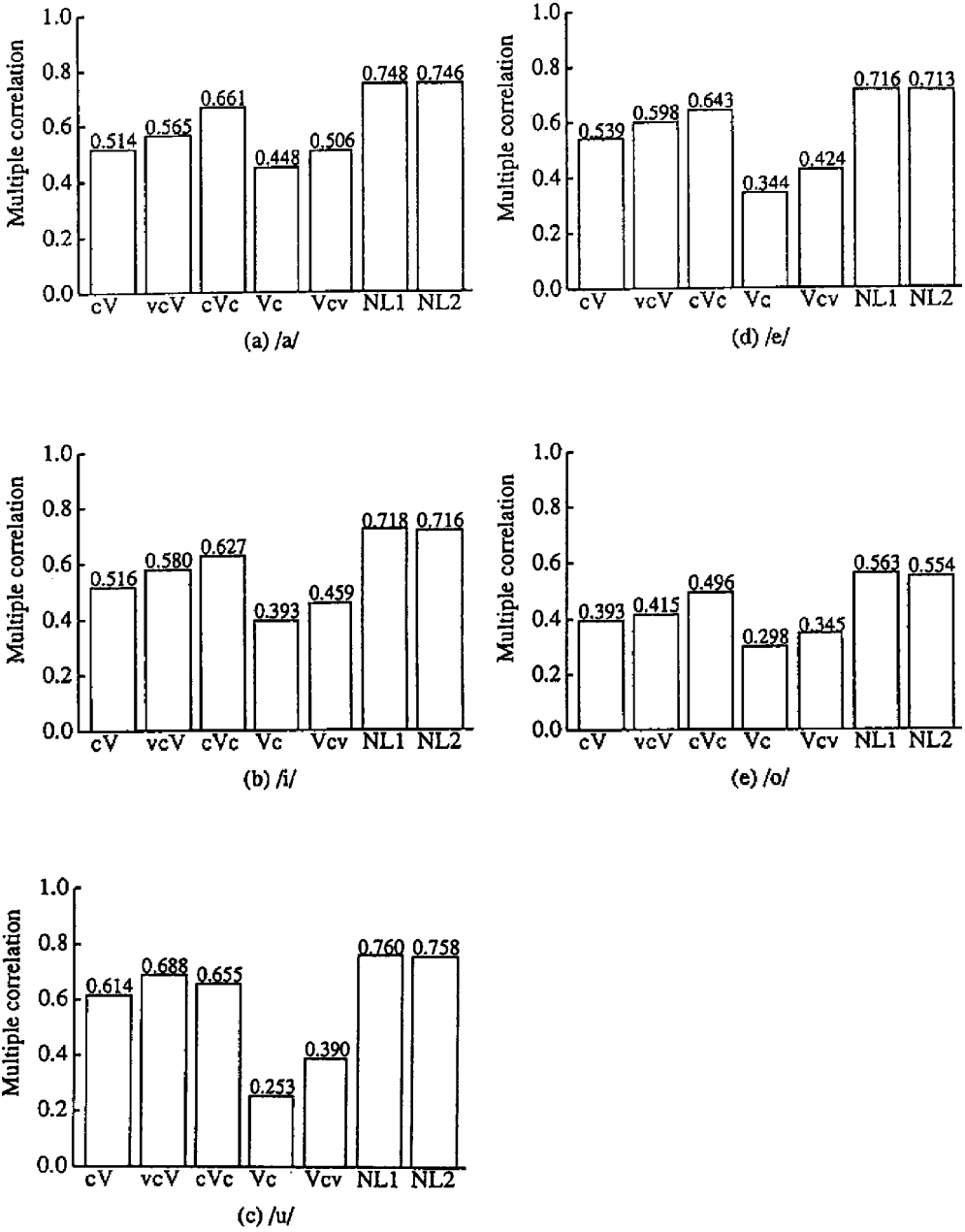
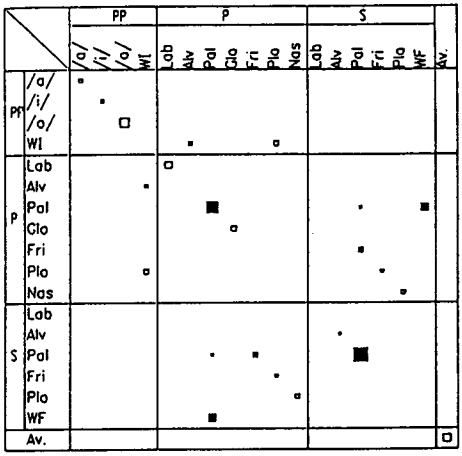
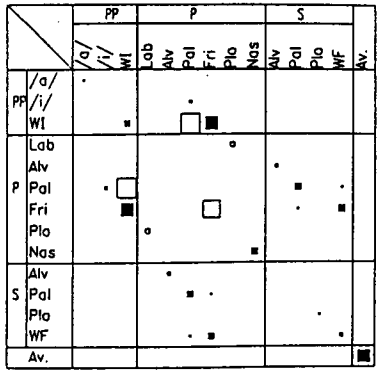


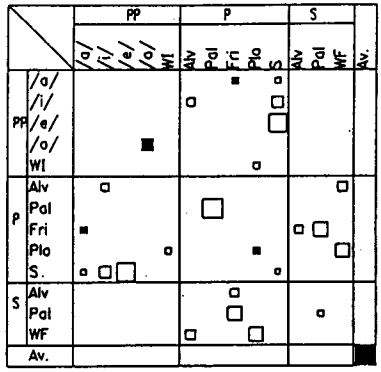
Figure 2. Multiple correlation coefficients for /a/ (a), /i/ (b), /u/ (c), /e/ (d), /o/ (e) for the different models.



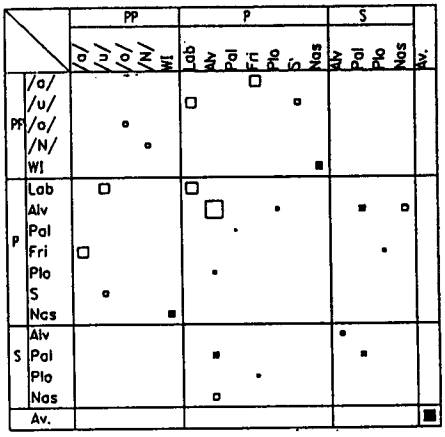
(a) /a/



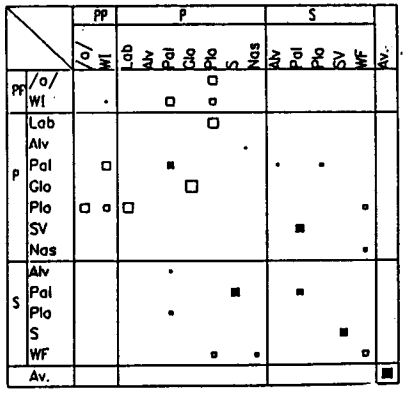
(b) /i/



(c) /u/



(d) /e/



(e) /o/

Figure 3. Category weights for the 1st principal component of /a/ (a), /i/ (b), /u/ (c), /e/ (d), /o/ (e) obtained using the NL1 model.

the second best, and cV model is the third. Only for the vowel /u/, the value of the vcV model is better than that of the cVc model. As for /u/, the influence of the next preceding sound is more important than that of succeeding sounds. As compared to the cV model, the cVc model improves the values of the multiple correlation coefficients by 14.7, 11.1, 4.1, 10.4 and 10.3% for /a/, /i/, /u/, /e/ and /o/, respectively. The model vcV improves the values by only 5.1, 6.4, 7.4, 5.9 and 2.2% as compared to the cV model. These results show that it is impossible to improve the multiple correlation in the framework of a linear model except for the vowel /a/. While, as for the nonlinear models, they improve the values by 14.6 - 23.4% as compared to the cV model. These results show the effectiveness of considering nonlinear factors. The values for the NL2 model are comparable to those for the NL1 model. Only a slight difference can be seen between them. Since the NL1 model is computationally very expensive, the NL2 model is the best of all models which are tested in these experiments.

Figure 3 shows the category weights obtained using the NL1 model for the first principal components of the vowels /a/ and /u/. In this figure, the color of the box indicates the sign : white for positive and black for negative. The size of the box corresponds to the absolute value. The diagonal entries in the figures indicate the category weights for corresponding categories and the other entries are the weights for the cross terms of two corresponding categories. From Figure 3 (a), it is found that /a/'s whose preceding or succeeding sound is palatal, are distributed far from other /a/'s. As for /a/, any nonlinear factor is not essential. From Figure 3 (b), it is fond that /u/'s with a succeeding palatal are distributed far from other /u/'s. In case that the next preceding sound is a front vowel (/i/ or /e/), the distribution tends to be similar to /u/'s with a succeeding palatal. However, this feature appears in only case that the preceding sound is sonorant.

6. CONCLUSION

A nonlinear categorical factor analysis method is proposed aiming at investigation of the influence of the phonemic context on the variation of vowel spectra. Compared with the conventional linear model, the multiple correlation coefficients are significantly improved.

In our nonlinear model, categories which contribute to reduce the deviation from the estimated target are selected automatically. Thus, it become possible to reveal some important categorical factors which affect vowel distribution.

References

1. C. Hayashi: "On the prediction of phenomena from qualitative data from the mathematico statistical point of view," *Annals of the Institute of Statistical Mathematics*, vol.3, pp.69-98.

2. T. Kobayashi and T. Matsuda: "A Categorical Factor Analysis of Vowel Distribution Based on the Modified Quantification Theory," *The second joint meeting of ASA and ASJ*, PP.1, 1988.

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

A Preliminary Study on a New Acoustic Feature Model for Speech Recognition

Masatake Dantsuji and Shigeyoshi Kitazawa

Faculty of Letters, Kansai University, 3-3-35 Yamate-cho, Suita-shi, Osaka, Japan
Faculty of Engineering, Shizuoka University, 3-5-1 Johoku, Hamamatsu-shi, Shizuoka, Japan

Abstract

The present study aims to develop a new model for the acoustic feature system for speech recognition by machines. The concept of acoustic features is modified and developed from that of distinctive features. It is usual for each feature to have a single correlation with some physical parameter in a general theory of distinctive features. Our model is, however, characterized by the following properties. (i) Each level of phonemes, allophones, distinctive features, acoustic features, acoustic parameters, and physical parameters is prepared. (ii) Every feature is related to some subset of the set which consists of the the limited number of the acoustic parameters. (iii) Some acoustic parameters can be extracted from statistical analysis and several acoustic features can be detected using neural networks. (iv) Acoustic parameters of each subset and the acoustic features are organized in a hierarchical structure. The acoustic feature in a higher node should be applied earlier than lower features. Previous findings as well as a new acoustic feature model are discussed in this paper.

1. INTRODUCTION

The purpose of our project is to develop an effective model for automatic speech recognition systems by machines. One of the main concerns of our research is to develop fundamental concepts for phonetic and phonemic decoding from speech signals. We have been engaged in this field focusing our interest on taking advantage of distinctive features (1-10). A new acoustic feature model was developed to address this item and the aim of this paper is to present a brief discussion of the nature of our model. A brief sketch of the total model is as follows. The levels of physical parameters, acoustic parameters, acoustic features, distinctive features, allophones, and phonemes were prepared for speech analysis and recognition. The physical parameters are extracted from speech signals first and are then transformed into acoustic parameters. The acoustic parameters are detected by consulting the articulatory and auditory mechanism. Acoustic features consist of the acoustic parameters. Linguistic information such as phonological rules are treated in

higher levels of the distinctive features, allophones, phonemes, etc. The relationships between these levels are illustrated in Figure 1.

2. THE LEVELS OF PHYSICAL PARAMETERS AND ACOUSTIC PARAMETERS

Physical parameters are pure physical elements and are first extracted directly from speech signals without special phonetic knowledge. These parameters are fundamental elements such as the FFT spectrum, the LPC parameters, the Cepstrum coefficient, etc. On the other hand, acoustic parameters are rather abstract elements which are transformed from relevant physical parameters with phonetic knowledge by expert systems, neural networks, statistical analysis, etc. The acoustic parameters are organized in a multilayer structure by consulting the articulatory and auditory mechanism.

For example, some acoustic parameters can be examined by discriminant analysis. Selections of physical parameters for reduction of the dimensionality are necessary. From the heuristic point of view, a too fine description of the spectrum is noisy and harmful for discrimination. This can be interpreted in the Cepstrum coefficient domain as that lower coefficients are useful but higher ordered coefficients are not. Also, in the time domain, the coefficients of the frame near the burst point change quickly, but those at the transitory part change slowly between adjacent frames; therefore the former need to be evaluated in greater detail than the latter, or less frames can be placed at the transitory part than at the burst point.

3. THE LEVELS OF ACOUSTIC FEATURES AND DISTINCTIVE FEATURES

The concept of distinctive features stems from quite old days and it is possible to find a similar and parallel concept of distinctive features as nowadays used in early works of not only Western writers but also in Oriental literature. One of the most important and epoch-making works for distinctive features is, however, ascribed to Jakobson, Fant and Halle's "Preliminaries to Speech Analysis (hereinafter PSA)" (1952). In the framework of PSA, distinctive features are defined in terms of the auditory aspect as well as the acoustic and articulatory aspect. Later, the auditory definitions are omitted in Jakobson and Halle's "Fundamentals of Language" (1956). The notion of distinctive features has had essential influence on the generative phonology. In the early works of generative phonologists, they adopted the distinctive features of PSA. Later, they revized the distinctive features in many respects. The standard notion of generative phonology is well described in Chomsky and Halle's "The Sound Pattern of English (hereinafter SPE)" (1968). In the framework of SPE, the distinctive features are mainly described from an articulatory point of view, and the same inclination has been maintained in current approaches. This, however, does not mean that the acoustic and auditory aspects have lesser importance, but rather that it was difficult to give an exact and precise description of the acoustic and auditory characteristics of distinctive features at that time.

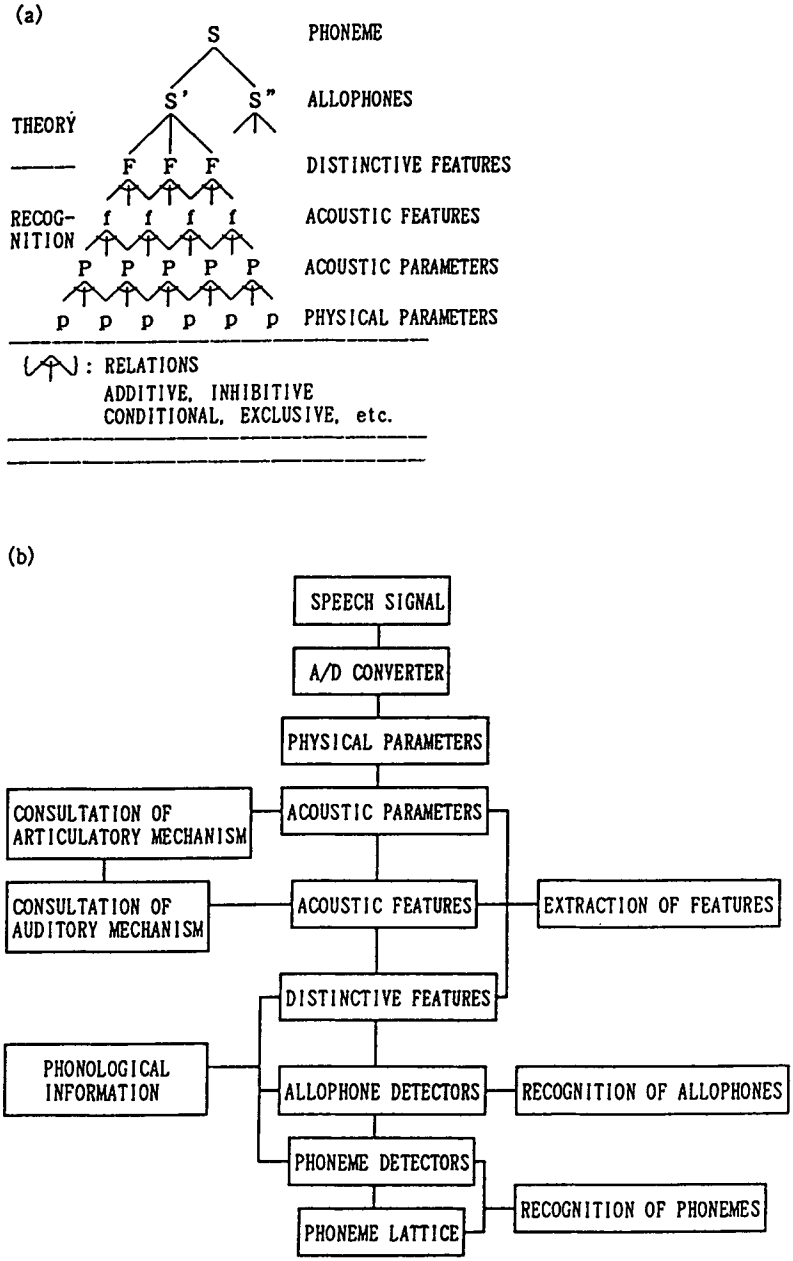


Figure 1. A new acoustic feature model, proposed in this paper (a), and speech recognition system using the acoustic feature model (b).

With regard to the Jakobsonian feature systems, one of the most important principles is to reduce the number of distinctive features as much as possible, for instance by using the same features for vowels and consonants. We would like to take the position, however, that for more phonetically oriented studies, such as speech recognition, it is more effective to increase the number of features. Fant has already pointed out this problem with regard to speech recognition. (14)

Distinctive features are constructed from articulatory, acoustic, and auditory features. Acoustic features consist of acoustic parameters and are organized in a hierarchical structure by consulting the articulatory and auditory mechanism. The concept of acoustic features is modified and developed from that of distinctive features. One of the main purposes of the present study was to develop a new acoustic feature system for machine recognition. The acoustic aspect has much importance, as well as the articulatory and auditory aspect. Therefore, we decided to develop an acoustic feature system based mainly on the Jakobsonian framework rather than on the generative phonology framework, since Fant discussed and improved the acoustic definitions and gave discussions of the fundamental problems of distinctive features in various articles. These acoustic features, however, are not defined as phonologically motivated features. Phonological information is treated at higher levels, such as the distinctive features, allophones, phonemes, etc. Therefore, the acoustic features are not directly dominated by phonemes in this model.

Acoustic features are considered to be highly multidimensional vectors in acoustic space, not only in linear space but also in nonlinear space. We expect, at the present stage, that several acoustic features can be extracted using statistical analysis or neural networks. Once a sufficient number of examples is observed, the established statistical analysis automatically leads to the solution and the solution is a physical realization of the acoustic features. The only necessary thing in this process is to decide on the observation method and the object to obtain the acoustic features.

For example, automatic detection of the feature [burst] can be obtained by means of artificial neural networks (ANN). ANN is supposed to have advantages over HMM for treating the temporal relationships among acoustic events because of a self-organizing mechanism. At first, a given number of input patterns, each of which includes the burst point at a different point, should be presented to the network. The network is trained to discriminate between those patterns in which the burst point occurs in the left half of the pattern ("left") and those patterns including the burst point at the right half of the pattern ("right"). In order to generate the input patterns, the following two structures are defined:

(1) Detection window (DW): this is the portion of the input speech wherein the search for the burst point is to be performed. This window is centered around the burst point.

(2) Shifting window (SW): this is the portion of the input speech which will serve as input to the network at any one time. It has the structure of an input pattern.

Each input pattern is obtained by one shift of the SW within the DW. Consequently, a number of shifted input patterns, p, equal to the length in number of the analysis frames of the DW minus the length of the SW can be obtained. The generation process for

the training patterns starts at frame p/2 of the DW and continues alternating between a right shift of the SW and a left shift of the SW until p patterns are produced. For the construction of the test patterns the process starts from the leftmost frame of the DW and continues by right shifts until p patterns are generated. During the detection process these patterns are fed one by one to the network after training and the output of the network is observed for every pattern. The feature [burst] is assumed to be detected when the output of the network switches from "right" to "left". The burst point is then in the first frame of the right half (m/2) of the pattern. In the case when the detection is in the exact frame, this pattern should be the input pattern number p/2.

4. THE LEVELS OF ALLOPHONES AND PHONEMES

With regard to Japanese consonant systems, the difference of the phonetic values between allophonic variations of certain phonemes are quite large; we should, therefore, prepare the level of allophones in addition to that of phonemes. For example, a dental stop /t/ includes [t, tɕ/t , ts] as positional variants.

[t] : voiceless dental/alveolar plosive,
[tc/tʃ] : voiceless alveolo-palatal/palato-alveolar affricate,
[ts] : voiceless dental/alveolar affricate.

(When we take the position that permits the affricate phoneme /c/, /t/ includes [t], and /c/ includes [tɕ/tʃ] and [ts] as allophones.)

These allophones have, therefore, not only a different place of articulation but also a different manner of articulation. It would be very difficult to find the common acoustic features among these allophones. It is effective, therefore, to prepare the level of allophones rather than to detect phonemes directly.

Double strata of allophones were prepared. The main stratum consists of a broad transcription of allophones, which are so-called positional variations. The sub-stratum consists of a narrow transcription of allophones. The term "broad transcription" in this model is defined as a transcription which makes use of the main chart of the international phonetic alphabet (IPA) plus diacritics such as "palatalized", "voiceless", and "long" symbols. The narrow transcription is prepared for contrastive study to other languages, free variations such as individual differences, etc. Phonemes are defined as a set of allophones. The correspondence between allophones and phonemes would be adjusted by rules.

Examples: phonemes	allophones (main str.)	allophones (sub str.)	contexts
/ t /	[tʃ]	[cɕ]	/ - [i], [j]
		[tɕ]	
	[ts]	[ts]	/ - [ɯ]
	[t]	[t]	/ other contexts
/ d /	[d]	[d]	/ - [a], [e], [o]
	[d]	[z]	/ - [i], [j]
		[dʑ]	
	[dz]	[dz]	/ - [ɯ]

5. THE HIERARCHY OF ACOUSTIC FEATURES

So far, several kinds of feature hierarchies have been proposed. For example, Fant (1973) referred to the feature hierarchy depending on the economy of description in terms of the smallest number of features (14). Clements (1985) and Sagey (1986) discussed a feature hierarchy founded on phonological and phonetic aspects, mainly based on the articulatory point of view (15, 16). We have introduced another hierarchy, based on the auditory distance.

Recent experiments on speech perception and speech recognition have revealed the fact that there exists a different tendency between human perception and machine recognition (Kitazawa and Tubach, 1987; Dantsuji and Kitazawa, 1988; Harada and Kawarada, 1988; Kitazawa, 1988, Kitazawa and Dantsuji, 1989, etc.). It is possible that a tendency to confuse voiced and voiceless plosives occurs quite frequently in the case of human perception. It is also possible that a tendency to confuse dentals (alveolars) and bilabials or velars and dentals (alveolars) occurs quite frequently, but confusion between bilabials and velars seldom occurs. Therefore, it would be possible to set up a kind of auditory distance based on these perceptual experiments. On the other hand, in the case of machine recognition confusion seldom occurs between voiceless plosives and voiced plosives. Confusion also occurs quite frequently between bilabials and velars.

In order to do further and finer investigations, perceptual tests making use of sounds misjudged by machines as stimuli for human perception have been carried out. By these experiments, the tendency observed in the former experiments has been confirmed, viz. there occurs misjudgement between voiced and voiceless plosives with high frequency and there seldom occurs confusion between bilabials and velars. Therefore, it can be hypothesized that a kind of auditory distance should be based on the human perception confusion matrix. These properties indicate the possibility of establishing an auditory triangle in which the side between bilabials and velars is longer than the other two sides, between bilabials and dentals/alveolars and between dentals/ alveolars and velars. This can be contrasted to the equilateral auditory triangle typically used by many phonologists. Furthermore, there is a possibility that we can hypothesize a different plane including velars in addition to the plane including bilabials and dentals/alveolars.

These properties can also be illustrated in a tree diagram. A lower (closer to the terminal) node indicates a shorter auditory distance and easier confusion. An upper (closer to the root) node indicates a longer auditory distance and less confusion. Therefore, in the case of human perception, the feature [compact] is located closer to the root, and the dimension differentiated by this feature is more essential. Velars are distinguished from both bilabials and dentals (alveolars) in this stage. The feature [voiced] is located closer to the terminal, and this implies that confusion between voiced and voiceless sounds takes place quite frequently.

On the other hand, in the case of machine recognition, hitherto, the distinction between voiced and voiceless sounds would be more sensitive than the distinction between places of articulation. This implies that another hierarchy would be set up in which the feature [voiced] is located closer to the root node than the features [compact] and [acute] in machine recognition. Therefore, we focus our research on finding useful cues to distinguish places of articulation by reiterating experiments on human perception and machine

recognition, and are planning to improve the system by means of developing the data processor simulating the auditory mechanism.

6. MULTILAYER REPRESENTATION AND THE COMPENSATORY INTERACTION OF ACOUSTIC PARAMETERS

It is usual for each feature to have a single correlation with some physical parameter in a general feature theory. However, some researchers suggest new models with a complex, many-to-many relationship between features and physical parameters. In addition to this relationship, we have introduced a multilayer representation of acoustic parameters (5, 7, 8) (Dantsuji and Kitazawa, 1987). In the case of an ordinary linear representation, even though it is possible to comprehend the order of the significance weight of each parameter, it is difficult to make manifest the correlation between parameters.

One of the advantages of the multilayer structure is that we can express the correlation between parameters as well as the weight of the element. Acoustic parameters in the higher rows represent more important parameters for the relevant feature. The leftmost elements in the same row have greater weights than the rightmost elements. The length of the line indicates the magnitude of the correlation coefficient. The shorter the line between two parameters, the stronger the correlation between them. The apparent significancy can be controlled by considering the correlation between parameters. Every feature is generally assumed to be independent and no correlations between features are taken into consideration. The correlation between features can be managed by adjusting the correlation between acoustic parameters.

The compensatory interaction of acoustic parameters for the multilayer structure can be illustrated as follows. Pi is the primary and the most important parameter, but sometimes, for a variety of reasons such as phonetic environments, the rate of speech, etc., this parameter is missing. In that case, another parameter Pj fills up this position and organizes a new multilayer structure. When this model is applied to a higher level, such as the phonemic level, it would be possible to recognize speech even if some segments are not clearly pronounced. We have also introduced a parameter sharing system (PSS). Some acoustic parameters are shared by several features. It is the significance weight of the parameters for the relevant feature that is different from each other. For example, features [Fi] and [Fj] share the same subset of the parameters Pi, Pi+1, ..., Pn. Pi is the most significant parameter for [Fi]. On the other hand, Pn is the most important parameter for [Fj]. The two features also differ in significance weights and hierarchies for other parameters.

7. ADAPTATION OF THIS MODEL

As we have mentioned earlier, acoustic features are organized in hierarchical structures. Several features are ranked in the higher node and should be applied and extracted earlier than others. For example, we set up an acoustic feature [burst] for the distinctive features [-continuant]. This acoustic feature [burst] should be detected earlier than the place of the articulation features in the case of stop consonants.

Simpler types of this model for automatic speech recognition have been already examined. Kawahara et al. (1988) and Doshita et al. (1989) reported the results for speaker-independent recognition of Japanese consonants in isolated syllables of /CV/(18),(19). In these reports the examined samples of stop consonants were followed by one of the five Japanese vowels. Each token was uttered by 17 - 84 male speakers. The total utterances were 3633 tokens for stop consonants, including both plosives and nasal stops. Speech samples were digitized in 12 bits at 18.5 kHz sampling rate.

After detecting the acoustic feature [burst] (in this case manually), seven consecutive frames were picked out for analysis. For each frame the spectrum envelope was calculated by 26th-order LPC analysis and then transformed or merged into 28 variables corresponding to the critical bandwidth. Thus each frame was analyzed to produce 29 variables (28 plus the mean square prediction error of the LPC analysis).

At the level of phonemes these stop consonants were /ʔ, p, t, k, b, d, g, m, n,/. In this experiment, since each syllable was pronounced separately, it was assumed that the glottal stop /ʔ/ was placed at the beginning of the syllable's initial vowel. In addition to the level of phonemes, following allophones of [p, p^j, t, tʃ, ts, k, k^j, ʔ, b, b^j, d, g, g^j, m, n, ɲ] were tentatively prepared at the level of allophones. The recognition results using the pair-wise discrimination method are shown in Table 1. The recognition rate is 92.1% for voiceless bilabial plosives (/p/), 91.1% for voiceless dental plosives (/t/), 94.9% for voiceless velar plosives (/k/), 92.5% for glottal stops (/ʔ/), 90.1% for voiced bilabial plosives (/b/), 87.9% for voiced dental plosives (/d/),88.1% for voiced velar plosives (/g/), 95.0% for bilabial nasals (/m/), and 96.3% for dental/alveolar nasals (/n/). The average recognition rate for all the stop consonants including both plosives and nasals reached more than 91%. It can be said that these phonemes are effectively recognized.

Table 1. Speaker Independent Recognition Rate for Japanese Consonants from 3633 Utterances.

CATEGORY	Voiceless Plosives				Voiced Plosives			Nasal stops		Total
PHONEME	p	t	k	ʔ	b	d	g	m	n	
ALLOPHONE	p p^j	t tʃ ts	k k^j	ʔ	b b^j	d	g g^j	m	n ɲ	
RATE (%)	92.1	91.1	94.9	92.5	90.1	87.9	88.1	95.0	96.3	91.3

8. SUMMARY

The properties of our model for acoustic features have been clarified as follows. (i)Each level of phonemes, allophones, distinctive features, acoustic features, acoustic parameters, and physical parameters is prepared. (ii)Every feature is related to some subset of the set which consists of the limited number of the acoustic parameters. (iii)Some acoustic parameters and several acoustic features can be extracted and detected using statistical

analysis or neural networks. (iv) Acoustic parameters of each subset and acoustic features are organized in a hierarchical structure. The acoustic feature in a higher node should be applied earlier than lower features. In order to examine this model, we have evaluated the speaker-independent recognition for Japanese consonants as preliminary research and have been able to obtain satisfactory results.

Acknowledgement

We would like to thank Prof. Tatsuo Nishida and Prof. Shuji Doshita of Kyoto University, who have supported our research in many ways. Part of this study was supported by Advanced Telecommunications Research Institute International and Grant-in-Aid for Scientific Research on Priority Areas, the Ministry of Education, Science and Culture, Japan.

References

1. S. Kitazawa: "Statistical discrimination of French initial stops and nasals," *SUP TELECOM ENST*, Paris, pp.1-21, 1987.

2. S. Kitazawa and J. P. Tubach: "Statistical discrimination of French initial stops," *Proc. Eurospeech*, vol.1, pp.91-94, 1987.

3. S. Kitazawa: "Mechanical discrimination and human perception," *IEICE Technical Report*, SP88-32, pp.9-16, 1988.

4. S. Kitazawa: "Burst point location in stop consonants using back propagation neural networks," *J. Acoust. Soc. Am.* 184, 1988.

5. M. Dantsuji and S. Kitazawa: "A study on an acoustic feature model for speech recognition by machines," Research Report PASL No.63-15-1, pp.1-20, 1988.

6. M. Fourati and S. Kitazawa: "An artificial neural networks model for the burst point detection in stop consonants," unpublished report, Shizuoka University, pp.1-6, 1989.

7. M. Dantsuji: "Phonetics and phonology, Japanese Language and Japanese Language Education," vol.11, pp.21-59, *Meiji Shoin*, Tokyo, 1989.

8. M. Dantsuji: "a tentative approach to the acoustic feature model," *Revue de Phonétique Appliquée*, #91,92,93, pp.147-159, 1989.

9. S. Kitazawa and M. Dantsuji: "A study on speech recognition based on fine phonetic features," Research Report PASL No.01-3-4.

10. M. Dantsuji and S. Sagayama: "A Study on Acoustic Aspects of Phoneme Environment Clustering and Distinctive Features," *IEICE Technical Report*, vol.89, No.340, pp.25-32, 1989.

11. R. Jakobson, C. G. M. Fant and M. Halle: "Preliminaries to Speech Analysis: The Distinctive Features and their Correlates," *MIT Technical Report*, 1969.

12. R. Jakobson and M. Halle: "Fundamentals of Language," 1956.

13. N. Chomsky and M. Halle: "The Sound Pattern of English," Harper and Row.

14. G. Fant: "Speech Sounds and Features," Cambridge, MA, MIT Press, 1973.

15. G. N. Clements: "The geometry of phonological features," *Phonological Year Book 2*, pp.225-253, 1985.

16. E. C. Sagey: "The Representation of Features and Relations in Non-linear Phonology, " MIT Diss.

17. T. Harada and H. Kawarada: "recognition of stop consonants by augmented feature space method," *IEICE Technical Report*, vol.88, No.91, pp.23-30, 1988.

18. T. Kawahara, Y. Mizutani, S. Kitazawa and S. Doshita: "Application of pair-wise discrimination model to Japanese consonant recognition," *Studia Phonologica, XXII*, pp.83-93, 1988.

19. S. Doshita, T. Kawahara, Y. Mizutani, H. Kojima, M. Ishikawa and S. Kitazawa: "Speaker-independent discrimination of Japanese consonant in isolated syllables using pair-wise discrimination method," *J. Acoust. Soc. Japan*, 45, pp.827-836, 1989.

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

A Hybrid Code for Automatic Speech Recognition

Renato DeMori

School of Computer Science, McGill University
3840 University Street, Montreal, Quebec, Canada, H3A 2A7

Abstract

A hybrid coder is introduced for obtaining descriptions of speech patterns. This coder uses popular Vector Quantisation (VQ) techniques on mel-scale cepstral coefficients and their derivatives together with a Recursive Network (RN) for describing suprasegmental features of speech. These features have a purpose of focussing the search when Hidden Markov Models (HMM) are used for unit or word models. Preliminary experiments of speaker-independent connected digit recognition (unknown string length) using a popular data-base (TI) have given 0.4 error rate on words and 1.1 error rate on strings.

1. INTRODUCTION

The best results achieved so far in terms of absolute performances (overall word recognition rate on a widely used data-base) for speaker-independent Automatic Speech Recognition (ARS) of connected words belonging to a vocabulary used in many applications have been obtained by L.R.Rabiner [1] and G.Doddington [2] on connected digits. The best reported performances with strings of unknown length are 0.5% word error rate and 1.5% string error rate. The two systems use multiple word models and different types of acoustic parameters.

The purpose of this paper is to describe a system in which similar performances on the same task have been obtained using a single model is built using models of smaller units that can be phonemes, diphones or triphones. In order to achieve high performances with a unit-based model, a number of features had to be added to the basic acoustic parameters and coded with them. These features have been obtained with Recurrent Networks (RN) trained by examples. The acoustic parameters used were Mel-scaled Cepstral Coefficients (MCC), their time derivatives (indicated in the following as DMCC), the signal energy (e) and its time derivative (∂e). MCC were computed every 10 msecs, and coded in order to produce a string π_1 with symbols belonging to an alphabet Σ_1. DMCC were computed every 10 msecs, and coded in order to produce a string π_2 with symbols belonging to an alphabet Σ_2. The improvements are due to the way the output of RN, e and e have been coded in order to produce a string π_3 with symbols of an alphabet Σ_3. The three descriptions are processed by a recognizer that uses Hidden Markov Models (HMM).

The details of RN as well as of the coder that produces the symbols of Σ_3 will be described in the paper. This introduction will continue with a discussion on the motivations for the solutions that will be described.

In designing units for ASR a number of choices have to be made by the designer. The first choice is the set of units. Triphones have been recently used for general models [3-4]. The problem with triphones is that there are at least tenths of thousands of such units and it is not clear what type of corpus should be used in order to train them for speaker-independent ASR. For a limited vocabulary (like in the case of digits) actually available data bases provide enough data to train any type of units relevant for the task. In spite of that, it is interesting to consider units (like the central part of fricative sounds) that can be assumed to have characteristics that are independent from the context. With such an assumption it is acceptable to train these units with samples from different contexts and assume that the model obtained in such a way can be used in contexts non available in the training set.

Another interesting aspect is related to allophones, that are different model of the same phoneme or diphone. The models are different because the acoustic parameters of a unit may be affected by the surrounding acoustic or phonetic context (for example certain acoustic characteristics may appear weak or absent in a phoneme at the end of a word). The need for introducing allophones may be suggested by recognition errors but also by the analysis of a sentence with general or **primary phonetic features (PPF)** that are detected by RN. Some of the features can be used for recognition.

Another important design choice concerns the **topology** of the HMM model and the tying of probability distributions. Information about model duration can be embedded into a model topology. Transitions between two speech pattern configurations can be modeled more accurately with subtle topologies and coding of the time evolutions of acoustic parameters. The use of DMCC and æis just one simple way for representing speech signal dynamics. A far better way consists in coding speech with RNs that allow to analyze a long speech interval and detect significant changes in it by remembering the past with distributed internal memories providing information about signal history whose strength can be leaned automatically. These representations may not be used for recognition (especially if the task is already well accomplished without them) but appeared to be very useful in grouping errors into classes that correspond to systematic deficiencies of the existing models and in suggesting topological improvements.

A third important choice is in the **type** of probability distributions for the acoustic parameters. The use of a symbolic representation of a speech frame does not impose any constraint on the distribution but implies another choice which is that of the coder(s). Furthermore, probabilities for symbols that have never been seen on a transition during training should be assigned a **default value** in order to prevent the model to fail because symbols never seen in a model transition during training appear in the testing data. Default values can be lower than usual for those symbol which correspond to features reliably recognized by RN and that are contradictory for a certain model (for example the feature silence in the model of a diphthong).

Other problems are encountered when HMMs are used for ASR. One of them is that probability distributions can be affected by segmentation errors. Some of the misrecognized cases can be recuperated by reducing segmentation errors especially during training.

Segmentation is improved by performing a sort of focus by inhibition. This is accomplished by presenting extremely low initial probability values for the symbols that are detected very reliably by RN and are incompatible with a model. This setting makes it very difficult for a model to accept a description that contains symbols for which the model has inhibitors.

The availability of features on a description of a speech may be very useful but also very dangerous especially if the descriptions are extracted by a device that is not error free. This probably explains why the attempt of recognizing connected digits with HMMs receiving inputs from Neutral Networks (NN) producing scores for phoneme hypotheses at their output has not been as successful so far as the use of pure acoustic parameters [5]. Nevertheless, a parsimonious use of features together with pure acoustic parameters is definitely beneficial because it allows to take into account events covering a large speech interval and to reduce errors due to statistical estimation with parameters produced by a wrong segmentation during training.

The possibility of using knowledge in a coder or just for reasoning about model topology, allophones and tying suggests the introduction of a **learning strategy** in which the training set of a data base is subdivided into training subsets TS1, TS2,...,TSi...,TSI. TS1 is used as first training set and TS2 is used as first test set. An analysis of the errors in TS1 and TS2 may suggest modifications on the model topology, the addition or deletion of new symbols in the coder, the introduction of some inhibitors in certain models (symbols for which initial values are extremely low) before retraining with a new training set TS12 made by the union of TS1 and TS2. Retraining may not be necessary if all the errors in TS2 can be justified only in terms of topologies or inhibitors.

2. THE HYBRID SPEECH CODES

Many schemes have been proposed for ASR systems. The one shown in Figure 1 is one of them. It uses HMMs for representing speech units and emphasizes the fact that a strategy (Vitervi like for example) generates hypotheses by chaining units under the constraints imposed by a lexical representation. There are three components used by this strategy, namely its **knowledge** (the HMMs), the **constraint knowledge** (represented by the portion of the lexicon made active by the lexical strategy) and a **search algorithm** that uses probabilities as hypothesis measures. A module applies its strategy to descriptions stored into a Short Term Memory (STM). The strategy uses constraints from knowledge of other units or from data written by other modules into the STM.

The lexical module operates in a similar way but uses a different strategy based on fast word hypothesization and subsequent detailed scoring. Word hypothesization is constrained by the language model. The type of constraint imposed by the sentence strategy is usually made of stochastic predictions based on the language model and the STM content while the constraint imposed by the lexicon on the unit model strategy is often deterministic.

The language model maintains a data structure of sentence hypotheses based on scored lexical hypotheses under the control of a strategy that uses a stochastic language model and may also use constraints of other units, like a dialog model which dynamically modifies probabilities of expectations.

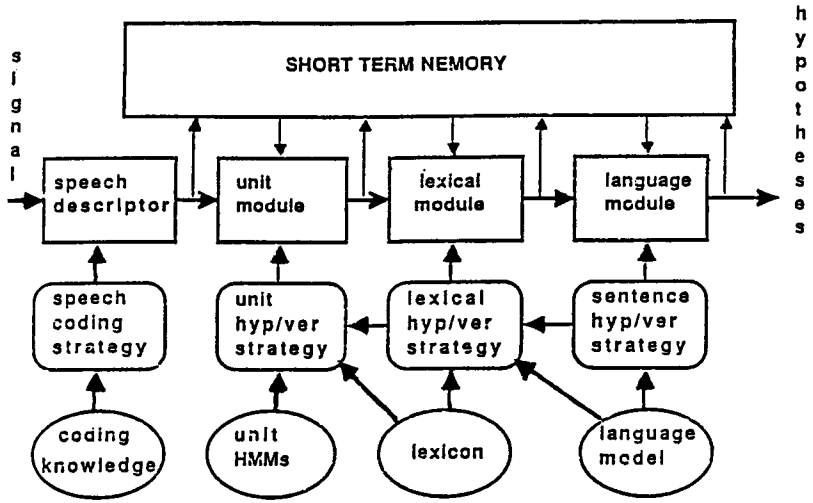


Figure 1. Automatic SPeech Recognition Model.

The focus of this paper is on the speech descriptor which uses a speech coding strategy and a coding knowledge.

The need of a speech descriptor is motivated by the large size of the pattern space of the speech signal. It is so large because it contains information that is redundant for recognition. A large pattern space size is not suitable to be effectively medelied by HMMs and to be used by the associated strategy. Furthermore, HMMs have to be **trained** and this is not practically feasible if the pattern space is large.

In order to be processed by the unit module, speech descriptions are usually generated at fixed time frames (typically every 10 msecs). Descriptions may be of various nature, they can be represented by vectors of parameters or by symbols. Each time frame can be represented by more than one vector or more than one symbol. Using symbols corresponds to quantizing the parameters. The loss of information incurred in quantization may be beneficial because it eliminates redundancies and focuses on distinctions of perceptual importance. Furthermore, HMMs can be trained on symbols without having to make any assumption on the type of probability distributions.

The coding knowledge can be determined by an estimation process combined with HMMs training [6]. Such an approach does not allow to discover properties that describe a given frame in the context of the history of a set of speech parameters. Relevant contexts can be decided based on a-prior knowledge (like the one required to compute the time derivative of a parameter) or can be learned automatically with RNs. The two approaches for determining the context of a description are used in the system described in this paper.

An interesting problem to be addressed when coding speech for ASR with symbols is whether coder knowledge learning should be **supervised** or **unsupervised**. In the first case speech segments should be labelled with the symbols that the coder should provide.

In some cases these symbols can be generated by algorithms without any learning.

In the work described in this paper **a hybrid coder** has been used in which a first set of symbols belonging to an alphabet Σ_1 is produced by unsupervised learning on vectors of MCCs, a second set of symbols belonging to an alphabet Σ_2 is produced by unsupervised learning on vectors of DMCC and a third set of symbols belonging to an alphabet Σ_3 is produced by a mixture of algorithms and a RN trained by supervised learning to generate hypotheses and gross phonetic categories.

Figure 2 shows a block diagram of the hybrid coder. The signal $x(t)$ is processed in order to obtain a sequence of spectra Ω. This is done with Fast Fourier Transformation (FFT) but for other projects in our research an ear model is used [7]. Another module extracts a sequence A of acoustic Properties like zero-crossing rates, energy ratios and energy contour profiles as described in [8]. Two vectors of parameters are extracted for every frame, namely M (a vector of 8 MCC) and ∂M (a vector of 8 DMCC). The time sequences of these two vectors are coded using classical Vector Quantization (VQ) techniques to produce two strings π_1 and π_2 of signal descriptions [9]. Properties A and spectra are sent to the input of a RN which generates degrees of evidence for three basic features in a set F:

$$F : \{\text{sonorant (S), fricative (FR), plosive/silence/buzz-bar (PL)}\} \quad (1)$$

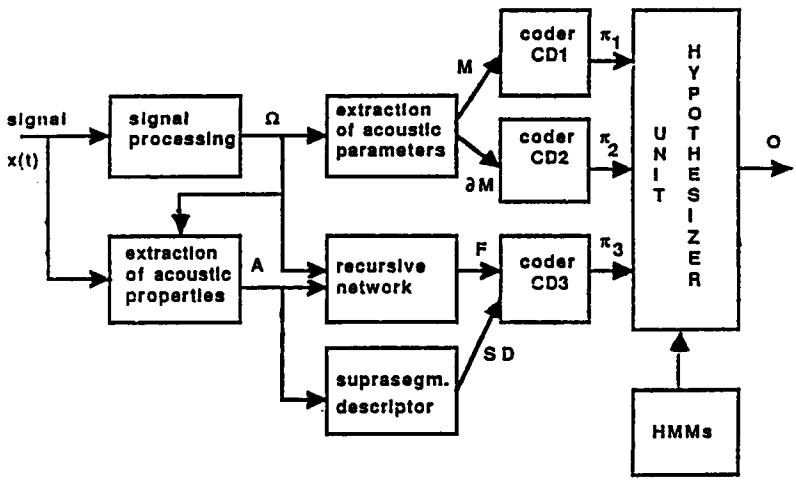


Figure 2. Hybrid coder.

A suprasegmental descriptor produces also a description SD of the energy contour. There are two versions of this descriptor. One (DESCR1) is based on vector quantization of the signal energy e and its time derivative œ. The other (DESCR2) uses syntactic pattern recognition techniques for characterizing the energy contour in terms of long and short peaks and valleys [8]. The coder CD3 produces a symbol FR or PL if the

corresponding output of RN has the maximum evidence for a segment (note that this evidence depends on previous history). If S is predominant, then CD3 generates a symbol in a set of 30 depending on the signal level, its slope, whether the frame is on a peak or a valley and the peak/valley duration. A symbol in the description π_3 generated by CD3 describes a frame in the context of its past history (taken into account by RN) and of the suprasegmental feature (described by SD) in which the frame is in.

The peculiar characteristics of the hybrid coder are that raw parameter vectors (M and ∂M) are used together with descriptions of the signal in terms of properties that are expected to be significant for certain speech units and cover a large time context. This knowledge is used to perform **a focus by inhibition**. For example, in the vowel of the word "two" the feature FR must be absent. When the corresponding speech unit is trained, the probability of the code for FR will be initialized to a much lower value than for other codes for which inhibition is not performed. Such a probability won't be increased by smoothing after training (or by other heuristic operations that are performed after training in order to compensate for the fact that some codes have never been observed just because the training set has a finite size), on the contrary it may be further lowered. Hybrid coding and focus by inhibition allow one to better **heuristics** than the ones that are usually applied for smoothing probability distributions.

For the cases which focus by inhibition is not used, hybrid coding allows to exploit the advantages of relevant properties when they are reliable (when they have high probabilities for the expected unit transitions) or give more emphasis to codes of row data when property probabilities have rather uniform distributions.

3. THE RECURRENT NETWORK FOR FEATURE RECOGNITION

The recurrent network used for the hybrid coder has the architecture shown in Figure 3. The input is made of four successive frames from twenty channels (the duration of each frame is 10 msecs.). Three channels have at the output of the logarithm of the broad band energy at low frequencies (200-900 Hz), intermediate frequencies (1-3 kHz) and high frequencies (3-5 kHz). Other three channels provide the signal energy, the energy of the signal derivative and the zero-crossing densities of the signal and its derivative. The remaining 13 channels provide the energy in mel-scaled frequency bands.

All the inputs to RN are normalized to vary between zero and one. The motivation for choosing these inputs relies on the fact that the features the network is supposed to hypothesize depend on known acoustic properties in the time domain as well as in the frequency domain. The delays on the input modes allow the network to see four frames at a time which constitute an acoustic context.

There are five hidden units receiving stimuli from all the 80 input nodes. The output of all the hidden units is fed back through a delay unit. The weight of the feed-back connection represents the characteristics with which each hidden unit has a memory of its past outputs. The three outputs of RN produce degrees of evidence for the three classes defined by the (1).

The outputs of RN are displayed on neurograms like the one shown in Figure 4. Each

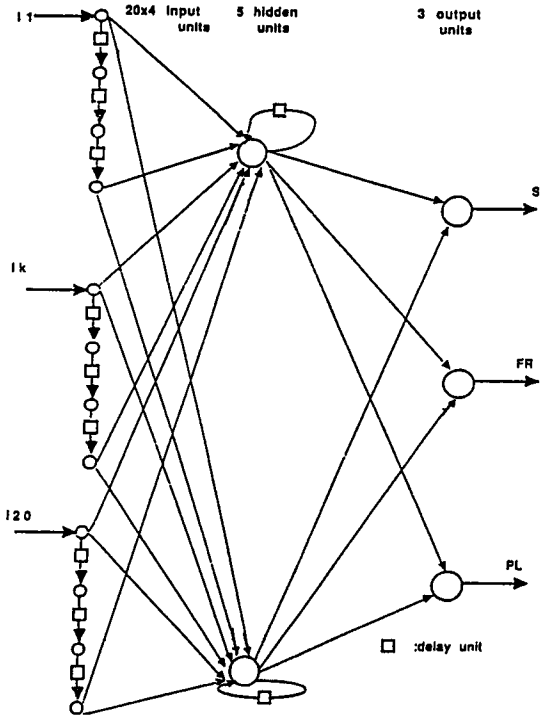


Figure 3. Structure of the recursive Network.

vertical bar corresponds to a time frame. Its height is proportional to the degree of evidence of the corresponding feature. For each time frame, the evidence of the highest feature is always displayed, the other evidences are not displayed if they are below a threshold (i.e. .6).

Details on the training algorithm for the RN described in this Section can be found in [10].

The network was trained on 10 speakers (used also for training the HMMs) with a total of 77 digit strings per speakers. It was tested on 4 new speakers and then introduced in the coder for the HMMs. The frame error rate for the test set was 6.9%, it was, 6.7% for the training set.

4. THE HIDDEN MARKOV MODELS AND THEIR USE

Each word of the lexicon is represented by a sequence of units as shown in Table 1. Each unit was initially represented by a Hidden Markov Model as shown in Figure 5. Symbols pij represent transition probabilities between states while pij represents a set of three vectors of probabilities (one vector per code-book). Each element during the transition from state B_i to state B_j.

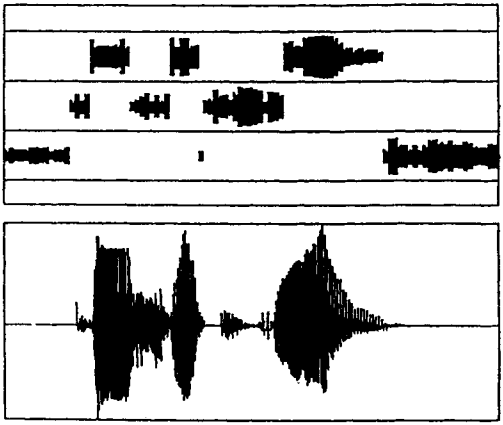


Figure 4. Example of a neurogram.

Table 1. Digit vocabulary.

1	WAH	AHN	N		
2	T	TUW	UW		
3	TH	THR	RIY	IY	
4	FAO	AOU	R		
5	FAA	AA	AAYV		
6	S	SIH	IHK	KS	S
7	S	SEH	EHV	VAX	AXN
8	EY	EYYT	T		
9	NAA	AA	AAY	YN	
OH	AO	OW			
ZERO	Z	ZIY	IYR	RAO	OW

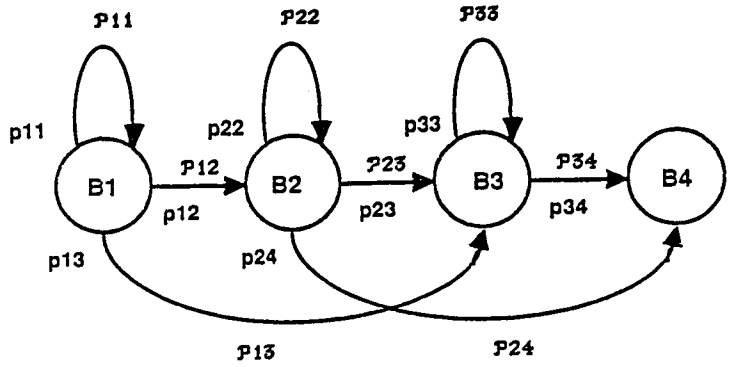


Figure 5. Structure of the Hidden Markov Model used for a unit.

Tying of probability distribution is performed in such a way that the distributions associated to all arcs going to the same state are the same. This means, for example, that $P_{23} = P_{33} = P_{13} = P3$.

5. EXPERIMENTAL RESULTS

The experiments reported in this Section are based on a subset of the TI database [11] used as training set (TS1) and another subset (TS2) used as a test set. TS1 contains 616 strings of digits (unknown length) pronounced by 14 speakers, while TS2 contains 300 strings of digits (unknown length) pronounced by 14 speakers.

The first experiment (EXP1) consisted in training the unit models with TS1 by using the coders shown in Figure 2 with CD3 just performing VQ on the signal energy e and its derivative œ. Training was accomplished with 0.2% word error rate and 0.8% string error rate. Recognition gave 2.7% word error rate and 7.3% string error rate.

A second experiment (EXP2) was conducted with CD3 performing vector quantization on the outputs of RN and using these symbols in the description π_3. Only the third codebook was retrained and the same number of errors although not always the same errors in the test set as in EXP1 were found.

A third experiment EXP3 was performed by coding e and œby vector quantization only when the predominant output of RN was S and by generating a code corresponding to FR or PL When these outputs of the RN were predominant. A word error rate of 1.6% and a sentence error rate of 5.4% were found in TS2.

In a fourth experiment, EXP4, inhibition was introduced by setting the probability of having FR or PL on the units /N/, /AO/, /WAH/, /IY/, /EY/ and /NAA/ extremely low (10^{-32}) in order to reduce the possibility of error alignments in recognition due to error alignments in training. The system was not retrained but the units containing stressed vowels were made longer by adding two more internal states B21 and B22 with probabilities associated to arcs reaching them equal to the probabilities of the corresponding

arcs reaching B2. Furthmore, an allophone T1 was introduced for the last phoneme of EIGHT. Lengthening the models reduced the error rate to 1.4% for wards and 4.4% for strings. Inhibition and the introduction of T1 allowed to reduce the word error rate to 0.45% and the string error rate to 1.3%.

Just for the sake of comparison, Maximum Mutual Information Estimation (MMIE) was used for training an eight state word model for each digit. Three codebooks were used, one for MCC a second for DMCC and a third for e and œ. Error rates on TS2 were 0.9% for words and 2.1% for strings. Experiments are in progress with other types of feature coders.

6. CONCLUSIONS

The feasibility of hybrid coders has been demonstrated. The addition of acoustic properties to speech descriptions has produced improvements when HMMs are used for recognizing speech from descriptions. Simple properties and features have been used in the experiments described in this paper. Other properties and features more useful for making fine distinctions between vowels, nasal and plosive sounds have been introduced and tested on isolated letters and digits pronounced by many speakers [10]. Their effectiveness has now to be tested with units used for the recognition of large vocabularies.

Acknowledgements

This work has been carried out at Mc Gill University and at the Centre de Recherche en Informatique de Montreal (CRIM). We wish to thank CRIM for making human and computer resources available to us and the Natural Sciences and Engineering Council of Canada (NSERC) for supporting students and equipment.

References

1. L. R. Rabiner, C. H. Lee, B. H. Huang and J. G. Wilpon, "HMM clustering for connected word recognition," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp.405-408, (1989)

2. G. Doddington, "Phonetically sensitive discriminants for improved speech recognition," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp.556-559, (1989)

3. L. Bahl et al., "Large vocabulary natural language continuous speech recognition," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp.465-467, (1989)

4. K. F. Lee, H. W. Hon, M. Y. Wang, S. Mahajan and R. Reddy, "The SPHINX speech recognition system," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp.445-448, (1989)

5. M. A. Franzini, M. J. Witbrock and K. F. Lee, "A connectionist approach to continuous speech recognition," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp.425-428, (1989)

6. J. Bellagarda and D. Nahamoo, "The mixture continuous parameter models for large vocabulary isolated speech recognition," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp.13-16, (1989)

7. P. Cosi, Y. Bengio and R. De Mori, "On the generalization capability of multi-layered networks for automatic speech recognition," Proceedings of the International Joint Conference on Artificial Intelligence, pp.1531-1536, (1989)

8. R. De Mori, P. Laface and Y. Mong, "Parallel algorithms for syllable recognition in continuous speech," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-7, no. 1, pp.289-305.

9. R. M. Gray, "Vector quantization," IEEE ASSP Magazine, pp.4-29, (1984)

10. Y. Bengio, R. De Mori and M. Gori, "Experiments on automatic speech recognition using BPS," Proc. International Joint Conference on Neutral Networks, (1989)

11. R. G. Leonard, "A database for speaker-independent digit recognition," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, paper 42.11., (1984)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Complementary Approaches to Acoustic-Phonetic Decoding of Continuous Speech

Jean-Paul Haton

CRIN/INRIA, B.P. 239, 54506 Vandoeuvre les Nancy Cedex, France

Abstract

The acoustic-phonetic decoding (i.e. the transformation of the acoustic-phonetic continuum of speech into a description under the form of discrete, linguistic units) represents an important step and a major bottleneck in the overall process of automatic speech recognition.

This paper presents the problem and its difficulties together with the different families of solutions proposed so far. After a recall of the methods based on pattern matching techniques and stochastic models we introduce a class of methods based on artificial intelligence knowledge-based techniques. Such methods make an explicit use of all available types of knowledge that intervene in phonetic perception. We then present the use of neural connectionist models and discuss their interest for the problem. The presentations will be illustrated by practical examples drawn from different systems.

1. INTRODUCTION

The process of understanding a spoken sentence can be considered as a sequence of steps from the acoustic level up to the semantic level. In this overall organization the level of phonetic decoding (PD) represents an important step and a major bottleneck in the design of speech recognizers even for single speaker applications. We refer to phonetic decoding as to the different processes involved in the transformation of the continuous, acoustic data coming from a microphone into a description under the form of discrete, linguistic units such as phonemes, syllables, diphones, etc.

A large number of methods have been proposed so far in order to solve this problem. They will be briefly recalled in section 3, after a general presentation of the speech recognition paradigm given in section 2. In order to illustrate this presentation two particular approaches will then be considered in more details, one based on knowledge-based reasoning (section 4) and the other using a connectionist model (section 5).

2. POSITION OF THE PROBLEM

2.1. The Speech Recognition Paradigm

The speech signal can be regarded as the result of a hierarchical encoding process at successive levels among which the most important ones are pragmatic, semantic, syntactic, phonological, articulatory levels. This signal presents specific characteristics that make its interpretation difficult, especially: continuity (which necessitates to take segmentation decisions at one time or another), variability (which complexifies the multi-speaker phonetic decoding of speech) and redundancy.

One possible approach to the automatic recognition of speech consists in referring to the above mentioned encoding process and in taking into account the various corresponding knowledge pieces. This knowledge-based approach, though not the only one possible, has been widely used (1), (2). In this framework the importance of the acoustic-phonetic level is presently widely admitted (3).

2.2. Variability of Speech

A language such as French or English has a relatively low number of phonemes (about some tens) but the variety of corresponding acoustic speech patterns is very high and difficult to characterize for several reasons:

- factors like speaking rate, loudness, prosody, etc. have an influence on speech sounds,
- the acoustic patterns depend on the speech production manner,
- coarticulation effects, context and speaker dependencies make the diversity of acoustic realization very large.

The contextual variability in the acoustic structure of phonetic segments is particularly important, due to several major processes (4):

- unvoicing of voiced glides in contact with a voiceless fricative or stop,
- labial coarticulation: Figure 1 shows that the low frequency cut-off of /s/ in labial context of /y/ is brought down to the level of /ʃ/,
- tongue coarticulation,
- nasal coarticulation,
- differences in vocal tract length,
- variations in the vocal source spectrum.

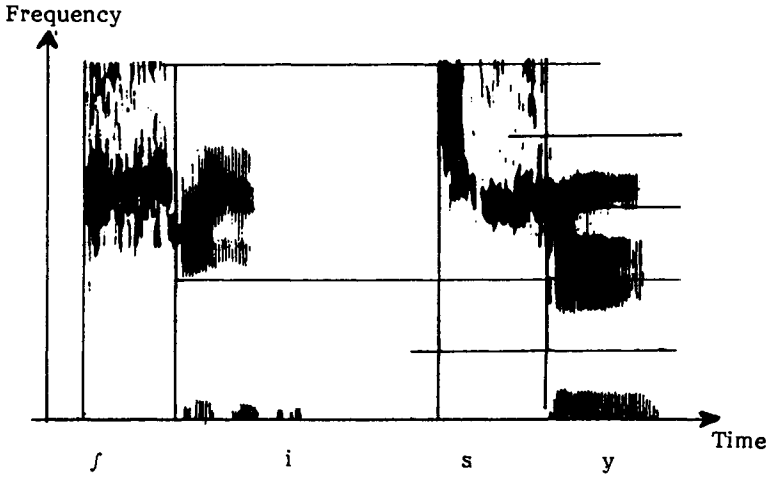


Figure 1. Example of phonological alteration (/i/ vs /sy/)

2.3. Acoustic-Phonetic Decoding of Speech

The causes and examples of variability just described illustrate the difficulty of the acoustic-phonetic decoding of speech. However it is of primary importance to design robust and efficient PD techniques since the overall performances of any sentence recognizer rely heavily on the quality of this decoding.

The activity of PD comprises two complementary subtasks that can be carried out sequentially or concurrently, according to the approach:

- a task of segmentation of the speech wave into acoustic segments,

- a task of phonetic identification or labelling of these segments.

An important point in the design of a phonetic decoder concerns the choice of a phonetic unit. Several units have been concurrently used so far: syllables, demi-syllables, diphones, triplets phonemes, allophones.

Syllables, diphones and demi-syllables present the interest of integrating information about transitions that constitute the most difficult parts to identify in the speech signal. The use of allophones can to a certain extent simplify the process of automatic labelling but a trade-off has to be found since the number of allophones is almost unlimited (several thousand are necessary for obtaining good performances). Many systems use at one stage or another phonemes as a decoding unit, for several reasons. The phoneme is the basic unit of language production. It requires less training and storage space since the number of phonemes in a given language is rather low (less than 100). The phonological variations of phonemes can be predicted by contextual rules within and between sounds.

3. TECHNIQUES OF ACOUSTIC-PHONETIC DECODING

3.1. Feature Selection and Extraction

Signal processing techniques (FFT, LPC, cepstrum, filter bank, auditory models, etc.) applied to a portion of speech signal yield a set of parameters which are specific to each technique. These sets or vectors of acoustic parameters can usually not be directly related to some phonetic knowledge: an interpretation of these vectors is requested. Phonetician experts have a long experience for relating acoustics to articulatory phenomena based on the study of speech spectrograms; position and trajectory of formants, bursts, VOT, etc. Such correlations must be computed in order to interpret the speech signal in terms of acoustic-phonetic elements.

Phonetics has proposed models for phoneme classification from acoustic and articulatory models. The classification system based on distinctive features is well-known (5). The distinctive features of this model constitute a minimal system for all languages but, unfortunately, they are not directly related to the acoustic reality and, therefore, cannot be practically implemented for automatic PD.

However a feature-based approach to automatic PD has been very often used in all languages. We will briefly review in this section the different techniques that make it possible to effectively combine features for labelling speech. Some of these acoustic-phonetic are redundant. Since speakers do not always articulate carefully all features are not always present in the speech signal. It is therefore important to use a reasonable number of parameters (as far as it is computationally tractable) even though these parameters are not statistically independent.

3.2. Segmentation

Segmentation is a fundamental process of PD. A major issue concerns the choice of the segmental unit. The different units already mentioned have been used for this purpose:

- phones (6): phones are subphonemic units which are then merged together in order to produce larger units;

- phonemes (7), (8): a human listener is able to identify phoneme-like segments in continuous speech. However, segmentation at phoneme level is very difficult and introduces errors of under - and over - segmentation;

- diphones (9), (10), demi-syllables (11), triplets (12) and syllables (12), (13), (14), (15), (16): such supra-phonemic units make it possible to incorporate coarticulatory phenomena but their identification is complex and recognition errors can have dramatic consequences in the understanding process.

Segments can be obtained synchronously by analysing fixed-length samples of speech ("centisecond" samples) (17) or asynchronously (18).

The segmentation process is usually based on the study of the variations of a function measuring the discontinuities of the speech wave or its spectra. Articulatory criteria have

also been used (19). Segmentation refers to acoustic-phonetic knowledge that can alternatively been explicited in a declarative way. This representation leads to segmentation systems based on parsers using rewriting rules (20), (21), and also to knowledge-based segmentation (22). This latter approach makes it easier to embed heuristic knowledge that a phonetician has and to enforce the segmentation algorithms through symbolic manipulation.

Most speech segmenters use a bottom-up strategy going from acoustic data to abstract linguistic representations but some attempts have also been done toward a top-down segmentation based on a prediction of the phonetic string.

3.3. Segment Identification Techniques

We have seen in paragraph 3.1 that the identification or labelling of phonetic segments is often based on a set of acoustic-phonetic features extracted from a parametric representation of the speech wave. We will now review the different techniques, more or less related to classification and/or pattern recognition algorithms, that are usually used in PD systems:

- **vector quantization:** this technique consists in taking into account the statistical properties of sounds in a represen tation space. It is widely used for speech coding and synthesis but it does also present some interest in PD for carrying out a first, rough classification of segments into broad classes (23).

- **statistical pattern recognition:** a large number of PD systems are classical pattern recognition systems. The identification of a segment is made by comparing this segment to a set of reference segments described by their acoustic and phonetic features with some statistical information. These techniques necessitate to define an efficient distance measure between patterns. Dynamic programming algorithms can be used in order to compensate for non-linear time distortions, especially for long units (diphones or syllables). The major importance of contextual phenomena in the speech production process (cf. section 2) makes it necessary to use an extremely large number of reference patterns (at least 10,000). That constitutes a basic limitation of the approach, especially for multi-speaker applications since the reference patterns are essentially speaker-dependent.

- **structural pattern recognition:** structural pattern recognition is concerned with the description of complex patterns in terms of simple, primitive patterns. This technique has been used in PD (20), (21) but the limitations encountered in statistical pattern recognition still exist. Once again the problem of multi-speaker recognition is solved by clustering prototypes among a large number of different speakers. This solution is not satisfactory and drastically limits the use of pure pattern matching techniques in PD.

- **stochastic modelling:** the acoustic-phonetic decoding of speech formally consists in finding the best string or lattice of speech units by optimally matching an input utterance against every possible concatenation of reference patterns or speech unit production models. This can be expressed in terms of stochastic processing,

especially in the framework of Markov sources, more precisely with Hidden Markov Models (HMM) (24), (25). Initially proposed for larger units (i.e. words) the method can be generalized to phonetic decoding (26), (27). An important advantage of this approach is the possibility to capture in a statistical way broad speech and speaker variances. This is carried out automatically by processing huge amounts of speech data coming from a large variety of speakers. HMMs provide one of the most efficient framework for multi-speaker PD. However these models are purely mathematical, without any explicit use of phonetic knowledge.

- **knowledge-based reasoning:** the explicit use of linguistic knowledge constitutes an alternative solution to the PD problem that will be presented in section 4.

- **connectionist modelling:** a new class of models, based on networks of large numbers of neuron-like processing elements, have recently appeared in various fields of automatic perception (28). Such connectionist models have encountered some success in limited applications of speech recognition, including phonetic decoding even though they have not yet definitively proved their superiority to more classical models used so far. We will present in section 5 a novel connectionist model based on cortical columns together with its application to various tasks in phonetic decoding.

4. KNOWLEDGE-BASED REASONING FOR PHONETIC DECODING

4.1. Position of the Problem

We have just seen that the acoustic-phonetic decoding of a sentence can be carried out by purely mathematical models such as Markov models. However the interpretation of the speech signal necessitates to take into account knowledge and information that are not present in the signal itself. Therefore it is interesting to design PD methods which allow for the integration of some knowledge in the recognition process, possibly in conjunction with other models. This knowledge will usually be coded simultaneously under a procedural form (e.g. procedures for detecting a feature in the speech signal) and a declarative form (e.g. the knowledge used by some control structures for reasoning). A knowledge-based approach to PD makes it possible not only to take into account relevant knowledge (phonotactic constraints, allophonic variations, phonological constraints, etc.) but also to implement more realistic decoding strategies.

The knowledge used by a human listener for decoding speech is mostly unconscious and implicit, and therefore quite impossible to formalize. It is then necessary to consider activities in which there exists conscious knowledge related to acoustics and phonetics. Speech spectrogram reading by skilled phoneticians represents a typical example of such activities. Early attempts at spectrogram reading were rather pessimistic and usually concluded on the enormous difficulty of the task due to the contextual variations of the acoustic signal (29). In fact recent studies have shown that the phonetic decoding was feasible without any high level linguistic knowledge with an average accuracy of 80-85%, i.e. far better than the one of present automatic systems (30), (4). Of course spectrogram

readers do not proceed in a way similar to the auditory system but the interest for an AI approach to PD is that the visual features used by these experts are clearly specifiable and that the knowledge involved in their reasoning is explicit.

The knowledge and decoding strategies used by experienced spectrogram readers are not easy to elucidate, in-as-much as spectrogram decoding involves visual cues, and therefore, requires perceptive knowledge and competence, which are by nature difficult to explain and/or to convey. This led us to develop specific observation and analysis methods that are fully reported in (31).

4.2. The APHODEX System

Overview of the system

In order to illustrate some of the issues related to knowledge-based phonetic decoding we will now give functional and experimental data about the APHODEX system that we have been developing in Nancy for the past 4 years (32).

APHODEX, Acoustic PHonetic Decoding EXpert, is an experimental tool designed for improving our knowledge about PD by the in-depth study of spectrogram reading activity. The knowledge gained from this study is coded in the system under three forms:

- procedures for segmentation and labelling of segments into gross phonetic classes,
- contextual production rules. Presently the knowledge base is made up of about 400 rules and it is regularly augmented.

A typical rule is as follows:

IF LEFT CONTEXT	/y u ɛ œ o a œ o u w ʃ z/
RIGHT CONTEXT	/ã ɛ̃ d e a p t k b d g s v z m n r l/
AND NOISE-LIMIT IS INCREASING	
AND FRICTION THRESHOLD	|2800 – 3300|
THEN	*s*1/*z*1.

The value 1 associated with the two conclusions of the rule represents the certainty factor (ranging from -1 (false) to 1 (true)) of these conclusions.

- decoding strategies that operate at two levels:
  - a global strategy which is roughly bottom-up from the speech signal to the gross phonetic labelling of a sentence and then mixed top-down and bottom-up for the refinement of the precise labelling.
  - a strategy used for propagating constraints during the reasoning process.

The overall architecture of APHODEX is given in Figure 2. The inference engine of the system uses both forward and backward chaining and carries out approximate reasoning by combination of certainty factors. It is completed by a constraint propagation algorithm which controls that the constraints that appeared in most production rules under the form of left and right contexts are satisfied.

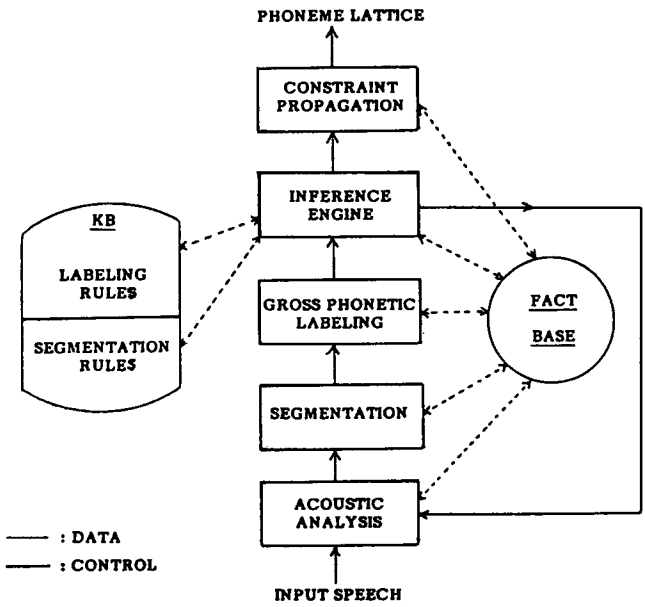


Figure 2. The architecture of APHODEX

Experimental results

The practical example shown in Figure 3 summarizes the different operations involved in the decoding of a sentence.

The overall performances of APHODEX presently range around 70% of phoneme recognition for any male native French speaker. The recognition accuracy depends on the class of phoneme (it is far better for plosives, fricatives and vowels) since the knowledge base is not yet completed.

5. CONNECTIONIST MODELS

5.1. Overview

The use of neural networks in ASR is still at an early stage but the intrinsic properties of such models are already attractive solutions for problems such as speaker variability or the incorporation of contextual information and of speech knowledge in the understanding process.

Most of the work done so far in the use of neural networks is based on two models, i.e. Boltzmann machines (33) and multi-layer perceptrons (34). Both models are based

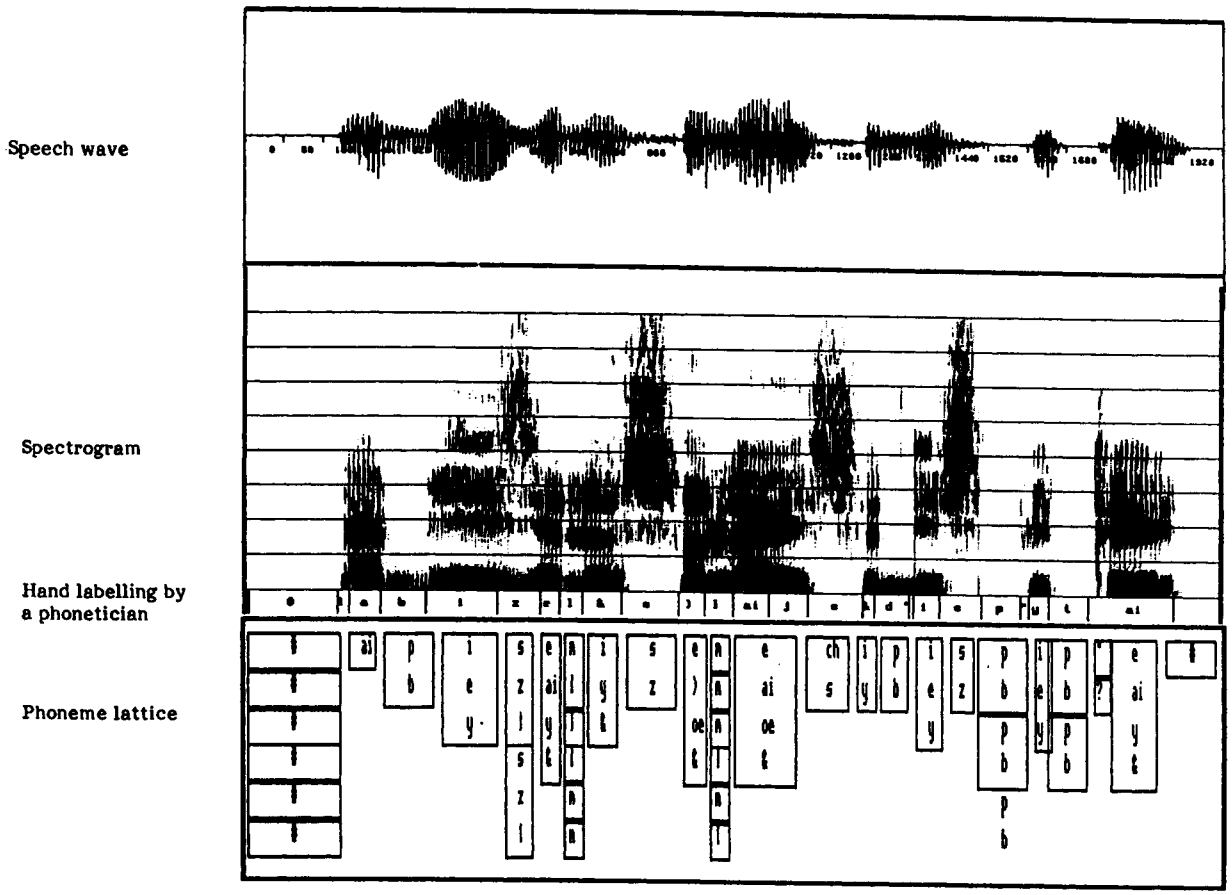


Figure 3. Example of phonetic decoding by APHODEX for the French sentence "la bise et la soleil se disputaient"

on sophisticated learning algorithms that consist in determining the most appropriate network configuration for a given set of data.

These models have sometimes been used for front-end processing, e.g. noise reduction, voicing and F_0 determination, recognition of place of articulation (35), (36). However, the most common use is for the recognition of speech units, either whole words, isolated or connected (37) or phonetic units (38), (39) or else gross phonetic classes (40).

An important problem arises in the use of such models for taking into account the temporal sequentiality of speech. A partial solution consists in introducing some amount of feedback in the networks. It is also possible to take into account contextual information during the recognition. For instance, Bourlard (37) introduces the left and right neighbours of a speech vector in the labelling decision concerning this vector.

The various experiments carried out with these models have given good, sometimes excellent results, but not yet exceeding results obtained with other classical pattern recognition techniques.

Another approach to connectionism is related to the definition of adaptive, associative networks. This approach, more closely related to ours, allows for a more natural use of time in the model. Such models have been used preliminary experiments in word recognition (41) and also for the implementation of a complete phonetic typewriter (42).

Boltzmann machines and multi-layer perceptrons are made up of elementary neuron-like elements of very small size and limited processing power. We propose a new approach that consists in taking as elementary processors cortical columns. A cortical column corresponds to the association of about a hundred of neurons having a specific, functional activity and it was designed according to neurobiological data (43).

Two architectural principles describe the cortical column:

1. a column is organized in three layers, each layer performing specific operations on specific data. The upper layer (layers I to III of the cortex) carries out all the reciprocal but not the symmetrical relations with the other columns, including memorization abilities. The intermediate layer (layer IV) receives the information flow which codes the sensory stimuli from the external world, while the lower layer (layers V and VI) effects output toward the external world by various motor actions (speech production, etc.).

2. the connectivity between columns can be summed up by compact neighboring connections and 2 long range connections, where 2 is the number of areas implicated in the stimulated process. Information propagates within areas with hypercubic connections, or inside an area from place to place with local connections.

The operation of the model takes into account three levels of activity, representing inhibition (E0), active research (E1), or action (E2). The output function of a column is expressed in a table where the states of activation of the upper and lower layers of the column are defined by its internal and external outputs (44).

5.2. Application of the Cortical Column to Speech Recognition

We have applied the model of cortical column to various problems of pattern recognition, including printed character recognition and isolated word recognition. We will now

describe some experiments in acoustic-phonetic decoding of speech using the phoneme as processing unit.

The decoding network that was built for this purpose is made up of some thousands of columns grouped into three areas (45):

- a sensory area which performs several preprocessings of the input acoustic data,
- a "motor" area which simply displays the phoneme that is recognized at a given instant in time,
- an associative area which establishes the link between the sensory and motor functions and is therefore crucial for the recognition process.

A central issue in the design of the system concerns the learning phase during which the associative area has to be differentiated into sub-areas corresponding to the different phonemes to be recognized. At the very beginning the associative area is not differentiated (i.e. it is made up of a single area). The learning algorithm consists of a differentiation mechanism which recursively splits a zone of the area into two zones of opposite activity (E0/E2) whenever the original zone is in state E1 when a certain phoneme is presented to the sensory area (46).

5.3. Experimental Results

Two different experiments were carried out in order to test the architecture and the learning algorithm just described. The first one concerns the recognition of 7 French vowels / a i o y e /, whereas the second one concerns the recognition of the 6 fricative consonants / f s v z z /.

Recognition rates obtained for continuous speech pronounced without any precaution by a single speaker are respectively 87% for vowels and 96% for fricatives. The action of the differentiation algorithm on the associative area is illustrated in Figure 4 for vowels.

Work is presently in progress in order to generalize the system and enhance the performances of the learning algorithm.

6. CONCLUSIONS

We have presented in this paper different techniques of acoustic-phonetic decoding of speech, a fundamental problem in the framework of automatic speech recognition.

The various approaches proposed to solve the problem have been first briefly recalled, including the Hidden Markov Model which appears to be particularly efficient. We have then proposed to consider the PD as a knowledge intensive process. The APHODEX phonetic decoding expert system has been used in order to illustrate this approach. Results obtained show that the knowledge acquired from expert phoneticians can substantially improve the performances of the PD systems. The incorporation of this knowledge into efficient operational models such as HMM represents a good compromise for further developments in the field (47), that will also largely gain from the use of new, parallel models such as connectionist nets. A particular example of such models, the cortical column was also presented as well as its application to phonetic decoding.

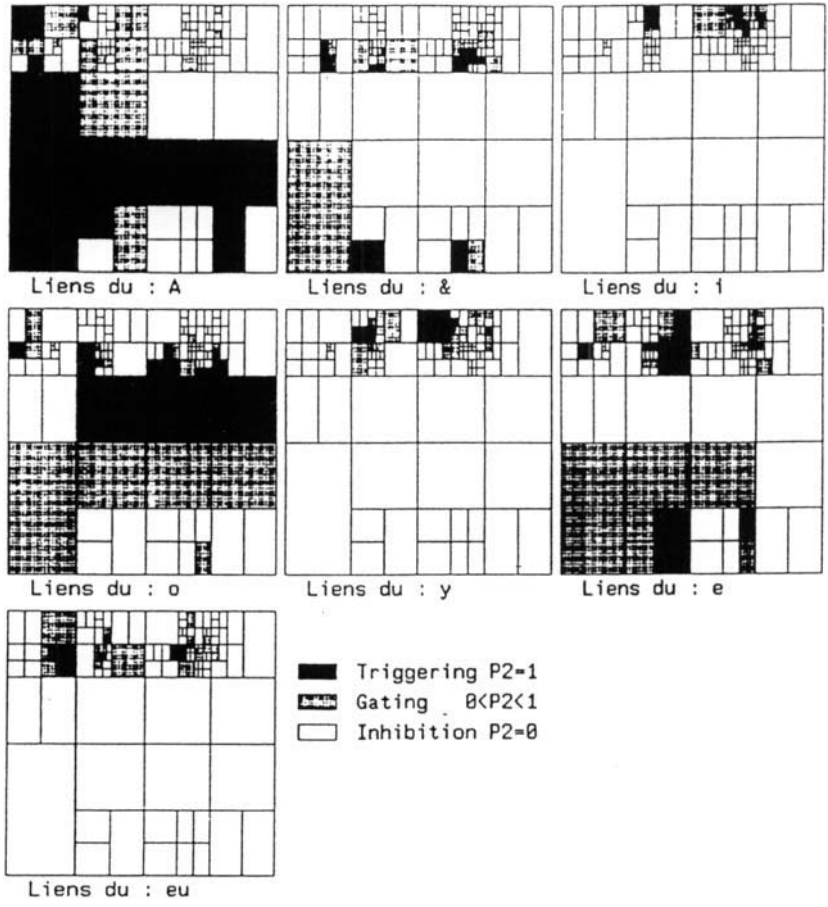


Figure 4. Division of the associative area for 7 French vowels

References

1. W. A. Lea, "Trends in Speech Recognition," Prentice-Hall (1980)

2. J. P. Haton, "Intelligence articielle en comprehension autom atique de la parole: Etat des recherches et comparaison avec la vision par ordinateur," T. S.I., 4, No.3, pp.265-287 (1985)

3. D. Klatt, "Review of the ARPA Speech Understanding Methods," JASA, 62, pp.1346-1366 (1977)

4. F. Lonchamp, "Reading Spectrograms: The View of the Expert," (in Fundamentals in Computer Understanding: Speech and Vision), J.P. Haton, editor, Cambridge University Press (1987)

5. R. Jakobson, G. Fant and M. Halle, "Preliminaries to Speech Analysis," MIT Press, Cambridge MA (1951)

6. J. Caelen, N. Vigouroux and G. Pérennou, "Structuration des informations acoustiques dans le projet ARIAL," Speech Com., 2, No.2-3, pp.219-222 (1983)

7. J. P. Haton, "Contribution a l'analyse, la parametrisation et la reconnaissance automatique de la parole," Doctorat d'Etat Thesis, University of Nancy (1974)

8. H. Meloni, "Etude et realization d'un systeme de reconnaissance automatique de la parole," Doctorat d'Etat Thesis, University of Marseille (1982)

9. J.S Liénard, "Analyse, synthese et reconnaissance automatique de la parole," Doctorat d'Etat Thesis, University of Paris (1972)

10. C. Scagliola, "Continuous Speech Recognition Without Segmentation: Two Ways of Using Diphones as Basic Speech Units," Speech Com., 2, No.2-3, pp.199-201 (1983)

11. G. Ruske, "Automatic Recognition of Syllabic Speech Segments Using Spectral and Temporal Features," IEEE ICASSP, Paris (1982)

12. O. Fujimura, "Syllables as Concatenated Demisyllables and Affixes," 9th Meeting of ASA (1976)

13. G. Perennou, "The ARIAL II Speech Recognition System," (in Automatic Speech Analysis and Recognition, J.P. Haton, editor, D. Reidel Publishing Company) (1982)

14. R. De Mori, "Extraction of Acoustic Cues Using a Grammar of Frames," Speech Com., 2, No.2-3, pp.223-225 (1983)

15. G. Mercier, "The KEAL Speech Understanding System," (in Spoken Language Generation and Understanding, J.C. Simon, editor, D. Reidel Publishing Company) (1983)

16. D. Fohr, J. P. Haton, F. Lonchamp and L. Sauter, "Methodes de segmentation syllabique en reconnaissance de la parole," XIV JEP, GALF, Paris (1985)

17. R. Bakis, "Continuous Speech Recognition via Centisecond Acoustic States," 91st Meeting ASA (1977)

18. H. G. Goldberg, "Segmentation and Labelling of Speech: A Comparative Performance Evaluation," Ph. D. Thesis, Carnegie Mellon University (1975)

19. T.G. Von Keller, "An On-Line Recognition System for Spoken Digits," JASA, 49, pp.1288-1296 (1971)

20. M. Baudry, "Etude du signal vocal dans sa representation amplitude-temps. Algorithme de segmentation et de reconnaissance de parole," Doctorat d'Etat Thesis, University of Paris (1978)

21. R. De. Mori and G. Giordano, "A Parser for Segmenting Continuous Speech into Pseudo-Syllabic Nuclei," Proc. ICASSP, Denver (1980)

22. R. Mizoguchi and O. Kakusho, "Continuous Speech Recognition Based on Knowledge Engineering Techniques," Proc. ICASSP, pp.638-640 (1984)

23. S. Roucos et al., "Vector Quantization for Very-Low-Rate," Proc. Global Telecom. Conf., Miami (1982)

24. J. K. Baker, "Stochastic Modelling for Automatic Speech Understanding," (in Speech Recognition, R. Reddy, editor, New York, Academic Press) (1975)

25. F. Jelinek, "Continuous Speech Recognition by Statistical Methods," Proc. IEEE, 64, No.4 (1976)

26. M. Cravero, R. Pieraccini and F. Raineri, "Definition and Evaluation of Phonetic Units for Speech Recognition by Hidden Markov Models," Proc. ICASSP, Tokyo, pp.2239-2242 (1986)

27. S. E. Levinson, L. R. Rabiner and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," The Bell System Technical Journal, 62, No.4, pp.1035-1074 (1983)

28. R. P. Lippmann, "An Introduction to Computing with Neural Nets," IEEE ASSP Magazine (1987)

29. R. A. Cole et al., "Speech as Patterns on Paper," (in Perception and Production of Fluent Speech, R.A. Cole, editor, Lea) (1980)

30. V. W. Zue and R. A. Cole, "Experiments on Spectrogram Reading," IEEE-ICASSP, pp.116-119, (1979)

31. N. Carbonell, M. O. Cordier, D. Fohr, J. P. Haton, F. Lonchamp and J. K. Pierrel, "Acquisition et formalisation du raisonnement dans un systeme expert de lecture de spectrogrammes vocaux," Colloque ARC, Orsay (1984)

32. N. Carbonell, D. Fohr and J.P. Haton, "APHODEX, an Acoustic-Phonetic Decoding Expert System," Int. Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), 1, No.2, pp.31-46, (1987)

33. G. E. Hinton, T. J. Sejinowski and D. H. Ackley, "Boltzmann machines: Constrainst Satisfaction Network that Learn," Technical Dept., CS-84-119, Carnegie-Mellon University (1984)

34. S. M. Peeling, R. K. Moore and M. J. Tomlinson, "The Multi-Layer Perceptron as a Tool for Speech Pattern Processing Research," Proc. IOA Autumn Conf. on Speech and Hearing (1986)

35. I. S. Howard and M. A. Huckvale, "Training Feature Detectors for Use in Automatic Speech Recognition," FASE Speech'88, Edinburgh (1988)

36. Y. Bengio and R. De Mori, "Use of Neural Networks for the Recognition of Place of Articulation," Proc. ICASSP-88, New York (1988)

37. H. Bourlard and C. Wellekens, "Multi-Layer Perceptrons and Automatic Speech Recognition," Proc. IEEE First Int. Conf. on Neural Networks, San Diego (1987)

38. N. A. McCulloch and W. A. Ainsworth, "Speaker-Independent Vowel Recognition Using a Multi-Layer Perceptron," FASE Speech'88, Edinburgh (1988)

39. A. Waibel et al., "Phoneme Recognition Using Time-Delay Neural Networks," Proc. ICASSP, New York (1988)

40. J. L Elman and D. Zipser, "Learning the Hidden Structure of Speech," JASA, 83, No.7, pp.1615-1626, (1988)

41. J. Leboeuf, "Un systeme connectionniste applique au traitement automatique de la parole," Thesis, University of Paris (1988)

42. T. Kohonen, "The 'Neural' Phonetic Typewriter," Computer, pp.11-22 (1988)

43. Y. Burnod, "An Adaptive Neural Network:The Celebral Cortex," Masson, Paris (1988)

44. F. Guyot, F. Alexandre, J. P. Haton and Y. Burnod, "The Cortical Column, a New Processing Unit for Cortex-Like Networks," (in From the Pixels to the Features, Elsevier) (1989)

45. F. Guyot, F. Alexandre and J. P. Haton, "Toward a Continuous Model of the Cortical Column: Application to Speech Recognition," Proc. ICASSP, Glasgow (1989)

46. C. Dingeon, F. Alexandre, F. Guyot and J. P. Haton, "un autre apprretissage cortical: differencier pour generaliser," Proc. Neuro-Nimes'89 (1989)

47. J. P. Haton, N. Carbonell, D. Fohr, J. F. Mari and A. Kriouille, "Interaction Between Stochastic Modelling and Knowledge-Based Techniques in Acoustic-Phonetic Decoding of Speech," Proc. ICASSP, Dallas (1987)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Is Rule-Based Acoustic-Phonetic Speech Recognition a Dead End ?

P. V. S. Rao

Tata Institute of Fundamental Research, Bombay, India

Abstract

Speech recognition using Sub-Word Units and Hidden Markov Models has become very popular in the recent past in view of their versatility and good performance. Acoustic-phonetic rule based recognition, which has been the main stay for many years, appears no longer to be so attractive. In this paper, we attempt a study of all three approaches and their implications to determine whether the acoustic-phonetic approach has a future at all. We conclude on the basis of our analysis that it holds a potential for the future. We also visualize the prospect of an integrated approach which combines the strategies of all three.

1. INTRODUCTION

Speech recognition consists in assigning to the tune domain acoustic signal a sequence of labels taken from a label bank or vocabulary consisting of a finite number of distinct labels. The labels could be sentences, words, phonemes and so on: any type of linguistically significant units. Recognition is feasible if there is some degree of consistency in the correspondence between the labels and their acoustic manifestations. This correspondence is complicated due to noise, articulatory laxity, speaking rate fluctuations, inter speaker differences and context dependencies in the acoustic signal. Segmentation of the continuous signal into units that correspond to any convenient type of units is therefore a non-trivial problem. Also, the contextual variability in the speech signal is much larger than can be conveniently dealt with as statistical variations.

Choosing long segments as units of recognition would have the advantage that such context effects become relevant only near the segment boundaries and therefore can essentially be neglected. The number of labels, however, would be quite large. To reduce this number, it is essential to choose short segments. Context effects become very significant in such a case; it therefore becomes necessary to incorporate knowledge concerning such context effects into the segmentation and recognition processes so that they can be effectively dealt with.

Word based recognition systems fall in the former category while phoneme based acoustic phonetic systems fall into the latter. Word based systems perform well for compact

vocabularies but problems arises as the vocabulary expands: the training procedure becomes complicated and recognition accuracy falls due to confusion between similar words. Acoustic-phonetic systems perform reasonably well, but require higher level information to achieve high enough recognition scores.

2. TYPICAL SCHEMES

The difficulties cited above arises mainly due to the fact the units of recognition (phonemes, diphones, syllables or words) are defined apriori for reasons of linguistic significances; they are not necessarily ideal from the points of view of segmentation or classification. There would therefore appear to be an advantage in choosing units using ease of segmentations and classifications as the main interia. It would be convenient if, for the units chosen, context sensitivity is minimal and the number of distinct prototypes is small.

2.1. Sub-Word Units

The problem of segmentation is solved here by adopting intuitively self-justifying but pre-specified criteria for segmentation, based on some kind of dissimilarity measure. The resulting segments are subject to clustering and classification. Alternatively, segmentation, clustering and classification are carried out in an integrated iterative loop for best results. This constitutes the training phase, which provides the sub-word unit prototypes.

A lexicon is then built, which stores for each word in the vocabulary, one or more SUB-WORD unit strings (one for each variant of the spoken word). These are then used as templates for recognition.

2.2. Hidden Markov Models

While the speech signal is visualized as a sequence of locally stationary segments, actual segmentation is circumvented in this approach. The model represents each segment as a state; it permits a probabilistic spread in the duration of each segment as well as in the values of the parameters measured for each of the segments. Each 'word' in the vocabulary has an associated model of this type. Recognition consists in determining which among all the word models has the highest likelihood of producing the given word sample.

2.3. Phonetic Feature Analysis

The acoustic signal is visualized as being composed of a number of (poorly defined) segments each corresponding to one (or more) phonemes. The segments here are neither stationary nor invariant.

The phonemes (or phoneme strings) are identified using the approach that experts use to read speech spectrograms. In other words, specific acoustic ones are sensed and these provide the means for inferring the presence (or absence) of phoneme types or individual phonemes. Early systems of this type primarily used quantitative statistical methods to

capture and use such (expert) knowledge. The more recent trend has been to use AI and expert systems techniques for achieving speech recognition.

3. DISCUSSION

The beginnings of both the sub-word and HMM schemes can be traced to the earliest speech recognition schemes ever presented (the Phonetic Typewriter of Olson and Blar which used hardware to decompose syllables into minimally distinguishable states with variable duration and a semi-manual training scheme).

The performance of both sub-word as well as HMM approaches has been very good (>95%) for small vocabularies (10 – 20 words). HMM recognition systems and sub-word systems where the sub-word are represented as HMM's perform well even for larger vocabularies. The scores improve further if higher level redundancies are made use of. Acoustic feature based systems, on the other hand, require to be augmented by strategies using higher level information before they can achieve performance levels comparable to the others.

Acoustic feature based techniques have been in vogue since the early days of speech recognition; of late, however, there has been a growing feeling – at least among enthusiasts for the more recent techniques (such as sub-word unit and HMM approaches) – that such feature based techniques are unlikely ever to be able to match the performance of the modern schemes. Our attempt here is to examine whether there is a case in support for the feature based schemes in the current context.

Sub-word units form an effective basis for recognition in the sense that individual words can be represented as mutually distinct strings of (sub-word unit) symbols. It is not clear, however, that such sub-word units can constitute an effective and universal (speaker-independent even if the language is specific) alphabet for speech. It is only then that it will be possible to train the system, say for new speakers, using a subset of the words in the vocabulary.

HMM provides (a) a powerful system of representing words (or even sub-word units) and their variability (along the time axes as well as along the dimensions of the acoustic parameters measured) and (b) computationally efficient techniques for building up such models and using them for recognition. It is compatible with and can incorporate probabilistic (e.g. digram and trigram) models which capture the higher level properties of language. It can capture temporal pattern in the variation of the measurement parameters, but only if these are more pronounced than the statistical variations in their values. The model in fact treats the statistical fluctuations in the measured parameters during each sampled interval as independent of each other and seems to ignore the time-wise continuity constraint on the parameter values. HMM uses transitions between adjacent and proximal states and explicit duration information t represent systematic variations in the speech signal; its representational effectiveness might tend to decrease for short segments, as the number of states (per model) reduces.

The individual states in each model are only locally defined and remain specific to that model, there is no convenient way of comparing states globally to form, say, an underlying common repertoire of states valid across words in the vocabulary.

Both sub-word and HMM approaches are statistical in nature and require very elaborate training. They do not utilize speech specific knowledge which is available; they have been termed ignorance models.

Rule-based approaches, being inspired by human spectrogram reading, are based on formant tracking. The human listener, nevertheless, uses this information very successfully. Spectrogram reading by experts employs strategies which are amenable to introspection. Even lower organisms are known to have 'feature extractors' which are sensitive to formant trajectories. These are strong 'existence theorems' which favour the rule-based approach.

The acoustic-phonetic rule-based approach captures in a transparent way the correspondence between articulatory processes and their manifestations at the acoustic level. By providing for rules which can be quantitative or in-between, it can be much more flexible than any statistical system can be. It can capture and incorporate properties which might escape, say, the HMM approach. A common set of rules can, for instance, capture the wide variations between the speech of men, women and children.

It uses a hierarchic approach in using available information, this is intuitively appealing and also known to be practically effective in introspective reading of spectrograms.

It is amenable to easy integration with schemes for representing and utilizing higher level features of language and speech.

A major objection to using expert systems for cognitive tasks arises from the fact that they do not work well in areas where the human uses tacit (as opposed to formally-required) knowledge. This objection is not applicable for spectrogram reading by experts which is subject to introspection and explication.

It has become clear in recent times that word length units are inconvenient (and perhaps necessary) for accurate recognition even of continuous speech. It might seem that rule-based systems are restricted to phoneme-level recognition and the consequent limitations of variability and context effects. This is not so. Feature-based lexicons are possible even at word level. These lexicons can contain 'skeletons' for each word, which consist only of 'strong cue' features. Using these, one can perform a first-shot recognition of individual words. (This would be comparable to the process humans utilize during rapid reading, for recognizing words, even phrases.) This would yield either a single word or a short-list of words. Recognition at a more detailed level using additional (weaker) cues can then be done to deal with the short-list.

4. CONCLUSIONS

The above discussion seems to indicate that rule-based approaches still have a future. In fact, rather than visualize the three approaches as being mutually exclusive, there is merit in recognizing their complementarity. In fact, they lay emphasis on three different (but important) facets of speech recognition: the units of representation (sub-word unit approach), modeling of the process (HMM) and the representation of speech specific knowledge (acoustic-phonetic rule-based approach). The sub-word and HMM approaches have already yielded the benefits deriving from a measure of integration. There is in existence, a system which uses a combination of HMM and acoustic cue-based approaches to improve performance. Integrating all three to take advantage of their combined strength

would be a very tempting prospect.

Even short of integration, the phonetic feature-based work could provide the speech related input for the sub-word units and the HMM the mathematical models needed. As well as by formal training, HMM's can be extended even by incorporation of speech specific knowledge.

Chapter 4

SPEECH RECOGNITION

This Page Intentionally Left Blank

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Speaker-Independent Phoneme Recognition Using Network Units Based on the *a posteriori* Probability

Jouji Miwa

Faculty of Engineering, Iwate University, 4-3-5 Ueda, Morioka, 020 Japan

Abstract

This paper describes a method of speaker-independent phoneme recognition using network units based on the *a posteriori* probability. The method is called a model using network units for recognition of phonemes (NEUROPHONE).

In this method, the convex characteristics of the time pattern of the *a posteriori* probability is adopted for the elimination of speaker individuality and coarticulation. The usage of the time pattern of the *a posteriori* probability is more suitable for the elimination than that of the distance.

In the first stage of the method, the *a posteriori* probability for all phonemes is calculated frame by frame from a 5 channel spectrum of 5 speech frames using Bayes rule. In the next stage, the convex part of the time pattern of the *a posteriori* probability is decided on as phoneme. The decision using the dynamic characteristics is more suitable for speaker-independent recognition than that with the static threshold.

In the network units, phonemes are discriminated with a nonlinear function, such as a quadric function. The weight coefficients in the units consist of statistical values such as the mean vector, the eigenvector, the eigenvalue and so on, so that the calculation time of the weight is smaller than that of the neural networks. The outputs of the units are analog values and not deterministic values such as those of the neural networks.

Recognition experiments are conducted with about 5300 phoneme samples in 166 Japanese city names uttered by 5 male speakers. These experiments are carried out under the condition of automatic phoneme spotting and without knowledge of the following vowels. The recognition scores obtained are 70% for the speaker-dependent case and 66% for the speaker-independent case.

1. INTRODUCTION

A speech recognition system, especially when it has a large vocabulary and if of the speaker-independent type, is useful as man-machine interface. A recognition system based on the unit phoneme is more extensive for continuous speech recognition and is easier to use with changing dictionaries than one based on the unit word. But the realization

use with changing dictionaries than one based on the unit word. But the realization of a system based on the unit phoneme is difficult because of speaker individuality and coarticulation.

For the elimination of speaker individuality and coarticulation, phoneme recognition using the dynamic feature, i.e. the convex time pattern of the parameter, is more suitable than recognition using the static feature. Some methods for phoneme recognition using the dynamic feature have been proposed [1-3]. But a dynamic feature such as the distance is sometime not suitable because of the folding effect at the portion of most likely being a phoneme in time.

In this paper, the convex characteristics of the time pat tern of the *a posteriori* probability is adopted for the elimination of speaker individuality and coarticulation for phoneme recogni tion [4-9]. A model of network units based on the *a posteriori* probability is applied for speaker-independent phoneme recognition.

2. SCALES FOR PHONEME RECOGNITION USING DYNAMIC PROCESSING [10]

Figure 1 shows an example of the acoustic parameter on the time axis. In the figure, two positions of the convex pattern are two phonemes; a difference of values of the parameter is effected by speaker individuality or coarticulation. At static processing for phoneme recognition, the portion over the threshold is only detected as phoneme so, that the first portion of the phoneme is correctly detected but the second portion of the phoneme since it causes a sake of the lower value than the threshold. At dynamic processing for phoneme recognition, the portion of the convex pattern is only detected as phoneme, so that the first and second portions of phonemes are correctly detected.

Three scales of a measurement are compared for the dynamic feature, i.e. the distance, the conditional probability density, and the *a posteriori* probability. These are defined as follows.

$$d = (x-\mu)^t C^{-1}(x-\mu)/2 + \ln(2\pi)^{\frac{N}{2}}|C|^{\frac{1}{2}}, \tag{1}$$

$$p(x|\omega) = \exp(-d) = \frac{\exp(-(x-\mu)^t C^{-1}(x-\mu)/2)}{(2\pi)^{\frac{N}{2}}|C|^{\frac{1}{2}}}, \tag{2}$$

$$p(x|\omega) = \frac{p(\omega)p(x|\omega)}{p(x)} = \frac{p(\omega)p(x|\omega)}{\sum_{k=1}^{K} p(\omega_k)p(x|\omega_k)}, \tag{3}$$

where x is the N dimensional vector of the acoustic parameters, μ is the N dimensional mean vector, C is the covariance matrix of the parameters, $p(\omega)$ is the *a priori* probability, and K is the total number of phoneme categories.

Figure 2 shows the characteristics of the distance, the conditional probability density and the *a posteriori* probability on the parameter axis. From the figure, the decision surface of the three cases is the same at static processing.

Figure 3 shows the time pattern of the distance, the conditional probability density and the *a posteriori* probability on the time axis. In the case of under-shooting, all time

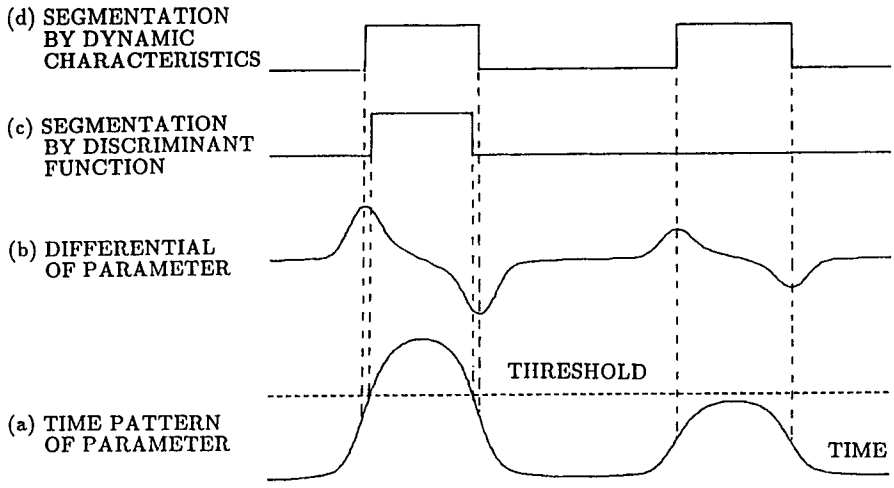


Figure 1. An example of segmentation using the threshold or the convex characteristic.

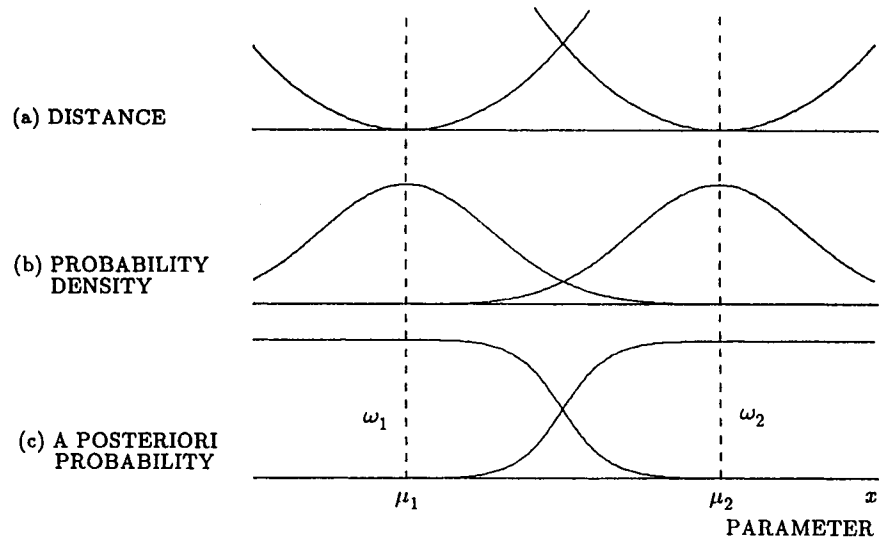


Figure 2. Characteristics of the distance, the conditional probability density and the *a posteriori* probability on the parameter axis.

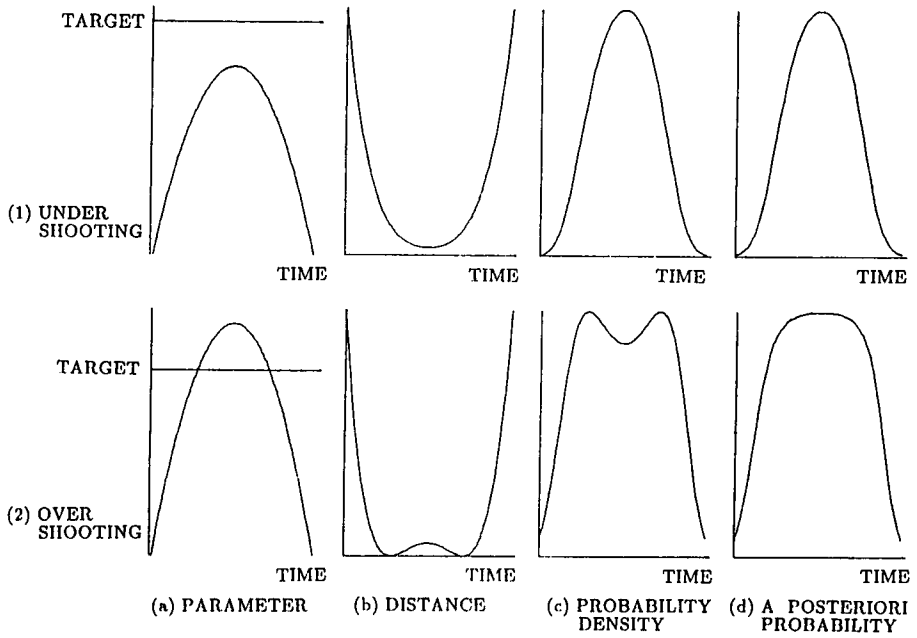


Figure 3. Time pattern of the distance, the conditional probability density and the *a posteriori* probability on the time axis.

patterns are convex at the phoneme segment. In the case of over-shooting, the time pattern of only the *a posteriori* probability shows a convex pattern at the phoneme segment but the time pattern of the distance and the probability density show a concave pattern. So the *a posteriori* probability is a suitable measurement for the dynamic processing for phoneme recognition.

3. MODEL OF THE NETWORK UNITS FOR RECOGNITION OF PHONEMES

The covariance matrix in eqs. (1) and (2) is composed of the eigenvectors and the eigenvalues as.

$$C = \sum_{i=1}^{N} \lambda_i \Phi_i \Phi_i^t, \tag{4}$$

where λ_i and Φ_i are i-th eigenvector and eigenvalue, respectively. The inverse matrix and the determinant are:

$$C^{-1} \sum_{i=1}^{N} \Phi_i \Phi_i^t / \lambda_i, \tag{5}$$

$$|C|^{-1} = \prod_{i=1}^{N} 1/\lambda_i. \tag{6}$$

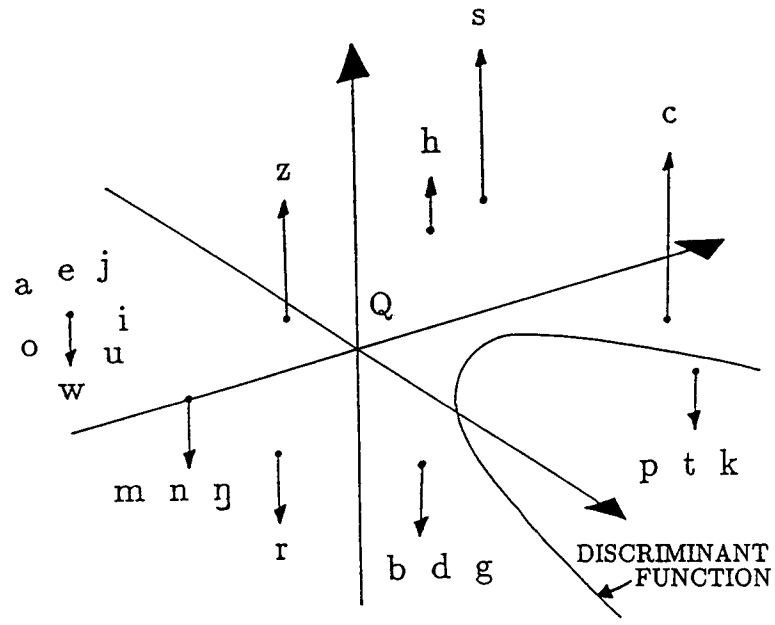


Figure 4. An example of phoneme space and the discriminant function.

From eq.(2), the conditional probability function is calculated as.

$$P(x|\omega) = \frac{\exp(-\sum_{i=1}^{N}((x-\mu)^t\Phi_i)^2/(2\lambda_i))}{(2\pi)^{N/2}|C|^{1/2}}. \tag{7}$$

The modified conditional probability function is represented by the truncation of the matrix to order M:

$$P(x|\omega) = \frac{\exp(-\sum_{i=1}^{M}((x-\mu)^t\Phi_i)^2/(2\lambda_i))}{(2\pi)^{M/2}|C|^{1/2}}. \tag{8}$$

A model of the network units for recognition of phonemes (NEURO PHONE) is represented by eq. (8).

In the model, the nonlinear discriminant function is nonlinear i.e. quadric. Figure 4 shows an example of phoneme space and the discriminant function. Figure 5 shows a model of the network units for phoneme recog nition using the *a posteriori* probability.

4. PHONEME RECOGNITION SYSTEM

Figure 5 shows a schematic diagram of the recognition system of phonemes using the convex time pattern of the *a posteriori* probability. The recognition system consists of

four stages, i.e. the feature extraction, the calculation of the *a posteriori* probability, smoothing of the probability, and the spotting and decision of the phoneme.

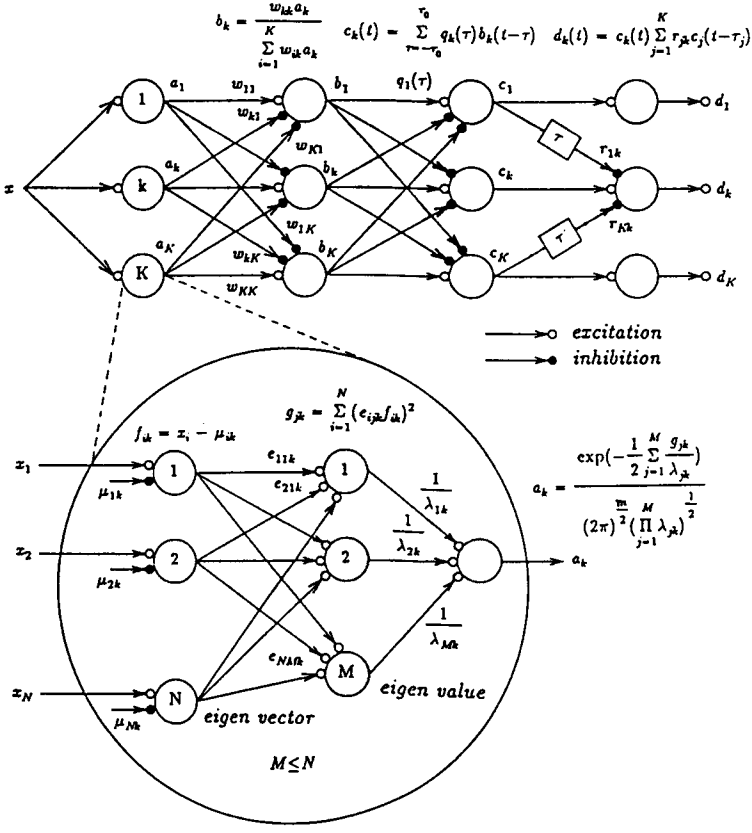


Figure 5. A schematic diagram of the network units for recognition of phonemes.

In the feature extraction stage, the input speech wave is digitized at 10 kHz and the input signal is analyzed every 10 ms frame and then the logarithmic spectrum using 5 channel BPF is calculated from the 12 order LPC spectrum in every frame. The frequency range [10] of the 5 channel BPF is shown in table 1. Figure 6 shows as example the 5 channel BPF spectrum and the LPC spectrum for /morioka/ uttered by a male speaker.

Figure 7 shows the examples of the mean vectors of /p/, /t/, /k/ and silence for the standard pattern. The mean vectors consist of 5 channel spectrum of 5 frames. The patterns of the mean vectors of /p/, /t/ and /k/, discussed [11], are diffuse- falling, diffuse-rising and compact, respectively.

In the phoneme spotting and decision stage, the portion of the phoneme is detected by peak picking of the smoothed *a posteriori* probability, i.e. the convex pattern. The segments of the phoneme are detected by picking of the convex and concave peaks of the

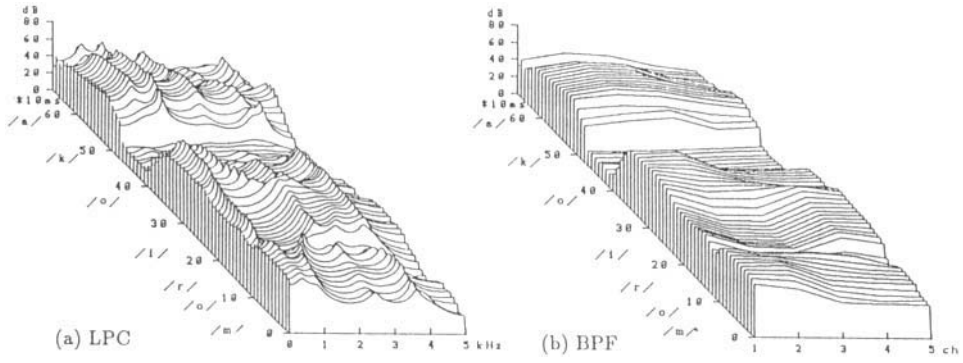


Figure 6. An example of the LPC spectrum and the 5 channel BPF spectrum for /morioka/ uttered by a male speaker.

Table 1. Frequency range of 5 channel BPF.

Channel	Frequency range (Hz)	/a/	/o/	/u/	/i/	/e/
1	250 - 600		F1	F1	F1	
2	500 - 1000	F1	F2			F1
3	900 - 1800	F2		F2		
4	1500 - 3000				F2	F2
5	2500 - 4500	F3	F3	F3	F3	F3

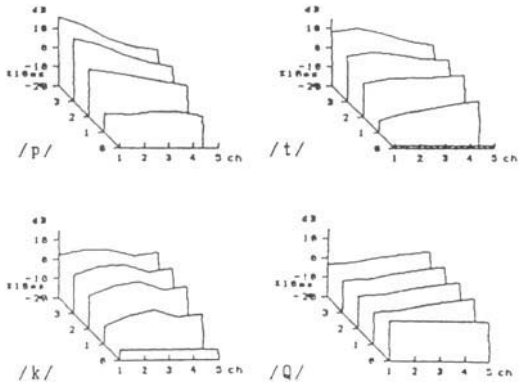


Figure 7. An example of the mean vectors for /p/, /t/, /k/ and silence.

difference of the smoothed *a posteriori* probability. And the phoneme is decided on by the category of the convex pattern of the *a posteriori* probability under the conditions of maximum probability and duration of segment.

Figure 8 shows an example of phoneme recognition using the *a posteriori* probability. In the figure, speech is the word /morioka/ uttered by a male speaker and all phonemes are correctly recognized.

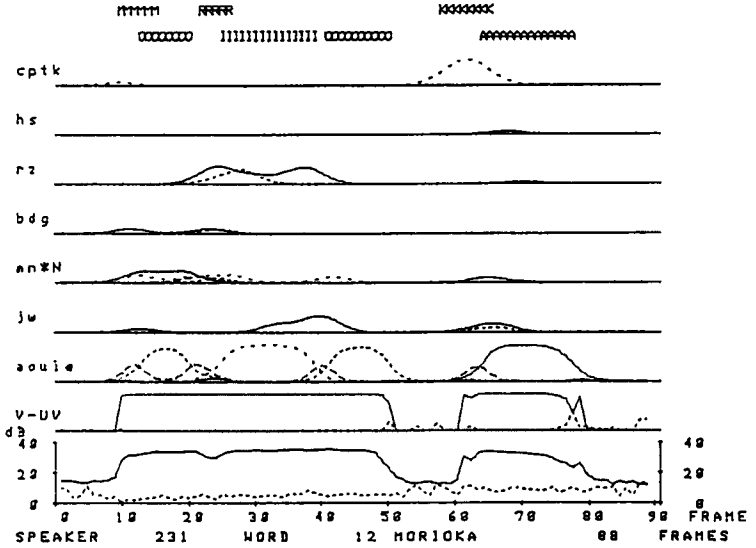


Figure 8. An example of phoneme recognition using the *a posteriori* probability.

5. RECOGNITION EXPERIMENTS

The recognition experiments are conducted with about 5300 phoneme samples in 166 Japanese city names uttered by 10 male speakers. The samples of 5 speakers are used for closed test and the samples of the other 5 speakers are used for open test. These experiments are carried out under the condition of automatic phoneme spotting and without knowledge of the following vowels. The recognition scores obtained are 70% for the speaker-dependent case and 66% for the speaker-independent case. From the experiments, phonemes are automatically detected and are highly recog nized. Table 2 shows a percentage of the correct, insertion and deletion phonemes every phonemes for speaker-independent case.

Table 2. Experimental results of phoneme recognition. (5 male speakers, Speaker-independent case)

Phoneme	Correct	Insertion	Deletion	Samples
/a/	95%	5%	5%	1255
/o/	86%	12%	11%	500
/u/	71%	19%	24%	328
/i/	85%	9%	11%	551
/e/	76%	14%	15%	105
/j/	48%	12%	20%	194
/w/	76%	8%	15%	109
/m/	24%	36%	47%	311
/n/	35%	51%	48%	157
/N/	30%	37%	35%	138
/b/	66%	19%	24%	68
/d/	59%	17%	15%	54
/r/	36%	53%	44%	166
/z/	25%	40%	58%	84
/h/	27%	44%	60%	151
/s/	56%	14%	32%	279
/c/	10%	23%	44%	164
/t/	55%	21%	43%	210
/k/	64%	22%	31%	485
/Q/	60%	35%	40%	20
Mean	66%	18%	23%	5329

6. CONCLUSION

In this paper, a method of speaker-independent phoneme recognition using network units based on the *a posteriori* probability is proposed. In the method, the convex characteristics of the time pattern of the *a posteriori* probability is adopted for the elimination of speaker individuality and coarticulation. The usage of the time pattern of *a posteriori* probability is more suitable for the elimination than that of the distance.

The recognition experiments are conducted with about 5300 phoneme samples in 166 Japanese city names uttered by 5 male speakers. These experiments are carried out under the condition of automatic phoneme spotting and without knowledge of the following vowels. The recognition scores obtained are 70% for speaker-dependent case and 66% for speaker-independent case.

[Work supported by Grant-in-Aid for Scientific Research on Priority Areas, The Ministry of Education, Science and Culture of Japan]

References

1. K. Kido, S. Makino, J. Miwa and Y. Niitsu, "Spoken word recognition system for unlimited speakers," Proc. IEEE ICASSP, pp.735-738, Tulsa (1978)

2. M. Yokota, K. Akizawa and H. Kasuya, "Automatic Identification of Vowels in Connected Speech Uttered by Multiple Speakers," Trans. Inst. E. C. E. Japan, J65-D, 1, pp.134-135, (1982)

3. Y. Kobayashi, Y Ohmori and Y. Niimi, "Recognition of vowels in continuous speech based on the dynamic characteristics of WLR distance," J. Acoust. Soc. Japan, E7, 1, pp.29-38, (1986)

4. T. Saito and J. Miwa, "Phoneme recognition using convex characteristic of a posteriori probability," Trans. Com mittee Speech Research, Acoust. Soc. Japan, S85-56, (1985).

5. J. Miwa and T. Yamazaki, "Recognition of Unvoiced Plosives Using Time Pattern of a Posteriori Probability," Trans. Committee Speech Research, Acoust. Soc. Japan, S87-56, (1987)

6. J. Miwa and T. Yamazaki, "Automatic Detection and Recognition of Phoneme Using Phoneme Inhibition Mechanism by a Posteriori Probability," Trans. Committee on Speech Research, Acoust. Soc. Japan, SP88-101, (1989)

7. J. Miwa and T. Yamazaki, "On the Model of Network Units for Recognition of Phoneme Based on Probability," Trans. Committee on Speech Research, Acoust. Soc. Japan, SP89-27, (1989)

8. J. Miwa, "Speaker-independent recognition of unvoiced plos- ives using convex time pattern of a posteriori probability," 2nd Joint Meeting Acoust. Soc. America and Japan, Hawaii, PPP-3, (1988)

9. J. Miwa, "Speaker-independent recognition of unvoiced plos- ives using convex time pattern of a posteriori probability," Preprints 2nd Symposium Advanced Man-Machine Interface Through Spoken Language, 25-1-8, Hawaii, (1988)

10. K. T. Kim, J. Miwa and K. Kido, "Recognition of isolated Korean digits using band pass filters based on FFT," J. Acoust. Soc. Japan, (E)4, 4, pp.179-187, (1983)

11. S. E. Blumstein and K. N. Stevens, "Acoustic invariance in speech production: Evidence from measurements of the spec tral characteristics of stop consonants," J. Acoust. Soc. America, 66, 4, pp.1001-1017 (1979)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Unsupervised Speaker Adaptation in Speech Recognition

Hiroshi Matsumoto and Yasuki Yamashita

Deparment of Electrical and Electronic Engineering, Shinshu University
500 Wakasato, Nagano-shi, 380 Japan

Abstract

This paper presents an unsupervised speaker adaptation method from short utterances based on a minimized fuzzy objective function. In this method, the code spectrum for the templates is adapted to that of an input speaker by interpolating the estimated speaker-difference vectors at given points in the spectral space of the templates. The speaker-difference vectors are estimated so as to minimize the fuzzy objective function for the adapted reference codebook under some constraints. The fuzziness and constraint parameters are examined using 28-vocabulary SPLIT-based word recognition tests with reference patterns from a male speaker. The result showed that this method with a fuzziness of 1.5, gives a 9.0% higher recognition rate for the four male speakers with 1.8 s of training samples than that based on minimum VQ distortion ($F = 1.0$). Under the best conditions with 16 speaker-difference vectors, this method improved the average recognition rate for 20 male speakers from 92.5% with no adaptation to 97.5% with 3.6 s of training samples and to 98.5% with 28 word training samples. Furthermore, a sequential speaker adaptation, using input speech itself, attained a recognition rate of 97.4% for the male speakers. Finaly, a speaker normalization scheme based on fuzzy mapping with the adapted codebook was found to be effective for a non-SPLIT based recognition system.

1. INTRODUCTION

Inter-speaker variation of speech is one of the important issues in designing a speech recognition system for use by unrestricted speakers, particularly for a large-vocabulary recognition system. Of several approaches to this problem, speaker adaptation or normalization is a practical solution to alleviate the speaker variabilities. Speaker adaptation algorithms are classified into supervised (or text-dependent) and unsupervised (or text-independent) ones. The former method requires users to speak specified texts, whereas the latter does not impose such a restriction upon users. Since the goal of speaker adaptation is to realize a speech recognition system which operates as a speaker independent one, it is desirable to dynamically adapt to a new speaker during recognition without any

information about the input speaker nor the text spoken. This paper aims to realize such an adaptive speech recognition system.

The inter-speaker differences of speech that affect the recognition performance comprise the static and dynamic characteristics. The approach studied here intends to eliminate the static spectral differences, which are mostly caused by the idiosyncrasies in a speaker's vocal apparatus. Although a speaker's learned characteristics, such as dialect, are also reflected in the spectra, this study will not be concerned with this problem since it might require phoneme information for adaptation.

For supervised speaker adaptation, vector quantization codebook mapping techniques have been successfully applied for speaker adaptation [1-5]. In the unsupervised cases, recent studies have proposed VQ-based speaker adaptation techniques as well [6-7]. These studies have represented the adapted spectra in terms of estimated speaker-difference vectors in several spectral subspaces. These adaptation algorithms are not based upon any explicit criterion on adaptation. Recently, in the area of speechcoding, Shiraki and Honda [8] proposed a minimum vector quantization distortion criterion for speaker adaptation. While this method is very effective to reduce VQ distortion, it requires even longer training samples to obtain a stable solution. This is because the standard VQ distortion criterion discards any information other than the distance between the input vector and the closest codeword. So, in order to realize a rapid speaker adaptation, this paper proposes an adaptation algorithm based on a minimized fuzzy objective function [9], which utilizes the relative position of the input vector to all the codewords. Furthermore, this method is applied in a sequential speaker adaptation scheme which successively modifies the reference spectra in parallel with recognition [10].

Following the algorithm, this method is applied to a small-vocabulary VQ-based word recognition system [9]. The reference pattern is created by a prototype male speaker, and only the spectral codebook is adapted to input speakers, keeping the word dictionary (code sequences) unchanged. The several parameters involved (e.g., fuzziness) and the effectiveness are examined using recognition test with 21 male and 4 female speakers. Finally, a speaker normalization method based on fuzzy mapping with the adapted codebook is compared with the speaker adaptation method [11].

2. FORMULATION OF SPEAKER ADAPTATION

2.1. Spectral Adaptation Model

At text-independent speaker recognition, the recognition score has been found to be improved by utilizing the phoneme-dependent speaker characteristics in addition to the phoneme-independent ones[12]. This suggests that speaker differences of spectra need to be modeled depending on the phonemes or spectra. Thus, as shown in fig. 1, the present method represents the spectral differences between an input and the prototype speaker in terms of a small number of estimated speaker-difference vectors, which will be called "adaptation vectors", $\{\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \cdots, \boldsymbol{\Delta}_M\}$, at typical points, $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_M\}$, in the spectral space of the prototype speaker. With these adaptation vectors, all the code vectors, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_L\}$, are adapted to the input speaker by the following weighted sum of $\{\boldsymbol{\Delta}_k\}$;

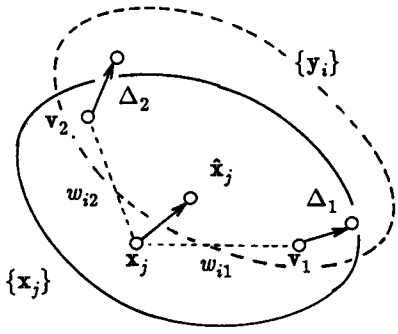


Figure 1. Conceptual illustration of spectral adaptation.

$$\hat{\boldsymbol{x}}_j = \boldsymbol{x}_j + \sum_{k=1}^{M} w_{jk} \boldsymbol{\Delta}_k, \tag{1}$$

where the weighting coefficient w_{jk} is determined by the distance $d(\boldsymbol{x}_j, \boldsymbol{v}_k)$ between $\boldsymbol{x}_j$ and $\boldsymbol{v}_k$ as

$$w_{jk} = \frac{d(\boldsymbol{x}_j, \boldsymbol{v}_k)^{-p}}{\sum_{k=1}^{M} d(\boldsymbol{x}_j, \boldsymbol{v}_k)^{-p}}, \tag{2}$$

where

$$d(\boldsymbol{y}_i, \hat{\boldsymbol{x}}_j) = \|\boldsymbol{y}_i - \hat{\boldsymbol{x}}_j\|. \tag{3}$$

This interpolation formula with $p = 1.0$ was first proposed by Niimi et al [4] and Shiraki et al [13]. The parameter p, which will be called the interpolation parameter, was introduced in the present study in order to adjust the continuity of the adapted vectors $\hat{\boldsymbol{x}}_j$ in spectral space[6]. Furthermore, since p is included in the objective function for estimation, which will be described later, it also controls the smoothing region from which speaker-difference vectors will be estimated. Thus, a smaller p might result in smoother speaker-difference vectors.

2.2. Criterion for Spectral Adaptation

In order to estimate the adaptation vectors from short utterances, the proposed method evaluates the goodness of adaptation in terms of all the distances from the input vector to the neighboring adapted code vectors $\{\hat{\boldsymbol{x}}_j\}$ instead of the nearest distance in vector quantization [8]. That is, the unknown adaptation vectors, $\{\boldsymbol{\Delta}_k\}$, are estimated so as to minimize the following weighted sum of all the distances between the input vector and every adapted code vector over all the training vectors, $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\}$, i.e., the fuzzy objective function [14];

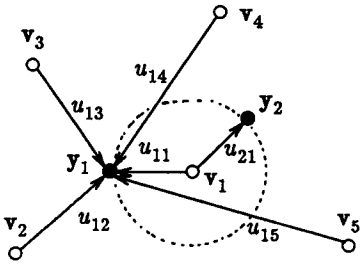


Figure 2. Comparison of distortions for two different vectors y_1 and y_2 between hard and fuzzy decisions.

$$J_F(u_{i1}, \cdots, u_{iL}; \Delta_1, \cdots, \Delta_M) = \sum_{i=1}^{N} \sum_{j=1}^{L} u_{ij}^F \cdot d(\boldsymbol{y}_i, \hat{\boldsymbol{x}}_j), \tag{4}$$

under the constraint,

$$\sum_{j=1}^{L} u_{ij} = 1. \tag{5}$$

In eq. (4), F is called the degree of fuzziness and controls the relative contribution of $d(\boldsymbol{y}_i, \hat{\boldsymbol{x}}_j)$ to J_F through the membership functions, u_{ij}. $F = 1$ corresponds to a hard decision, i.e., the standard vector quantization. Figure 2 illustrates an example of the difference between hard and fuzzy decisions. Whereas vector quantization results in the same distortion by two different vectors $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, the fuzzy decision can take their different positions into account using every distance to the neighboring code vectors.

In addition to the above constraint, the mean norm of $\{\Delta_k\}$ is bounded by the following inequality to avoid an excessive spectral modification, especially for short training samples,

$$R = \frac{1}{M} \sum_{k=1}^{M} \|\Delta_k\|^2 \leq \eta(E_y^2 - E_x^2). \tag{6}$$

In this inequality, η is a norm constraint parameter to adjust the degree of restriction, and E_y^2 is the vector quantization distortion of the training samples, $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\}$, with the prototype speaker's codebook, and E_x^2 is that produced in making the prototype speaker's codebook.

3. ADAPTATION ALGORITHM

3.1. Basic Algorithm

The fuzzy objective function is minimized with respect to $\{\Delta_k\}$ and $\{u_{ij}\}$ under the norm constraint (6). Although this minimization problem is difficult to solve for both

$\{\boldsymbol{\Delta}_k\}$ and $\{u_{ij}\}$ simultaneously, we can derive the optimum solution if one of these variable sets is fixed.

With fixed $\{\boldsymbol{\Delta}_k\}$, the membership functions are given by

$$u_{ij} = \frac{d(\boldsymbol{y}_i, \hat{\boldsymbol{x}}_j)^{\frac{-1}{F-1}}}{\sum_{r=1}^{L} d(\boldsymbol{y}_i, \hat{\boldsymbol{x}}_r)^{\frac{-1}{F-1}}}. \tag{7}$$

On the other hand, with fixed $\{u_{ij}\}$, the vectors $\{\boldsymbol{\Delta}_k\}$ are given by the normal equations:

$$\sum_{r=1}^{M}(W_{rk} + \lambda \cdot \delta_{rk})\boldsymbol{\Delta}_r = \sum_{i=1}^{N}\sum_{j=1}^{L} u_{ij}^{F} \cdot w_{jk} \cdot (\boldsymbol{y}_i - \boldsymbol{x}_j), \tag{8}$$

$$(k = 1, \cdots, M)$$

where

$$W_{rk} = \sum_{i=1}^{N}\sum_{j=1}^{L} u_{ij}^{F} \cdot w_{jr} \cdot w_{jk}, \tag{9}$$

and δ_{rk} is Kronecker's delta; and λ denotes a Lagrange multiplier associated with the constraint (6). Since R is proved to be a monotonic decreasing function of λ, the norm constraint can be satisfied by adjusting the magnitude of λ[15].

Thus, given a set of typical points $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_M\}$ in the prototype speaker's spectral space, a local minimum solution for $\{\boldsymbol{\Delta}_k\}$ is derived by the successive approximation procedure below, using the separate minimization steps above. In the subsequent iterative algorithm, the superscript (l) represents iteration number l, starting with $l = 0$ for the initial guess.

(a) Set $\hat{\boldsymbol{x}}_j^{(0)} = \boldsymbol{x}_j$ (i.e., $\boldsymbol{\Delta}_k^{(0)} = 0$) and $\boldsymbol{v}_k^{(0)} = \boldsymbol{v}_k$.Then at step $l : l = 1, \cdots$:

(b) Calculate $u_{ij}^{(l)}$ by substituting $\hat{\boldsymbol{x}}_j^{(l-1)}$ for $\hat{\boldsymbol{x}}_j$ in eq. (7) and the vector quantization distortion E_y.

(c) Put $\lambda = \lambda_0 \sum_{k=1}^{M} W_{kk}$.

(d) Compute $\{\boldsymbol{\Delta}_k^{(l)}\}$ for $\{u_{ij}^{(l)}\}$ fixed from eq. (8).

(e) If $R \leq \eta(E_y^2 - E_x^2)$, go to (f); otherwise, return to (d) with $\lambda = \lambda \cdot \beta(\beta > 1)$.

(f) Calculate the adapted code vectors $\hat{\boldsymbol{x}}_j^{(l)}$ from eq. (1) and (2), and update the $\boldsymbol{v}_k^{(l)}$'s by

$$\boldsymbol{v}_k^{(l)} = \boldsymbol{v}_k^{(l-1)} + \boldsymbol{\Delta}_k. \tag{10}$$

(g) Compare $J_F^{(l)}$ to $J_F^{(l-1)}$: if $(J_F^{(l-1)} - J_F^{(l)})/J_F^{(l-1)} < \delta$, stop; otherwise, return to (b) with $l = l + 1$.

In the above algorithm, β represents the step size to decrease the mean norm $R(\lambda)$ and δ is a convergence criterion. These parameters will be determined experimentally. Although there is no assurance that $J_F^{(l)}$ converges to the global minimum, the above iteration ensures that $J_F^{(l)}$ is monotonically decreasing with l.

The above algorithm, which will be referred to as the "single-step" adaptation, may fall into an undesirable local minimum for a larger M because of too much flexibility. Therefore, as in the stage-wise procedure in regression analysis[16], the single-step adaptation with a fixed number of typical points can be applied in a step-by-step manner, increasing the number of typical points from 1 to M. This algorithm is called the "stage-wise" adaptation. A similar procedure was first applied in a speaker adaptation method based on a clustering technique by Shiraki et al [13] and Furui [7].

3.2. Relationship to the piecewise VQ-error Averageing method

When p is set to infinity, w_{jr} is equal to one if $\boldsymbol{x}_j$ is nearest to the typical point $\boldsymbol{v}_k$ and otherwise zero;

$$\lim_{p \to \infty} w_{jr} = \delta_{rk}, \tag{11}$$

where

$$k = \arg\min_r \, d(\boldsymbol{x}_j, \boldsymbol{v}_r). \tag{12}$$

Furthermore, in the case of $F = 1.0$ and $\eta = \infty$, eq. (8) becomes

$$\boldsymbol{\Delta}_k = \frac{\sum_{i=1}^{N} \delta_{jk} \cdot (\boldsymbol{y}_i - \boldsymbol{x}_j)}{\sum_{i=1}^{N} \delta_{jk}}, \tag{13}$$

where $\boldsymbol{x}_j$ is the best match codeword to the input $\boldsymbol{y}_i$;

$$j = \arg\min_j \, d(\boldsymbol{y}_i, \boldsymbol{x}_j). \tag{14}$$

Thus, the adaptation vector for the k-th typical point is equal to the average vector quantization error over the training vectors, $\boldsymbol{y}_i$, quantized to $\boldsymbol{x}_j$'s which are closest to the k-th typical point. As a result, the piecewise VQ-error averageing method previously proposed [6] is a special case of the present method.

3.3. Sequential Speaker Adaptation

Since the fuzzy objective function method is effective even for short training samples, it might be possible to implement a sequential adaptation algorithm which dynamically

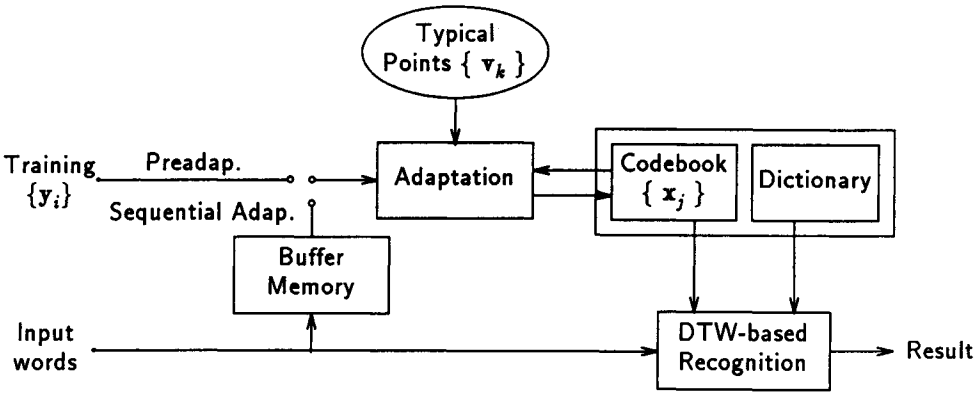


Figure 3. Word recognition system with preadaptation or sequential adaptation to a new speaker.

adapts to a new speaker using the input speech itself as he speaks to the system[10]. As illustrated in fig. 3, the spectral codebook for the reference patterns is successively updated by a new codebook at every short period which is adapted to a block of past input frames accumulated in a buffer. This scheme will be called sequential speaker adaptation, and the previous scheme a preadaptation which adapts a new speaker before recognition. In the sequential speaker adaptation, the number of typical points in the reference spectral space, the norm constraint, and the convergence threshold should be appropriately set depending on the duration and variation of buffered training samples in order to avoid an unreasonable modification of the codebook.

4. DATA BASE AND RECOGNITION PROCEDURE

The speech data consist of two repetitions of 28 city names uttered by 21 male speakers and four female speakers. The speech signals were digitized at a sampling frequency of 10 kHz with the frequency band limited to 4.2 kHz. The linear predictive autocorrelation method with 12 poles was applied to these speech data with a constant frame of 25.6 ms, a frame shift of 12.8 ms, first-order backward differences for preemphasis, and a Hamming window. The 1st to 15th LPC cepstral coefficients were used as the components of the feature vector.

Figure 3 shows a SPLIT-based word recognition system [17] with preadaptation or sequential speaker adaptation processors [10]. The codebook with 128 codewords of the prototype speaker was obtained by the standard LBG algorithm [18] using the cepstral distortion measure. The set of typical points, $\{v_k\}$, consisted of the nearest codeword in the codebook to those in a small-size codebook. For time alignment, an unconstrained-endpoint DTW algorithm was used. In the subsequent experiments, the reference patterns were created by one repetition uttered by a prototype speaker, whose vocal tract length is estimated to be close to the average for male speakers. The first repetition of each word

from other speakers was used as the training sample and the other as a test sample, and the training and test samples were also exchanged. In order to prepare training sets of different durations, all of the 28 word data (22 s on average) for each speaker were divided into 1/12-, 1/6- and 1/3-subsets (1.8, 3.6, and 7.2 s in average duration, respectively). The performance of the spectral adaptation is evaluated by recognition rates and/or the DTW distance (per frame) between the same words uttered by the reference and test speakers.

5. EXPERIMENTS

5.1. Effect of Fuzziness

First, the effects of fuzziness in the stage-wise adaptation was examined using the twelve 1/12-subsets and three 1/3-subsets of the training samples for the four female and four male speakers who result in the lowest recognition scores without adaptation. Figure 4 shows the average recognition rates for the male and female speakers as a function of fuzziness. In this experiment, the number of typical points, interpolation parameter, and the norm constraint parameter were set to 16, 1.0, and 1.0, respectively. From these results, it is clearly shown that the fuzzy objective function criterion with $F = 1.5$ gives the highest recognition rates. In particular, for 1.8 s of training samples, this adaptation method improved the recognition rate for the male speakers from 87.5% for the standard vector quantization distortion criterion ($F = 1.0$) to 96.5% for $F = 1.5$.

Furthermore, in order to reduce the computational cost, the effect of pruning the membership functions is examined. The membership functions whose value is less than a threshold are removed from the summation in eqs. (8) and (9). Figure 5 shows the average recognition rates and the percentage of used membership functions with respect to the threshold under the above-mentioned best condition with 1/12-subsets of training samples. As seen this figure, when the threshold is set to the mean value of u_{ij} (i.e., 1/L), the amount of computation can be reduced to about 20% of the full calculation without any performance degradation.

5.2. Effect of the Interpolation Parameter

In order to find an optimum value of the interpolation parameter, p, recognition experiments with several values of p were carried out using the same speakers and 1/3-subsets of the training data, as in section 5.1. In this experiment, stage-wise adaptation with $M = 16$, $F = 1.5$, and $\eta = 1.0$ was applied. Figure 6 shows the average recognition scores and DTW distances for the four male and four female speakers as a function of p. As seen in this figure, the effect of p on the recognition scores is not clear, especially for male speakers. However, the DTW distances tend to be smallest around $p = 4.0$. In the subsequent experiments, the value of p will be set to 2.0 for computational simplicity.

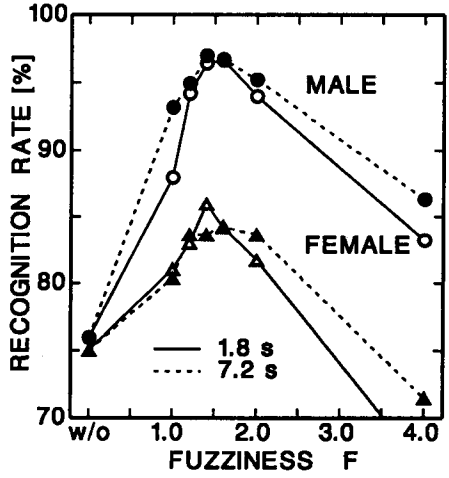


Figure 4. The average recognition rates for the 4 male and 4 female speakers as a function of the fuzziness for stage-wise adaptation with $M = 16$, $p = 1.0$, and $\eta = 1.0$ using 1.8s of training samples.

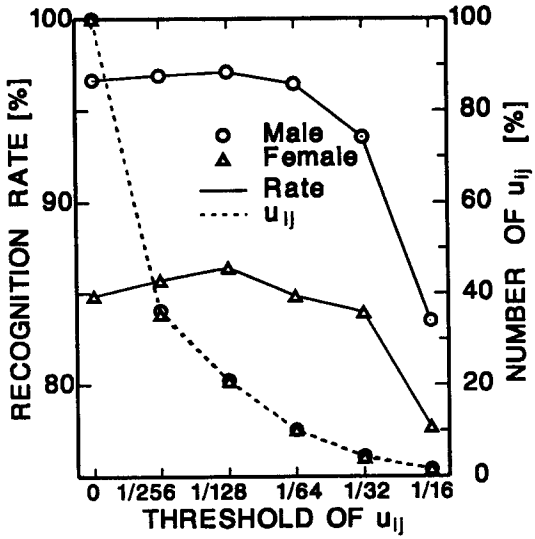


Figure 5. The average recognition rates and the percentage of membership functions as a function of the threshold on membership functions under the same experimental conditions as in fig. 4.

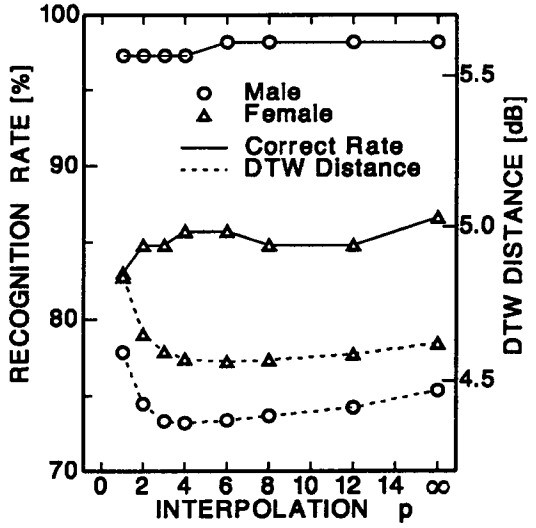


Figure 6. The average recognition rates and DTW distances for the 4 male and 4 female speakers as a function of the interpolation parameter for stage-wise adaptation with $M = 16$, $F = 1.5$, and $\eta = 1.0$.

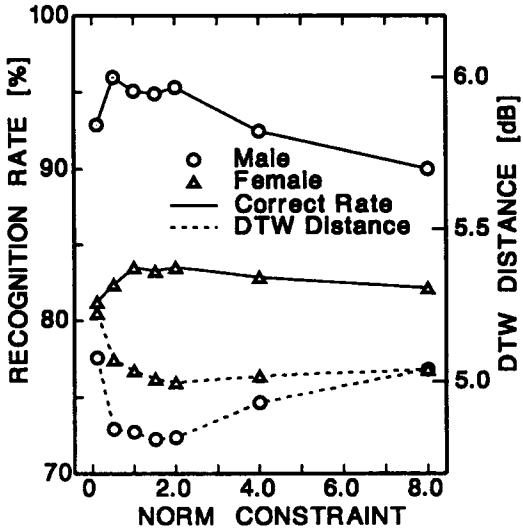


Figure 7. Effects of the norm constraint for the 4 male and 4 female speakers on recognition rates and DTW distances as a function of the Lagrange multiplier λ for single-stage adaptation with $M = 16$, $F = 1.5$, and $p = 1.0$.

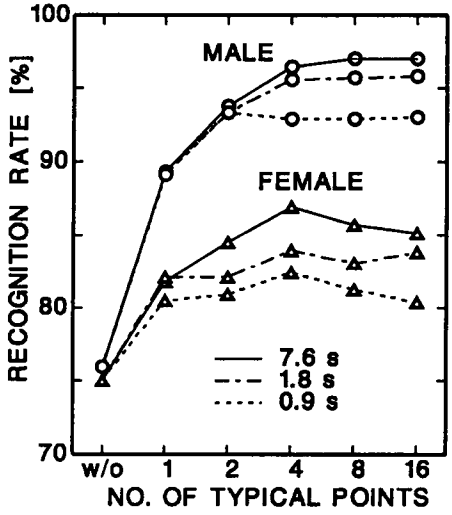


Figure 8. The average recognition rates for the 4 male and 4 female speakers as a function of the number of typical points for stage-wise adaptation with $M = 16$, $F = 1.5$, $p = 1.0$, and $\eta = 1.0$ for three training durations.

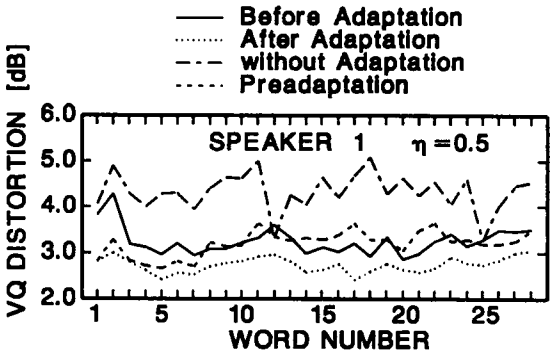


Figure 9. VQ distortions for each of 28 words with sequential speaker adaptation, with preadaptation, and without adaptation.

5.3. Effect of the Norm Constraint

In order to examine the effect of the norm constraint, an adaptation experiment was carried out using the same data as in section 5.1. In this experiment, since the norm constraint is expected to be effective for short training samples, single-stage adaptation with $M = 16$, $p = 1.0$, and $F = 1.5$ was applied. The DTW distance and the recognition rate as a function of η are shown in fig. 7. As seen in this figure, the norm constraint, η, is most effective for values from 1.0 to 2.0. Thus, this value will be set to 1.0 in the following experiments. It should be noted here that the normal equations for the hard decision, i.e. $F = 1.0$, were singular without the norm constraint.

5.4. Effect of the Number of Typical Points

First, the single-stage and stage-wise adaptation algorithms were compared. The result showed that the stage-wise adaptation provides a slightly higher recognition rate than the single-stage adaptation. Therefore, the effect of the number of typical points was examined under stage-wise adaptation with $F = 1.5$, $p = 1.0$, and $\eta = 1.0$ using the 1/24-, the 1/12- and 1/3-subsets of the training samples. Figure 8 shows the average recognition rates as a function of the number of typical points for the four male and four female speakers. As seen in this figure, the recognition rate is improved as the number of the typical points increases, but tends to saturate or decrease beyond a small number of points for the shorter training samples. Thus, the number of typical points should be determined depending on the amount of training samples.

5.5. Experiments on Sequential Speaker Adaptation

In these experiments, the codebook was sequentially adapted before recognition of each test word using the past two words as training samples. According to the experiments in section 5.4, the number of typical points in the spectral space is set to eight if the number of training frames is less than 100, and otherwise it is set to 16. The convergence threshold, δ, is set to the slightly larger value of 0.1. The experiment was carried out for the four worst male speakers with a norm constraint, η, of 0.5 and 1.0. Since the recognition rate depends on the order of test words, the recognition tests were conducted using three different sequences of test words.

As a result, the average recognition rate for the four speakers, three word sequences, and two repetitions was found to be 96.3% for $\eta = 0.5$, which was 5% higer than that for $\eta = 1.0$. The lower rate for $\eta = 1.0$ is caused by excessive spectral modification, which depends on each test word. Thus, the smaller norm constraint of 0.5 is appropriate for sequential adaptation. Furthermore, in order to gain insight into the sequential adaptation process, fig. 9 shows an example of VQ distortions of each word produced by the following four codebooks: (1) the codebook without adaptation (short dashed line), (2) the preadapted codebook obtained by the 1/3-subset (dotted line), and (3) the codebooks before and after sequential adaptation (solid and dashed line, respectively). As seen in this figure, although the difference between VQ distortions before and after adaptation for each word does not converge to zero, the VQ distortion for each word before adaptation approaches that of the preadaptation codebook after about the fifth word.

Table 1. Average correct scores for the three speaker groups in various recognition experiments with a prototype male speaker.

Experimental Condition	20 male Speakers	4 male Speakers	4 female Spekaers
Without Adaptation	92.7%	79.4%	76.8%
Preadaptation (3.6s)	97.5	97.3	85.8
Preadaptation (22s)	98.5	98.2	87.1
Sequential Adap. (2 words)	97.4	96.3	82.9
Supervised Adap. (22s)	98.2	98.6	96.4
Speaker Dependent	99.9	99.5	99.6

5.6. Comparative Experiments

First, the recognition experiments with speaker adaptation proposed here were carried out for the 20 male speakers as well as the four male and four female speakers under the following conditions: (1) without adaptation, (i.e., "speaker independent" recognition using references from a prototype speaker) (2) with preadaptation using the 1/6-subsets and 28 words, (3) with sequential adaptation with $\eta = 0.5$ as in section 5.5, (4) with a supervised speaker adaptation similar to the method by Shikano et al. [1], where a mapped codebook was created without iteration using all of the 28 word data, and (5) with speaker dependent reference. Table 1 compares the average recognition rates for the above experiments: the average recognition rate for the 20 male speakers increased from 92.5% for no adaptation to 98.5% for the unsupervised adaptation with all of the 28 word data, which is almost the same as the average score of 98.2% for the case of supervised adaptation. Furthermore, only the 1/6-subset of training data (3.6 s) attained an average score of 97.5%, close to the above score. For the female speakers, the unsupervised speaker adaptation also improved the average recognition score from 76.8% without adaptation to 85.8% for the 1/6-subsets and 87.1% for all the training data. However, these scores are still much lower than that for supervised adaptation. Finally, sequential adaptation with $\eta = 0.5$ improved the average recognition rate for the 20 male speakers to 97.4%, which is close to that obtained by preadaptation with the 1/6-subsets of training samples.

5.7. Speaker Normalization by Fuzzy Mapping

The speaker adaptation method presented above is the most suitable for the VQ-based recognition systems, such as the SPLIT or HMM method. However, this method can be applied to any other recognition scheme, such as a neural network, continuous HMM, or statistical methods by mapping the input spectra onto the reference spectral space utilizing a fuzzy mapping technique [5] with the adapted codebook [11]. First, in this mapping or "normalization" scheme an input vector is encoded by a fuzzy vector quantizer with the adapted codebook, $\{\hat{\boldsymbol{x}}_j\}$, using eq. (7), generating a set of membership

functions, $\{u_{ij}\}$. Second, since each adapted codeword has a one-to-one correspondence with the original reference codeword, the mapped vector, $\hat{\boldsymbol{y}}_i$, is derived by decoding, $\{u_{ij}\}$, with the reference codebook:

$$\hat{\boldsymbol{y}}_i = \frac{\sum\limits_{u_{ij}<u_T} u_{ij}^F \cdot \boldsymbol{x}_j}{\sum\limits_{u_{ij}<u_T} u_{ij}^F}, \tag{15}$$

where the summation is restricted to the terms whose u_{ij} are less than a threshold, u_T.

Actually, this normalization scheme was compared with the adaptation scheme under the same experimental conditions as in section 5.6. As a result, the normalization scheme provided slightly higher recognition scores than the adaptation scheme. This impovement might be attributed to the fact that the mapped vectors are composed of the reference spectra themselves. Thus, although the normalization scheme needs an extra amount of computation for the fuzzy encoding and decoding, it is superior to the adaptation scheme in performance and applicability to various recognition methods.

6. CONCLUSION

This paper has presented an unsupervised speaker adaptation method based on a minimized fuzzy objective function. As a result of SPLIT-based word recognition experiments with 28 vocabulary words, the adaptation with 1.5 of fuzziness attained higher recognition scores than that with a minimum VQ distortion criterion. Particularly, using as short as 3.6 s of training samples, this method attained high scores, close to those when using all of the training samples for male speakers.

Furthermore, the sequential speaker adaptation by two successive words under the norm constraint of 0.5 was found to be as effective as a preadaptation method using about 10 words as training samples. Finally, it was shown that the speaker normalization scheme based on a fuzzy mapping technique with the adapted codebook is superior to the adaptation scheme.

In future work, it is necessary to improve the adaptation accuracy for large speaker differences, such as those between male and female speakers. In addition, it is important to develop a method to eliminate speaker differences in dynamic characteristics due to coarticulation.

References

1. K. Shikano, "Speaker adaptation through vector quantization," IEEE Int. Conf. Acoust., Speech and Signal Processing, Tokyo, Japan, pp.2643-2646, Paper No.49.5.1. (1986)

2. R. Schwartz, Y. L. Chow, F. Kubala, "Rapid speaker adaptation using a probabilistic spectral mapping," IEEE Int. Conf. Acoust., Speech and Signal Processing, Dallas, TX. pp.633-636, Paper No.15.3.4. (1987)

3. H. Bonneau and J. L. Gauvain, "Vector quantization for speaker adaptation," IEEE Int. Conf. Acoust., Speech and Signal Processing, Dallas, TX. pp.1434-1437, Paper No.34.6.1. (1987)

4. Y. Niimi and Y. Kobayashi, "Speaker-adaptation of a code book of vector quantization," Proc. of European Conference on Speech Technology, (1987)

5. S. Nakamura and K. Shikano, "A comparative study of spectral mapping for speaker adaptation," IEEE Int. Conf. Acoust., Speech and Signal Processing, Albuquerque, NM. pp.157-160, Paper No.S3.7 (1990)

6. Y. Yamashita and H. Matsumoto, "Speaker adaptation for word recognition based on error vectors in VQ," Trans. Committee Speech Research Acoust. Soc. Japan, SP87-118, pp.35-42 (1988)

7. S. Furui, "Unsupervised speaker adaptation algorithm based on hierarchical spectral clustering," Trans. Committee Speech Research Acoust. Soc. Japan, SP88-21, pp.1-8, (1988)

8. Y. Shiraki and M. Honda, "Piecewise-linear adaptive vector quantization and its application to speaker adaptation," Acoust. Soc. Japan Meeting in Spring, No.1-1-4 (1988)

9. H. Matsumoto and Y. Yamashita, "Unsupervised speaker adaptation of spectra based on a minimum fuzzy vector quantization error criterion," The Second Joint Meeting of Acoustical Society of America and Acoustical Society of Japan, G19, (1988)

10. H. Matsumoto, "An unsupervised dynamic speaker adaptation method based on fuzzy vector quantization,' Acoust. Soc. Japan Meeting in Spring, No.2-P-12, (1989)

11. Y. Nakatoh, Y. Kondo and H. Matsumoto, "An aplication of fuzzy-VQ-based unsupervised speaker adaptation to statistical recognition method," Acoust. Soc. Japan Meeting in Spring, No.2-P-13, (1989)

12. H. Matsumoto, "Text-independent speaker identification from short utterances based on piecewise discriminant analysis," Computer Speech and Language 3, pp.133-150, (1989)

13. Y. Shiraki and M. Honda, "Speaker adaptation algorithms for segment vocoder," Trans. Committee on Speech Research, Acoust, Soc., Jap, SP87-67, (1987)

14. J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms," Plenum Press, New York, (1981)

15. D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," J. Soc. Indust. Appl. Math., vol.11, No.2, pp.431-439, (1963)

16. N. R. Draper and H. Smith, "Applied Regression Analysis," Jhon Wiley & Sons, Inc., New York, (1966)

17. N. Sugamura, K. Shikano, and S. Furui, "Isolated word recognition using phoneme-like templates," IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, (1983)

18. Y. Linde, A. buzo and R. M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Comm., COM-28, 1, pp.84-95 (1980)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

A Japanese Text Dictation System Based on Phoneme Recognition and a Dependency Grammar

Shozo Makino, Akinori Ito, Mitsuru Endo and Ken'iti Kido

Research Center for Applied Information Sciences, Tohoku University, Sendai, 980 Japan

Abstract

A Japanese text dictation system has been developed based on phoneme recognition and a dependency grammar. The phoneme recognition is carried out using the modified LVQ2 method which we propose. The linguistic processor is composed of a processor for spotting Bunsetsu-units and a syntactic processor with semantic constraints. In the processor for spotting Bunsetsu-units, using a syntax-driven continuous-DP matching algorithm, the Bunsetsu-units are spotted from a recognized phoneme sequence and then a Bunsetsu-unit lattice is generated. In the syntactic processor, the Bunsetsu-unit lattice is parsed based on the dependency grammar. The dependency grammar is expressed as the correspondence between a FEATURE marker in a modifier-Bunsetsu and a SLOT-FILLER marker in a head-Bunsetsu. The recognition scores of the Bunsetsu-unit and phoneme were 73.2% and 86.1% for 226 sentences uttered by two male speakers.

1. INTRODUCTION

A number of continuous speech recognition systems[1-6] have been reported. However, there still remain several problems in developing a continuous speech recognition system for ordinary Japanese text utterances. The traditional continuous speech recognition systems only dealt with particular linguistic information in a specified domain. As necessary techniques for the construction of a Japanese text dictation system, we should develop the following methods:

(1) A phoneme recognition method with high accuracy,

(2) A Bunsetsu-unit spotting method with high accuracy and with a small amount of computation, where the Bunsetsu-unit is a unit which is uttered with one breath and is composed of a conceptual word followed by several functional words, and

(3) An efficient parsing method taking into account syntactic and semantic constraints.

In order to construct the system, we propose a modified LVQ2 method for the phoneme recognition, syntax-driven continuous-DP for spotting Bunsetsu-units and a CYK-based parsing method using semantic constraints for the syntactic processing. Finally, we will describe the perfomance of the system when text speech is uttered Bunsetsu by Bunsetsu.

2. OUTLINE OF THE JAPANESE TEXT DICTATION SYSTEM

Figure 1 shows a schematic diagram of the Japanese text dictation system. The system is composed of an acoustic processor[7], a processor[8-10] for spotting Bunsetsu-units, and a syntactic processor with semantic constraints[11-13]. In this research the speech to be recognized includes spoken sentences whose syntax and semantic structures are syntactically and semantically reasonable. We use sentences from a scientific paper, where the sentences contain 843 conceptual words and 431 functional words.

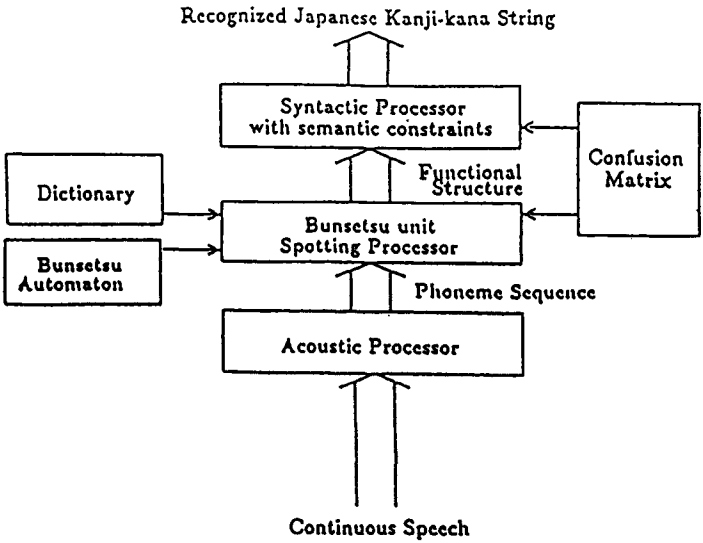


Figure 1. A schematic diagram of the Japanese text dictation system.

The input speech is analyzed using a 29 channel band-pass filter bank. In the acoustic processor a phoneme sequence is recognized from the input speech using the modified LVQ2 method[7] which we propose.

The structure of Japanese sentences is effectively described by a two-level grammar which consists of an intra-Bunsetsu grammar[14] and an inter-Bunsetsu grammar[15]. Accordingly, the analysis of the Japanese sentences is divided into two stages. The first stage is the extraction of the Bunsetsu-unit candidates from the recognized phoneme sequence. The second one is the analysis of the dependency structure between the Bunsetsu-unit candidates. The Bunsetsu-unit can be modeled by a finite-state automaton, which is convenient to describe the syntactic structure. The test set perplexity[16] of the finite-state automaton is 230.

The proposed linguistic processor can be extended so as to deal with ordinary Japanese text utterances, even if the number of conceptual words is increased.

3. PHONEME RECOGNITION USING A MODIFIED LVQ2 METHOD

The learning vector quantization(LVQ,LVQ2) methods were proposed by Kohonen et al.[17]. McDermott et al.[18] developed a shift-tolerant phoneme recognition system based on the LVQ2 method. In the LVQ2 algorithm proposed by Kohonen, two reference vectors are modified at the same time if the first nearest class to an input vector is incorrect and the second nearest class to the input vector is correct. We propose a modified training algorithm for the LVQ2 method. In the modified LVQ2 algorithm, n reference vectors are modified at the same time if the correct class is within the N-th rank where N is set to some constant.

Figure 2 shows the process of the modified LVQ2 algorithm. In step 1, reference vectors are chosen using a K-Means clustering method from each class. In step 2, the nearest reference vector of each class to an input vector is selected. In step 3, the rank of the correct class is computed. When the rank of the correct class is n, we assume that the reference vector of the correct class is m_n. In step 4, n is checked to see whether or not n falls in the range of $2 \leq n \leq N$. In step 5, the check is made to see whether or not the input vector falls within the small window, where the window is defined around the midpoint of m_1 and m_n. In step 6, the i-th reference vector is modified according to the following equations.

$$\begin{aligned} [m_i]^{t+1} &= [m_i - \alpha(n)(x - m_i)]^t \quad (i = 1, 2, \cdots, n-1), \\ [m_n]^{t+1} &= [m_n + \alpha(n)(x - m_n)]^t . \end{aligned}$$

The phoneme recognition system is similar to the shift-tolerant model proposed by McDermott et al.[18]:

(1) 8 mel-Cepstrum coefficients and 8 Δ mel-Cepstrum coefficients are computed for every frame from the 29 chanel BPF spectrum. Each reference vector is represented by 112 coefficients(7 frames $\times$ 16 coefficients). Each class was assigned 15 reference vectors chosen by the K-Means clustering method.

(2) A 7-frame window is moved over the input speech and yields a 112(16$\times$7) dimensional input vector every frame.

(3) In the training stage the modified LVQ2 method is applied to the input vector as described above.

(4) In the recognition stage we compute the distances between the input vector and the nearest reference vector within each class.

(5) From this distance measure, each class is assigned an activation value a_w as follows:

$$a_w(c,t) = 1 - d(c,t) / \sum_i d(i,t),$$

where d, c, and t are distance, class, and time, respectively.

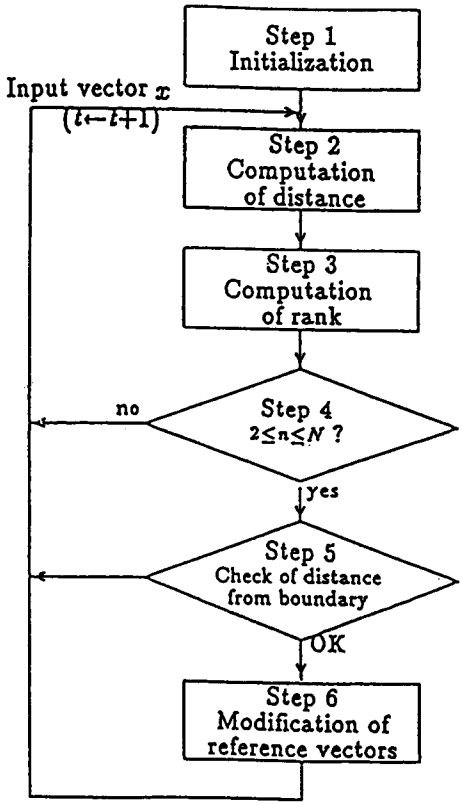


Figure 2. Algorithm of the modified LVQ2 method.

(6) The final activation a_f is computed by summing 9 activation values as follows:

$$a_f(c,t) = \sum_{j=-4}^{4} a_w(c,t+j)g_w(j),$$

where g_w is the weight of the Gaussian type window.

(7) The class with the maximum activation value is regarded as phoneme candidate of each frame. The activation value is regarded as a posteriori probability $P(C_k|t_k)$ of the phoneme C_k at the t_k-th frame.

(8) The optimum phoneme sequence is computed from the phoneme candidate sequence using dynamic programming and the duration constraints[19].

4. PROCESSOR FOR SPOTTING BUNSETSU-UNITS

There are two traditional methods(methods 1 and 2) for extraction of the Bunsetsu-units, as shown in fig. 3. The first method(method 1) spots the Bunsetsu-units using all possible Bunsetsu-unit reference patterns. However, the method needs a large amount of storage and computation, since the number of Bunsetsu-units is huge in the Japanese text dictation system. On the other hand, the second one(method 2) detects the conceptual words and the functional words independently. However, the Japanese language has many functional words with short lengths such as copulae, endings of conjugated words, and auxiliary verbs. The current spotting method shows poor performance in spotting words having short duration and therefore insertion and deletion errors are common, although the amount of computation for this method is very small. The method(method 3) which we propose[8-10] is an intermediate one. This method spots the Bunsetsu-units based on a finite-state automaton representing the Japanese Bunsetsu-unit structure. We call this method a syntax-driven continuous-DP matching algorithm.

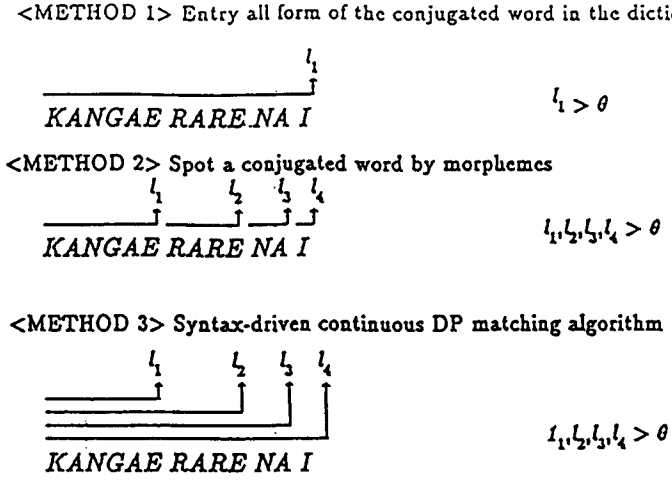


Figure 3. An example of the Bunsetsu-unit spotting method.

Because the word order of functional words is fixed, the intra-Bunsetsu grammar can be expressed as a finite-state automaton. Figure 4 shows the Bunsetsu-unit structure model for the Japanese text dictation system. The four arcs, "adverb", "verb/adjective ", "noun", and "adnominal", represent conceptual words and the other arcs represent functional words (or null transitions). Double circles in the figure indicate a terminal state.

Figure 5 shows an example of processing with syntax-driven continuous-DP. The processing with syntax-driven continuous-DP starts with a conceptual word. If the probability of the final phoneme of the stem of the conceptual word exceeds a threshold, the automaton generates the next word. The calculation of the probability for the next word is carried out using the final probability obtained at the previous stage as the initial

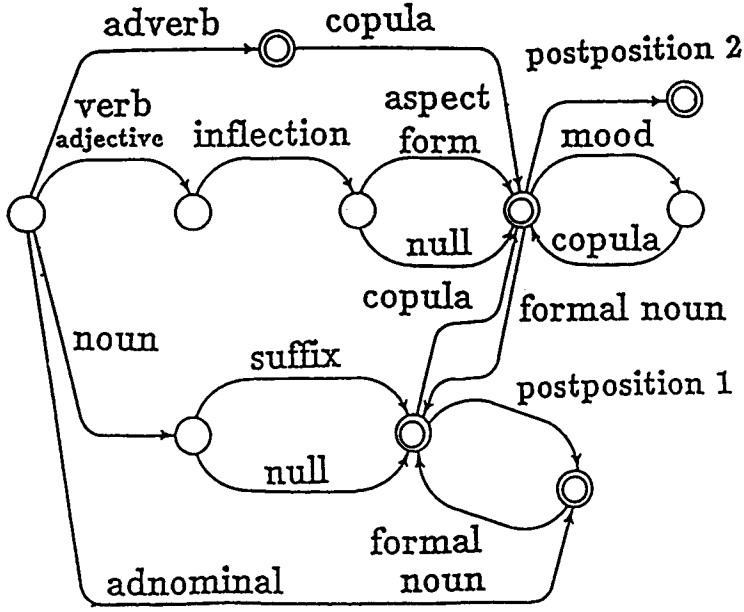


Figure 4. The outline of the structure of a Bunsetsu-unit.

value. In the same manner the calculation of the probability continues until each path reaches its terminal state, as shown in fig.4. If the probability at the terminal state exceeds a threshold, the Bunsetsu-units of every valid path to that terminal state are recognized as candidates, and thus a Bunsetsu-unit lattice is made from an input phoneme sequence. Using syntax-driven continuous-DP the Bunsetsu-units are spotted from an input phoneme sequence and simultaneously the morpheme analysis is carried out.

The results for conjugated word spotting[10] can be seen in fig. 6. This figure shows the relation between the number of candidates per 100 input phonemes and the detection score. Method 3 shows a performance similar to method 1. Method 2 detects 20 times more candidates compared to method 3 when the detection score is 90%.

5. SYNTACTIC PROCESSOR WITH SEMANTIC CONSTRAINTS

Syntactic processing is applied to the candidates of the Bunsetsu-unit in the Bunsetsu-unit lattice detected by syntax-driven continuous-DP. The inter-Bunsetsu grammar is implicitly expressed as the correspondence of the markers in the functional structure of the Bunsetsu-unit candidates. The two partial trees are merged when the modifier's FEATURE marker set and the head's SLOT-FILLER marker set both contain the same syntactic markers. Figure 7 shows an example of the merging of the two partial trees. All

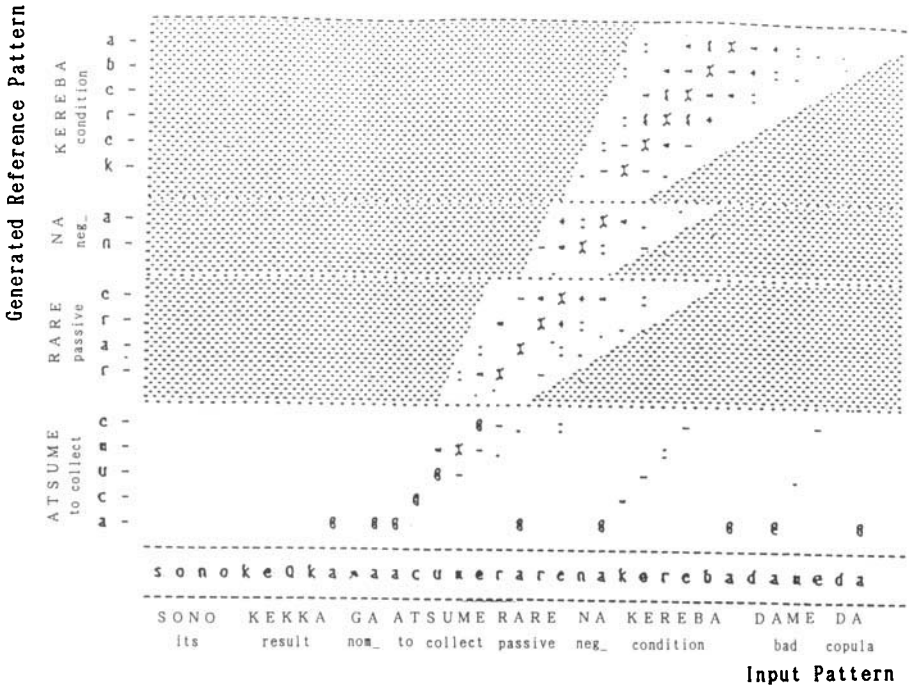


Figure 5. An example of the processing with syntax-driven continuous-DP.

syntactic dependency, including modification, complement, object and subject are treated in the same framework. We use 95 functional features[11] as the FEATURE markers and the SLOT-FILLER marker.

The algorithm[12,13] for parsing is based on the Cocke-Younger-Kasami(CYK) algorithm using the beam search. Multiple candidates for a sentence are obtained for one input phoneme sequence using this algorithm. The computation amount for the parsing is $O(N^3D^2)$, where N is the length of the input phoneme sequence and D is the maximum number of the stored candidates.

6. EXPERIMENTAL RESULTS

The training based on the modified LVQ2 method was carried out with speech samples of the 212 word vocabulary uttered by 7 male and 8 female speakers. The recognition experiments of 30 phonemes were carried out with speech samples of the 212 word vocabulary uttered by another 3 male and 2 female speakers. Table 1 shows the phoneme recognition scores. The result for $N = 2$ corresponds to the original LVQ2 method. The recognition scores for $N \geq 3$ are higher than the score for $N = 2$. This indicates superiority of the modified LVQ2 method to the original LVQ2 method.

We applied this method to a multi-speaker dependent phoneme recognition task for continuous speech uttered Bunsetsu by Bunsetsu. Table 2 shows the phoneme recognition

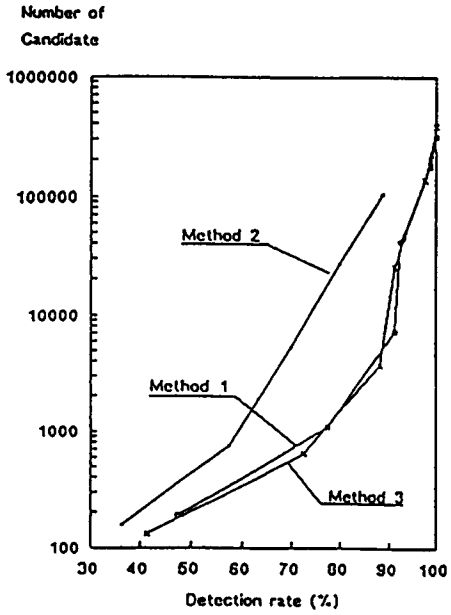


Figure 6. The relation between the number of candidates per 100 input phonemes and the detection score when the phoneme recognition score is 85%.

Table 1. Speaker-independent phoneme recognition scores for spoken words using the modified LVQ2 method and the method for selecting the optimum phoneme sequence.

Rank of reference vector for training	Phoneme recognition score	Deletion score	Insertion score
N=2	83.1	2.0	11.3
N=3	85.6	1.9	9.8
N=7	86.5	1.7	9.0

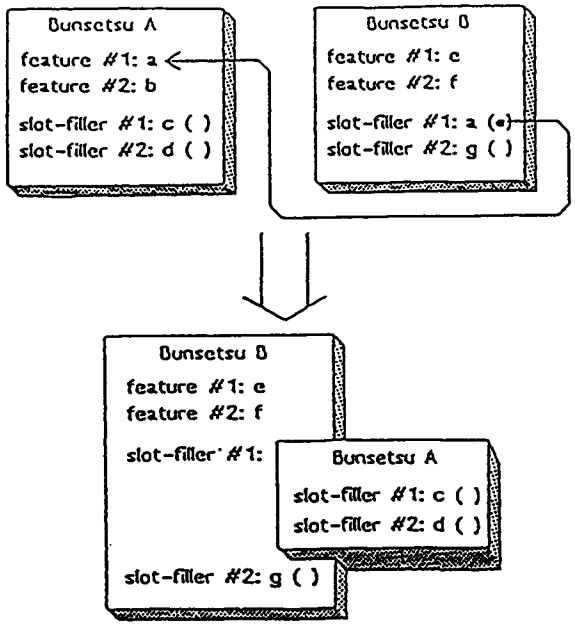


Figure 7. An example of merging two Bunsetsu-units.

scores for 2 male speakers. The training, based on the modified LVQ2 method, was carried out using 70 sentences uttered by the two male speakers, where each of two speakers uttered 35 sentences. The recognition experiments were carried out with the other 113 sentences uttered by each of the two speakers. The average phoneme recognition score was 86.1%. The average insertion and deletion scores were 7.7% and 3.9%. Table 3 shows the recognition scores of the conceptual word, the Bunsetsu-unit, and the sentence. The average recognition scores of the conceptual word, the Bunsetsu-unit, and the sentence were 85.7%, 73.2%, and 32.6%. Figure 8 shows examples of sentence recognition. Most sentence recognition errors were due to errors in recognition of functional words and in recognition of phonemes at the end of the sentence.

Table 2. Multi-speaker-dependent phoneme recognition scores for continuous speech uttered Bunsetsu by Bunsetsu.

Speaker	Phoneme recognition score	Deletion score	Insertion score
A	84.4	4.7	5.8
B	87.8	3.0	9.5

Table 3. Multi-speaker-dependent Bunsetsu-unit recognition scores for continuous speech uttered Bunsetsu by Bunsetsu.

Speaker	Conception word	Bunsetsu-unit	Sentence
A	84.8	70.9	28.4
B	86.7	75.6	36.7

Sentence: 第2の節目は音響技術に対する電気の応用である
dainino husimewa oNkyogizyucunitaisuru deNkino oyodearu
Phoneme: taiNnido husinewa oNkyoNgizyukinitaisiu teNkino moayodeha
Recognized:第2の節目は音響技術に対する電気の応用だった
dainino husimewa oNkyogizyucunitaisuru deNkino oyodaQta

Sentence: 特に音の測定には電気はなくてはならない
tokuni otono sokuteniwa deNkiwa nakutewanaranai
Phoneme: hokuN otono sokukuenya teNkiwa nakutewadaranai
Recognized:送る音の測定には電気はなくてはならない
okuru otono sokuteniwa deNkiwa nakutewanaranai

Sentence: 音響に電気はつきものである
oNkyoni deNkiwa cukimonodearu
Phoneme: oNkyoni teNzio cukinonodeae
Recognized:音響に電気もつきものである
oNkyoni deNkimo cukimonodearu

Sentence: コンピュータもディジタル技術も斬しすぎる
konpyutamo dizitarugizyucumo atarasisugiru
Phoneme: koNpyukamo piziuterunizyucumo tatarasuri
Recognized:コンピュータもディジタル技術も斬しい
konpyutamo dizitarugizyucumo atarasii

Figure 8. Examples of sentence recognition.

7. CONCLUSION

We have developed a prototype of a Japanese text dictation system which is composed of an acoustic processor, a processor for spotting Bunsetsu-units, and a syntactic processor. We constructed the acoustic processor using the modified LVQ2 method. The modified LVQ2 method achieves a high phoneme recognition performance of 86.1%. The syntax-driven continuous-DP matching algorithm is used for spotting Bunsetsu-units. This method greatly reduces the computation amount and storage capacity necessary for spotting the Bunsetsu-units. Analysis of the dependency structure between the Bunsetsu-unit candidates is effectively carried out using the syntactic and semantic information.

References

1. Y. Sekiguchi and M. Shigenaga, "Speech Recognition System for Japanese Sentences, " J. Acoust. Soc. Jpn., vol.34, No.3, pp.204-213, (1978)

2. K. Shikano and M. Kohda, "A Linguistic Processor in a Conversational Speech Recognition System," Trans. IECE Jpn., vol.J61-D, No.4, pp.253-260, (1978)

3. S. Nakagawa, Y. Ohguro and Y. Hashimoto, "Syntax Oriented Spoken Japanese Recognition/Understanding System -SPOJUS-SYNO-," Trans. IEICE Jpn., vol.J72-D-II, No.8, pp.1276-1283, (1989)

4. T.Tsuboi, N.Sugamura, A.Tomihisa and F.Obashi, "Japanese Conversation Method fro Voice Activated Japanese Text Input System," Trans. IEICE Jpn., vol.J72-D-II, No.8, pp.1284-1290, (1989)

5. S. Matsunaga, "Candidate Prediction Using Dependency Relationships Rules Combined with Transition Rules for Minimal Phrase Speech Recognition," Trans. IEICE Jpn., vol.J72-D-II, No.8, pp.1299-1306, (1989)

6. M.Shigenaga, Y.Sekiguchi, T. Hanagata, M. Taki and T. Yamaguchi, "On Prediction Possibility of Predicates and Noun Phrases for Continuous Speech Recognition," Trans. IEICE Jpn., vol.J72-D-II, No.8, pp.1307-1312, (1989)

7. M. Endo, S. Makino and K. Kido, "Phoneme Recognition Using a LVQ2 Method," Trans. IEICEJ, SP89-50, (1989)

8. M. Okada, A. Ito, H. Matsuo, S. Makino and K. Kido, "Analysis of Japanese Dictation System," Trans. Speech IEICEJ, SP86-33, (1988)

9. M. Okada, S. Makino and K. Kido, "A Study of Morphemic and Syntactic Processing Sub-system for Japanese Dictation System," Trans. IEICEJ, SP86-71, (1986)

10. A. Ito, Y. Ogawa, S. Makino and K. Kido, "Refinement and Evaluation of Bunsetsu Automaton in Japanese Dictation System," Proc. ASJ meeting, pp.135-136, (1987)

11. Y. Ogawa, A. Ito, M. Okada, S. Makino and K. Kido, "Refinement of Syntactic Processor in Japanese Dictation System Using Semantic Information," Proc. ASJ meeting, pp.137-138, (1987)

12. A. Ito, S. Makino and K. Kido, "A Parsing Algorithm Based on CYK Algorithm for Continuous Speech Recognition," Proc. ASJ meeting, pp.91-92, (1988)

13. A. Ito, S. Makino and K. Kido, "Syntactic Processing Using the Principle of Least Bunsetsu's Number Method for Continuous Speech Recognition," Proc. ASJ meeting, pp.93-94, (1988)

14. K. Shudo, T. Narahara and S. Yoshida, "A Structural Model of Bunsetsu for Machine Processing of Japanese," Trans. IECE Jpn., vol.J62-D, No.12, pp.872-879, (1979)

15. S. Yoshida, "Syntax Analysis of of Japanese Sentence Based on Kakariuke Relation between Two Bunsetsu," Trans. IECE Jpn., vol.55-D, No.4, pp.238-244, (1972)

16. F.Jelinek, R. L. Mercer, L. R. Bahl and J. K. Baker, "Perplexity-A measure of difficulty of speech recognition tasks," presented at the 94-th Meet. Acoustical Society of America, Miami Beach, FL, (1977)

17. T. Kohonen, G. Barna and R. Chrisley, "Statistical Pattern Recognition with Neural Networks: Benchmarking Studies," IEEE Proc. of ICNN, vol.1, pp.61-68, (1988)

18. E. McDermott and S. Katagiri, "Shift-invariant Phoneme Recognition Using Kohonen Networks," Proc. ASJ meeting, pp.217-218, (1988)

19. S. Moriai, S. Makino and K. Kido, "A Method for Selecting an Optimum Phoneme Sequence Using a Posteriori Probabilities of Phonemes," Journal of ASA, supplement No.1, PPP5, (1988)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Word Recognition Using Synthesized Templates

Mats Blomberg, Rolf Carlson, Kjell Elenius, Björn Granström and Sheri Hunnicutt

Department of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH), Stockholm, Sweden

Abstract

This is an expanded version of a paper presented at the FASE SPEECH'88 meeting in Edinburgh. This paper also includes some new experiments along the same lines.

1. INTRODUCTION

With the ultimate aim of creating a knowledge based speech understanding system, we have set up a conceptual framework named NEBULA. In thin paper we will start by briefly describing some of the components of this framework and also report on some experiments where we use a production component for generating reference data for the recognition. The production component in the form of a speech synthesis system will ideally make the collection of training data unnecessary. Preliminary results of an isolated word recognition experiment will be presented and discussed. Several methods of interfacing the production component to the recognition / evaluation component have been pursued.

2. NEBULA

During the last years, many experiments have been carried out at our department concerning different aspects of speech recognition and speech perception. At the same time work on speech synthesis has been pursued. The speech recognition scheme, NEBULA, combines results and methods from these efforts into coherent system. The system is presented in Figure 1.

2.1. The front end

Using conventional signal processing techniques, we have earlier tried some of the proposed auditory representation in the context of a speech recognition system, Blomberg et al. [1]. Based on one of these models, the DOMIN model, we are currently working on a new primary analysis module. This peripheral auditory model explores the possibility for synchrony effects that will enhance spectral peaks and suppress valleys. At the same time, wide band effects will be taken into account.

motoric disability or in aphasia rehabilitation. These algorithms are currently being complemented with syntactic and semantic components, Hunnicutt [4].

2.5. The identification component

There are presently two types of recognition techniques available for NEBULA. One if a whole-word pattern-matching based on filter bank analysis, cepstral transformations and non-linear time warping, described in more detail elsewhere, Elenius & Blomberg [5]. As in other systems of this kind, a separate training session is needed to establish acoustic reference data. In our system, the reference material is provided by the rule synthesis system. It will be possible to gave the reference generated during the recognition process and then take into account word juncture and word position effects, which is not easily achievable in conventional word-based speaker_trained systems. This is the method used in the present experiments.

The second method is based on phonetic recognition using a network representation of possible realizations of the vocabulary. The acoustic analysis is the same as in the previously described method, but the phonetic decisions are based on comparisons to a library of synthetic allophones. The network approach enables handling of optional pronunciations. On the other hand, non-stationary parts of the speech wave may be better represented by a more detailed description of the time evolution of the utterance, as in the first method. A combination of the two methods would enable the advantages of both techniques to be used.

2.6. Word references from text-to-speech system

The phonetic component of a text-to-speech system is used to create references from the cohort. The synthesis system has been described elsewhere, Carlson et al. [6]. It is based on rules and has a formant synthesizer as output module. These references are sent to the identification and verification part of NEBULA.

3. USING SYNTHETIC TEMPLATES: PREVIOUS WORK

Use of synthetic speech as reference for aligning natural speech with dynamic programming techniques has been reported by many authors, i. e. , Woods et al [7], Chamberlain & Bridle [8], Höhne et al. [9], Hunt [10]. The papers by Chamberlain and Höhne are mainly concerned with the time-warping aspects of mapping long utterances (sentences) to each other. Hunt cites four reasons besides the obvious one of improving speech synthesis for the reach in this field. First he mentions the analysis by synthesis based technique as a good method for extracting formant frequencies, which seem to be better for indicating phonetic identity than the gross spectral shape, often used in speech recognition. Another property of synthetic speech is that it can be modified to match the voice of the current speaker. Synthesis can also be used to exploit knowledge that is available about natural speech such as duration and the context of a word. Finally he discusses the positive effect of the perfect consistency of synthetic speech. It can be used for speaker verification, where the speaker characteristics can be related to synthetic speech. In recognition the

consistency may be used to improve separation between words such as stalagmite and stalactite that have phonetically identical parts. After having compared synthetic speech (MITalk) to natural speech in some recognition experiments, he notes that aligning speech between natural speakers gives considerably better results than using synthetic speech. His conclusion is that the synthesis rules must be improved before the synthesis can give comparable results.

4. SOME EXPERIMENTS WITHIN NEBULA

In 1987 we reported on the ongoing work inside some of the clouds in the NEBULA scheme, Blomberg et al. [11]. At that time the experiments were run on three different computers and two different kinds of special hardware. This created practical problems and slowed down the continuation of the work. During the last year, both the synthesis and the recognition software have been implemented on our Apollo work stations. This has opened up new possibilities to make additional experiments along the same lines as before. The interactions between the different modules in NEBULA is now fast and easy. We will in this paper review the earlier work and report some results from a new test series. This paper also includes additional experiments that were not included in the written version that is published in the proceeding of the SPEECH'88 FASE symposium, Blomberg et al. [12].

4.1. Test vocabulary and subjects

In all experiments reported in this paper, the lexical search was stimulated. In this case, the suggested preliminary analysis only discriminates between vowel and consonant and identifies the stressed syllables. A 26-word cohort was chosen which was of the type "VCVCC'VC. It was drawn from a corpus consisting of the most frequent 10, 000 words of Swedish. Ten male subjects were asked to read the 26 words from a list with little instruction except to pronounce each word separately. The vocabulary was recorded in a normal office room with additional noise from a personal computer.

The word structure, in most cases, is a compound word with a disyllabic initial morph. The structure is rich enough to expose a variety of deviations among the subjects and a synthesized reference. These deviations generally occur across the compound boundary. Both deletions and insertions and hypercorrect pronunciations occur and 37 such deviations from the norm were identified among the total of 26 * 10 words recorded. Within the cohort there are many examples of morphological overlap as can be seen from the list of words in Table 3. One word pair (14 '*äventyrs*' and 15 '*äventyr*') differs in only one consonantal segment.

4.2. Preliminary recognition results; experiment 1

The recorded speech material was used as input to the pattern-matching verification component of Nebula, which in the first experiment consisted of the special hardware recognition system developed at KTH. The output from our hardware text-to-speech

system was recorded and used to train this system. No adjustments were done to the synthesis in this first stage. 74.6% of the test words were identified correctly.

In addition to the synthesis, each speaker was used to create references for the other speakers. All the human speakers served better as reference speakers than the synthesis, and the correct result ranged from 79.1% to 93.6% with an average value of 89.5%.

At an early point we noticed discrepancies in the durational structure of the synthetic and the human speech. Differences in segment duration will cause spectral differences that cannot be eliminated by a time warping procedure, since time dependent co-articulation and reduction rules are active in the synthesis system. The segmental durations for one speaker were measured and the durational framework for each word was imposed on the synthesis. The result showed an increase to 81.5% correct identification, which is slightly better than our worst human speaker. The results from our experiments are summarized in Table 1 and 2, and Figure 2.

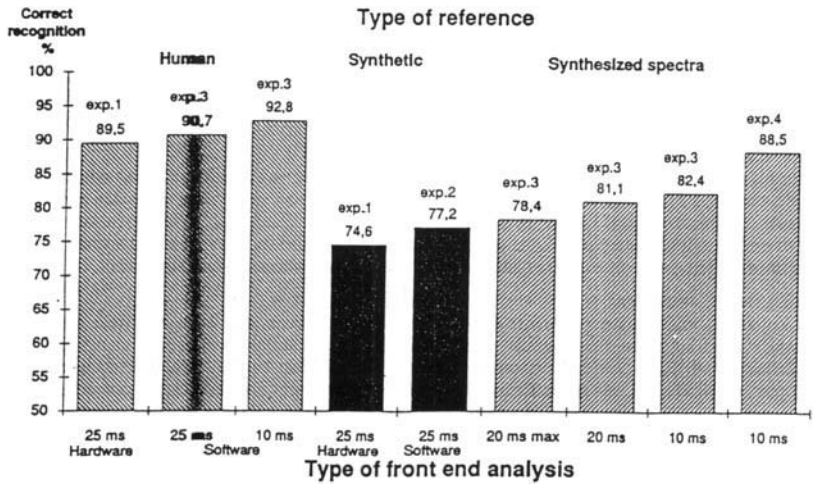


Figure 2. Recognition results in the experiments using different kinds of reference.

4.3. A new environment; experiment 2

The new series of experiments use the same recordings as before, but different methods are used to create the spectral templates for the identification components.

As a start, the old recordings of both the human speech and the synthesis were digitized, using 16kHz sampling frequency. The filter bank was simulated with the help of an FFT analysis followed by pooling of the spectral components into 16 bands, from 200 to 5000Hz, equally spaced along the Bank scale. The recognition system, now running in the Apollo work station, was used to repeat the same experiment as before. The result can be seen in

Table 1. The increased accuracy in the analysis gives a slightly better recognition result for the human speakers and the non-normalized synthesis.

In the following experiment, all parts of the text-to-speech system were running in the same computer including the synthesizer. The result of 76.4% correct identification is presented in Table 1 under the name 'software synthesizer'.

Table 1. Result from the recognition experiment 1 and 2.

Reference pattems from	Experiment 1	Experiment 2
Human Speakers	89.5	90.7
Synthesis	74.6	77.2
Synthesis, Duration Adjusted	81.5	81.2
Software Synthesizer	—.-	76.4

4.4. Parameter generated spectrum; experiment 3

The next experiment in this series included a different method to generate spectral frames. The control parameters to the synthesizer were used directly to generate the spectral shape. We can then by-pass several problematic areas in the analysis. The interaction between harmonics and formant peaks in the vowel spectrum can be avoided and the fricative noise spectrum is stable. Figure 3 gives an example of this method. The control frames to the synthesizer are used to generate the spectral representation in Figure 3a. These spectral slices are transformed into 16 channels corresponding to the output from the filter bank used in the identification part of the recognition system, Figure 3b. To the left in Figure 3c is the synthetic reference for the first word, 'obekvăm', in the vocabulary and to the right is the analysis of a speaker saying the same word. Several observations can be made. The noise level has a considerable influence on the spectrum. The speaker uses an unvoiced labiodental fricative instead of the voiced counterpart in the synthesis. We will return to a more detailed analysis later in the paper.

The frame rate in the synthesis is 10 ms while the recognition system uses 25 ms between each observation. The down-sampling was achieved by simply taking the maximum of two sequential spectral slices. The recognition result with this method was 78.4% correct identification. An alternative method to use every other spectral slice gave a better result, 81.1%. Interpolation between two static spectra will give unwanted effects. If a resonance is moving too fast between two frames a double peak will be stored in the spectral representation. As an alternative, we used all synthesis frames and increased the frame rate in the verification part by two. The synthesis gave a slightly higher value 82.4% correct. A repetition of experiment 2 using each of the human speaker as reference gave a mean of 92.8%. The two worst cases were 84.7% and 88.9%. These results are presented in Table 2 and Figure 2.

In Figure 4 the distance between the correct and the best incorrect match is shown for both a typical human reference and the synthesis. We can observe that the distribution is different is at least two aspects. The data points are closer to the diagonal in the synthesis case compared to the human case. This means the synthesis gives a less confident answer

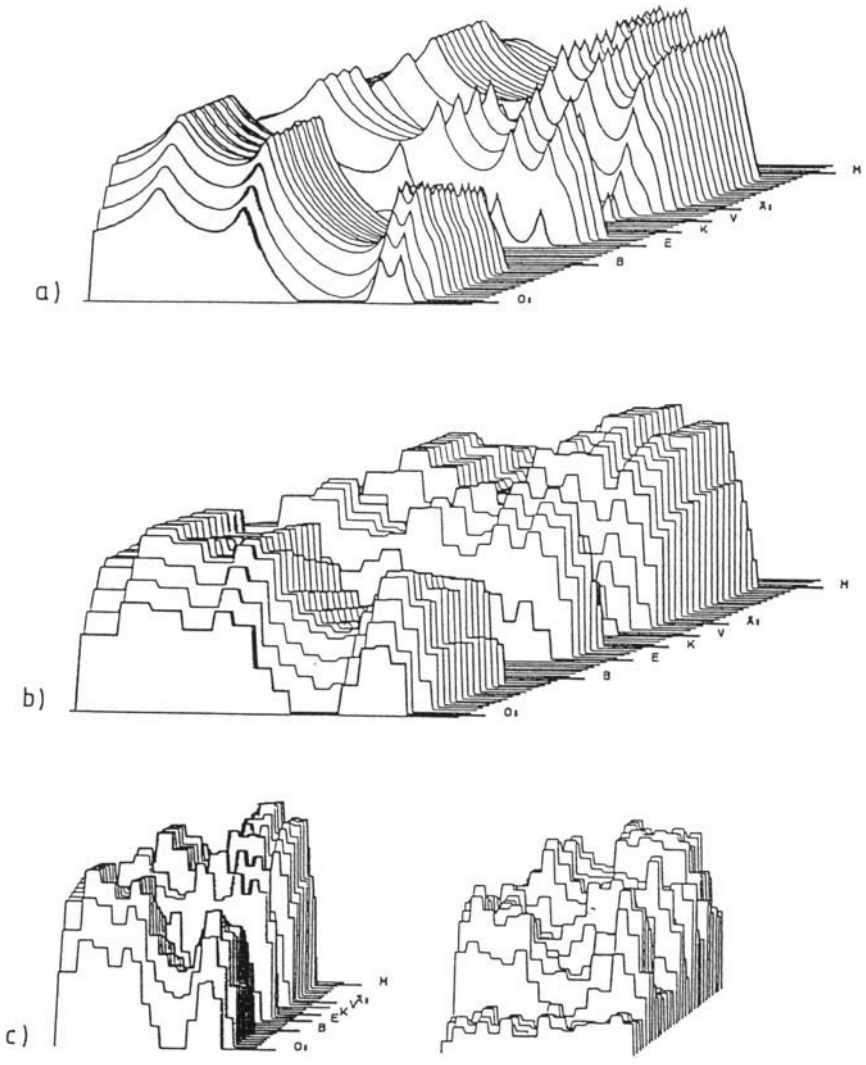


Figure 3. Creation of special slices a)and transformation to filter bank representation b). Comparison between synthetic (left) and natural (right) spectral templates for the word 'obekvam' c).

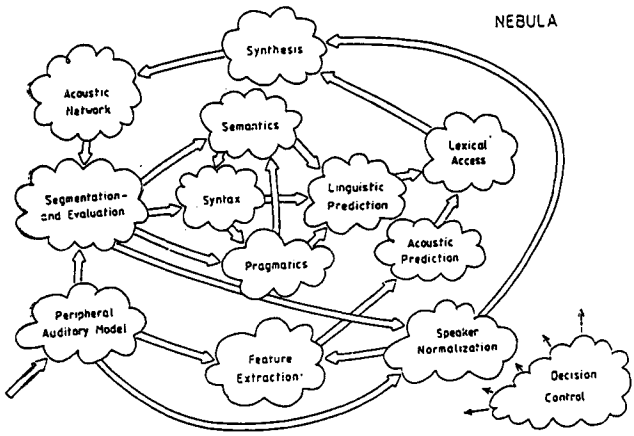


Figure 1. The speech recognition scheme, NEBULA.

2.2. Feature extraction

At the output of the auditory model, the speech is represented as a continuous flow of information in multiple channels, Carlson et al. [2]. This makes it possible to use diverse analysis mechanisms which can be simple but should work in a coordinated structure. The process could include spectral transformations, lateral inhibition, temporal onset / offset effects, and a variety of phonetic-cue detectors.

2.3. The lexical component

The low levels of NEBULA explore the descriptive power of cues, and uses multiple cues to analyze, classify, and segment the speech wave. These classifications are used during the lexical search, Carlson et al[3]. Additional information from a prediction system is also used in the lexical selection part of NEBULA. As a result of this component, we get a selection of possible words, a cohort.

2.4. High level linguistic processes

The mid-portion of NEBULA is currently represented by a syntactic component of the text-to-speech system, morphological decomposition in the text-to-speech system, and a concept-to-speech system. A special phase structure grammar is employed which takes account of word order, phrase order, and grammatical information. These parts were originally developed for a different purpose than speech recognition, but we expect them to be applicable in this area as well.

One of several word prediction algorithms has been designed to find cohorts of possible words from partial information generated by a word recognition scheme. Other prediction algorithms are being used in handicap applications, to help persons with a speech or

Table 2. Result from the recognition experiment 3.

Reference pattern		
Human Speakers		
25 ms and 10 ms sampling:	90.7	92.8
Parameter Generated Spectrum		
20 ms, max of two frames:	78.4	
Parameter Generated Spectrum		
20 ms and 10 ms sampling:	81.1	82.4

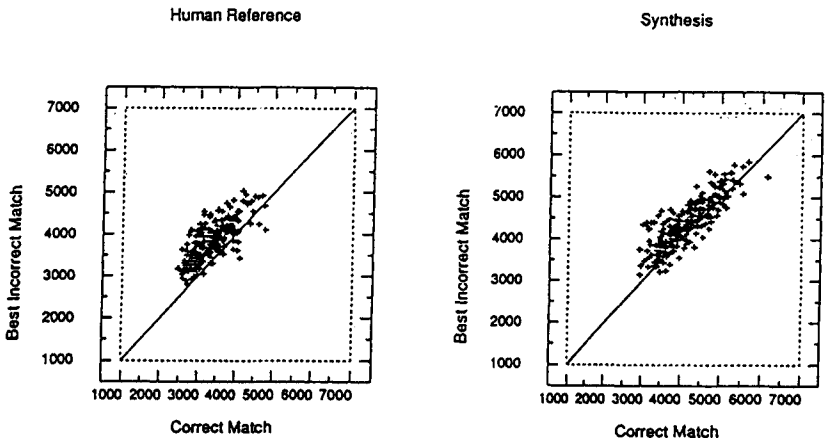


Figure 4. The distance between the correct stimulus and the correct reference, and between the correct stimulus and the best incorrect match. Human reference a) and synthetic reference b).

even if it is correct. Also the mean distances between the synthesized reference and the test vocabulary are 17% larger compared to the corresponding distances using references by human speakers. To reduce the distances to a comparable or smaller value and to push it away from the diagonal is a challenge for our continued work.

A confusion matrix of the experiment 3 (20 ms frames) is shown in Table 3. We can observe that the word '*ingenting*' has been over-represented in the responses. A comparison is made in Figure 5 between the tense vowel /e:/ and the lax /I/ for the same speaker pronouncing words 13 : *enighet* and 3 : *ingenting*. It is obvious that the spectral shape of the main stressed vowel in this cannot be used as a distinguishing cue. However relative duration, coarticulation and diphthongization can give supportive information for vowel discrimination.

As a complement to the case study of the /e:/ and /I/ mentioned above we made a statistical analysis of the energy distribution, see Figure 6. the figures are created by making a dimensional histogram of the energy/frequency distribution of the observed material. This distribution is then divided in 10% intervals, which are drawn in the graph.

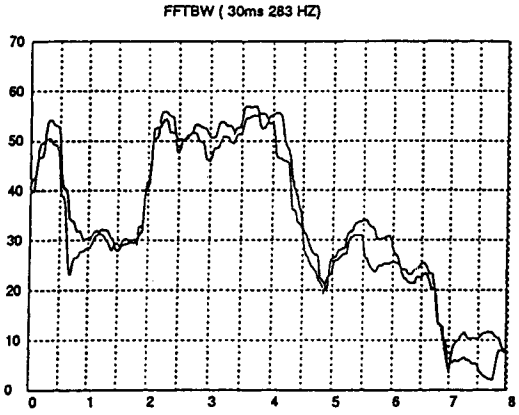


Figure 5. Spectral section of /I/ in the world 'ingentin' and /e:/ in the world 'enighet' for one of the speakers.

Table 3. Confusion matrix for Experiment 3 (20 ms frames).

1	obekväm	10
2	uppenbar	6 2 2
3	ingenting	10
4	ovanför	10
5	innanför	10
6	inifrån	7 2 1
7	inomhus	9 1
8	utanför	10
9	utifrån	10
10	utomhus	2 8
11	uppifrån	7 3
12	egendom	4 6
13	enighet	3 6 1
14	äventyrs	10
15	äventyr	1 18
16	överskott	1 9
17	övergick	51 1 3
18	öppenhet	4 14 1
19	uteblev	64
20	uteslöt	10
21	övergår	1 1 8
22	övergett	12 2 5
23	återstod	1 9
24	återstår	19
25	återger	10
26	återkom	8

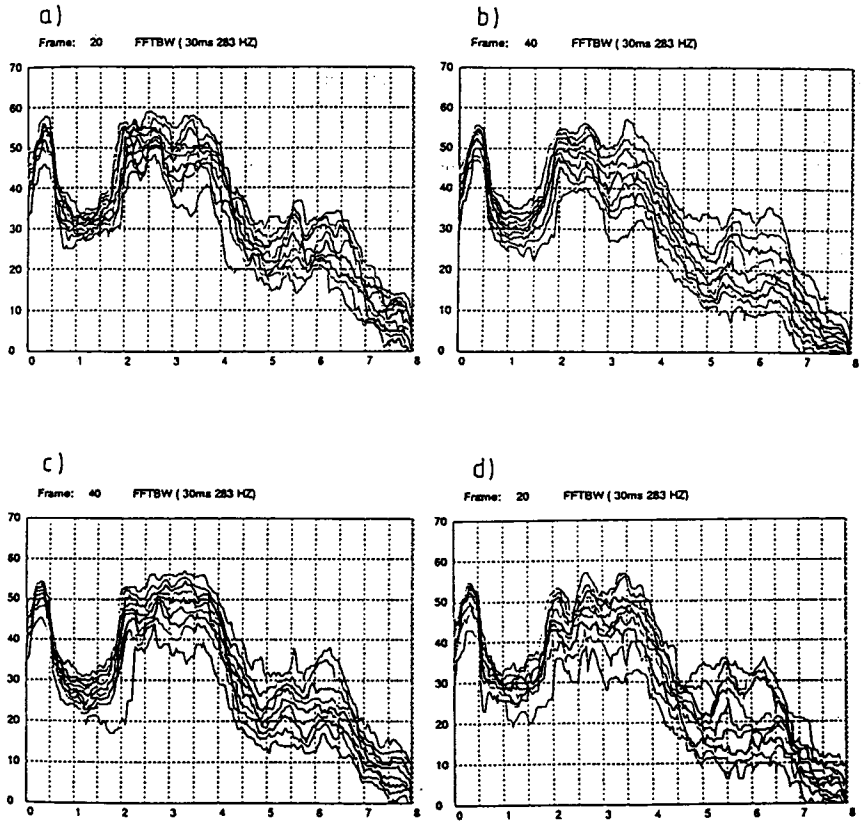


Figure 6. Energy distribution for initial and final /e:/ (top) and initial and final /I/ (bottom).

The 30%, 50% and 70% are marked with a thicker line. This method is used in our general work with our speech data-base [13]. The two phonemes were analyzed in initial and final position in the same recorded materials. A small difference in the higher formants can be seen as expected. If we compare the spectrum in initial and final position, we find a small difference in spectral slope which can be referred to the glottal source. Several studies at our laboratory have been dealing with these types of variations, [14].

4.5. Improved synthesis; experiment 4

Detailed analysis of the results gave good information on basic errors in the synthesis. First the phonetic transcription was not according to the typical pronunciation of the speakers. The /g/ in 'enighet' was for example deleted by most speakers. This created most of the errors for this word.

The synthesis system used so far is based on smoothed square waves for most of the parameters. This has proven to be a good method for interpolation in many cases. It will automatically create reduction effects when the duration of a segment is short. The method is related to the thinking that the production of speech can be simulated by control step functions smoothed by the muscular/mechanical system of articulators. However, the frequency domain is probably not the correct domain for this smoothing. Articulatory parameters are more natural in this respect. As a alternative we are currently building a slightly different synthesis system where the parameters are specified by target values and the time it takes to reach this value. The movement towards a target can be interrupted by a new target. Unfortunately this method adds new demands on the system. Reduction does not come automatically as before. It has to be described in a more explicit way. On the other hand phonetic knowledge can be more accurately specified.

As a final experiment word references from this new system was tested in a recognition experiment. Several new synthesis aspects were considered. Differrent allphones for the /r/ phoneme was used in the 'CV' or 'VC' position. The diphthong /e:/ was adjusted. Figure 7 shows the synthesis spectrum for /e:/ compared to the energy histogram from Figure 6a. This experiment gave a result of 88.5% accuracy, which is better than the worst human reference and close to the second to last one.

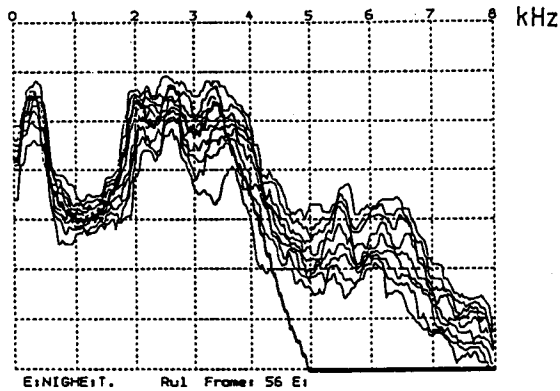


Figure 7. Comparison between synthetic /e:/ and energy histogram.

4.6. Analysing identification results

In order to analyze the recognition results we can use a program giving a display as seen in figures 8a and 8b. Before time aligning words by dynamic programming they are linearly normalized to a nominal length of 800 ms - or 80 cepstral frames, since frames in this experiment are calculated at 10 ms intervals. At the top we see the warping function as a thin line. The bold line is a cepstral difference function between the matched words calculated along the warping function. Below this plot we see the three energy functions displaying: 1) the rest word (bold), 2) the time warped reference word, and 3) the reference

word. Below these we see two 16 band spectral section of corresponding frames at a point along the time warp. The bold section belongs to the rest word and the other is from the reference. The time point is marked by a vertical bar in the warped difference function at the top.

The display makes it possible to interactively analyze why words are misrecognized and what part of words are mismatched. It also gives a means of understanding what makes the test word more similar to an incorrect reference than to the correct one. In this case we are analyzing the test word '*äventyr*' by speaker GF what was erroneously identified as the synthesized reference '*ingenting*'. In Figure 8a we see result of matching '*äventyr*' by GF to the synthesized version of the same word and in Figure 8b we see it matched to the synthesized word '*ingenting*'. The special sections are from the first vowel as marked by the vertical bar in cepstral difference plot. The cepstral difference is larger between 'ae' and the synthetic 'i'. The corresponding plots of spectral sections show that in this case the synthetic 'ae'(8a) has too little low energy compared to the natural voice (the bold line). It should be stressed that during identification the matching is done in the cepstral domain, not in the frequency domain, and that one should be careful about conclusions drawn from looking only at spectral sections.

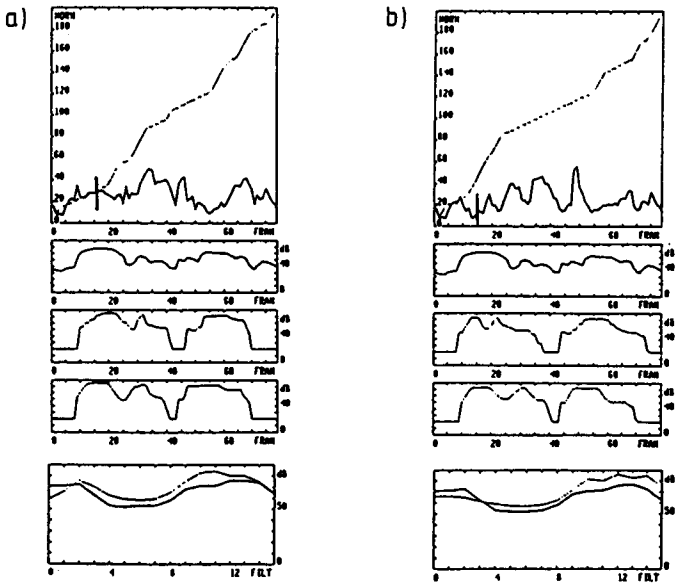


Figure 8. Plots showing word matching in the identification component. In 8a the test word 'aventyr' by speaker GF is matched to the same synthesized reference word and in 8b the same word 'aventyr' is matched against the same synthetic word reference word 'ingenting'. See section 4. 6 for more information.

4.7. Comparison between the experiments

Figure 2 shows the results from all experiments. It can be seen that various improvements have successively raised the recognition accuracy. The increased frame rate gave a bigger improvement for the human references compared to the synthetic.

5. CONCLUDING REMARKS

In our paper we have reported on some experiments, which are part of a long term project towards a knowledge based speech recognition system, NEBULA. We have taken the extreme stand in these experiments of comparing human speech to predicted pronunciations on the acoustic level with the help of straightforward pattern matching technique. The significantly better results when human references are used was not a surprise. It is well known that text-to-speech systems still need more work before they reach human quality. However, the results can be regarded as encouraging.

In the last experiments we reached an important goal in our work strategy. We have created an experimental system that gives us control of each separate module of the system. We can easily do a special comparison between synthesis and human speech. We can adjust the spectral shapes in order to adapt the synthesis to a specific speaker. This will give us valuable feed-back on both the prediction/synthesis component and the matching algorithm, and some information on how these components should interact when exposed to a variety of speakers.

References

1. M. Blomberg, R. Carlson, K. Elenius and B. Granström, "Auditory models in isolated word recognition," Proceedings IEEE-ICASSP, San Diego, (1984)

2. R. Carlson, B. Granström and S. Hunnicutt, "A parallel speech analyzing system," STL-QPSR, 1, (1985)

3. R. Carlson, K. Elenius, B. Granström and S. Hunnicutt, "Phonetic properties of the basic vocabulary of five European languages: implications for speech recognition," Proceedings IEEE-ICASSP, Tokyo, (1986)

4. S. Hunnicutt, "Lexical Prediction for a Text-to-Speech System," Communication and Handicap: Aspects of Psychological Compensation and Tecchnical Aids, E. Hjelmmquist and L. G. Nilsson, editors, Elsevier Science Publisher B. V. (North Holland), (1986)

5. K. Elenius and M. Blomberg, "Voice Input for Personal Computers," Electronic Speech Recognition, G. Bristow, editor, William Collins Sons & Co. Ltd, London, p.361 (1986)

6. R. Carlson, B. Granström and S. Hunnicutt, "A multi-language text-to-speech module," Proceeding IEEE-ICASSP, Paris (1982)

7. W. Woods, M. Bates, G. Brown, B. Bruce, C. Cook, J. Klovstad, J. Makhoul, B. Nash-Webber, R. Schwartz, J. Wolf and V. Zue, "Speech Understanding Systems - Final Technical Progress Report," Report No.3438, BBN, Cambridge, MA, USA (1976)

8. R. M. Chamberlain and J. S. Bridle, "ZIP: A Dynamic Programming Algorithm for Time Aligning Two Indefinitely Long Utterances," Proceedings IEEE-ICASSP 83, p.816 (1983)

9. H. D. Hohne, C. Cooker, S. E. Levinson, L. R. Rabiner, "On Temporal Alignment of Natural and Synthetic Speech," IEEE Trans. on Acoust. Speech and Sign. Proc. vol.ASSP-31, 4 p.807 (1983)

10. M. J. Hunt, "Time Alignment of Natural Speech to Synthetic Speech," Proceedings IEEE-ICASSP 84, p.2.5.1 (1984)

11. M. Blomberg, R. Carlson, K. Elenius and B. Granström, "Speech Recognition Based on a Text-to-Speech System," Proceedings European Conference on Speech Technology, Edinburgh, p.369 (1988)

12. M. Blomberg, R. Calson, K. Elenius, B. Granström and Hunnicutt, "Word recognition Using Synthesized Templates," Proceedings SPEECH'88 7th FASE Symposium, Edinburgh, (1988)

13. R. Carlson, B. Granström, "Rule-Controlled Data Base Search," STL-QPSR, No.4, pp.29-45, KTH, Stockholm, (1988)

14. G. Fant, J. Lijencrants and Q. Lin, "A Four-Parameter Model of Glottal Flow," STL-QPSR, No.4, pp.1-13, KTH, Stockholm, (1985)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

A Cache-Based Natural Language Model for Speech Recognition

Renato De Mori and Roland Kuhn

School of Computer Science, McGill University
3480 University Street, Montreal, Quebec, Canada, H3A 2A7

1. INTRODUCTION

A type of system popular today for Automatic Speech Recognition (ASR) consists of two components. An acoustic component matches the acoustic input to words in its vocabulary, producing a set of the most plausible word candidates together with a probability for each. The second component, which incorporates a language model, estimates for each word in the vocabulary the probability that it will occur, given a list of previously hypothesized words. Our work focuses on the language model incorporated in the second component. The language model we use is based on a class of Markov models identified by Jelinek, the "n-gram" and "Mg-gram" models. These models produce a reasonable non-zero probability for every word in the vocabulary during the speech recognition task. Our combined model incorporates both a Markov 3g-gram component and an added "cache" component which tracks short-term fluctuations in word frequency. The addition of the cache component and the evaluation of its effects are the original contributions of this paper.

We adopted the hypothesis that a word used in the recent past is much more likely to be used soon than either its overall frequency in the language or a 3g-gram model would suggest. The cache component of our combined model estimates the probability of a word from its recent frequency of use. The model uses a weighted average of the 3g-gram and cache components in calculating word probabilities, where the relative weights assigned to each component depend on the Part of Speech (POS). For purpose of comparison, we also created a pure 3g-gram model, consisting of only the 3g-gram component of the combined model.

Our research was greatly facilitated by the availability of a large and varied collection of modern texts, in which each word is labelled with an appropriate POS. This is the Lancaster-Oslo/Bergen (LOB) Corpus of modern English. Part of this corpus (391, 658 words) was utilized as a training text which determined the parameters of both models: the standard 3g-gram model, and our combined model consisting of the same 3g-gram model along with a cache component.

We required a yardstick with which to compare the performance of the two models.The measure chosen is called "perplexity"; it was devised by F.Jelink, R.L.Mercer,

and L.R.Bahl[4]. The perplexity of a model can be estimated by the success with which it predicts a sample text (which should NOT be the one used to train the model). The better the model, the higher the probability it will assign to the sequence of words that actually occurs in the sample text, and the lower the perplexity.

Once the parameters of the two models, the pure 3g-gram and the combined, had been calculated from part of the LOB Corpus, we could have used any sample text from any source whatsoever to compare the perplexity of the models. We chose to use part of the remaining portion of the LOB Corpus because of the wide range of different types of text represented therein. The sample text we constructed (like the training text)includes such diverse types of written English as press reports, religious literature, love stories, and government documents.

The results of the comparison between the two models exceeded our expectations. The pure 3g-gram model, as expected, had a high estimated perplexity:332. The estimated perplexity of the combined model, on the other hand, was 107. This more than three-fold improvement indicates that addition of a cache component to a 3g-gram language model can lead to dramatic improvement in the performance of the model, as measured by its perplexity.

2. MARKOV MODELS FOR NATURAL LANGUAGE

2.1. Mathematical Background

An Automatic Speech Recognition (ASR) system takes an acoustic input, A, and derives from it a string of words $W_1, W_2, ..., W_n$taken from the system's vocabulary, V.Formally, let $WS =< W_1, W_2, ..., W_n >$denote one of these possible word strings and $P(WS \mid S)$ the probability that it was uttered, given the acoustic evidence A. Then the speech recognizer will pick the word string WS satisfying,

$$P(\hat{W}S \mid A) = \max_{WS} P(WS \mid A) \tag{1}$$

i.e., the most likely word string given the evidence. From the Bayes Formula we have

$$\hat{W}S = \{WS\text{such that}P(WS) \cdot P(A \mid WS)\text{is a maximum}\} \tag{2}$$

In this paper, we are concerned with the model that estimates $P(WS)$, the probability of a given word string independent of the acoustic input.

2.2. Jelinek's Trigram Model

The trigram model is a Markov model, approximating $P(W_i = W \mid < W_1, ..., W_{i-1} >)$ by $P(W_i = W_{i-2}, W_{i-1})$. The latter, in turn, is estimated from the training text as the ratio of the number of times the word sequence $< W_{i-2}, W_{i-1}, W >$ occurred to the number of times the sequence $< W_{i-2}, W_{i-1} >$ occurred:

$$P(W_i = W \mid W_{i-2}, W_{i-1}) \cong f(W \mid W_{i-2}, W_{i-1}) = \frac{N(W_{i-2}, W_{i-1}, W)}{N(W_{i-2}, W_{i-1})} \tag{3}$$

In practice many trigrams that do not occur in the training text show up during the recognition task, and should therefore not have the zero probability assigned them by this formula. One way of dealing with this problem is to use a weighted average of trigram, bigram, and individual word frequencies:

$$P(W_i = W \mid W_{i-2}, W_{i-1}) \simeq q_2 f(W_i = W \mid W_{i-2}, W_{i-1}) + q_1 f(W_i = W \mid W_{i-1}) + q_0 f(W_i = W) \tag{4}$$

where $q_0 + q_1 + q_2 = 1$ and
$f(W_i = W \mid W_{i-2}, W_{i-1}) = N(W_{i-2}, W_{i-1}, W)/N(W_{i-2}, W_{i-1})$,
$f(W_i = W \mid W_{i-1}) = N(W_{i-1}, W)/N(W_{i-1})$, and $f(W_{i-1} = W) = N(W_i = W)/NT$,
where NT =total number of words in training text.

If $q_0 \neq 0$, this smoothed trigram model guarantees that any word W that occurs at least once in the training text is assigned a non-zero probability. The values for q_0, q_1,and q_2 are chosen in order to meet the maximum likehood criterion.

2.3. The 3g-gram Model

The 3g-gram model (terminology of A. Martelli and of Derouault and Merialdo [1, 2]) is analogous to the trigram model; it employs grammatical parts of speech-henceforth abbreviated "POS".

Let $g(W_i) = g_i$ denote the POS of the word that appears at time i, let G be the set of POSs recognized by our model, and let g_j be a particular POS whose probability of occurring we wish to predict. The model will give us an estimate $\hat{P}(g_i = g_i \mid g_{i-2}, g_{i-1})$ of that probability based on the identity of the two preceding POSs. Note that many words belong to more than one POS category. For example, the probability that "light" will occur is the probability that it will occur as a noun plus the probability that it will occur as a verb plus the probability that it will occur as an adjective. Thus, the general 3g-gram formula is:

$$\begin{aligned} P(W_i = W \mid < W_1, ..., W_{i-1} >) &\simeq \sum_{g_j \in G} P(W \mid g_j) \cdot P(g_i = g_j \mid g_{i-2}, g_{i-1}) \\ &\simeq \sum_{g_j \in G} f(W \mid g_j) \cdot \hat{P}(g_i = g_j \mid g_{i-2}, g_{i-1}) \end{aligned} \tag{5}$$

Given a sufficiently large training text, $\hat{P}(g_i = g_j \mid g_{i-2}, g_{i-1})$ could be estimated for every POS g_i in G as $f(g_i = g_j \mid g_{i-2}, g_{i-1})$. In practice, existing training texts are too small-many POS triplets will never appear in the training text but will appear during a recognition task. If we do not modify the procedure to prevent zero probabilities, a particular g_j that actually occurs may have zero estimated probability.

Recall that an analogous problem occurred with the trigram model. The solution we described was the "weighted average" approach, which uses bigram and singlet frequencies to smooth out the trigram frequencies.This solution is also applicable to the 3g-gram model-Derouault and Merialdo [1, 2] employed a variant of the weighted average 3g-gram approach.

Their corpus consisted of 1.2 million words of French text tagged with 92 POSs. Only 5 percent of the possible triplets occurred. Thus, the doublets were tabulated as well; this

time half of the possible pairs occurred. Instead of using individual POS frequencies as the third component of a weighted average, these researchers chose to add an arbitrary small value $e = 10^{-4}$ to the overall probability estimate of each word in order to prevent zero estimates for the probability of occurrence of any given word. Thus, they approximated the probability of occurrence of a word W at time i, given that W has part of speech g_j, the two preceding parts of speech are g_{i-2} and g_{i-1}, and vocabulary size is n, as

$$\begin{aligned} & P(W_i = W \mid g(W) = g_j, g_{i-2}, g_{i-1}) \\ \simeq\ & (1 - ne)f(W \mid g_j) \times [l_1 f(g_i = g_j \mid g_{i-1}) l_2 f(g_i = g_j | g_{j-1})] + e, \\ & e = 10^{-4}, l_1 + l_2 = 1 \end{aligned} \tag{6}$$

They experimented with two different ways of calculating l_1 and l_2. Derouault and Merialdo's first let l_1 and l_2 be a function of the count of occurrence of $< g_{i-2}, g_{i-1} >$. Each possible history $< g_{i-2}, g_{i-1} >$ was assigned to one of ten groups, depending on how often it had occurred in the training text. Each of the groups had different values of l_1 and l_2, with the highest value of l_2 occurring in the group for histories $< g_{i-2}, g_{i-1} >$ that never occurred in the training text. The other way in which these researchers calculated l_1 and l_2 was to allow them to depend on g_{i-1}.

Let $h(< g_{i-2}, g_{i-1} >)$ denote the parameter on which l_1 and l_2 depend. For Derouault and Merialdo's first approach, $h = N(< g_{i-2}, g_{i-1} >)$ = the number of occurrences of $< g_{i-2}, g_{i-1} >$ in the training text;for the second approach, $h = g_{i-1}$ = the POS of the preceding word. They calculated $l_1(h)$ and $l_2(h)$ by the same algorithm in both cases, called the deleted interpolation method [5]. Having split the training text into two portions in the ratio 3 : 1, they used the larger portion to calculate $f(g_i \mid g_{i-2}, g_{i-1})$ and $f(g_i \mid g_{i-1})$. They then set $l_1(h)$ and $l_2(h)$ to arbitrary values such that $l_1(h)$ and $l_2(h) = 1$, and iteratively reset them from the remaining portion of the corpus. Summing over all triplets $< g_{i-2}, g_{i-1}, g_i >$ in this portion, they defined

$$S_1(h) = \sum l_1(h) f(g_i \mid g_{i-2}, g_{i-1}) / [l_1(h) f(g_i \mid g_{i-2}, g_{i-1}) + l_2(h) f(g_i \mid g_{i-1})] \tag{7}$$

$$S_2(h) = \sum l_2(h) f(g_i \mid g_{i-1}) / [l_1(h) f(g_i \mid g_{i-2}, g_{i-1}) + l_2(h) f(g_i \mid g_{i-1})] \tag{8}$$

They then redefined

$$\begin{aligned} l_1(h) &= S_1(h)/(S_1(h) + S_2(h)), \\ l_2(h) &= S_2(h)/(S_1(h) + S_2(h)) \end{aligned} \tag{9}$$

Iteration continued until $l_1(h)$ and $l_2(h)$ converged to fixed values. Derouault and Merialdo found only a small difference between the performance of the model in which l_1, l_2 depend on the count $N(< g_{i-2}, g_{i-1} >)$ and that in which they depend on the POS g_{i-1}. Both models were superior to one in which coefficients were arbitrarily set to $l_1 = 0.99, l_2 = 0.01$ for all POS.

2.4. Perplexity: A Measure of the Performance of a Language Model

We can view a language as a source of information whose output symbols are words. Unfortunately, we cannot know the probabilities $P(w_1, w_2, ..., w_n)$ for strings of a language. However, each language model provides an estimate $\hat{P}(w_1, w_2, ..., w_n)$ for such strings.

where $k_{M,j} + k_{C,j} = 1$,
instead of by $f(W_i = W \mid g_i = g_j)$ alone. The values of $k_{M,j}$ and $k_{C,j}$ are found by the method mentioned in 2.4.

We must also specify how we estimated the POS component $P(g_i = g_j \mid g_{i-2}, g_{i-1})$ of both the 3g-gram and the combined models. This was done in almost the same way as was done by Derouault and Merialdo. We chose to use the variant of their model in which the l-values depend on the previous POS g_{i-1}. To ensure that no POS g_j is ever assigned a probability of zero, we added an arbitrary small number 0.0001. We thus made the approximation

$$P(g_i = g_j \mid g_{i-2}, g_{i-1}) \simeq l_1(g_{i-1}) f(g_i = g_j \mid g_{i-2}, g_{i-1}) + l_2(g_{i-1}) f(g_i = g_j \mid g_{i-1}) + 0.0001 \tag{13}$$

where $l_1(g_x) + l_2(g_x) = 0.9847$ for all x (where $0.9847 = 1 - (no.of\ POSs) \times 0.0001$).

The above description ignores the case where a word will be encountered in the sample text that is not in the system's vocabulary V. We estimated the probability that such a word will occur by Turing's formula [9], which uses the frequency of unique words among all words in the training text;this yielded 0.035.

We can now give the overall formula that we used:

$$\begin{aligned} P(W_i = W \mid g_{i-2}, g_{i-1}) \quad &= \\ \text{if W in V} \quad & \\ & (1-d) \sum_{g_j \in G} [[k_{M,j} \times f(W_i = W \mid g_i = g_j) + k_{C,j} C_j(W,i)] \\ & \times \quad [l_1 f(g_i = g_j \mid g_{i-2}, g_{i-1}) + l_2 f(g_i = g_j \mid g_{i-1}) + 0.0001]] \\ \text{else d,} \quad & \\ \text{where } d \quad &= \quad 0.035, k_{M,j} + k_{C,j} = 1, l_1 + l_2 = 0.9847 \end{aligned} \tag{14}$$

Only one major modification to this model proved to be necessary in practice. We were faced with severe memory limitations, which required that we economize on the amount of data stored. For this reason, we decided to restrict the number of POSs for which 200-word caches were maintained. To be given a cache, a POS had to meet two criteria. It had to

1. comprise more than 1% of the total LOB Corpus
2. consist of more than one word(for instance, the LOB category BEDZ was excluded because it consists of the single word "was")

Only 19 POSs met these two criteria;however, these 19 together make up roughly 65% of the LOB Corpus. Thus, for POSs other than these 19, there is no cache component in the combined model; the estimated probability is identical to that of the pure 3g-gram model.

Another problem was what to do when the recognition task is biginning and the cache for g_j, containing the previous words that belong to POS g_j, is nearly empty, i.e. the number of words on which our estimate is based is far less than N.Arbitrarily, we set $k_{C,j} = 0$ until the corresponding cache has 5 words in it; at that moment $k_{C,j}$ attains its maximum value.

4. IMPLEMENTATION AND TESTING OF THE COMBINED MODEL

4.1. The LOB Corpus and Texts Extracted from It

The Lancaster-Oslo/Bergen Corpus of British English consists of 500 samples of about 2000 words each. The average length per sample is slightly over 2000, as each sample is extended past the 2000-word mark in order to complete the final sentence. Each word in the corpus is tagged with exactly one of 153 POSs. The samples were extracted from texts published in Britain in 1961, and have been grouped by the LOB researchers into 15 categories spanning a wide range of English prose [7, 8, 9]. These categories are shown in Table 1.

Table 1.

Distribution of L O B Categories				
Symbol	Descriprion	Corous	Trainig Text	Para Setting & Testing Texts
A	press reportage	44	15	9
B	editorials	27	9	5
C	press reviews	17	6	3
D	religion	17	6	3
E	skills and hobbies	38	13	8
F	popular lore	44	15	9
G	biography and essays	77	25	15
H	misceilaneous	30	10	6
J	learned writings	80	27	16
K	general fiction	29	10	6
L	Mystery fiction	24	8	5
M	science fiction	6	2	1
N	adventures and westerns	29	10	6
P	love stories	29	10	6
R	humour	9	3	2

The table above shows the 15 text categories. The column labelled "Corpus" gives the number of samples in each category in the original LOB Corpus. We extracted three different, non-overlapping collections of samples from the tagged LOB Corpus, and used each for a different purpose. All three were designed to reflect the overall composition of the LOB Corpus as closely as possible. The column labelled "Training Text" shows the number of samples in each category for the first of these collections; the last column applies to both remaining collections.

The training text for our models was used for further parameter setting, including calculation of the 1-values in the Derouault-Merialdo formula (subsection 2.4), which give the relative weights to be placed on triplet and dcublet probability estimates for the POS-prediction portion of both models. It was also used to calculate the k-values, which give the relative weights to be placed on the cache component and the Markov component in the combined model. It contained 100 samples altogether.

The third collection of samples formed the testing text. It was used to compare the combined model with the Markov model. It contained 100 samples distributed among the

LOB categories in exactly the same say as in the parameter setting text. Note, however, that only the categories and not the samples themselves are the same.

4.2. Implementing the Combined Model

Because of memory limitations, it proved impossible to implement a cache for every one of the 153 POSs in the LOB Corpus. As was mentioned in 3.2, two criteria were used to select the POSs which would be assigned a cache:

1. the POS had to constitute more than 1% of the LOB Corpus
2. the POS had to contain more than one word or symbol

The second criterion is obvious-if only one vocabulary item has a given POS, the cache component yields no extra information. The first criterion is based on the premise that rate POSs will be more spread out in time, so that the predictive power of the cache component will be weakened.

4.3. Testing the Combined Model

Two parts of the LOB Corpus were used to find the best-fit parameters for the pure 3g-gram model and the combined model, made up of the 3g-gram model plus a cache component. These two models were then tested on 20% of the LOB Corpus 100 samples as follows. Each was given this portion of the LOB Corpus word by word, calculating the probability of each word as it went along. The probability of this sequence of 230, 598 words as estimated by either model is simply the product of the individual word probabilities as estimated by that model.

Note that in order to calculate word probabilities, both models must have guessed the POSs of the two preceding words. Thus every word encountered must be assigned a POS. There are three cases:

1. the word did not occur in the tagged training text and therefore is not in the vocabulary
2. the word was in the training text, and had the same tag whenever it occurred
3. the word was in the training text, and had more than on tag (e.g.the word "light" might have been tagged as a noun, verb, and adjective)

The heuristics employed to assign tags were as follows:

1. in this case, the two previous POSs are substituted in the Derouault-Merialdo weighted average formula and the program tries all 153 possible tags to find the one that maximize the probability given by the formula
2. in this case, there is no choice;the tag chosen is the unique tag associated with the word in the training text

3. when the word has two or more possible tags, the tag chosen from them is the one which makes the largest contribution to the word's probability

Thus, although the portion of the LOB Corpus used for testing is tagged, these tags were not employed in the implementation of either model; in both cases the heuristics given above guessed POSs. A separate part of the program compared actual tags with guessed ones in order to collect statistics on the performance of these heuristics.

5. RESULT

5.1. Calculation of the L-Values

The first results of our calculations are the values $l_1(g_{i-1})$ and $l_2(g_{i-1})$, obtained iteratively to optimize the weighting between the POS triplet frequency $f(g_i \mid g_{i-2}, g_{i-1})$ and the POS doublet frequency $f(g_i \mid g_{i-1})$ in the estimation of $P(g_i = g_j \mid g_{i-2}, g_{i-1})$. As one might expect, $l_1(g_{i-1})$ tends to be high relative to $l_2(g_{i-1})$ when g_{i-1} occurs often, because the triplet frequency is quite reliable in the case. For instance, the most frequent tag in the LOB Corpus is "NN", singular common noun;we have $l_1(NN) = 0.57$. The tag "HVG", attached only to the word "having", is fairiy rars; we have $l_1(HVG) = 0.17$.

5.2. Caluculation of the K-Values

For each part of speech g_j, we calculated the weight $k_{C,j}$ given to the cache component of the combined model and the weight $k_{M,j}$ given to its Markov component. Recall that we originally created a different cache for each POS because we had hypothesized that the cache component would be more useful for prediction of content words than for function words.

The optimal weights, calculated by means of the Forward-Backward Method and shown in Table 2, decisively refute this hypothesis.

The pattern in Table 2 is just the opposite of what we had expected, with function POSs having significantly higher optimal weights for the cache component of the combined model than content POSs. This intriguing result is discussed in the Conclusion.

5.3. Performance of Both Models on the Testing Text

We calculated the performance of the various models on the testing text of 100 samples from the LOB Corpus (230, 598 words); the most important results will be given first. The pure Markov model gives perplexity equal to 332 (average probability per word is 0.003008). This compares unfavourably to Jelinek's value of 128. On the other hand, the combined model gives perplexity equal to 107 (average probability per word is 0.009341). This dramatic, more than three-fold, improvement can only be attributed to the inclusion of a cache component in the combined model.

There were 230, 598 words in the testing text. Of these, 14, 436(6.2%) had never been encountered in the training text and were thus assumed not to be in the vocabulary (not recognized). Of the remaining 216, 162 words that had occurred at least once in the training text, 202, 882(93.8%) had tags that were guessed correctly (6.2% incorrectly).

Table 2.

Oprimal Weights by POS			
POS	Descriprion	Cache Component	Markov Component
AT	singular article	0.999	0.001
ATI	sing. or pl. art.	0.998	0.002
BEZ	is, 's	0.999	0.001
CC	cocrd. conjunction	0.997	0.003
CD	cardinal	0.783	0.217
CS	subord. conjunction	0.973	0.027
IN	preposition	0.919	0.081
JJ	adjective	0.402	0.598
MD	modai auxillary	0.989	0.011
NN	sing. noun	0.403	0.597
NNS	pl. noun	0.498	0.502
NP	sing. proper noun	0.592	0.408
PPS	possessive det.	0.997	0.003
PP3A	pers. pron. 3rd pers. nom	1.000	0.000
RB	adverb	0.660	0.340
VB	verb base form	0.456	0.544
VBD	verb past tense	0.519	0.481
VBG	present part., gerund	0.518	0.482
VBN	past part	0.325	0.673

The 14, 436 words that never occurred in the training text were assigned the correct tag only 3676 times (25.4% correct, 74.6% incorrect).

6. CONCLUSIONS

The results listed in the previous chapter seem to strongly confirm our hypothesis that recently-used words have a higher probability of occurrence than the 3g-gram model would predict. When a 3g-gram model and a combined model resembling it but containing in addition a cache component were used to calculate the perplexity of a testing text, the perplexity of the combined model was lower by a factor of more than three. The importance of our results is in the trend the show, not in the precise values we obtained; these depend on the size and origin of both the training text and the testing text.

Several ideas for improvement have occurred to us. It would make sense for the weighting of the cache component to depend on the number of words in the cache in a more sophisticated way than our current step-function heuristic. Another idea would be to extend the idea of a model that dynamically tracks the linguistic behaviour of the speaker or writer from the lexical to the syntactic component of the model. In other words, recently employed POSs would be assigned higher probabilities. One might also explore the possibility of building a morphological component so that the occurrence of a word would increase the estimated probability of morphologically related words.

The line of research described in this paper has more general implications. Perhaps if we followed an individual's written of spoken use of language through the course of a day, it would consist largely of time spent in language "islands" or sublanguages, with brief periods of time during which he is in transition between islands. One might attempt

to chart these "islands" by identifying groups of words which often occur together in the language. If this work is ever carried out on a large scale, it could lead to pseudo-semantic language models for speech recognition, since the occurrence of several words characteristic of an "island" makes the appearance of all words in that island more probable.

References

1. A. M. Derouault and B.Merialdo, "Natural Language Modeling for Phoneme-to-Text Transcription," IEEE Trans.Pattern Anal.Machine Intell., vol.PAMI-8, pp.742-749.

2. A. M. Derouault and B.Merialdo, "Language Modeling at the Syntactic," (1984)

3. F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," Proc.IEEE, vol.73, No.11, pp.1616-1624, (1985)

4. F. Jelinek, R. L. Mercer and L. R. Bahl, "A Maximum Lidehood Approach to Continuous Speech Recognition," IEEE Trans.Pattern Anal.Machine Intell., vol.PAMI-5, pp.179-190, (1981)

5. F. Jelinek and R. L. Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data," Pattern Recognition in Practice, edited by E. S. Gelsema and L. H. Kanal, pp.381-397, (1981)

6. S. Johansson, E. Atwell, R. Garside and G. Leech, "The Tagged LOB Corpus Users Manual," Norwegian Computing Centre for the Humanities, Bergen, Norway, (1986)

7. S. Johansson, "Some Observations on Word Frequencies in Three Corpora of Present-Day English Texts," ITL Review of Applied Linguistics, Vol.67-68, pp.117-126, (1985)

8. S. Johansson, "Word Frequency and Text Type: Some Observations based on the LOB Corpus of British English Texts," Computers and the Humanities, vol.19, pp.23-36, (1985)

9. S. Katz, "Recursive M-gram Modeling Via a Smoothing of Turing's Formula," forthcoming paper.

10. E. M. Muckstein, "A Natural Language Parser with Statistical Applications," IBM Research Report RC7516 (#38450), (1981)

11. A. Nadas, "Estimation of Probabilities in the Language Model of the IBM Speech Recognition System," IEEE Trans.Acoust., Speech, Signal Processing, vol.32, pp.859-861, (1984)

12. J. Peterson and A. Silberschatz, "Operating System Concepts," Addison-Wesley Co., (1983)

13. L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models," IEEE ASSP Magazine, pp.4-16, (1986)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

On the Design of a Voice-activated Typewriter in French

J.-J. Mariani

LIMSI-CNRS, BP 30, 91406 Orsay Cedex, France

Abstract

Designing a Voice-Activated typewriter in French necessitates a study both on how to realize the acoustic level recognition, and on how to obtain a model of the French language. Such a project was initiated at LIMSI 10 years ago. This paper will present the different steps that have been completed since the beginning of this project.

First, a study on the phoneme-to-grapheme conversion, for continuous, error-free phonemic strings, using a large vocabulary and a natural language syntax has been completed, and published in 1979. The corresponding results of this work has then been improved, with some attempts to convert phoneme strings containing (simulated) errors, while the methodology was adapted to the case of stenotype-grapheme conversion.

Second, LIMSI is participating in the ESPRIT project 860 "Linguistic Analysis of the European Languages", In this framework, the approach for language modeling developed at LIMSI in the first project mentioned above has been studied, and compared width other approaches, that are closely related. This study has been conducted on 7 different European languages. This 4 year project is now approaching its end, and should be followed by ESPRIT Phase II project, with the goal of designing a speech-to-text and text-to-speech system, for the same 7 languages.

Third, the link between the acoustic recognition and a language model close to the one used in the above studies was made, and resulted in a complete system ("HAMLET"), for a limited vocabulary (2000 words), pronounced in isolation. This work was conducted during my sabbatical year (1985-1986) at the IBM t.j. Watson Research Center.

The part of this work concerning the design and realization of the Hamlet system has been conducted as a World Trade Visiting Scientist at the IBM TJ. Watson Research Center, in the Speech Group, from September 1985 to August 1986. The part of the work concerning the ESPRIT 291/860 project has been financed partially by the Commission of the European Communities. The part of the work concerning the MuPCD has been financed by the French Ministry of Telecommunication (Contract DGT/DAII 86.35.053).

Fourth, a parallel development at LIMSI, with a similar approach, resulted in a VAT on a 5000 word vocabulary pronounced in isolation. This system takes advantage of the existence of a specialized DTW chip (MuPCD) that has been designed at LIMSI, with the BULL and the Veesys companies. The acoustic recognition has been presented

with the chip emulator in March 1987, and the complete system with the chip itself was demonstrated in Spring 1988. This work is presently extended to continuous speech.

1. INTRODUCTION

From the very first successful attempts, speech recognition systems have been improved independently along three axes: from speaker-dependent to speaker-independent, from isolated word to connected word, and, more recently, from small vocabularies (ten to fifty words) to large vocabularies of up to 20,000 words (L. Bahl, 1983, J. Baker, 1986, W. Meisel, 1986, R. Kurzweil, 1986, J.L. Gauvain, 1986, W. Drews, 1987, C. Vincenzi, 1987, A. Averbuch, 1987). The last results concern medium size vocabularies (1000 words), speaker independent, continuous speech recognition (k.F.Lee, 1988, D.B. Paul, 1988).

At LIMSI, the idea of realizing a voice-activated typewriter for a very large dictionary was initiated in the early 70s. The very first experiment on this topic concerned the phoneme-to-grapheme conversion (that is the segmentation into words and the correct orthographic translation of those words) of an error-free continuous phoneme string (the text was "La chèvre de Monsieur Seguin"), using a simple heuristic: the choice should be made on the solution giving the smaller number of words for a given sentence (J. Mariani, 1977). The conclusion of this work was to say that there was a need for a syntactic filter eliminating the incorrect word successions.

This conclusion was followed by a cooperative project with a group working at CEA (Center for Nuclear Research) on Data Base query in natural language, using positional grammars trained on text corpora (A. Andreewski, 1972). As the first experiments demonstrated the effectiveness of the approach (A. Andreewski, 1978), the work was continued at LIMSI and resulted in an operational system for phoneme-to-grapheme conversion (A. Andreewski, 1979).

Since that time, the system has been improved. It has been adapted to stenotype-to-grapheme conversion (G. Adda, 1988), and experimented on phoneme strings containing errors (D. Bellity, 1984). The basic methodology has also been used in the ESPRIT project 860 "Linguistic Analysis of the European Languages", and extended to different European languages (L. Boves, 1987). On the other hand, a similar approach for language modeling has been applied on real speech recognition, with a reduced lexicon, and isolated word pronunciation (J. Mariani, 1987, J.L. Gauvain, 1988).

Presently, our Voice-Activated TypeWriter project forms with our Dialog project the two goals around which are organized our research in Speech Processing.

2. SOME PROBLEMS RELATED TO TEXT DICTATION IN FRENCH

2.1. General Problems

General problems concern the phoneme-to-grapheme conversion of the homophones in French, which seems to be more difficult than for other languages, even for isolated words.

On a general point of view, a basic dictionary of 22,000 words will gibe a full-form dictionary of 170,000 graphemic words. Doing the grapheme-to-phoneme translation of those words gives a dictionary of 90,000 phonemic forms. This means that, for a very large dictionary, as a mean, a phonemic word corresponds to two different graphemic words.

The main problems are related to verb conjugation; this gives an average of 40 forms, and up to 3 different spellings of the same pronunciation, for all verbs.

The mark of the plural of most of the substantives, most of the adjectives, and all the past participles (an -s at the end of the word) is never pronounced in isolation. The mark of the feminine for some substantives, most of the adjectives and the past participles (-e at the end of the word) is not pronounced in fluent speech, not even in isolation.

The demonstrative adjective "ces" (*those*) and the possessive adjective "ses" (*his*) have the same pronunciation /se/, but different spelling.

If we consider now the case of continuous speech, the problem of segmenting the continuous phoneme string into words seems to be especially difficult in French. We conducted experiments on a simple sentence containing 9 phonemes, with the 170,000 word (full-form) lexicon. We obtained more than 32,000 possible transcriptions (segmental and orthographic translation) at the lexical level. Using phonological rules, syntax and semantics will still allow for two acceptable sentences that need a pragmatic analysis in order to get the right graphemic transcription (Figure 1).

2.2. Problems due to the Pronunciation "In Isolation"

Some other problems are arising, if the pronunciation is made "in isolation". There are in French "liaisons" ("links") between words, i.e. phonemes that are pronounced at the junctions between two words, but wouldn't be pronounced at the end of the first word, or at the beginning of he second one, if the words were pronounced in isolation. A possibility is not to pronounce the liaison at the beginning of the following word, but it increases the size of the vocabulary, as all the possible liaisons at the beginning of the word should be included. A third one is to pronounce it as three words, but the pronunciation of the liaison in isolation will be quite unnatural. However, the liaisons help for certain graphemic conversions (such as deciding whether the form is plural or singular.

In the same way, a vowel at the end of some words can be omitted, if the next word is beginning by a vowel, and this will result in an apostrophe. The possibilities here are also to pronounce the first word as it was not modified, but it will sound unnatural. Another one is to pronounce the two wards together, but here also it will enlarge the size of the vocabulary. A third one is to pronounce it as three words, with the word "*apostrophe*" in the middle, but it is also quite unnatural.

2.3. Other Problems

As for other languages, the pronunciation of numbers is a problem, as recognizing the numbers 0 to 9999 is already recognizing a vocabulary of 10000 words!

	%Words	Number of Words	Mean Number of forms/Word	Number of Forms
Verbs	10-15%	3000	40	120000
Substantives	60%	12000	2	24000
Adjectives	25%	5000	2,5	12500
Adverbs and Other	5%	1000	1	1000
Total	100%	21000		157500
		22000		170000
Phonemic				90000

Figure 1. Problems related to the phoneme-to-grapheme conversion in French. (a) Information on the transcription of the basic dictionary into the full-form dictionary.

Verbs:

/kas / : casse,casses,cassent *(break)*

Substantives

Masculine/feminine:

/ami/ : ami *(friend (he))*

/ami / : amie *(friend(she))*

singular/plural

/tasə/ : tasse,tasses *(cup, cups)*

/kanal/ : canal *(canal)*

/kano/ : canaux *(canals)*

Adjectives

Masculine/feminine

/ene/ : aîné *(older(he))*

/eneə/ : aînée *(older(she))*

/grã/ : grand *(big (he))*

/grãdə/ :grande *(big (she))*

singular/plural

/grãdə/ : grande,grandes *(big)*

/mãtal/ : mental *(mental)*

/mãto/ : mentaux *(mental)*

Past Participies

/kase/: cassé, cassés, cassée, cassées *(broken)*

Figure 1. (b) Some usual homophones heterographs in French.

Phonemic string:	/ʒemalopje/	*(My foot hurts)*
Possible segmentations	J'ai mal au pied	
at the lexical level:	Geai mâle au pied	
	Geais ma lot pieds	
	J'hait mât l'eau piller	
	J'aime allo pillé	
	J'aimes allo pillé	
	Jet malles hop y est	
	Gemme halles hopi et	
	....	
Remaining possibilities	J'ai mal au pied	*(My foot hurts)*
with phonology,syntax	J'ai mal aux pieds	*(My feet hurt)*
and semantics		

Figure 1. (c) Some problems related to lexical segmentation for continuous speech.

Des:/de/*(some)*
amis:/ami/*(friends)*
Des amis:/deʒami/*(some friends)*
Mon ami:/mõŋami/*(my friend)*
Petit ami:/pətiţami/*(boy friend)*
Petits amis:/pətiʒami/*(boy friends)*
Le ami:L'ami*(the friend)*
de ami:d'ami*(from a friend)*

Figure 2. Problems of the Haisons and apostrophe in French.

3. PHONEME-TO-GRAPHEME CONVERSION

As a consequence of the difficulty of phoneme-to-grapheme conversion for continuous error-free phoneme strings, illustrates in Figure 1, we experimented the use of a natural language syntactic parser (A. Andreewski, 1979).

3.1. The Lexicon

The lexicon is composed of 170,000 words, It is a full-form dictionary, as all the different forms corresponding to the conjugation of a verb, for example, will be considered as different inputs in the dictionary. It has been obtained from a 22,000 word basic dictionary. Each graphemic word has been converted into its phonemic form by using an automatic phoneme-to-grapheme conversion software designed at LIMSI (B. Prouts, 1980). It also gives the grammatical category of each word, its gender and number for the substantives and adjectives, the mode, time, person, group, transitivity, and root for the verb...

3.2. The Syntax

The syntax is positional.It is given by a 2D Boolean matrix giving the possibility of the succession of two grammatical categories and a 3D frequential matrix giving the frequency of the succession off three grammatical categories. 150 grammatical categories have been chosen, based both on linguistics, and on the results of experiments. The matrixes were trained iteratively on a set of texts.

3.3. The Test

The test was obtained by segmenting into words a text of 1800 words, converting each word into its phonemic representation by the same grapheme-to-phoneme conversion software used for the lexicon. In that way, the liaisons are not taken into account. All punctuation marks are kept. Then all blanks were deleted, in order to get the error-free continuous phoneme string.

3.4. The Conversion Process

The phonemic string is processed in the following way: all possible segmentation into wards, with regards to the lexicon, are tried and filtered by the 2D and 3D matrixes, without using the frequencies. When several possibilities are still existing, the one with the smaller number of words is kept. If there are still several possibilities, the frequency in the 3D matrix is taken into account. The one of lexical frequency was also mentioned, but wasn't actually implemented.

3.5. Results

The results of the experiment were the following (Figure 3): on the 1,800 word test, 75 errors occurred (that is less than 5%): 13 homophones heterographs (plan / plant (*map* /

plant), Heures / heurts (*hours / collisions*), ère / air / erre / hère / aire (*era / air / wanders / wretch / area*)...), 21 for singular / plural, some of them being impossible to distinguish (plans / plan d'èxècution (*maps / map for execution*), demande /demandes de permis (request / *requests for permission*)...), 10 concerning the number of posterior adjectives (périmètre de protection des monuments historique / historiques (*area of protection of historical / historical monuments*)), 13 syntax parsing errors (les baisses ont équipé / les baies sont équipées (*the falls hove equipped / the windows have equipped*), et celles situées / et sels situés (*and those situated / and salts situated*)...), 4errors due to the heuristic of the 'smaller number of words" (et décors étrangers / et des corps etrangers (*and foreign sceneries / and some foreign bodies*), un temps froid éventé / un temps froid et venté (*a fanned out cold weather / a cold and windy weather*)). The processing time was 90 words/minute on the IBM 370/168 of the CNRS computer center, functioning in time sharing.

The conclusion of the experiment was to recommend the use of lexical frequency, and that acoustic and linguistic aspects should be processed all together, The automatic semantic analysis based on word co-occurence was another recommendation.

LES EXEMPTIONS PREVUES PAR LE PRESENT ARRETE NE SONT PAS APPLICABLES AUX TRAVAUX CONCERNANT LES CONSTRUCTIONS FRAPPEES D'ALIGNEMENT ET SEL/SELLE/SELS (CELLES) SITUE/SITUES (SITUES) DANS LE PERIMETRE DE PROTECTION DES MONUMENTS HISTORIQUES/HISTORIQUE (HISTORIQUES) ET DES SITES CLASSES.

The exemptions allowed by the present decree are not applicable to the works concerning the buildings that have to be aligned and salt/saddle/salts (those) situated/situated (situated) in the area of protection of historical/historical (historical) monuments and landmarks.

POUR LES CONSTRUCTIONS EDIFIEES SUR LE TERRITOIRE DE LA VILLE DE PARIS, LA CONSULTATION S'EFFECTUE AU LIEU, JOUR/JOURS (JOUR) ET HEURE/HEURES/HEUR/HEURT/HEURTS (HEURE) FIXE/FIXEES/FIXES (FIXES) PAR ARRETES DU PREFET DE LA CENE/SCENE.

For the constructions built on the district of the town of Paris, the consulting is made at the place, day/days (day) and hour/hours/fortune/collision/collisions (hour) fixed/fixed/fixed (fixed) by decree of the prefect of the Holy Communion/scene(Seine).

Figure 3. Some examples of phoneme-to-grapheme conversion. *The error of ambiguities are underlined, and followed by the right wording inside parenthes.*

3.6. Extensions of the Work

From those results, some improvements were introduced, such as increasing the size of the vocabulary to 270,000 forms, taking into consideration the liaisons and the elisions (apostrophe) and the corresponding phonological rules, and introducing a better algorithm for gender and number marking (J. Avrin, E. Bsaiis, 1983). Another attempt was to process phonemic isolated word containing errors, that were obtained by confusion matrixes, or by using existing phoneme recognizer prototypes, with the 270,000 word dictionary. It was found that the way of accessing the dictionary was critical, and that severe errors, or an overall phoneme error rate of more than 15%, would not allow the selection of the right word (D. Bellilty, 1984).

Pour ce rôle militaire qu'elle a à assumer, la Syrie et (est) aidait (aidée) par l'Union Soviétique. l'armement vièn̂t de là-bas. Un rôle militaire prédominant qui a modelé l'économie comme le montrent (montre) ce raportage de Dominique Bromberger au moment où l'on s'appréte à tout revoir au plan gouvernemental. Tel un château-fort, le balai (palais) du peuple, comprenez le balai (palais) présidentiel, surplombe Damas et ces (ses) casernes.

For this military role that it has to assume, Syria and (is) helped (helped) by Soviet Union. The arms come from there. A major military role that shaped the economy as it appear (appears) in this report from Dominique Bromberger at a time where everything is to be changed at the govemmental level. As a fortress, the broom (palace) of the people, you should understand the presidential broom (palace), is overlooking Damas and those (its) barracks.

Figure 4. Some results on Stenotype-to-grapheme conversion from real data. *The errors are underlined and followed by the right wording inside parentheses.*

3.7. Adaptation to Stenotype-to-Grapheme Conversion

Another use of those results was to adapt the system to stenotype-to-grapheme conversion. The goal was to realize a real time TV subtitles editor. The stenotype method allow an experienced typist to take dictation in real time of what is said. The keyboard has a number of keys corresponding to phoneme-like units, the difference between voiced and unvoiced phonemes (such as b and p) being absent. A syllable is typed at a time. Those peculiarities of the keyboard make it quite similar to a speech input, and some errors will correspond to what could be expected from a speech recognition device. The system has been realized and demonstrated. The 270,000 graphemic word dictionary has been translated in a stenotypic dictionary. It has 520,000 stenotypic forms due to the ambiguities of the keys. The error rate varies from 5 to 20%, on actual broadcast TV news, depending on the number of unknown words (proper names...). Some problems arise from the fact that the subtitles cannot appear exactly in real time as the syntactic parser needs a window of a few words to decide on the most likely parsing, and the corresponding wording (G. Adda, 1988).

4. THE ESPRIT 291/860 PROJECT

This project was launched in 1984, for two years (291), and was followed in 1985 by a three year contract (860), for a total of 4 years. The goal is the Linguistic Analysis of the European Languages, in view of their oral recognition or synthesis. This project involves 8 laboratories (Olivetti (Italy) as a prime partner, LIMSI (France), Nijmegen University (The Netherlands), University of Bochum (FRG), Acorn Computer (UK), University of Madrid (Spain), University of Patras (Greece), CSATA (Italy)) studying 7 languages (Italian, French, Dutch, Spanish, Greek, German). An important part of the project was the building of a language model for the different languages. The approach that was chosen is the Markovian approach using 2D and 3D frequency matrixes on grammatical categories that was developed at LIMSI. As an alternative, LIMSI also experimented a different Markovian Approach called Binary-Rules, which focuses the language model on the ambiguities to solve.

The main results of this project are statistics on phoneme clusters, grapheme-to-phoneme conversion software, phoneme-to-grapheme conversion software, language models and syntactic parsing, the integration of those different elements under a blackboard structure, and the definition of the quality of a language model (or of the difficulty of the language) (Boves, 1987, Vittorelli, 1987).

On this last issue, some interesting results have been found. One, for example, is related to the experiments made to the phoneme-to-grapheme conversion using context dependent rewriting rules, at the phonological level (Figure 5). It shows that for Italian, a set of 67 rules is able to transcribe the words with 0.5% graphemic words not existing in the language, and 0.5% graphemic words missing in the ones that were transcribed. On the contrary, for French, a set of 586 rules will conduct to have 98% of the word generated not existing in the vocabulary, and 30% of the correct words missing in the resulting graphemic cohorts. Although this measure illustrates also the quality of the rules, it seems obvious that the Italian Language will require less linguistic process than the French language in order to achieve its translation from a phoneme string.

This 4 year project is now approaching its end, and should be followed vy an ESPRIT Phase U project, with the goal of designing a speech-to-text system, for the same 7 languages, using the Language Models developed in the 860 project.

Lang.	Nw/Nph	Nfalse/Nw	Nmiss/Nph	+Ndif/Nph	#Rules
Dutch	6	90%	20%	+0%	289
English	10	90%	6%	+1%	530
French	250	98%	30%	+40%	586
German	400	99%	10%	+0%	551
Greek	100	100%	2.5%	+0%	394
Italian	1	0.5%	0.5%	+0%	67
Spanish	1	7%	6%	+0%	845

Figure 5. Phoneme-to-grapheme translation for different languages. (*Nph is the number of phonemic words as input. Nw is the number of graphenic words that were generated Nfalse is the number of graphemic words that are not in the language. Nmiss is the number of graphemic words that were not found in the resulting cohort. Ndif is the number of graphemic words corresponding to a phonemic word. Ndif/nph represents the increase of the size of the lexicon after phoneme-to-grapheme translation.*)

5. "THE HAMLET"SYSTEM

The goal of that work was to integrate the different parts of an Isolated Word, Speaker-Dependent Voice Activated Typewriter (VAT) (lexical decoding, graphemic translation) on a stand-alone personal computer without using specialized Integrated Circuits to carry on the recognition task. The target vocabulary size was one to several thousand words.

Another point of interest was to measure the ability of a "natural language" linguistic model to correct "acoustic" word recognition errors, as well as achieving phoneme-to-grapheme conversion.

The project was conducted in three steps: First, a language model was constructed. Then, speech compression and recognition techniques were tested in order to obtain acceptable memory size and response time. Finally, the language model was introduced in the recognition process, and the whole system was tested.

5.1. Why "Hamlet"

The name of the system was chosen as an illustration of the typical problem in the phoneme-to-grapheme translation of the phrase "To be or not to be" (that could be translated as "2B or not 2B" in an office dictation task (F. Jelinek, 1986)). As a second reference, the Guinness Book of Records has reported successful attempts to pronounce the 262 words of Hamlet soliloquy in less than 24 seconds. That is a pronunciation rate of 655 words/minute (to be compared with the world record for fastest typewriting by A. Tangora, at a rate of 147 words/minute). Unfortunately, however, the resulting text is completely undecipherable!...

5.2. Building Up the Language Model (Figure 6)

The semantic universe is related to the dictation of a research report in the field of speech technology in French. The training data is made up of an existing report of 20 pages in length (15,000 words).

During the training of the linguistic model, a page of this text corpus is analyzed using the language model built from the previous pages (for the first page, it will start from scratch). The text is first segmented into words, and each word is looked up in the lexicon. If it is found, its phonemic representation and grammatical category are given. If not, its phonemic representation is obtained by using an automatic grapheme-to-phoneme conversion software, and its grammatical category is inferred inductively by using a stochastic syntactic parsing method. The result of this analysis is "processed" (or "verticalized" (L. Boves, 1987)) text, where each graphemic word of the text is followed by its phonemic translation, its grammatical category, and the type of inference (lexical or syntactic) that was used to get the information. This text is manually corrected, and is used to update the lexicon, and the syntax, that will be used to process the next page in the same way.

5.3. Grapheme-to-phoneme conversion

Many grapheme-to-phoneme conversion programs have been written for speech synthesizers. The one we have developed here is based on phonological rules. It uses a set of declarative rules, the exceptions being considered as longer rules (Example: "eur" is pronounced / œr /, but "monsieur" is pronounced / mesjø/). Those rules are then compiled in a tree structure, in order to accelerate the conversion process. This conversion system has been adapted to the dictation task. For example, punctuation marks are not pronounced explicitly in speech synthesis, but will be pronounced in text dictation. 520 rules have been defined, and tested on a set of 6 topics (on very different texts), giving a total of 5,000 words.

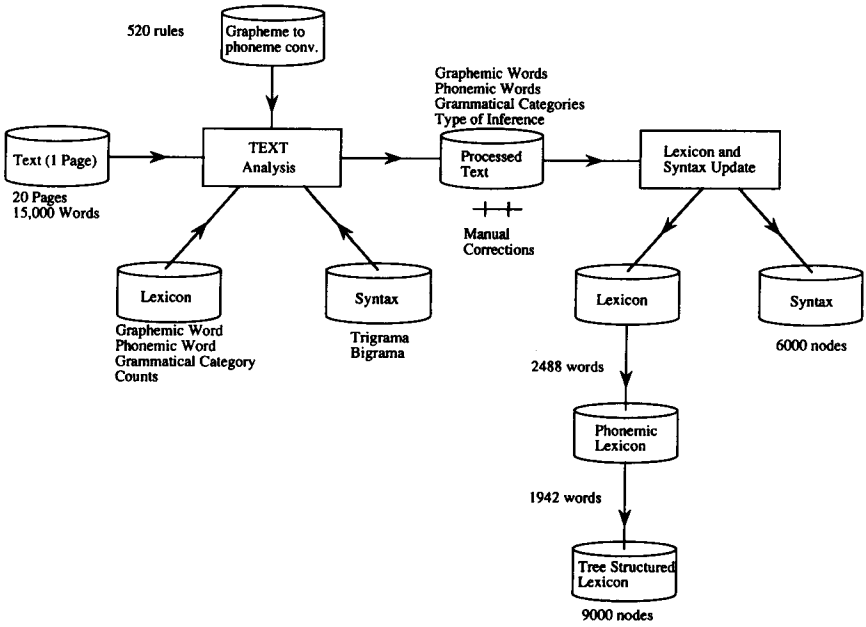


Figure 6. Building up the Language model.

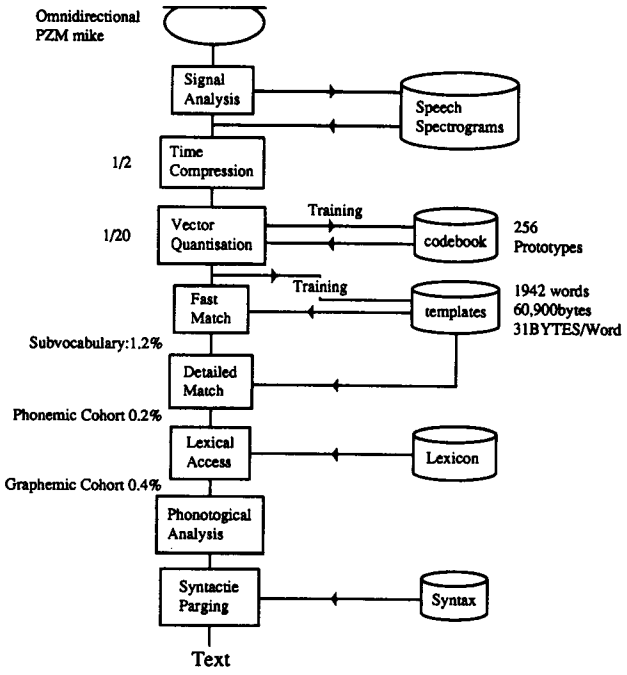


Figure 7. The template training and recognition processes. *At each stage of the decoding the compression rate from the whole template of word dictionary is given.*

5.4. Pronunciations Rules

Following the peculiarities of French presented in 2., the following choices have been made: numbers are usually pronounced as digit strings, unless they are included in a word. The apostrophe in French is pronounced as a word. The "liaisons" between words are not pronounced.

5.5. Phonological Rules Used in the System

From the previous findings, some phonological rules have been defined and are used in the system. For example, a rule says that "an apostrophe must be followed by a vowel", or that "the possessive adjective "mon" cannot be followed by a feminine word beginning by a consonant". 12 different general rules have been defined.

5.6. The Lexicon

The words contained in the lexicon are defined by their graphemic form, their phonemic form, and their grammatical category. The 15,000-word text corpus, including a special set of "grammatical" words, gives about 2,55 different graphemic words, and 2,000 different phonemic words. Inorder to accelerate the look-up, the phonemic lexicon is translated into a free structure, as a "cohort model" (W. Marslen-Wilson, 1980). The analysis of the lexicon shows that about 10% of the words are proper names, acronyms, or foreign words. This shows that VAT systems must give the user the possibility of adding easily new words, that will not be found in the general dictionary of a given language.

5.7. The Grammatical Categories

160 grammatical categories have been defined, closely related to the categories used by other authors (A. Andreewski, 1972, A.M. Deroualt, 1985). They are obtained from 55 basic categories, by adding gender or number information (for example, SMS: "Substantive Masculine Singular", LMS: "Article Masculine Singular", JOU: "Weekday"...). They are differentiated as "close categories" (such as JOU, or LMS), which already have all of their elements in the lexicon, since they are easy to find, or j"open categories" (as SMS). This differentiation is used in all the inductive inference process during the training.

5.8. The Language Model

The language model is given by a Markov chain that gives the possibility of the occurrence of two (bigram) or three (trigram) successive grammatical categories (A. Andreewski, 1972, L.R. Bahl, 1978, A.M. Deroualt, 1985). Those probabilities are obtained from the count of those occurrences in the training data (for example, if the grammatical category "article masculine singular" (LMS) has been found N times in the training text, and was followed P times by the "substantive masculine singular" (SMS) category, the corresponding probability of the Bigram LMS*SMS is computed as P/N). They are stored as a tree structure: each node of the 3-level tree contains a grammatical category, with the count of how many times this node has been accessed during the linguistic training.

5.9. Syntactic Parsing

The syntactic parsing is done by using a Viterbi algorithm (G.D. Forney, 1973) applied on a lattice of the possible grammatical categories of each word of a sentence, obtained from the lexicon. This process can be extended to default parsing, if a word is unknown in the lexicon, or to phoneme-to-grapheme conversion through syntactic parsing (A. Andreewski, 1979, A.M. Deroualt, 1985, L. Boves, 1987).

5.10. Signal Acquisition

The microphone is an omnidirectional Crown PZM microphone, which is placed on the keyboard. The ambient noise is that of a typical calm office environment. The signal processing is done on the IBM Personal Instruments card set, based on an IBM VLSI DSP chip. The sampling rate is 20 kHz, and each sample is coded on 12 bits. The speech spectrogram is obtained, after 512 point FFT, by a 20-filter Critical-Band Filterbank (E. Zwicker, 1981). Each value is logarithmically coded, the spectrogram is smoothed. The noise spectrum, obtained from the "silent" zones, is subtracted, and the amplitude is normalized. Endpoint detection is done by using several thresholds.

5.11. Speech Compression

Different speech compression algorithms have been tested on difficult vocabularies (minimal pair syllables and words) in order to measure the loss of recognition quality that is supposed to occur, due to the loss of information. On the contrary, it has been found that best recognition results were obtained with the highest compression rate. First, a non-linear time compression (Variable Length Trace Segmentation (j.L. Gauvain, 1872, M.H. Kuhn, 1983) is applied, and gives a typical compression rate of 2. The goal of this algorithm is to compress the steady parts of the signal. Then, Vector Quantization is applied (A. Buzo, 1979, J. Mariani, 1981,...). The Codebook has been built by using a covering method, and it contains 256 prototypes. Using those prototypes to encode the spectrograms gives a compression rate of 20. That is a total compression rate of 40.

5.12. Reference Templates Training

During the training phase, all the words of the phonemic vocabulary are pronounced once. They are compressed and stored in memory, with their phonemic label. The 2,000 templates are stored in 60 KBytes of RAM memory (about 31 Bytes/word, on the average). Speaker adaptation on a small amount of speech data could be obtained through vector quantization (K. Shikano, 1986, H. Bonneau, 1987).

5.13. Fast Match

The first recognition step is used to eliminate most of the words of the lexicon. As the vocabulary is large, this step must be carried out rapidly. Two parameters are used: First the length of the word to be recognized is compared with the length of the templates (both after time compression). This process can only be achieved after complete pronunciation

of the whole test word. The second parameter is an "extended similarity" function, which is computed synchronously with the pronunciation of the test word, that gives the distance between segments of the test word and each of the prototypes in the Codebook. This similarity function is then used to compute a gross similarity measure between the test word and the templates, This Fast Match gives an average preselection of 24 words (1.2%). For both matching processes, thresholds are are used to do word preselection.

5.14. Detailed Match

A more accurate recognition process is then applied to the preselected words. This matching uses an asymmetrical, test synchronous, Dynamic Time Warping algorithm, with slope constraints, and local and global rejection thresholds. The recognition score for each word is normalized in (0, 1). The average size of the phonemic cohorts is 4 words (0.2%).

5.15. Use of the Language Model

The phoneme-to-grapheme conversion for each of the words in the phonemic cohorts uses the tree structured phonemic lexicon. The resulting graphemic cohort contains the graphemic words, with the corresponding recognition score, and the corresponding grammatical category. The average size of the graphemic cohort is 9 words (0.4%). The choice of the graphemic word strings is made by using a Viterbi algorithm, and combining the acoustic distance measure and the probabilities of grammatical bigrams and trigrams obtained from the language model.

5.16. Results

Tests have been conducted first with a 100-word text dictation test. At the acoustic level, 9 errors are reported (91%). The complete (trigram) language model corrects 6 errors, and does not introduce any errors in the grapheme-to-phoneme conversion thus improving the recognition score to 97%. The incomplete language (bigram + trigram) model, where the pronounced text is not included in the training text data, corrects 5 errors, giving a recognition score of 95%. This difference with the results obtained on the complete model shows that training needs more text data. Figure 8 gives some examples of graphemic transcriptions when the right word wasn't recognized in the first position.

Further tests were conducted three months later, in order to test the variation of the reference templates, on the following 200 words of the text. Acoustic recognition results were 92.5% correct. Introducing the language model improves the results to 95%. Complete results for the 300-word test data are given in Figure 9.

The average recognition time is about 2 seconds, (1.8 s for the acoustic match, 0.2 s for the linguistic process).

Correct Sentence(CS):	Titres	et	Travaus	de
Recognized Cohort(RC):	titres	clé	travaux	de(3)
	titre	clés	avons(2)	deux
		et		2
		...		te(3)
				peut
				peu

(*Title and works of ...*)

Figure 8. (a) Example giving the different graphemic words. *The number of possible grammatical categories for each graphemic word is given.*)

CS:	dans	le	cadre	de	ma	thèse	de	docteur
RC:	dans	le	4	te	ma	thèse	ou	docteur
	tant	me	carte	de	bas		au	laquelle
	banc	eux	par	plus	pas		...	
	en	peut	capot	peut	la		...	
	...	...	car		...		de	
			cadre					

(*have been conducted on the problems related...*)

Figure 8. (b) Another example with two consecutive errors. (*The number of possible grammatical categories for each word doesn't appear here.*)

CS:	se	sont	portés	vers	les	problèmes	relatifs
RC:	se	sont	portés	air	les	programme	relatifs
	ce	son	porter	faire	liè	problèmes	relatif
	ceux	sons	portèes	heure	clè	problèes	
	CEE	soit		vers	clès		
	seul	sans		...	mes		
	...	ont			...		
		...					

(*In the framework of my thesis doctor...*)

Figure 8. (c) An example where the language model does not succeed in correcting the error: three recognition errors have been made in a close interval, and the first error is made on a word having the same grammatical category as the correct word. (*Here, a single graphemic word is given for each phonemic word.*)

Figure 8. Examples of errors corrected by the language Model, or not. (*The word-candidates are ordered following their recognition score. The words finally recognized are in bold characters and underlined.*)

Recognition mode:	% correct
Phonemic words	92%
Phonemic words with complete LM	98%
Graphemic words with complete LM	95.7%
Graphemic words with imcomplete LM:	
-Trigrams	92.7%
-Trigrams+bigrams	95%

Figure 9. Recognition results. The average recognition time is about 2 seconds, (1.8'*s for the acoustic match,* 0.2'*s for the linguistic process*).

5.17. Discussion of the Results

Analysis of the results in the first experiment (on 100 test-words) show that, if the recognition rate of the correct word in the first position is 91%, it goes up to 98%, if we consider the first 5 word-candidates.

All the recognition errors have been made on 1 or 2-syllable words. As the shortest words, which seem to be more difficult to recognize, are also the most frequent ones, it should be noted that recognition rates on text dictation will be worse than when using the pronunciation of a lexicon. It should be also noticed that, as those short words are very common, they will be well represented in the language model, and thus, the language model will greatly help in correcting the "acoustic" recognition errors which are more frequent on those short words.

The recognition rate doesn't vary when the size of the lexicon goes from 1500 words to 2000 word. Here also, it may be thought that, ass the lexicon is built up incrementally from successive pages of text, the shortest word which bring most errors and are most common ones will be rapidly included in the lexicon. Thus, one may think that the error rate will not increase linearly with the size of the lexicon.

Finally, we see that the best results with "incomplete" linguistic training are those obtained with the grammatical "bigram + trigram" model, as, if some trigrams may be absent in the linguistic model, corresponding bigrams may have been learned.

6. THE PRESENT VOICE-ACTIVATED TYPEWRITER PROJECT

The present state of the project demonstrates a 5,000-word Voice-Activated Typewriter (Isolated Words, Speaker Dependent). It takes advantage of a custom VLSI, that has been designed at LIMSI.

6.1. The MuPCD DTW Custom VLSI

The MuPCD DTW custom VLSI has been designed by a consortium including the BULL and Veesys companies, and LIMSI, on a contract of the French Ministry of Telecommunication (DGT/DAII) (G. Qu not, 1986). The goal was to design an integrated circuit

allowing for faster Dynamic Programming processes. This circuit is aimed at all applications involving a pattern matching operation (Speech and Character Recognition, Stereovision, Scene Analysis, Operational Research...). In the field of Speech Recognition, the goal was to increase the size of the vocabularies. The distance is a L-1 distance. The DP equation is programmable, and includes Isolated Word and Connected Word recognition algorithms. The circuit is available since December 1987. Its power is 10 MIPS (Million Instructions per Second), or 30 MOPS (Million Operations per Second), as its Pipe-Line structures allows to process up to 3 operations in parallel. Using an optimal DP-matching equation, it allows for the recognition of 1,000 words in isolation, or 300 words continuously, in real time. The technology used is CMOS 2 microns. It includes 127,309 transistors, It has been used for DTW-based recognition, and therefore, is usable for Viterbi alignment in discrete HMMs. The design of a Pattern Recognition board (RdF), at the PC format, has been realized. A project aiming at integrating 4 to 8 MUPCD on a single board at PC format is presently on its way.

6.2. The Methodology for Large Vocabulary Recognition

The system operates also in two-passes. After the vector quantization, a Fast Match is first applied to reduce the size of the vocabulary by selecting the words that are the most similar to the one that is to be recognized, and then a Detailed Match is used, that gives the list of word-candidates with their recognition score (F. Simon, 1985). Here, the Fast Match is obtained through the simple summation of the scores on the diagonal of the distance matrix. The Detailed Match is obtained by a classical DTW algorithm. Both are recognized in the MuPCD custom DTW VLSI, which allows for a vocabulary of 5,000 words in Real Time, using this non-optimal two-pass recognition process. The language model is then applied. A bigram model has been used, with 59 grammatical categories.

6.3. Results (Figure 10)

The system has been tested on the vocabulary corresponding to a text-book in French for foreigners. It has 5,127 phonemic words, corresponding to 6,7000 graphemic words. On a 1000 word text, the rough phonemic word recognition results were of 91%, and increased to 99% using the Language Model. The final results on graphemic words were of 75 errors (that is 92.5% correct), for one speaker. All tests were made on text data that were used for building the Language Model. The recognition time is 480 ms on the average (J.L. Gauvain, 1988). The system is now enlarged to continuous speech recognition, through syllables (J.L. Gauvain, 1986) and dissyllables as decision units, and HMMs. It will take advantage of the MultiMuped hardware under development, to achieve real-time.

7. CONCLUSION

We think that practical use of such systems in the future will necessitate easy speaker adaptation, and continuous speech recognition.

l' activité corporelle. 1 . les pommes (hommes) et les animaux peuvent remuer, se mouvoir, se donner du mouvement . les pommes (hommes) sont capables de faire des gestes de la tête et de la main . si on ignore un mot étranger , on peut se taire (faire) comprendre par des signes . 2 . monsieur Leclerc est fort , il a de la force , il est robuste . André Caron fait des courses de dix ou quinze kilomètres , il est résistant . madame Leclerc coud ; elle est adroite ; si elle était maladroite , le travail serait mal fait . la robe va bien ; madame Leclerc est habile ; la fillette veut coudre aussi ; elle a encore des gestes gauches . 3 . cette (cet) âme (homme) à (a) une jambe plus courte que l'autre ; il est boiteux ; il boite de la jambe gauche ; un accident l'as (a) rendu infirme . les mutilés ont perdu un bas (bras), une jambe ou un oeil dans un accident . 4 . le maître arrive . les enfants se lèvent . il (ils) reste (restent) debout . le maître Guy (crie) : " assis " . dans le fonds (fond) , quelque (quelques) élève (élèves) n' ont pas entendu . le maître répète : "asseyez vous " ... " acier (assied) toi , Daniel " .

the corporal activity . 1 . the apples (men) and the animals can move, stir, move themselves . The apples (men) are able to move their head and hand . if one ignores a foreign word , he can be silent (understood) using signs . 2 . Mister Leclerc is strong . he has strength, he is robust . André Caron runs ten or fifteen kilometers , he is tough . Mrs Leclerc is skilful ; the young girl also wants sewing ; she still has clumsy motions . 3 . this (this) soul (man) at (has) a leg shorter than the other ; he is wobbly ; an accident have (has) made him handicaped . the disabled persons have lost a stocking (arm), a leg or an eye in an accident . 4 . the teacher arrives . the children stand up . he (they) stays (stay) standing . the teacher Guy (shouts) : " sit down " . in the fund (back) , a certain (some) pupil (pupils) did not hear . the teacher repeats : "sit down" ... "steel (sit) down , Daniel " .

Figure 10. An example of text dictated. *The errors are underlined and followed by the right wording inside parentheses.*

References

1. G. Adda, "Reconnaissance de Grands Vocabularies: Une étude Syntaxique et Lexicale," These de Docteur-Ingenieur, Université Paris XI, (1987)

2. A. Andreewski and C. Fluhr, "Experiénces de constitution d'un programme d'apprentissage pour le traitement automatique du langage," Note CEA 1606, (1972)

3. A. Andreewski, F. Débili, C. Fluhr, J.S. Liénard and J. Mariani, "Une expérience D'aide linguistique à la reconnaissance de la parole," Note CEA 2055, (1978)

4. A. Andreewski, J.P. Binquet, F. Debili, C. Fluhr, Y. Hlal, J.S. Liénard, J. Mariani and B. Pouderoux, "Les dictionnaires en forme complète, et leur utilisation dans la transformation lexicale et syntaxique de chaînes phonétiques correctes," 10èmes "Journées d'Etudes sur la Parole" du "Groupement des Acousticiens de Langue Française," Grenoble, pp 285-294, (1979)

5. A. Averbuch, L. R. Bahl, R. Bakis et al., "Experiments with the TANGORA 20,000-Word Speech Recognizer," IEEE ICASSP, Dallas, pp.701-704, (1987)

6. J. Avrain and E. Bsalis, "Transcription orthographique d'une chaîne phonétique pour la reconnaissance de la parole," Rapport DEA, LIMSI, Juillet (1983)

7. L. R. Bahl, R. Bakis, P. S. Cohen, F. Jelinek, B. L. Lewis and R. L. Mercer, "Recognition of a Continuously Read Natural Corpus," IEEE ICASSP'78, Tulsa, pp.422-425, (1978)

8. J. K. Baker, "Automatic Transcription on a Personal Computer," Speech Tech '86, New York, p.193, (1986)

9. D. Bellilty, A. Lund, "Conversion Phonèmes-graphèmes de suites phonétiques entacheés d'erreurs," Rapport interne LIMSI, Juillet, (1984)

10. H. Bonneau, J. L. Gauvain, "Vector Quantization for Spiaker Adaptation," IEEE ICASSP, Dallas, pp.1434-1437, (1987)

11. L. Boves, M. Refice et al., "The Linguistic Processor in a Multi-Lingual Text-to-Speech and Speech-to-Text System," European Conference on Speech Technology, Edinburgh, pp.385-388, (1987)

12. A. Buzo, A. H. Bray, Jr., R. M. Bray and J. D. Markel, "A Two-Step Speech Compression With Vector Quantizing," IEEE ICASSP'79, Wachington, pp.52-55, (1979)

13. A. M. Deroualt, "Modélisation d'une langue naturelle pour la désambiguation des chaînes phonetiques," These de Doctorat d'Etat, Univ. Paris VII, (1985)

14. W. Drews, R. Lanois, J. Pandel and A. Sroelzle, "A 1000 Word Speech Recognition System using a Special Purpose CMOS Processor," European Conference on Speech Technology, Edinburgh, pp.218-221, (1987)

15. G. D. Forney, Jr., "The Viterbi Algorithm," Proc. IEEE, vol.61, pp.268-278, (1973)

16. J. L. Gauvain, J. Mariani, "A Method for Connected Word Recognition and Word Spotting on a Microprocessor," IEEE ICASSP, Paris, pp.891-894, (1982)

17. J. L. Bauvain, "A Syllable-Based Isolasted Word Recognition Experiment," IEEE ICASSP, Tokyo, (1986)

18. J. L. Gauvain, "Large Vocabulary Isolated Word Recognition Using the MuPCD Chip," I A/Speech Group-SFA/Speech Communication Group Seminar, Brighton, (1988)

19. M. H. Kuhn, H. H. Tomaschewski, "Improvements in Isolated Word Recognition," IEEE Trans. on ASSP, vol.31, No.1, (1983)

20. R. Kurzweil, "The Kurzweil Voice Writer. A Large Vocabulary Voice Activated Word Processor," Speech Tech '86, New York, pp.184-188, (1986)

21. K. F. Lee, "Speaker-Independent Phone Recognition Using Hidden Markov Models," Report CMU-CS-88-122, (1988)

22. J. Mariani, "Contriburion à la Reconnaissance de la Parole Continue utilisant la notion de Spectre Différentiel," Thèse de Docteur-Ingénieur, Université Paris VI, (1977)

23. J. Mariani, "Reconnaissance de parole continue par diphonèmes," Séminaire du "Groupement des Acousticiens de Langue Francaise:" "processus d'encodage et de decodage phonétique," Toulouse, (1981)

24. J. Mariani, "HAMLET: A Prototype of a Voice-Activated Typewriter," European Conference on Speech Technology, Edinburgh (1987)

25. W. D. Marslen-Wilson, "Speech Understanding as a Psychological Process," in Spoken Language Generation and Understanding, NATO/ASI series, D. Reidel, pp.39-68, (1980)

26. W. S. Meisel, "Implications of Large Vocabulary Recognition," Speech Tech '86, New York, pp.189-192, (1986)

27. B. Merialdo, "Speech Recognition with very large size Dictionary," IEEE ICASSP '87, Dallas, pp.364-367, (1987)

28. D. B. Paul, "Robust HMM Continuous Speech Recognition," DARPA Strategic Computing Speech Recognition Program Meeting, Pittsburgh, (1988)

29. G. Pirani, L. Fissore, A. Martelli, G. Volpi. "Experimental Evaluation of Italian Language Models for Large Dictionary Speech Recognition," European Conference on Speech Technology, Edinburgh, pp.159-162, (1987)

30. B. Proust, "Contribution à la Synthèse de la Parole à partir du Texte: Transcription Graphème-Phonème en temps réel sur Microprocesseur," Thèse de Docteur-Ingénieur, Université Paris XI, (1980)

31. G. Quénot, J. L. Gauvain, J. J. Gangolf and J. Mariani, "A dynamic Time Warp VLSI processor for Continuous Speech Recognition," IEEE ICASSP, Tokyo, (1986)

32. K. Shikano, K. F. Lee, R. Reddy, "Speaker Adaptation through Vector Quantization, " IEEE ICASSP, Tokyo, pp.2643-2646, (1986)

33. F. Simon, "Préclassification pour la reconnaissance de mots isolés," Rapport de DEA, LIMSI, (1985)

34. F. Simon, "preclassification pour la reconnaissance de mots isoles," Rapport de DEA, LIMSI, (1985)

35. C. Vincenzi, C. Favareto, D. Sciarra, A. Carossino, A.M. Colla, C. Scagliola and P. Pedrazzi, "Large Vocabulary Isolated Word Recognition: A Real-Time Implementation," European Conference on Speech Technology, Edinburgh, pp.214-217, (1987)

36. V. Vittorelli, "Linguistic Analysis of the european Languages," ESPRIT '87 Achievements and Impact, North-Holland, pp.1358-1366, (1987)

37. E. Zwicker, R. Feldtkeller, "Psychoacoustique, L'oreille récepteur d'informations," Masson, (1981)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Speech Recognition Using Hidden Markov Models: a CMU Perspective

Kai-Fu Lee, Hsiao-Wuen Hon, Mei-Yuh Hwang and Xuedong Huang

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract

Hidden Markov models (HMMs) have become the predominant approach for speech recognition systems. One example of an HMM-based system is SPHINX, a large-vocabulary, speaker-independent, continuous-speech recognition system developed at CMU. In this paper, we introduce hidden Markov modeling techniques, analyze the reason for their success, and describe some improvements to the standard HMM used is SPHINX.

1. INTRODUCTION

Hidden Markov models (HMMs) have recently become the predominant approach to automatic speech recognition. HMM-based systems make certain structural assumptions, and then try to learn two sets of parameters from training data. The forward-backward learning algorithm adjusts these model parameters so as to improve the probability that the models generated the training data. This seemingly simple technique has worked surprisingly well, and has led to many state-of-the-art systems [1-6].

At Carnegie Mellon, we have utilized the HMM technique and developed a state-of-the-art speech recognition system [5]. This system, SPHINX demonstrated the feasibility of accurate large-vocabulary speaker-independent continuous speech recognition.

In the next section, we will describe the fundamentals of hidden Markov modeling. We will also examine the advantages of HMMs, and explain why they are particularly suitable for modeling time-varying signals such as speech.

In Section 3, we will describe the three key factors for a successful HMM system: plentiful training data, a powerful learning algorithm, and detailed models. We will use SPHINX [5, 7], our large-vocabulary speaker-independent continuous speech recognizer, as an example to illustrate the contributions of each factor.

Finally, we believe that HMMs have not yet realized their full potential, and there are still many unexplored areas that could further advance the state of the art. In Section 4, we will identify some of these areas that we are currently exploring at Carnegie Mellon.

Speech recognition involves a search in a state-space for an optimal, or a near-optimal solution.

2. HIDDEN MARKOV MODELS

2.1. A Brief Introduction to HMMs

Hidden Markov models (HMM) were first described in the classic paper by Baum [8]. Shortly afterwards, they were extended to automatic speech recognition independently at CMU [9] and IBM [10, 11]. It was only in the past few years, however, that HMMs became the predominant approach to speech recognition, superseding dynamic time warping.

A hidden Markov model is a collection of states connected by transitions. Each transition carries two sets of probabilities: a transition probability, which provides the probability for taking this transition, and an output probability density function (pdf), which defines the conditional probability of emitting each output symbol from a finite alphabet, given that that the transition is taken. Figure 1 shows an example of a hidden Markov model with two output symbols, A and B.

There are several types of hidden Markov models. The simplest and most natural one is discrete density HMMs, which are defined by:

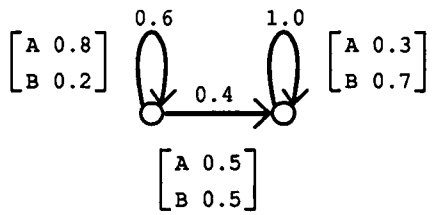


Figure 1. A simple hidden Markov model with two states, and two output symbols, A and B.

- $\{s\}$ — A set of states including an initial state S_I and a final state S_F.
- $\{a_{ij}\}$ — A set of transitions where s_{ij} is the probability of taking a transition from state i to state j.
- $\{b_{ij}(k)\}$ — The output probability matrix: the probability of emitting symbol k when taking a transition from state i to j.

The forward-backward algorithm is used to estimate a and b. We provide only a simplistic sketch here; details of the algorithm can be found in [7, 12]. The forward-backward algorithm adjusts a and b iteratively. For each iteration, the estimates from the previous iteration are used to count how frequently each symbol is observed for each transition, and how frequently each transition is taken from each state. These counts are then normalized into new parameters. Let $c_{ij}(k)$ represent the frequency (or count) that the symbol k is observed, and that the transition from i to j is taken, the new output probability $\bar{b}_{ij}(k)$ is given by the normalized frequency:

$$\bar{b}_{ij}(k) = \frac{c_{ij}(k)}{\sum_{k=1}^{K} c_{ij}(k)} \tag{1}$$

Similarly, transition probabilities are re-estimated by normalizing the frequency that a transition is taken from a particular state:

$$\bar{a}_{ij} = \frac{\sum_{k=1}^{K} c_{ij}(k)}{\sum_{\forall j'} \sum_{k'=1}^{K} c_{ij'}(k')} \tag{2}$$

Baum [8] showed that re-estimating a and b, as shown in equations 1 and 2, will increase the likelihood of generating the training data, unless a local maximum has been reached. Although the forward-backward algorithm guarantees only a local maximum, it efficiently produces an approximation to the maximum-likelihood estimates (MLE) of the HMM parameters.

2.2. Advantages of HMMs for Speech Recognition

In the previous section, we have introduced the basic mechanism of hidden Markov models. It has been, and probably still is, surprising to many that such a simple modeling technique has led to the highest performance systems in almost every speech recognition problem today. In this section, we will try to explain why HMMs have worked so well.

Hidden Markov models have a rich representation in their two sets of parameters. The output probabilities represent the acoustic phenomena. They could be based on either discrete densities, where speech is quantized into sequences of symbols from a finite alphabet (usually through vector quantization). Alternatively, they could be based on continuous mixture densities (usually Gaussian), where speech parameters are directly modeled. In either case, they have the power of modeling any arbitrary probability density function, given sufficient training data. The other set of parameters, the transition probabilities, represent timescale distortions. With a large number of states, duration of very fine phonetic events can be modeled. Yet, with the use of self-loops, the range of durations modeled is very large.

- It requires minimal supervision — only an orthographic transcription of the speech is needed.
- It has a mathematical basis, guaranteeing convergence to a critical point.
- It scales gracefully to increased training, requiring only linearly more computation.

Finally, the joint optimization of the two sets of parameters makes HMMs particularly suitable for modeling of time-varying signals.

Speech recognition involves a search in a state-space for an optimal, or a near-optimal solution. HMM-based searches differ form bottom-up approaches, which propagate errors and cannot integrate top-down knowledge, and from top-down approaches, which are often intractable. It is possible to represent sounds, phonemes, syllables, words, and even grammar states in terms of HMMs. By integrating many knowledge levels into a unifying HMM framework, the HMM search is a global, goal-driven strategy, where all knowledge

sources participate in every decision. Finally, by using a probabilistic framework, we have a consistent scoring mechanism.

In summary, hidden Markov models have a number of very powerful properties. The ability of HMMs to automatically optimize parameters from data is extremely powerful, the HMM integrated search that considers all of the knowledge sources at every step is very effective, and the absorption of faulty structural assumptions is most forgiving. By turning an *unknown structure problem* into an *unknown parameter problem*, and by automatically optimizing these parameters, HMM and maximum likelihood estimation are one of the most powerful learning paradigms available today.

3. THE SPHINX SPEECH RECOGNITION SYSTEM

In the previous section, we have explained why hidden Markov models are particularly suitable for modeling speech. The single greatest advantage of hidden Markov models is the existence of an automatic training algorithm. However, it does not imply that HMMs are completely self-organizing tools. In fact, the literature is full of examples where simple-minded HMMs produce very poor results. For example, the first recognition rate we attained on the 991-word, perplexity 60 task was only 58% [5] (compared to the recent result of 96%). In this section, we will examine a number of enhancements to the HMM paradigm. We will focus on the enhancements and effects on the speaker-independent SPHINX continuous speech recognition system.

We believe that these enhancements can be categorized into:

- Detailed speech models.
- Large training databases.
- Improved learning algorithms.

In the next three sections, we will discuss how these factors contributed to SPHINX.

3.1. Detailed Speech Models

By "detailed speech models," we mean the expansion of the HMM parameters or modification of the HMM structures selectively. Intuitively, both should be helpful to HMMs. Expansion of the parameters should improve performance, assuming sufficient training data are available. Improving the HMM structures should also be helpful, since the HMM learning process assume the correctness of fixed underlying structures.

However, we have found that the amount of improvement greatly depends on careful selection of the *right* parameters to expand, and the *right* structures to tune. In particular, we have found two types of improvements to be the most helpful:

- Detailed models that compensate for the weaknesses of HMMs.
- Detailed models that utilize our speech knowledge.

In this section, we will examine how these enhancements improved the SPHINX Speech Recognition System.

Multiple Codebooks

Discrete hidden Markov models model speech as a sequence of vector quantized symbols. In other words, every frame of speech is reduced to a symbol from a finite alphabet. Typically a single codebook using stationary coefficients (FFT, LPC, etc.) is used. However, it has been shown recently that the use of differential information and power information is extremely important [3]. Moreover, one of the serious problems with HMMs is that they assume that speech events are only dependent on the state, which causes the HMMs to have no vision of the past or the future. The inclusion of differential coefficients gibe HMMs more scope than a small 20msec frame.

One possible approach to incorporate the differential and power coefficients is to use a single monolithic codebook. However, in such a codebook with many dimensions, the VQ distortion is very large. Therefore, we used the multi-codebook approach [14]. Multiple codebooks reduce the VQ distortion by reducing the dimensions of the parameter space. We created three VQ codebooks, each with 256 prototype vectors, using:

1. 12 LPC cepstral coefficients.
2. 12 differential LPC cepstral coefficients.
3. Power and differenced power.

Because we use three VQ codebooks, our discrete HMM must produce three VQ symbols at each time frame. By assuming that the three output pdf's are independent, we can compute the output probability as the product of the three output probabilities. The use of multiple codebooks increased SPHINX's accuracy from 26% to 45% (no grammar), and from 58% to 84% (word pair grammar).

Duration Modeling

For recognition, we have used a Viterbi search [15] that finds the optimal state sequence in a large HMM network. At the highest level, this HMM is a network of word HMMs, arranged according to the grammar. Each word is instantiated with its phonetic pronunciation network, and each phone is instantiated with the corresponding phone model. Beam search [16, 17] is used to reduce the amount of computation.

One problem with HMMs is that they do not enforce any global durational constraints. For example, a 50-state word HMM may have reasonable state durations at all 50 states, but the word duration may be unreasonable. To add this higher-level constraint, we incorporated word duration into SPHINX as a part of the Viterbi search. The duration of a word is modeled by a univariate Gaussian distribution, with the mean and variance estimated from a supervised Viterbi segmentation of the training set. By precomputing the duration score for various durations, this duration model has essentially no overhead. This duration model resulted in a substantial improvement when no grammar is used (45% to 50 %), but not when a grammar is used.

Function Word and Phrase Modeling

One problem with continuous speech is the unclear articulation of function words, such as *a, the, in, of,* etc. Since the set of function words in English is limited and function words occur frequently, it is possible to model each phone in each function word separately. By explicitly modeling the most difficult subvocabulary, recognition rate can be increased substantially. We selected a set of 42 function words, which contained 105 phones. We modeled each of these phones separately.

We have found that function words are hardest to recognize when they occur in clusters, such as *that are in the.* The words are even less clearly articulated, and have strong inter-word coarticulatory effects. In view of this, we created a set of phone models specific to *function phrases,* which are phrases that consist of only function words. We identified 12 such phrases, modified the pronunciations of these phrases according to phonological rules, and modeled the phones in them separately. A few examples of these phrases are: *is the, that are,* and *of the.*

Modeling these frequently of occurring words and phrases increased the number of parameters by a factor of five, and improved SPHINX's accuracies from 50% to 59% (no grammar), and form 85% to 88% (word pair grammar).

Generalized Triphones

The function-word and function-phrase dependent phone models provide better presentations of the function words. However, simple phone models for the non-function words are inadequate, because the realization of phone crucially depends on context. In order to model the most prominent contextual effect, Schwartz, *et al.* [17] proposed the use of triphone models different triphone models is used for each left and right context. While triphone models are sensitive to neighboring phonetic contexts, and have led to good results, there are very large number of them, which can only be sparsely trained. Moreover, they do not take into account the similarity of certain phones in their affect on other phones (such as /b/ and /p/ on vowels).

In view of this, we introduced the *generalized triphone models.* Generalized triphones are created from triphone models using an agglomerative clustering procedure that clustering triphone models together using the following distance metric:

$$D(a,b) = \frac{(\prod_i (P_a(i))^{N_a(i)}) \cdot (\prod_i (P_b(i))^{N_b(i)})}{\prod_i (P_m(i))^{N_m(i)}} \tag{3}$$

where $D(a,b)$ is the distance between two models of the same phone in context a and b. $P_a(i)$ is the output probability of cord word i in model a, and $N_a(i)$ is the count of cord word i in model a. m is the merged model by adding N_a and N_b. This equation measures the ratio between the probability that the individual distributions generated the training data and the probability that the combined distribution generated the training data. Thus, it is consistent with the maximum-likelihood criterion used in the forward-backward algorithm.

This context generation algorithms enables us to empirically determine how many models could be trained given training set. Generalized triphones further increased the number of parameters by a factor of five, and improved the results from 59% to 79% (no grammar), and 88% to 94% (word pair grammar). Details of the context-dependent models used in SPHINX can be found in [5, 18]

Between-Word Coarticulation Modeling

Triphone and generalized triphone models are powerful subword model techniques because they account for the left and right phonetic contexts, which are the principal causes of phonetic variability. However, these phones-sized models consider only intra-word context. A simple extension of triphones to models between-word coarticulation is problematic because the number of triphone model grows sharply when between-word triphones are considered. For example, there are 2381 within-word triphones in our 997-word task. But there are 7057 triphones when between-word triphones are also considered.

Therefore, generalized triphones are particularly suitable for modeling between-word coarticulation. We first generated 7057 triphone models that accounted for both inter-word and inter-word tirphones. These 7057 models were then clustered into 1100 generalized triphone models. Few program modifications were needed for training, since the between-word context is always known. However, during recognition, most words now have multiple initial and final states. Care must be taken to to ensure that each legal sentence has one and only one path in the search. Details of our implementation can be found in [19]. The use of between-word coarticulation did not increase the number of parameters, since we felt that we could not reliably trained any more parameters using our training database. Yet, SPHINX's accuracies were improved from 73% to 78% (no grammar), and 94% to 95.5% (word pair grammar).

3.2. Large Training Database

A large database of 4200 sentences from 105 speakers were used to train SPHINX. Although this database is crucial to the success of SPINX, it is more important to derive a system configuration that has enough parameters to model the variabilities in the data, but not too many parameters that we cannot reliably estimate.For example, if we fix the system configuration and reduce the training data, results deteriorate much faster than if we use a configuration that is data-dependent (such as using fewer generalized triphones). This phenomenon is clearly demonstrated in Figure 2.

Therefore, while HMM system benefit greatly from increased training, it is inadequate to simply increase the training data. Instead, data-dependent system configuration is needed to optimize the performance. We have described some of these techniques in the previous section, and we will outline our future work in Subsection 4.1 and 4.2.

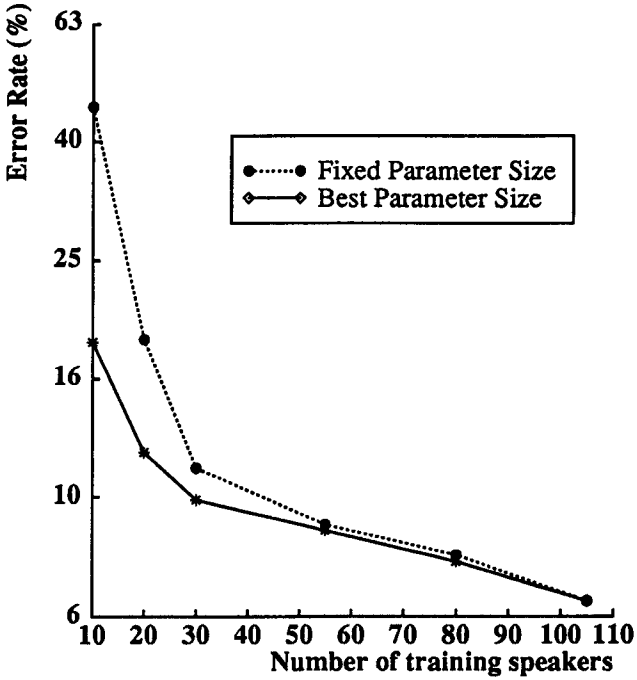


Figure 2. SPINX accuracies with different number of training speakers and parameters

3.3. Better Learning Algorithms

Corrective Training

Hidden Markov models with maximum-likelihood estimation (MLE) constitute the predominant approach to automatic speech recognition today. Indeed, the forward-backward algorithm is responsible for the success of SPHINX and many other systems. However, one of the problems with MLE is that it may produce inferior results when the underlying models are incorrect, which HMM's obviously are as models of *real*speech. Thus, alternate training algorithms that do not suffer from this problem may be desirable. We have only experimented with one variant — corrective training, which will be described below.

Bahl *et al.* [20] introduced the corrective training algorithm for HMMs as an alternative to the forward-backward algorithm. While the forward-backward algorithm attempts to increase the probability that the models generated the training data, corrective training attempts to maximize the recognition rate on the training data. This goal has a definite practical appeal, since error rate, not sentence likelihood, is the bottom line for speech recognition. This algorithm has two components: (1) *error-correction learning* — which improves correct words and suppresses misrecognized words, (2) *reinforcement learning* — which improves correct words and suppresses near-misses.

We extended this corrective training algorithm to speaker-independent continuous

speech recognition. We used a large training database and a cross-validated training procedure that fully made use of the training material. More importantly, we proposed a reinforcement learning method that hypothesized *near-miss sentences* by first formulating a list of *near-miss phrases* using a DTW algorithm, and then creating the near-miss sentences by substituting phrases from the near-miss phrase list.

Using this training algorithm, we were able to further improve our results from 78% to 82% (no grammar), and 95.5% to 96.2% (word-pair grammar). More details about this work are described in the proceeding [21].

Semi-Continuous Modeling

The methodology described previously are based on discrete HMMs. Another type of HMMs is the continuous density HMM [22-24], which models the acoustic observation directly by estimating continuous probability density functions without VQ. For speaker-independent speech recognition, a mixture of a large number of probability density functions [4, 25] is generally required to model the characteristics of different speakers. However, such mixtures of a large number of probability density functions considerably increase not only the computational complexity, but also the number of free parameters that can be reliably estimated. In addition, continuous density HMMs are more fragile, in that inaccurate assumptions about the parameter distribution will lead to poor results [24, 26]. Yet standard continuous mixture density HMMs have a large number of parameters, which gives us the unpleasant choice of simple but incorrect assumptions, or reasonable assumptions but untrainable parameters.

On the other hand, the semi-continuous hidden Markov model (SCHMM) [27, 28] is a general model that includes both the discrete and the continuous HMM as its special forms, which provides a way to unify both acoustic (VQ) and phonetic (HMM) sources. The SCHMM was motivated by the fact that each VQ codeword can be represented by a continuous probability density function, and these continuous probability density functions can be unified with the HMM. The semi-continuous output probability is a combination of model-dependent weighting coefficients (discrete output probability of VQ codeword) with continuous model-independent codebook probability density functions (probability distribution of the codeword that generates the observed coefficients) [26, 29]. The semi-continuous output probability can be used to reestimate the HMM parameters together with the VQ codebook. The feedback from HMM estimates to the VQ codebook implies that the VQ codebook is optimized based on the HMM likelihood maximization rather than minimizing the total distortion errors from the set of training data. This feedback provides a way to unify both acoustic (VQ) and phonetic (HMM) sources. Under such a unified framework, the VQ codebook could be iteratively adjusted according to HMM parameters that are closely associated with phonetic information (HMM); and the HMM parameters could be, in return, iteratively updated based on the acoustic-related VQ codebook.

In comparison to the conventional continuous mixture HMM, the SCHMM maintains the modeling ability of large-mixture probability density functions. In addition, the number of free parameters and the computational complexity is reduced because all of the probability density functions are tied together in the codebook. The SCHMM thus pro-

vides a good solution to the conflict between detailed acoustic modeling and insufficient training data. In comparison to the conventional discrete HMM, robustness is enhanced bye using multiple codewords in deriving the semi-continuous output probability; and the VQ codebook itself is optimized together with the HMM parameters in terms of the maximum likelihood criterion. Such a unified modeling can substantially minimize the information lost in conventional VQ [30]. With the SCHMM, we were able to reduce the error rate (word-pair grammar) of SPHINX by 15-20%.

3.4. The SPHINX System

By discussing various methods of generating detailed HMMs, we have uncovered most of the SPHINX System. SPHINX is trained by first using a set of context-independent models. Next, the context-dependent (function word/phrase dependent, generalized triphone, and between-word triphone) models are trained. Then, the parameters are smoothed using deleted interpolation [31], and corrective training is applied to improve discrimination. This training procedure is shown in Figure 3.

The SPHINX System was trained on about 105 speakers and 4200 sentences. It was tested on 150 sentences from 15 speakers. These sentences were the official DARPA test data for evaluations in March and October 1987. The word accuracies for various versions of SPHINX with the word-pair grammar (perplexity 60) and the null grammar (perplexity 997) are shown in Table 1. Here, word accuracy is defined as the percent of words correct minus the percent of insertions. These results do not include semi-continuous models, which are being incorporated at the time of this writing.

4. FUTURE DIRECTIONS FOR SPHINX AND HMM

While HMM-based systems have achieved record performance in many applications, they are still substantially worse than humans. The natural question to ask is: will HMMs approach human performance, or have they been pushed to the limit? We believe hidden Markov modeling is a powerful approach that has not yet realized its full potential. Our previous experience has indicated that HMMs benefited from: detailed speech models, large training databases, and powerful learning algorithms. We feel that many promising improvements lie ahead in all three areas. In this section, we will describe some of the promising areas that we are pursuing, and hoping to incorporate in the next generation of SPHINX-like HMM-based recognizers.

4.1. Detailed Speech Models

Subword Modeling

In the forseeable future, we expect to continue to use context-dependent phonetic models. Currently, context-dependency includes only left context, right context, and word boundary. In practice, there are many other causes of phonetic variability, which can be classified into three categories:

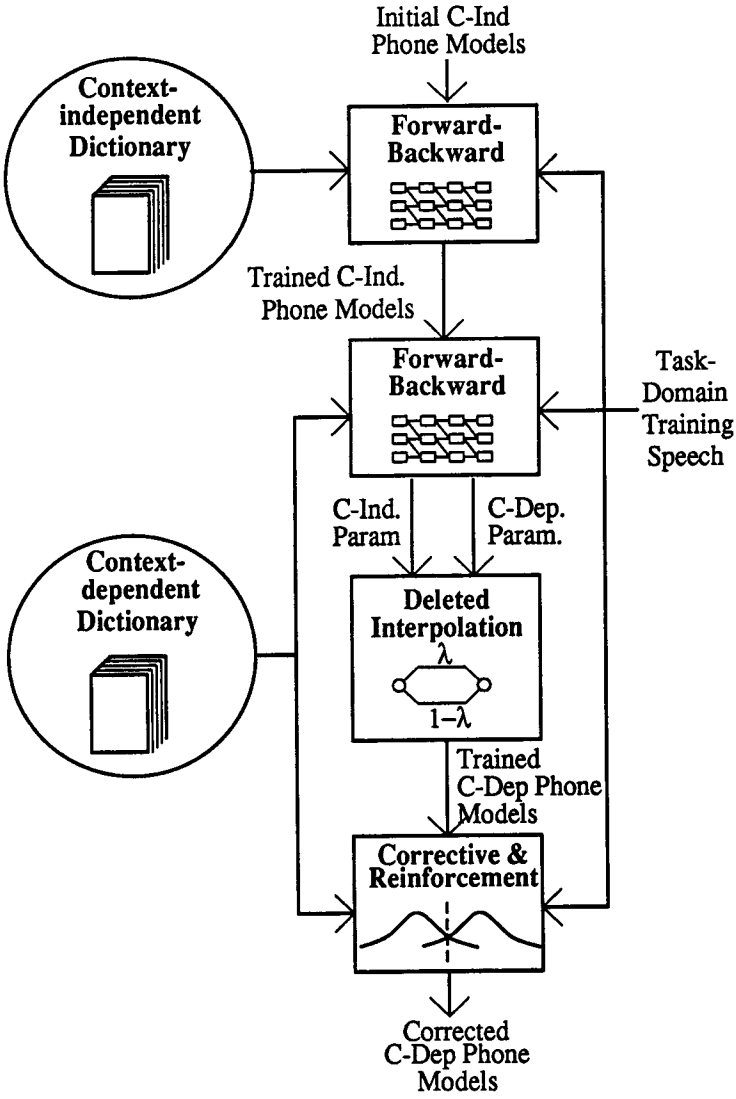


Figure 3. The SPHINX Training Procedure.

Table 1. Results of various versions of SPHINX

Version	No Grammar	Word Pair
1 Codebook	25.8%	58.1%
3 Codebook	45.3%	84.4%
+Duration	49.6%	83.8%
+Fn-word	57.0%	87.9%
+Fn-phrase	59.2%	88.4%
+Gen-triphone	72.8%	94.2%
+Between-word	77.9%	95.5%
+Corrective	81.9%	96.2%

- Articulation-related variabilities are caused by the fact that our articulators cannot move instantaneously, and that the locations and movements of the articulators affect the realization of a phone.
- Language-related variabilities are caused by attributes of specific languages.
- Speaker-related variabilities result from differences in anatomical features.

These three sources of variability modify various attributes and qualities of each phone, and we call the resultant of this transformation an *allophone*. The factors affecting the process of transforming a phone to an allophone is illustrated in Figure 4.

First, we will use our speech knowledge to identify and model only the most relevant contexts. For example, in order to collect a large database for this training, we will not model speaker-related variabilities. We will model immediate phonetic context, word/syllable boundary, and stress. We will also selectively model other contextual effects, such as non-neighboring phonetic contexts *only when they are relevant*. This strategy allows us to reduce an astronomical number of models into a more reasonable number (about 50,000).

50,000 subword models are still two orders of magnitude more than our current system can learn. Since we do not expect to have two orders of magnitude more training data in the near future, we must further reduce these models to a more manageable level. One possibility is to extend the notion of bottom-up subword clustering as used in generalized triphones [5, 18]. We call this set of phonetic models *generalized allophones*. The clustering procedure will combine allophone models in order to maximize the probability that they generated the training data. The precise number of generalized allophones is data-dependent, and can be determined empirically.

The bottom-up subword clustering process finds a good mapping for each of the 50,000 allophones. However, if a context is not covered by these allophones, the context-independent phone model must be used instead, which will lead to substantially degraded performance. In other words, bottom-up clustering does not facilitate generalization; therefore, its utility will be determined by the allophonic coverage in the training data.

Another approach that sacrifices some optimality to improve generalization is the use of decision trees [32-34] to cluster subword models. At the root of the decision tree is the

set of all allophones corresponding to a phone. The algorithm incrementally splits nodes in the tree by asking "questions." These questions might be general ones like "is the previous phone a front vowel," or specific ones like "is the next phone in the set /p,t,k/ or the set /b,d,g/." These questions will lead to a set of leaf nodes, which represent the contextual units to be used. This type of top-down subword clustering has two important advantages. First, if a new allophone is encountered, we might still be able to reach a leaf node, if all questions are sufficiently general. Even if unanswerable question is encountered at an internal node, we can still use that node as a subword unit, which should be much more appropriate than backing off the context-independent phone.

Second, since a child node must be somewhat similar to its parent node, we can improve trainability by interpolating each node with all of its ancestor nodes. One disadvantage of the that it improves recognition accuracy slightly [35].

We hope that by making subword models more consistent and detailed, we will not only improve performance, but also have models that are more *vocabulary-independent.* This issue is discussed in more detail in [36].

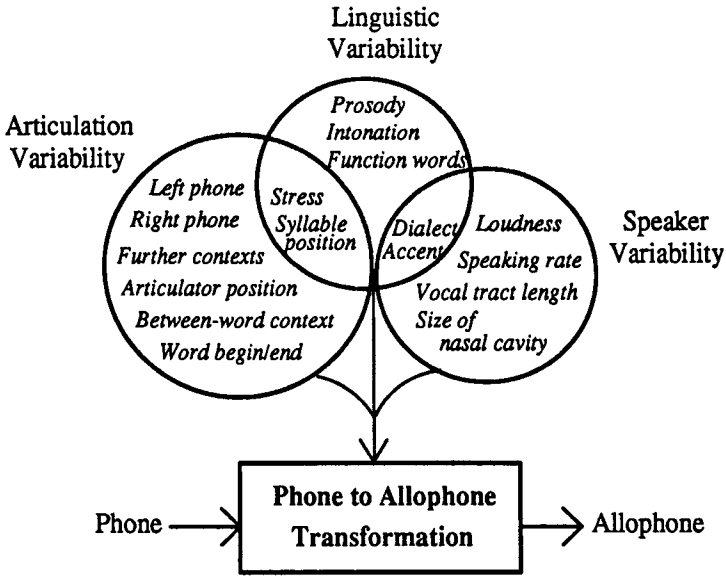


Figure 4. Sources of variability that affect the realization of a phone.

4.2. Large Training Database

It is well known that a fixed statistical learning system will improve with additional training, until all the parameters are well-trained. Thus, one might conclude that additional training will help a speech recognizer only until an asymptote is reached. However, this statement is only true for a recognizer with *fixed structures and parameter size.* In reality, we are free to improve the structures and to increase the number of parameters of

a recognition system, and we have seen that in so doing, the recognizer performance can be improved substantially.

In the previous sections, we have presented some ideas on how to make phonetic models more trainable with more data. However, we aimed our research given the amount of training that is likely to become available in the next few years, or about tens of thousands of sentences. In the future, with the use of computers that have voice capabilities, data collection will become much easier. In a decade from now, we expect to see several orders of magnitude more training data. These data can be utilized to further refine the speech models. Table 2 shows the types of models that might be trainable with these future speech databases.

Table 2. Types of models that might be trainable as the number of training sentences is increased in a speaker-independent database.

Number of Sentences	Type of Models
> 100	Phonetic models.
$> 1,000$	Phonetic models with simple contexts (e.g.triphones).
$> 10,000$	Phone models with more context (e.g.stress, syllable/word position).
$> 100,000$	Longer (e.g.syllable) models; Rough speaker cluster (e.g. gender)models.
$> 1,000,000$	Even longer (e.g.morph, word) models; Detailed speaker cluster (e.g. dialect) models.

4.3. Better Learning Algorithms

In our future research, we would like to improve HMM learning in three directions: (1) a more integrated learning framework, (2) use of discriminant learning, and (3) speaker adaptive learning.

One of the main advantages of HMMs is the integrated approach, where output probabilities and transition probabilities for *all* units are learned to improve a global measure. However, if we examine systems like SPHINX, there are at least two areas that are detached, and learning does not take place. First, the vector quantization process is a preprocessing stage that uses a distance metric not related to the MLE criterion. As introduced in the previous section, the SCHMM has been used to rectify this problem. One issue that remains to be resolved with the SCHMM (or any continuous density HMM) is the model-assumptions problem. We hope to find models that are self-organizing, and not depend on the correctness of model assumptions. The second detached element is the pronunciation dictionary, which maps words to phone sequences. This dictionary is created from phonetic knowledge alone *. The integration of dictionary learning with

*Although we map the phones to generalized triphones using a consistent distance metric, the original phone sequences are unrelated to the global optimization

HMM learning should lead to further improvements. We believe that unified modeling of acoustic and phonetic sources requires further exploration.

The second area of research involves the incorporation of discrimination in the HMMs. To that end, we have used the corrective training algorithm [21]. Many other promising techniques, such as maximum mutual information estimation [24], and linear discriminants [6] have been introduced. We are beginning to investigate the incorporation of above techniques into SPHINX. We are also investigating the possibility of integrating HMMs with neural networks. A preliminary study of that has been reported in [37].

Finally, we have only addressed the issue of HMM learning when presented with a large amount of multi-speaker training data for a one-time training process. In reality, few applications require *true* speaker-independence. There are usually opportunities to adapt on a small number of utterances from each speaker. Our previous work on speaker adaptation [5, 38] has only led to modest error reductions (about 10%) with substantial adaptation (30 sentences). In the future, we must explore alternative approached that can *incrementally* adapt more accurately on less data.

5. CONCLUSION

In this paper, we have presented the hidden Markov model methodology, and describe SPHINX, a large-vocabulary, speaker-independent, continuous speech recognition system. We discussed the key issues in designing SPHINX, and outlined areas of future research.

We believe that hidden Markov models have benefited greatly from the use of detailed subword models, large training databases, and powerful learning techniques. Further, we believe that HMMs have not yet realized their full potential, and that by expanding in each of the three areas, more advances are yet to come.

Acknowledgments

The authors wish to thank the CMU Speech Group for their support and contributions. This research was sponsored by Defense Advanced Research Projects Agency Contract N00039-85-C-0163

References

1. Averbuch, et al.: "An IBM PC Based Large-Vocabulary Isolated-Utterance Speech Recognizer," *Proc. ICASSP-86*, pp.53-56, (1986)

2. D. B. Paul, R. P. Lippmann, Y. Y. Chen, C. Weinstein: "Robust HMM-Based Techniques for Recognition of Speech Produced under Stress and in Noise," *Proc. the Speech Technology Conference*, (1986)

3. Y. L. Chow, R. Schwartz, S. Roucos, O. Kimball, P. Price, Kubala, F., Dunham, M., Krasner, M., Makhoul, J.: "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-Based Speech Recognition System," *Proc. ICASSP-86*, (1986)

4. L. R. Rabiner, J. G. Wilpon, F. K. Soong: "High Performance Connected Digit Recognition Using Hidden Markov Models," *Proc. ICASSP-88*, (1988)

5. K. F. Lee: *Automatic Speech Recognition: The development of the SPHINX System*, Kluwer Academic Publishers, Boston, (1989)

6. G. R. Doddington: "Phonetically Sensitive Discriminants for Improved Speech Recognition," *Proc. ICASSP-89*, (1989)

7. K. F. Lee: *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX SYSTEM*, PhD dissertation, Computer Science Department, Carnegie Mellon University, (1988)

8. L. E. Baum: "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes," *Inequalities*, vol. 3, pp. 1-8, (1972)

9. J. K. Baker: "The DRAGON System — An Overview," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-23, No. 1, pp.24-29, (1975)

10. R. Bakis: "Continuous Speech Recognition via Centisecond Acoustic States," *91st Meeting of the Acoustical Society of America*, (1976)

11. F. Jelinek: "Continuous Speech Recognition by Statistical Methods," *Proc. the IEEE*, vol. 64, No. 4, pp.532-556, (1976)

12. L. T. Bahl, F. Jelinek, R. Mercer: "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, No. 2, pp.179-190, (1983)

13. S. Furui: "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans. on Acoustics, Speech and Signal Process.*, vol. ASSP-34, No. 1, pp.697-700, (1987)

14. V. N. Gupta, M. Lenning and P. Mermelstein: "Integration of Acoustic Information in a Large Vocabulary Word Recognizer," *Proc. ICASSP-87*, pp.697-700, (1987)

15. A. J. Viterbi: "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Trans. on Information Theory*, vol. IT-13, No. 2, April 1967, pp.260-269.

16. B. T. Lowerre: *The HARPY Speech Recognition System*, PhD dissertation, Computer Science Department, Carnegie Mellon University, (1976)

17. R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, J. Makhoul: "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech," *Proc. ICASSP-85*, (1985)

18. K. F. Lee: "Context-Dependent Phonetic Hidden Markov Models for Continuous Speech Recognition," *Proc. ICASSP-90*, (1990)

19. M. Y. Hwang, H. W. Hon, K. F. Lee: "Modeling Between-Word Coarticulation in Continuous Speech Recognition," *Proc. Eurospeech*, (1989)

20. L. R. Bahl, P. F. Brown, P. V. De Souza, R. L. Mercer: "A New Algorithm for the Estimation of Hidden Markov Model Parameters," *Proc. ICASSP-88*, (1988)

21. K. F. Lee, S. Mahajan: "Corrective and Reinforcement Learning for Speaker-Independent Continuous Speech Recognition," *Proc. Eurospeech*, (1989)

22. L. R. Rabiner, B. H. Juang, S. E. Levinson, M. M. Sondhi: "Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities," *AT&T Technical Journal*, vol. 64, No. 6, pp.1211-33, (1985)

23. A. B. Poritz, A. G. Richter: "On Hidden Markov Models in Isolated Word Recognition," *Proc. ICASSP-86*, (1986)

24. P. Brown: *The Acoustic-Modeling Problem in Automatic Speech Recognition*, PhD dissertation, Computer Science Department, Carnegie Mellon University, (1987)

25. D. B. Paul: "The Lincoln Robust Continuous Speech Recognizer," *IEEE*, pp.449-452, (1989)

26. X. D. Huang, Y. Ariki and M. A. Jack: *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, (1990)

27. X. D. Huang, M. A. Jack: "Semi-Continuous Hidden Markov Models with Maximum Likelihood Vector Quantization," *IEEE Workshop on Speech Recognition*, (1988)

28. J. R. Bellegarda and D. Nahamoo: "Tied Mixture Continuous Parameter Models for Large Vocabulary Isolated Speech Recognition," *Proc. ICASSP-89*, pp.13-16, (1989)

29. X. D. Huang, H. W. Hon, K. F. Lee: "Speaker-Independent Continuous Speech Recognition with Continuous and Semi-Continuous Hidden Markov Models," *Proc. Eurospeech*, (1989)

30. X. D. Huang and M. A. Jack: "Semi-Continuous Hidden Markov Models for Speech Recognition," *Computer Speech and Language*, vol. 3, No. 3, pp.239-252, (1989)

31. F. Jelinek and R. L. Mercer: "Interpolated Estimation of Markov Source Parameters from Sparse Data," in *Pattern Recognition in Practice*, E. S. Gelsema and L. N. Kanal, ed., North-Holland Publishing Company, Amsterdam, the Netherlands, 1980, pp.381-397.

32. L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone: *Classification and Regression Trees*, Wadsworth, Inc., Belmont, CA., (1984)

33. S. Sagayama: "Phoneme Environment Clustering for Speech Recognition," *Proc. ICASSP-89*, (1989)

34. L. R. Bahl, et. al: "Large Vocabulary Natural Language Continuous Speech Recognition," *Proc. ICASSP-89*, (1989)

35. K. F. Lee, S. Hayamizu, H. W. Hon, C. Huang, J. Swartz, R. Weide: "Allophone Clustering for Continuous Speech Recognition," *Proc. ICASSP-90*, (1990)

36. H. W. Hon, K. F. Lee and R. Weide: "Towards Speech Recognition Without Vocabulary-Specific Training," *Proc. Eurospeech*, (1989)

37. M. Franzini, M. Witbrock and K. F. Lee: "A Connectionist Approach to Continuous Speech Recognition," *Proc. ICASSP-89* , (1989)

38. X. D. Huang, K. F. Lee and H. W. Hon: "On Semi-Continuous Hidden Markov Modeling," *Proc. ICASSP-90*, (1990)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Phonetic Features and Lexical Access

Kenneth N. Stevens

Research Laboratory of Electronics, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Abstract

In the past one or two decades, there have been significant advances in our understanding of the acoustic properties that distinguish one class of speech sounds from another. These advances have arisen in part because of a better grasp of acoustic mechanisms of speech production and in part because of new findings in the areas of auditory and speech perception and auditory physiology. Thus, for example, a few years ago relatively gross methods of analysis of stop and nasal consonants were used, and these led to the development of algorithms yielding errors in identification of place of articulation for stop and nasal consonants (in consonant-vowel syllables) of about 10-15 percent (Blumstein and Stevens, 1979; Kewley-Port, 1983). More refined procedures based on time-varying changes with finer time resolution in the vicinity of the consonantal release have recently been used in some pilot work, and indicate significant improvement in these scores. Similar refinements in procedures for measuring voicing for fricatives, the distinction between sonorant and nonsonorant consonants, and other properties have also been developed. As a consequence of this research, we are approaching the point where, on the basis of speaker-independent properties extracted from the speech signal, we are able to specify most of the relevant aspects of the aerodynamic and articulatory processes that produced speech sounds occurring in simple utterances. Independence of speaker in the specification of the properties is achieved for the most part by defining relational properties that are minimally dependent on a speaker vocal-tract size, laryngeal characteristics, and speaking rate.

Recent years have also seen advances in our understanding of the distinctive features or of other attributes that can be used to represent words in the lexicon. We have a clearer idea of the acoustic correlates of the features, the relations between the features, and possible structures for the representation of lexical units in terms of features. (See, for example, Clements, 1985; Sagey, 1986.) An important advantage of a feature-based representation over a segment-based representation is that it provides a notation for capturing in a simple manner many kinds of variability that occur in speech. Thus, for example, the sequence of words did you might be spoken as either [didyu] or [diju]. A representation of these two versions in terms of features shows that a change from the [dy] sequence to the affricate [j] keeps most features intact, the only difference being in two features (a

shift from [+ anterior] to [- anterior] and a shift from [+ sonorant] to [- sonorant], with a redundant introduction of stridency). (A more detailed discussion of an inventory of features is given below in Section 2.)

The study of the distinctive features and their acoustic correlates has led to a modification in the way we analyze and interpret the stream of sound that constitutes the acoustic manifestation of an utterance. A conventional approach to this analysis has been to attempt to segment the signal into stretches of sound and to assign labels to these pieces of sound. In the modified approach, particular types of events or landmarks are identified in the signal, and acoustic properties of the sound in the vicinity of these landmarks are detected. These acoustic properties are correlates of the distinctive features in terms of which the lexical items are represented. The modified approach to extracting acoustic data from the acoustic signal is event- orientated rather than segment-orientated.

Although researchers are developing a clearer understanding of the distinctive features and their acoustic correlates, the variability in the acoustic manifestation of words remains a significant stumbling block in developing speaker-independent continuous speech recognition systems and, indeed, to modeling the process of human word recognition. This variability in the acoustic pattern of a word arises from several sources, particularly the structural position of the word in a sentence, and the immediate phonetic context in which the word appears.

We are still very far from a quantitative theory that explains the transformations that occur in the acoustic properties and in their timing when a word is produced in different contexts within a sentence. Thus, although we might be able to extract appropriate acoustic properties from the speech stream to tell us how the sound sequence was produced, this pattern of acoustic properties and hence our inferences about the gestures used to generate the sound may deviate from the expected canonical patterns for the word. For example, the canonical lexical representation for the word did indicates that certain acoustic properties should appear near the end of the word indicating the manner and places of articulation of the stop consonant, but these properties are often not in fact present in sequences like *did you or did it come* (where [d] is often manifested as a flap without all the properties of an alveolar stop consonant). Or, to give another example, the intervocalic consonant in the word *legal* is presumably represented in the lexicon as a stop, whereas in fluent speech it often surfaces as a velar fricative. Until a theory that accounts for these types of variability is developed, attempts at speaker-independent speech recognition must rely on statistical methods involving training with large numbers of utterances. These methods are necessarily limited by the fact that, at least on the surface, the changes that occur are so variable and so pervasive that sufficient training data are difficult if not impossible to collect. On the other hand, closer analysis suggests that the changes are sufficiently regular that a non-statistical approach may be called for.

1. FEATURES AND LEXICAL REPRESENTATIONS

The approach we propose to follow in developing a framework for lexical representation and lexical access is based on a representation of lexical items in terms of features. A principal reason for using a feature-based representation is that, as we have indicated above, phenomena of assimilation and lenition can often be described as modification or

spreading of a limited number of features. Another reason is that appropriately defined acoustic properties that are observable in the speech signal bear a rather direct relation to the features that specify lexical items. Other evidence for the role of features in lexical representation comes from experimental data on speech production, speech perception, and language learning.

1.1. Inventory of Features

A list of features that are agreed upon by at least some phonologists is given in Table 1. A complete list should probably contain three or four additional features, and there may be some disagreement as to what these additional features are. The inventory of features in Table 1 is similar to that proposed in 1968 by Chomsky and Halle, which in turn is a modification of the features originally described by Jakobson, Fant, and Halle (1952). Some of the Chomsky-Halle features were originally defined in terms of articulatory attributes, although the requirement that the features have acoustic or perceptual correlates was always assumed by Chomsky and Halle. In recent years there has been a continuing effort to develop acoustic-perceptual as well as articulatory correlates of the features. This effort has led to some modification of some of the features originally proposed by Chomsky and Halle. Phonological considerations have also led to some adjustments of the feature inventory.

Thus, for example, our current view is that the feature [+ sonorant] should be defined in such a way that it is redundantly [+ voice], since this definition can lead to a well-defined and perceptually more reasonable acoustic correlate of the sonorant feature. Another modification of the Chomsky-Halle features involves describing the laryngeal configuration. We have included the feature *consonantal* on the list, even though the value of the consonantal feature appear to be predictable from other features. (The role of this feature will be discussed later.) However, as our work proceeds, we are prepared to delete this feature if it does not perform a useful function. We expect that there will continue to be some adjustments of the feature inventory, but that, for the most part, the features given in Table 1 will not undergo significant revision.

The features in Table 1 are organized into two sublists depending on the way they are implemented. The features listed in the left-hand column are identified as being represented in the sound when the vocal tract is relatively unconstricted and the acoustic source that gives rise to the generation of sound is at the glottis. The spectrum of the sound that is generated during time intervals when the voice tract is relatively unconstricted is characterized by several prominent peaks or formants, particularly in the midfrequency range 700 to 3000 Hz. An additional spectral maximum (or additional maxima) will occur at low frequencies, with a degree of prominence that depends on the glottal and velopharyngeal configuration.

In contrast to the features on the left of Table 1, the acoustic manifestation of the features in the right-hand column occur when the vocal tract is relatively constricted at some point along its length. The source of sound may be at the glottis or it may be in the vicinity of the constriction. The acoustic and articulatory correlates of these features are discussed in other publications (for example, Fant, 1973; Stevens, 1980; Stevens, 1983).

Table 1. List of distinctive features to be used as a starting-point in proposed research. The features in the left-hand column are implemented in the sound when the vocal tract is relatively open, and the features in the right-hand column are implemented when the vocal tract is relatively constricted.

open vocal tract	constricted vocal tract
high	continuant
low	sonorant
back	strident
round	voiced
tense	consonantal
nasal	labial
breathy voice (spread glottis)	coronal
pressed voice (constricted glottis)	anterior
high pitch (stiff vocal cords)	velar
low pitch (slack vocal cords)	lateral
	retroflex
	distributed

1.2. Lexical Representation as a Matrix of Features

A conventional way of specifying a lexical item is in terms of a matrix of features, as shown in Table 2 for the word *pawn*. In this representation, it is assumed that the lexical unit is specified as a sequence of segments, and that each segment is characterized by a bundle of features. (Only a partial list is given in Table 2.) At this level of representation, the features are assumed to be binary. A change in one feature (such as changing from [+ coronal] to [- coronal] in the third column of Table 2) can potentially change the representation to that of a different lexical item (in this case, the word *palm*). In this example, we have specified a value for every feature in each column, although not every feature is distinctive; that is, there are some features which, when changed, could not specify a new lexical item. For example, changing [+ anterior] to [- anterior] in the third column (under [n]) has no meaning, since palatalized nasals cannot occur in this position in English. Furthermore, there are some feature combinations which are not allowed due to the inherent properties of the vocal tract. An example is that a nasal consonant cannot be [+ strident].

The convention of organizing the features into bundles of the type shown in Table 2 is an abstraction that does not capture directly the way in which these features are realized

Table 2. A conventional partial lexical representation for the English word pawn, in terms of segments and features.

	p	ɔ	n
high	−	−	−
low	−	+	−
back	−	+	−
nasal	−	−	+
spread glottis	+	−	−
sonorant	−	+	+
voiced	−	+	+
strident	−	−	−
coronal	−	−	+
anterior	+	−	+
continuant	−	+	−

in the sequence of articulatory gestures or in the stream of sound. There are two types of constraints that are imposed on the manner in which the features are implemented: (1) the nature of certain features leads to a requirement that some groups of features be implemented in the sound wave more or less together but that other groups of features within a column of Table 2 may be implemented at different times; (2) there are phonotactic and other constraints governing the sequencing of groups of features.

2. MODIFIED FRAMEWORK FOR LEXICAL REPRESENTATION

A conventional way of specifying lexical items is in terms of a sequence of phonetic units, or, more precisely, in terms of a matrix of features, much like the representation in Table 2. This way of representing the lexicon, however, bears a complex relation to attributes that are observable in the acoustic signal. There are several reasons for the lack of a direct correspondence between acoustic attributes and the abstract phonemic or feature-matrix representation. One reason stems from the fact that the acoustic properties that signal the presence of features are not synchronous in time. That is, the acoustic properties associated with some features are not aligned with the acoustic properties for other features that belong to the same segment. A second reason is that particular implementations of a word may involve changes in some of the features, and hence changes

in the acoustic properties, depending on the context. A third source for the lack of direct correspondence between sound and feature or phoneme is that the lexical representation is inherently binary or quanta, whereas acoustic properties that are correlates of the features can often be present with varying degrees of strength. Still another reason is a consequence of the potential redundancy in the lexical representation, particularly if it includes specifications for all (or almost all) features of each segment. In view of this redundancy, it is to be expected that greater importance may be attached to some features than to others, and thus that the acoustic manifestations of different features may be permitted lesser or greater degrees of variability.

These and other reasons suggest two approaches that might be followed in developing procedures for accessing lexical items that are represented in a form similar to that in Table 2. These two approaches are schematized in Fig. 1. One method (Fig. 1a) requires that the output of the initial acoustic analysis be modified or transformed in a way that takes context-dependent variation into account, to yield a representation that is in a phonemic form similar to that in Table 2. This phoneme-like representation is then matched against the stored lexical items.

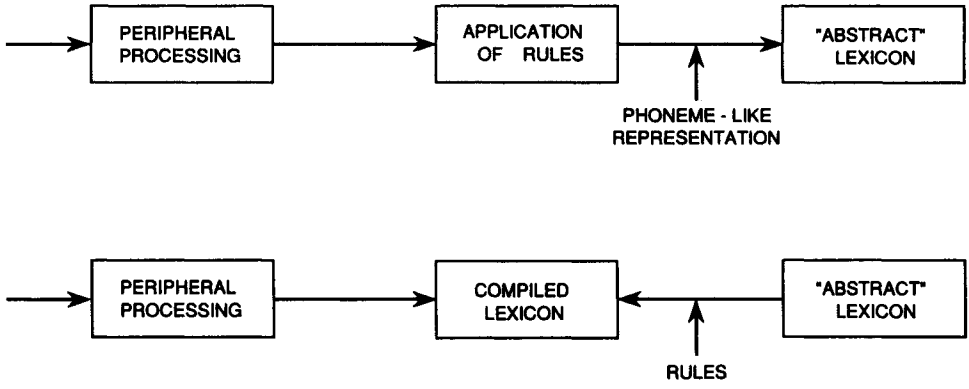


Figure 1. Schematization of two approaches to lexical access. (a) Rules operate to modify the peripheral representation to yield a representation of segment-like units that are used to access the lexicon. (b) Rules operate on the abstract lexical representation to yield a "compiled" lexicon, and properties extracted through peripheral processing are matched against features in the compiled lexicon. In the case of the model in (a), the lexicon is accessed at the "abstract" level, whereas in (b), access is achieved at the level of the compiled lexicon.

A second approach (Fig. 1b) involves compiling from the abstract lexicon a representation of each lexical item to yield a form that is closer to the acoustic representation that is the more abstract representation, and hence is more amenable to direct comparison between acoustics and lexicon. One version of this approach would represent the compiled lexical items in terms of sequences of spectral patterns (or of auditory repre-

sentations of these patterns), similar to the proposal of Klatt called Lexical Access From Spectra, or LAFS (Klatt, 1979). A modification of the LAFS approach would preserve in the compiled lexical item the feature-like aspects of the abstract lexicon, but would relax the requirements of time-synchrony of features and would provide some indication of the relative importance of different features (Stevens, 1986). Access to the lexicon would be achieved by extracting from the signal a set of acoustic properties that are correlates of the features, and matching these properties against the features in the compiled lexicon. Lexical access using such a modified feature-based lexicon has been called Lexical Access From Features, or LAFF (Klatt, in press). The proposal of Browman and Goldstein (in press) to represent lexical items in terms of a gestural score has attributes that are similar to the expanded feature-based representation in LAFF, although many details of the two representations are quite different.

The thrust of the proposal research is based on the modified approach in Fig. 1b, in which representations of lexical items in terms of features are compiled from the more abstract lexicon.

2.1. A Framework of Landmarks to Form the Skeleton of the Lexical Representation

Our current thinking regarding the form of the compiled lexicon and how it is accessed is based on a view of the production of speech and its acoustic manifestation as the generation of a sequence of acoustic landmarks or events or regions. These events are of several kinds, and they are identified on the basis of particular acoustic properties in the sound. Each of the events identifies a region of the acoustic signal around which certain additional acoustic properties are sampled to provide measures of the strength of particular features. The compiled lexicon contains pointers that indicate the relative positions of these landmarks in time, and specifies the features that are associated with the landmarks.

We postulate that there are four kinds of events in the speech stream that are salient, and that provide a skeleton for constructing a lexical representation. Two of these are (1) events that occur when a complete closure of the vocal tract is formed or released, and (2) events concomitant with the buildup or release of pressure above the vocal folds. The first is associated with the feature [- continuant], and the second is associated with a change from [+ sonorant] to [- sonorant] or vice versa. These two kinds of acoustic events can be detected in the sound more or less independently of the presence of other features, although the detectability of the events may be influenced by features that are being implemented concomitantly. The features continuant and sonorant, therefore, have a special status in the lexical representation and in the acoustic speech stream. Most approaches to labeling of databases identify these landmarks in the signal (Leung and Zue, 1984; Glass and Zue, 1988).

A third primary or acoustic event or landmark is (3) the occurrence of a sequence of one or more nonsyllabic segments, i.e., segments involving an articulatory movement that leads to a relatively constricted voice tract. The acoustic property associated with a nonsyllabic region is a reduction in low-frequency amplitude in relation to the adjacent vocalic region, but an optimal procedure for detecting such a region may involve a somewhat modified

technique, possibly based on the output of an auditory model. It should be noted that the interval between two sonorant boundaries in the lexicon (defined by a change from + to - on the left and - to + on the right) is always a nonsyllabic interval. Likewise, a [- continuant] event occurs only within or at the edge of a nonsyllabic interval. Thus nonsyllabic intervals are already identified in cases where continuant and sonorant events are marked, and it is necessary to independently specify a nonsyllabic interval only for [+ sonorant, + continuant] consonants (i.e., for the consonants /w y r/ in English) when they occur in positions that are not adjacent to continuant or sonorant markers. For example, in utterances like away, a nonsyllabic marker indicates the /w/, whereas in senate or in fashion the nonsyllabic interval internal to the word is identified by the continuant or sonorant markers.

Still another primary acoustic landmark (4) indicates the presence of a syllabic peak, when the vocal tract has a maximum opening. This landmark is detected on the basis of measures of low-frequency amplitude and of formant trajectories. It is always located in a region whose edges are defined by nonsyllabic or by sonorant or continuant markers.

The representation of a lexical item, then, has a basic structure that involves specification of the sequence of the four types of landmarks or regions: continuant, sonorant, nonsyllabic, and syllabic. Two of these (continuant and sonorant) designate events that occur over brief time intervals, whereas the other two (nonsyllabic and syllabic) designate locations where valleys or peaks occur in certain acoustic parameters.

Some examples of this basic structure are shown in Fig. 2. The continuant and sonorant pointers are indicated by vertical lines. Implosions and releases for [- continuant] consonants are designated by $C1$ and $C2$ respectively, and transitions from sonorant to nonsonorant and vice versa are designated by $S1$ and $S2$. We use a convention of placing a [- continuant] symbol immediately adjacent to the right of a $C1$ pointer and to the left of a $C2$ pointer, and showing the transitions between sonorant and nonsonorant by a +- or a -+ sequence. The designation [- syllabic] appears in the lexical representations only when there is no continuant or sonorant pointer to indicate a valley between syllabic peaks. The exact locations of the pointers [+ syllabic] and [- syllabic] are not crucial, and correspond roughly to maxima and minima in amplitude (as defined with appropriate frequency weighting).

In the examples in Fig. 2, the word rabbit illustrates situations where S and C pointers are coincident, and the aspirated stop consonant in pawn illustrates a sequence of a $C2$ and $S2$ pointer.

2.2. Assigning Features at the Pointers or in the Regions

In the speech signal, there are particular properties or events that correspond to each of the pointers or regions in the basic structure of a lexical item. The acoustic manifestations of other phonetic features relating to place and manner of articulation occur at locations in the sound wave defined by these pointers or regions.

In the case of regions that are [+ syllabic], we assume that there are no narrow constrictions in the vocal tract. The list of features that can be implemented during the syllabic interval is reasonably extensive, and these features may be specified at the edges of the region or more centrally within the region. It is basically the list in the left-hand

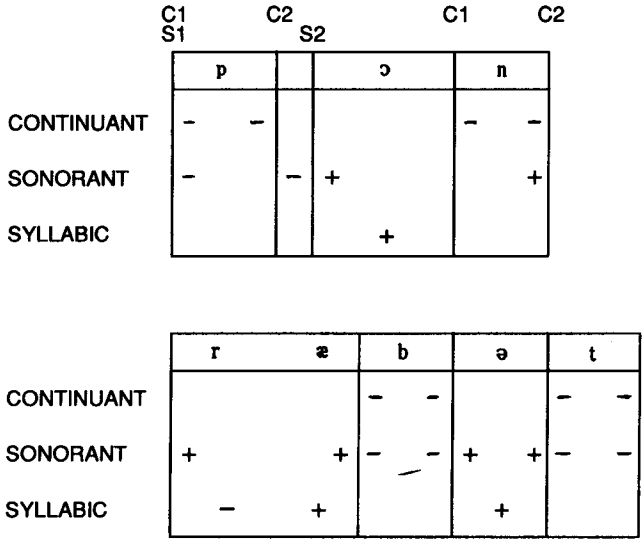


Figure 2. Framework of lexical representations for the words pawn (top) and rabbit (bottom). The phonetic transcription is indicated for reference. The labels for the events are C1 (stop implosion), C2 (stop release),S1 (transition from sonorant to nonsonorant), and S2 (transition from nonsonorant to sonorant). A nonsyllabic mark is indicated only when C or S events do not identify the presence of nonsyllabic interval.

column of Table 1 above. There are more constraints, however, on the location of these features within the syllabic region in English in the proposal lexical representation. For example, the features *breathy voiced* and *pressed voiced* are not distinctive for vowels in English, but can operate redundantly in some consonantal contexts. Thus, [+ breathy voice] can be indicated on the plus side of the sonorant boundary for a voiceless aspirated stop consonant or for an [h]. The acoustic correlate of this feature has been discussed in several publications (Bickley, 1982; Ringo 1988; Klatt and Klatt, submitted). Likewise, the feature *nasal* is only designated in the lexicon at a boundary of the syllabic region that is [- continuant] but remains as [+ sonorant]. That is, nasality in vowels only occurs adjacent to nasal consonants in the lexicon. Features such as *high* and *back* may also be designated at the boundary of a syllabic region, if a consonant adjacent to the syllabic region imposes such a constraint. For example, an alveolar consonant tends to be [- back], and the glides [w] and [j] are [+ high], with [w] designated as [+ back] and [hj] as [- back].

An example of the proposed lexical representation during the syllabic interval for the word *pawn* is given in Fig. 3. We observe the designation [+ breathy voice] at the beginning of the interval, and the specification [+ nasal] and [- back] just before to the postvocalic continuant boundary. These features are normally associated with the consonants /p/ and /n/, but their acoustic manifestation occurs within syllabic interval.

A number of constraints are associated with the continuant and sonorant events in the

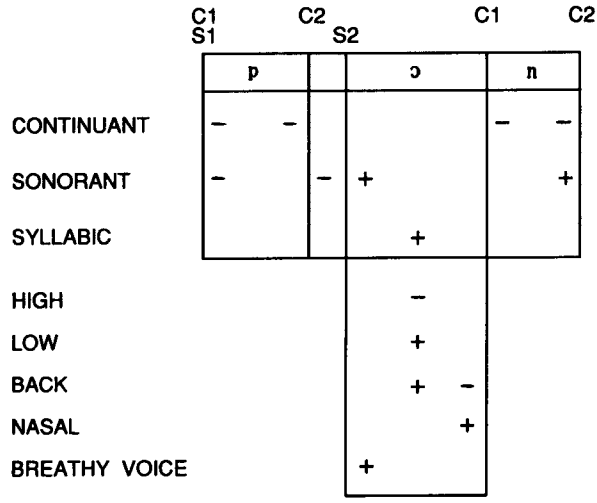


Figure 3. Proposed lexical representation during the syllabic interval is shown below the basic framework for the word *pawn*.

lexical representation, requiring particular features to be associated with each of these events. For example, the presence or absence of the feature *voice* is designated only at a sonorant boundary. The acoustic correlate of voicelessness is the absence of low-frequency energy due to glottal pulses during a portion of the nonsonorant region adjacent to the sonorant boundary. (A detailed discussion of issues associated with voicing is beyond the scope of this paper.) Likewise, the feature *strident* is only designated in the [- sonorant] region, and, except for [s], this designation is almost always at the sonorant boundary. (For example, [æks] is possible, but not [ækf], although [*wit*θ] is an exception.) A third feature that is represented in the sound at certain sonorant boundaries is *consonantal*. The acoustic correlate of this feature is a rapid change in the first-formant frequency on the sonorant side of the boundary–an indication of the formation of a narrow constriction in the midline of the vocal tract.

Place features for consonants (such as the features *labial, coronal* and *anterior*) are constrained to be designated at particular points or regions in the [- syllabic] region depending on the status of the continuant and sonorant boundaries. For example, at a continuant boundary these place features are designated adjacent pointer, whereas when there is neither a sonorant nor a continuant boundary, the designation of the place features is linked to the [- syllabic] designation, which would be in the vicinity of the point where the first-formant frequency (and also the amplitude of the first-formant peak) is a minimum.

Examples of the compiled representations for some lexical items are given in Fig. 4. The representation of each item has been divided into four sections to indicate the natural groupings of the features. In the upper section the three features specifying the pointers or landmarks are listed. Below this section, we specify features that are represented in the sound during the syllabic interval, when the source is at the glottis, whether these

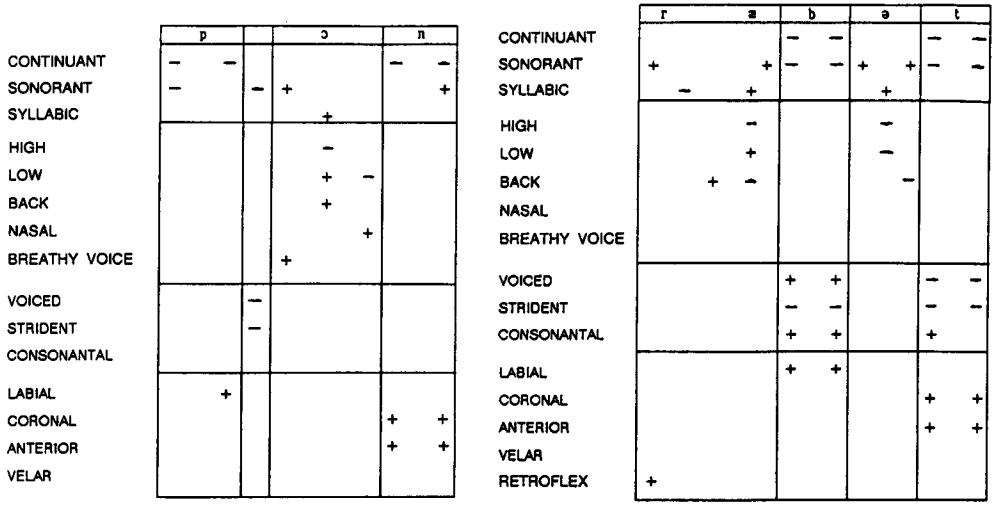


Figure 4. Tentative partial lexical representation of the items for which the framework was shown earlier in Fig. 2. The list of features is not complete, and some entries may need to be revised.

features are normally associated with vowels or with consonants. The features in the next grouping, consisting of voicing, stridency, and consonantality features, are usually specified only at a sonorant boundary. The consonantal place features are grouped at the bottom of each item in Fig. 4. The locations of the entries for these features are directly tied to the pointers in the uppers section, as has already been discussed, since the acoustic properties corresponding to these features are represented in the sound in time intervals that are defined by these landmarks.

A glance at the examples in Fig. 4 indicates that many positions in the matrix are left blank. Some of these blanks occur because the feature is irrelevant (e.g., the feature nasal is irrelevant during a nonsonorant interval, or the tongue body features for a reduced vowel are left unspecified). Others are left blank by convention (e.g., *breathy voice* is left blank, or is assumed to be negative, unless it is specifically marked "+"). Not all of the details of these conventions for marking features in the lexicon have yet been worked out. The examples in Fig. 4 are given to indicate the present status of our thinking about lexical representations.

It is important to observe that the compiled representation can be derived from the more abstract representation in terms of segments and features (i.e., of the type in Table 2) by application of a set of rules. Thus the same basic information is carried by both representations. In the case of the compiled items, some of the acoustic structure of the word is made more explicit, particularly the distribution of acoustic events and acoustic properties over time. An example of the detail contained in the compiled lexical representation is the different specification for consonants depending on syllable position. Thus for the word *spawn*, the continuant and sonorant events for the unaspirated stop consonant are simultaneous in the lexical representation, and the feature [+ breathy voice] is not marked (as it is in *pawn*). For a syllable-final voiceless stop (as in the word *soap*), the feature [+ breathy voice] is again not marked in the lexical item. These different fea-

ture specifications for segments depending on syllable position can help to resolve word boundary ambiguities that exist when only a phonetic representation of an utterance is available (Church, 1987; Harrington and Johnstone, 1987.). For example, one or more entries in the compiled lexical representation of the words in the sequence *gray train* would be different from those in *great rain* or *saw Sally* from *sauce alley.*

3. ACCESSING THE LEXICON FROM ACOUSTIC PROPERTIES

The compiled lexical representation of the type shown in Fig. 4 designates, in some sense, a canonical form for each lexical item. In order to access lexical items in this form from acoustic data derived from an incoming acoustic signal, the first step is to determine from acoustic measurements the four major types of landmarks or regions described above: the continuant, sonorant, and nonsyllabic events and the syllabic regions. Once these events and regions are identified, additional acoustic properties are measured in order to establish the strength with which other features are represented in the sound. These patterns of graded acoustic property values are compared with specifications in the lexicon (which are binary, but perhaps with some indication of relative importance, as discussed below) to determine the best-matching item.

As we have seen, a particular acoustic realization of a lexical item in a sentence context may show evidence for modification of some of the features and the landmarks. The changes are, of course, not random, but are characterized by certain principles. Thus, for a given lexical item, there is a tendency for the implementation of some features (for example, features for stressed vowels and for consonants in prestressed position) to remain invariant independent of context, and for others to undergo some change with context and with speaking style. Among the features that undergo change, some are redundant, and the lexical access process should not be impeded substantially by these variations. An example is the feature [- continuant] that characterizes the [g] in the word *legal.* For other features, some knowledge of the possible changes (as well as changes that are not allowed) may need to be built into the description of the compiled lexicon. These features may be subject to certain kinds of modification, and this fact should be noted in the lexical entry.

As an example of this second type of feature, consider the lexical entry for the word *about.* The compiled description of this word shows both an implosion and a release for the final stop consonant. Depending on the context in which it is produced, this consonant may not be released. On the other hand, if the speaker chooses to release this consonant, the information available in the release burst can contribute correct access of the word. This optional implementation of a word-final stop consonant is a rule that applies across many different lexical items.

In the compiled lexicon that we are considering here, we will need to mark in some way features that have this attribute of being subject to modification but also contributing information if they are implemented. In the case of the example *about*, the final coronal stop can undergo many other modifications, as we have seen, including a change in place of articulation or possibly glottalization (such as [əbáutpi:tər] for *about Peter*) or flapping ([əbáutəmáɪl]) for *about a mile.* In all of these realizations, the [- continuant]

feature is realized in the sound, and it is the place features *coronal* and *anterior* that are subject to modification. Similar changes (except for glottalization) occur in the consonant /n/, as in *can Peter* or *run a mile.* During the matching process that occurs when the lexicon is being searched, the criterion for matching is relaxed when a feature that is marked in this way is encountered. In this same lexical about, it is expected that the place and manner features for the consonant [b] and the features for the stressed vowel would usually be marked in the compiled lexicon as being represented in the sound of a robust manner. In a sense, then, the location of a stressed syllable is identified in the lexical representation by an indication that certain features within the syllable are not subject to variation.

Needless to say, it is a challenging task to find ways of labeling the compiled lexical items so that the features that are almost always well-represented in the sound and features that are prevented from occurring are distinguished from features that are subject to modification in particular phonetic environments or for particular speaking styles. The objective is to have a single entry indicating the phonetic description of each item in the lexicon (leaving aside for the time being the problem of dialectal variation). This entry should be derivable directly from the more abstract representation of the item by application of a set of rules. We hope to be able to account for so-called allophonic variation, or feature modification of various kinds, by appropriate notations on particular features in each lexical item.

4. CONCLUDING REMARKS

We have suggested an approach to lexical access in which the lexicon is searched directly from a representation in terms of features rather than phonetic segments. The lexical representation highlights the fact that some distinctive features are manifested in the sound as well-defined landmarks or events, and that the implementation of other features is tied in a systematic way to acoustic properties in the vicinity of these landmarks. The kinds of variability that occur in the acoustic manifestation of a word can usually be succinctly described in terms of modification of a few of the features that specify the word. Consequently, the match between the acoustic properties and the lexical representation deteriorates only partially when contextual influences modify the features that are used to produce the word. Clearly, many questions must be answered before all the details of the proposed approach to lexical access can be worked out.

References

1. Bickley, C. (1982), "*Acoustic analysis and perception of breathy vowels,*" Speech Communication Group Working Papers I, Massachusetts Institute of Technology, Cambridge MA, pp. 71- 81.

2. Blumstein, S.E. and K.N. Stevens (1979), "*Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants,*" J. Acoust. Soc. Am. 66, pp. 1001-1017.

3. Browman, C.P. and L. Goldstein (in press), "*Gestural structures and phonological patterns,*" In I.G. Mattingly and M. Studdert-Kennnedy (eds.), Modularity and the Motor Theory of Speech Perception. Hillside NJ: Lawrence Erlbaum Associates.

4. Chomsky, N. and M. Halle (1968), "The Sound Pattern of English," New York: Harper and Row.

5. Church, K.W. (1987), "Phonological Parsing in Speech Recognition," Norwell MA: Kluwer Academic Publishers.

6. Clements, G.N. (1985), "*The geometry of phonological features,*" Phonology Yearbook 2, pp. 225-252.

7. Fant, G. (1973), "Speech Sounds and Features," Cambridge MA: MIT Press.

8. Glass, J.R. and V.W. Zue (1988), "*Multi-level acoustic segmentation of continuous speech,*" Proc. ICASSP-88.

9. Harrington, J. and A. Johnstone (1987), "*The effects of word boundary ambiguity in continuous speech recognition,*" Proceedings of the 11th International Congress of Phonetic Sciences, Vol. 3. Tallinn, Estonia, pp. 89-92.

10. Jakobson, R., G. Fant, and M. Halle (1952), "*Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates,*" Acoustic Laboratory Technical Report No. 13, Massachusetts Institute of Technology, Cambridge MA.

11. Kewley-Port, D. (1983),"*Time-varying features as correlates of place of articulation in stop consonants,*" J. Acoust. Soc. Am. 73, pp. 322-335.

12. Klatt, D.H. (1979), "*Speech perception: A model of acoustic -phonetic analysis and lexical access,*" J. Phonetics 7, pp. 279-312.

13. Klatt, D.H. and L.C. Klatt (submitted),"*Analysis, synthesis and perception of voice quality variations among female and male talkers,*" J. Acoust. Soc. Am..

14. Leung, H.C. and V.W. Zue (1984), *A procedure for automatic alignment of phonetic transcriptions with continuous speech* 93 Proc. ICASSP-84.

15. Ringo, C.C. (1988), "*Enhanced amplitude of the first harmonic as a correlate of voicelessness in aspirated consonants,*" J. Acoust. Soc. Am. Suppl.. pp. 1, 83 570.

16. Sagey, E.C. (1986), "*The representation of features and relations in nonlinear phonology,*" Doctoral dissertation, Massachusetts Institute of Technology, Department of Linguistics.

17. Stevens, K.N. (1980), "*Acoustic correlates of some phonetic categories,*" J. Acoust. Soc. Am. 68, pp. 836-842.

18. Stevens, K.N. (1983), "*Acoustic properties used for the identification of speech sounds,* " Annals of the New York Academy of Sciences, Vol. 405, pp. 2-17.

19. Stevens, K.N. (1986), "*Models of phonetic recognition II: A feather-based model of speech recognition,*" In P. Mermelstein (ed.), Proceedings of the Montreal Satellite Symposium on Speech Recognition. Twelfth International Congress of Acoustics, pp. 67-68.

This Page Intentionally Left Blank

Chapter 5

SPEECH UNDERSTANDING

This Page Intentionally Left Blank

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

A Large-Vocabulary Continuous Speech Recognition System with High Prediction Capability

Minoru Shigenaga*and Yoshihiro Sekiguchi

Faculty of Engineering, Yamanashi University, 4 Takeda, Kofu, 400 Japan

Abstract

A large-vocabulary (with 1019 words and 1382 kinds of inflectional endings) continuous speech recognition system with high prediction capability, applicable to any task and aiming to have unsupervised speaker adaptation capability is described. Phoneme identification is based on various features. Speaker adaptation is done using reliably identified phonemes. Using prosodic information,phrase boundaries are detected. The syntactic analyzer uses a syntactic state transition network and outputs the syntactic interpretations. The semantic analyzer deals with the meaning of each word, the relationship between words, and extended case structures of predicates. Recognizing noun phrases first, predicates are predicted, and vice versa.

1. INTRODUCTION

For a continuous speech recognition system with a large vocabulary, some of the following problems, from an acoustic point of view, become more serious. (1) Most of the words in a sentence are pronounced rapidly and unclearly. (2) As a consequence, the effect of coarticulation becomes dominant. (3) Uncertainty of word or phrase boundaries. (4) These tendencies may appear even if a sentence is pronounced slowly and become more difficult when speaker independency is required.In order to cope with these difficulties, recently HMM techniques have been used intensively and good results have been obtained (for example ref.[1]).

On the other hand, we can easily understand spoken sentences. This may be due to the fact that, using various knowledge, we can predict not only the following words but also the meaning of the sentences and the context of the topic.However, it is very difficult for a mechanical speech recognition system to predict the following words, because the knowledge which the system has is poor and it must treat phoneme or word strings with erroneous phonemes or words.Therefore, an approach to these problems using syntactic and semantic aspects may be more fundamental and important, although language has a

*Now at School of Computer and Cognitive Sciences, Chukyo University, Kaizu-cho, Toyota, 470-03 Japan

statistical aspect.So, putting emphasis on linguistic processing and prosodic information, the authors have been trying to construct a fundamental continuous speech recognition system with high prediction capability and applicable to any task[2-6]. Since recently the authors are trying to construct a system with a vocabulary size of 1019 words, 1382 kinds of inflectional endings, and a perplexity of about 280 with the grammar by refining especially the semantic information of each word, the dependency relationships between words, the extended case structures, and adding an associative function[9,10].The semantic information is expressed in terms of semantic attributes and effectively used for prediction in universally applicable forms. The recognition procedures are syntactically and semantically excuted in parallel, and words or phrases are recognized after various kinds of linguistic prediction. The acoustic analyzer has a phrase boundary detector[11], and is aiming to have unsupervized speaker adaptation capability[12].

This paper describes mainly some attempts to detect phrase boundaries and to predict predicates and noun phrases using semantic information.

2. ACOUSTIC PROCESSOR

The acoustic processor consists of a phoneme identifier and a phrase boundary detector. Speech waves are sampled at 10 kHz, 12 bits, and after filtering with $(1 - z^{-1})$ or adaptive inverse filters, LPC and PARCOR analyses are executed every 10 ms.

2.1. Phoneme identification and characteristic phonemes

Before phoneme identification, taking the maximum slopes of the speech waveform envelope into consideration, speech waves are segmented globally into segments which contain usually one phoneme. In order to identify phonemes – the Japanese five vowels /a,i,u,e, o/, /s,h,r/, the unvoiced stop consonant group /p,t,k/(designated as /P/), the nasal group /m,n, ŋ/and N/(/N/), the buzz bar(/B/) and the silent part (/./) –, at first, various kinds of preliminary phoneme identification methods are carried out every 10 ms, speaker independently, using various characteristic features extracted from LPC spectra, vocal tract area functions, waveform envelopes, and numbers of zero-crossing. The phonemes identified with high reliability are marked with *. The characteristic phoneme string of each word, which consists of reliably identifiable vowels, /s/, and silence, is used effectively for pre-selection of candidate words.

2.2. Adaptation to a new speaker

After processing each input sentence spoken by a new speaker, the reliably identified phonemes, marked with *, are collected and the vocal tract area functions and 20 spectrum components or Cepstrum coefficients of these phonemes are used, respectively, as training data for two kinds of neural networks and reference patterns. Thus, the neural networks and the reference patterns for the new speaker are obtained. These networks and reference patterns may be revised sometimes by adding new reliable parameters of phonemes marked with *. Thus listening to several free sentences spoken rather slowly by a new speaker, the system gradually adapts its neural networks and reference patterns to a speaker.

2.3. Detection of Phrase Boundaries

One of the difficult problems in speech recognition is the fact that most boundaries between successive words or phrases are entirely unknown to the system. Division of a sentence into a sequence of words or phrases by detecting word or phrase boundaries is desired for obtaining good performance. So the authors have tried to take prosodic information into account for the detection of phrase boundaries in any sentence.

In order to detect phrase boundaries, first, CV (consonant-vowel) or V syllable boundaries are detected, since in Japanese phrase boundaries coincide with the boundaries of CV or V syllables. Extraction of dips in the waveform envelope is useful for detection of CV boundaries, since boundaries between CV syllables show dips in the waveform envelope. Also, when a syllable boundary in a vowel concatenation coincides with a phrase boundary, the syllable boundary usually shows a dip in the waveform envelope as a consequence of a somewhat loose coupling due to the phrase boundary. So, first, as candidate phrase boundaries, syllable boundaries are extracted. In order to decide on phrase boundaries, it may be useful to examine whether each syllable boundary has any of the following six features.

(1) A valley in the fundamental frequency contour

In general, the fundamental frequency contours in a declarative sentence display a sequence of figures with a "$\frown$" shape and phrase boundaries exist usually in the valleys of the sequence of such figures. But, by the effect of mutual relation of accent patterns between successive phrases or the modifying relation between phrases, there are cases where no apparent valley between phrases appears in the fundamental frequency contour.

(2) A local variation in the fundamental frequency contour

Even if a valley in the fundamental frequency contour is not detected at the phrase boundary, a phrase boundary may be found at the place where a local variation in the fundamental frequency contour has a characteristic variation, such as (i) the contour is almost flat, or (ii) the contour changes from decreasing to increasing or becomes flat.

(3) A gradual slope of the valley in the waveform envelope

When the slope of the valley in the waveform envelope is not steep the valley tends to be a phrase boundary.

(4) A long interval of silence

When the modifying relation between successive phrases is weak or when there is a pause for breath, there is a rather long interval. So when such a silent stretch is long, there is a high probability that it is a phrase boundary.

(5) A valley in the pseudo-fundamental frequency contour

A pseudo-fundamental frequency contour is defined as a contour which connects each fundamental frequency at a point having maximum value within the speech waveform envelope in an interval between successive dips in the waveform envelope. The place having the minimum value of the pseudo-fundamental frequency contour may be a phrase boundary.

(6) Long distance from the adjacent phrase boundaries

Any place far from a candidate phrase boundary and having a gradual slope in terms of the fundamental frequency contour, is possibly a phrase boundary.

If a syllable boundary has one of these six characteristic features, a score of "1" is

Table 1 Parts of speech.

Part of speech	Symbol	Part of speech		Symbol
Noun, Pronoun	Noun	Particle	Adnominal	PP
Pre-noun(Attribute)	PREN		Predicate	PT
Adjective	ADJ		Case	PC
Verb	VERB		Terminal	PF
Conjunctive	CONJ	Adverb	Conditional	CADV
			Statal	SADV
Auxiliary verb	AUX		Degree	DADV

given to the syllable boundary. Using linear discriminant functions composed of these six parameters, and the variation of the function's values, the detector expresses the possibility of a phrase boundary with scores independently from individuality. Scores which the detector has given to 710 syllable boundaries, contained in the total 66 sentences uttered by six adult males, have been examined. The total number of phrase boundaries is 186. For 6 out of the them syllable boundaries have not been detected and so also not the phrase boundaries. One phrase boundary had score 0. To 354 (63%) out of a total of 524 syllable boundaries a score 0 has been given.

It is impossible to detect all phrase boundaries accurately by this method, but most of the remaining ones may be at the syllables' boundaries next to the ones pointed out by the method; and even if the detector cannot point out all the correct phrase boundaries, there may be at most only two phrases between adjacent detected phrase boundaries, and this may happen in cases where a preceding phrase strongly modifies the following phrase, or concatenation of two specific accent patterns exists. But even in such cases correct phrases may be extracted by the matching process in the linguistic processor.

3. KNOWLEDGE REPRESENTATION

The system has a knowledge source which contains syntactic, semantic, and vocabulary information in universally applicable forms.

3.1. Syntactic Knowledge

Most of the fundamental grammar in the Japanese language is included in the system. However, some changes concerning categories and word inflections were made in order to facilitate machine recognition.

3.1.1. Words and phrases

Words are classified as shown in table 1. Pronouns are included in the category of nouns. In this classification, some parts of speech are divided into several groups with reference to their syntactic roles and semantic features.

In Japanese, a sentence can be divided into phrases, each of which consists of conceptual and functional words. The conceptual words convey a concept by themselves, such as

nouns, verbs, adjectives, etc. Functional words give a conceptual word or phrase a function to modify the following phrase. Auxiliary verbs, particles, and their combinations are included in this class. These two kinds of words usually appear as combined forms in a sentence. The combined form is called a phrase and some modifying relation is established between the phrases. The syntactic function of a phrase is classified into three types: (i) modification of a nominal word or phrase — adjective modification, (ii) modification of a verbal or adjective word or phrase — adverbal modification, (iii) termination (end of a sentence).

3.1.2. Classification of inflectional forms

In Japanese, there are many types of inflections, and further, each inflectional word has various inflectional forms and represents various aspects and meanings. In our system, these are classified as shown in table 2, and 1382 kinds of inflectional endings are prepared with semantic markers.

Table 2 Inflectional forms.

Form	Symbol	Form	Symbol
Adverbial modification	RY	Imperative	IMP
Adjective modification	RT	Interrogative	INT
Conclusive	END	Pause	PAUSE
Conditional	COND		

3.1.3. Syntactic state transition network

The conjunctive relations among the parts of speech in Japanese are represented by a syntactic state transition network as shown in fig.1.
This network is useful for an agglutinative language such as Japanese, which has loose constraints on the order of phrases. The perplexity of the system with this grammar is about 280 for the task domain of fairy tales, a vocabulary size of 1019 words, and 1382 kinds of inflectional endings. In fig. 1, SOC,SVP, etc., enclosed by large circles, represent syntactic states reached after recognition of a phrase. N, A, etc., enclosed by small circles, represent syntactic states reached after recognition of a conceptual word. SOS and EOS represent the beginning and the end of a sentence, respectively. The symbol along an arc from a large circle or SOS to a small circle shows the part of speech of a conceptual word. The functional word X in X/Y along an arc from a small circle to a large one or EOS shows a particle or nil(ϕ) when X is not an inflectional word, and Y shows a syntactic function. That is, MODN at Y means the modification of a noun, MODV that of a verb, MODA that of an adjective, and MODP that of a predicate, while EC and ES show the ends of a clause and a sentence, respectively. When X is an inflectional word, X shows its inflectional form category(table 2) and is enclosed in parentheses. In this case, X is an auxiliary verb, a particle PF or a combination of these. Y is the same as described above. Auxiliary verbs and particle PFs do not appear explicitly in fig.1, but are embedded in the Xs in parentheses.

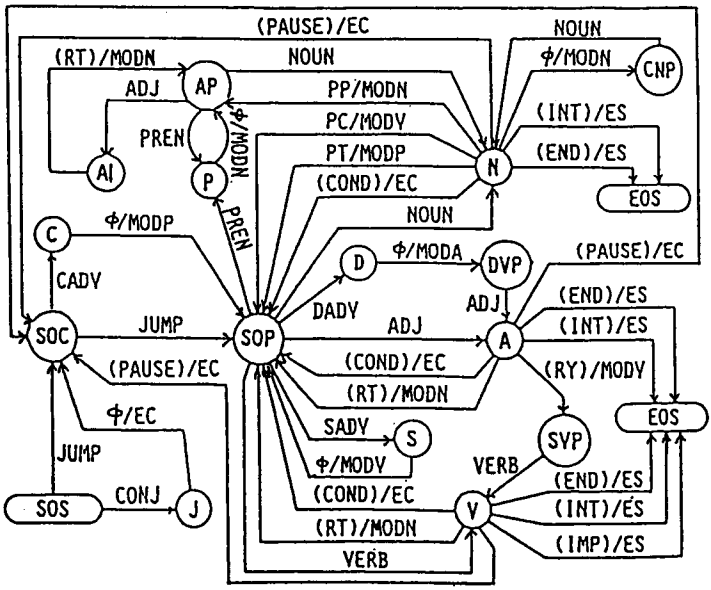


Fig.1 Syntactic state transition network.

Thus, this network is effective for processing sentences in terms of phrases, because it indicates modifying relations between phrases, has much freedom for the order of phrases and expansibility, can insert semantic, associative prediction anywhere, and the system can process input sentences syntactically and semantically in parallel from left to right, right to left, or any island. Thus, words or phrases are recognized after syntactic and semantic prediction.

3.2. Semantic Expression of Words

The meaning of each word and the case structures (with somewhat extended forms) of verbs, predicate adjectives, and predicate nouns are taken into consideration.

(i) Nouns

From a semantic point of view, nouns are classified globally into nine categories, which are the highest-level attributes of meaning. These are Life, Thing, Abstract, Action, Relation, Location, Time, and Quantity, and they are marked by *. A noun has certain attributes which represent a higher-level concept of the noun, such as *Human, *Conveyance, *Language, *Work. Therefore,nouns are classified with a hierarchical structure of attributes. To predict predicate nouns, case structures of nouns are also prepared, as shown in table 3. "- not " in the table means that "female, child" must be excluded from *Human.

(ii) Verbs

Table 3 Case structure of the noun "old man".

Noun	Case	Attributes of nouns that shoulda be attached to the case
old man	Agent	* Human (- not female, child)
	At time	* Time (at,ϕ(empty))

Each verb (in general, predicate) has an inherent case structure which plays a very important role in representing its meaning in a sentence. Normally, a case takes the form "noun + particle" in Japanese, which constitutes a phrase, and this case structure is prescribed for each verb. Thus, the concrete meaning of a verb is represented by the meaning of each case, or the meaning of the phrase adopted to each case, in addition to the meaning of the verb itself. Table 4 shows, as an example, the case structure of the verb "cut". Thus, the case "Instrument" in the case structure of "cut" should have a lower-level attribute "*Cutlery" with the preposition "with". Thus the verb "cut" is not predicted when in the case of an "Instrument", a vehicle, such as a car or an airplane, appears.

To each verb some of 31 attributes, such as "Action, Change (of state), Movement, Contact, Perception, Thinking, are also given and used to represent the modifying relation between an adverb and a verb.

Table 4 Case structure of the verb "cut".

Verb	Cases	Attributes of nouns that should be attached to the case
cut	Agent	* Human
	Object	* Thing, * Life
	Instrument	* Thing, * Cutlery (with)
	Somebody	* Human (with)
	At location	* Location (at,on,in,ϕ)
	At time	* Time (at,ϕ)

(iii) Adjectives

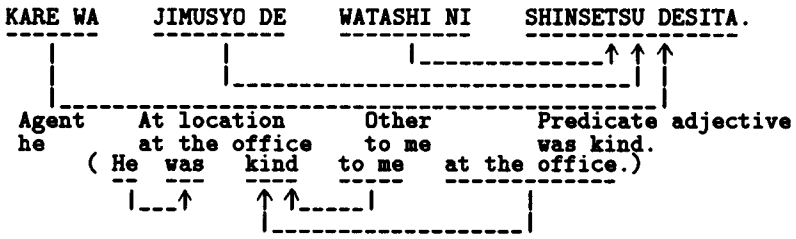
Adjectives are also classified into 25 kinds of semantic attributes, such as "Feeling, Character, Weather, Hight, Color", according to their meaning. These attributes are put to the attributes of nouns. For example, to "*Human"(an attribute of human beings) "Feeling", "Character", etc. are attached. Thus "gentle" and "kind" have "Feeling" and "Character" as attributes, respectively, and can modify " boy " which has " *Human " as an attribute, but "hot", having an attribute "Weather", cannot.

Table 5 shows the case structure of the predicate adjective "kind" which is used in a sentence as follows:

The modifying relationships with arrows in (S1) are indicated along arcs in the network shown in fig.1. Further explanations of meanings of other parts of speech are described in refs, [4] and [8].

Table 5 Case structure of the adjective "kind".

Adjective	Case	Attributes of nouns that should be attached to the case
kind	Agent	* Human, God
	Other	* Human (to)
	Comparison	* Human, God (than)
	At location	* Location (at, on, in,ϕ)
	At time	* Time (at,ϕ)



4. PREDICTION OF NOUNS AND OTHER WORDS BY THE RELATIONSHIP BETWEEN ADJACENT WORDS

Usually the number of nouns occupies 50% or more of a 0% vocabulary, and also nouns have an important roles in case structures, when there is some relationship between two adjacent words. In our system, for the following cases universal relationships are written using the semantic attributes of each word:

"noun + TO(and) + noun", "noun + YA (or) + noun", "noun + NO (of) + noun", "adjective + noun", "pre-noun + noun", " adjective form of a verb + noun", "adverb + verb", "pre-noun + pre-noun", etc.(TO,YA,NO:particles). For example, in the case of concatenation "noun+TO(and)+noun" both nouns may have the same semantic attributes, and the conjunction "adjective + noun" has the relationship described in section 3.2(iii).

Moreover, cooperation between adverbs and conflectional endings(in English, for example, "if ... then", a style appearing in subjunctive mode) is also taken into consideration. These relationships are written in semantically universal forms and may be used for predicting or checking each other.

5. PREDICTION OF PREDICATES BY CASE STRUCTURES

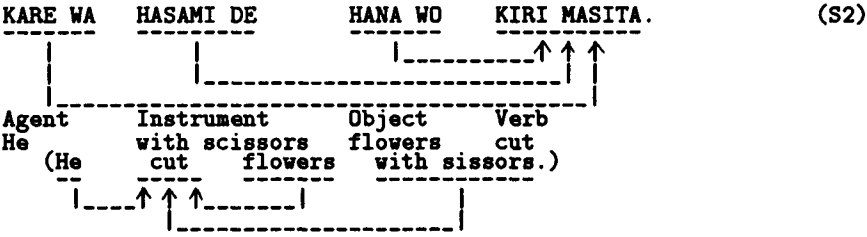
It is very important for a continuous speech recognition system to be able to predict following words and decrease the number of candidate words as much as possible. Espe-

cially in Japanese, it is essential to be able to predict the main predicates, because verbs, adjectives, and auxiliary verbs inflect, and in our system about 90 inflectional endings (which are composed of ending of a verb, and some auxiliary verbs, particles, or their combination) can follow a predicate verb, and about 23 inflectional endings may follow a predicate adjective, and also a predicate is usually placed at the end of a sentence.

The prediction of a predicate is done using the noun phrases which compose the case structure of the following predicate. That is, processing an input sentence from left to right, noun phrases are successively recognized and the kind of case for each noun phrase is decided. Thus, only the predicates that can have such a case structure are selected.

(i) Prediction of verbs

Processing an input sentence from left to right, noun phrases are successively recognized and the kind of case for each noun phrase is decided. Thus only verbs that can have such a case structure are selected. For example, in the sentence:



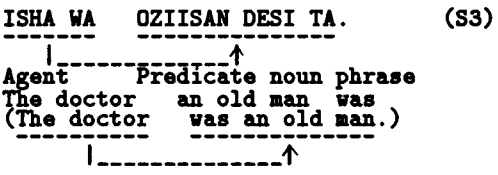
If "He" has been recognized and a verb should be predicted, then verbs such as "go, have, live, cut, think etc.", which can have "*Human" as an attribute of the case "Agent", are predicted. In this situation, many verbs may become candidates. But if the next phrase "HASAMI DE (with scissors)" has been recognized, verbs such as "go, have, cut, etc.", which have the case "Instrument", remain, and verbs such as "live" which do not have the case "Instrument" will be omitted. Moreover, "scissors" has an attribute "*Cutlery (with)", which is a kind of "Instrument (*Thing)", so verbs which have only the attribute "*Cutlery" with the preposition "with" (in Japanese DE (particle)) among the various kinds of instruments are selected. Therefore, verbs such as "go, have, think etc." will be rejected, and a small number of verbs such as "cut, trim" will remain.

(ii) Prediction of predicate adjectives

If in the sentence (S1) "he (Agent), to me (to Other) or at the office(At location)" have been recognized, adjectives such as "kind, cold-hearted" will be predicted and adjectives such as "white, hot" will be omitted.

(iii) Prediction of predicate nouns

Predicate nouns are also predictable by using case structures of nouns. For example, in the sentence (S3) "old man" may be predicted by hearing "doctor" from the case structure of "old man" shown in table 3, if the doctor is not female.



6. PREDICTION OF NOUN PHRASES BY PREDICATES

Noun phrases and predicates have very close relations, and by recognizing a predicate first, noun phrases may be predicted. In the case of languages such as English, in which a main predicate usually follows a subject at an early stage, following noun phrases are predictable. On the other hand, in the case of Japanese, in which a predicate is placed at the end of a sentence, it is difficult to recognize a predicate first. However, in our system, segmentation of the last predicate phrase is rather easy owing to the phrase boundary information, so it is possible to recognize predicates first and proceed the recognition process by stepping backward from the node EOS(end of sentences) in the syntactic state transition network.

7. RECOGNITION PROCEDURES WITH BOTH FORWARD AND BACKWARD PREDICTION

By starting one of two recognition processes from left to right (forward) and the other from right to left (backward), it may be possible to verify each recognition result mutually or to reject words or phrases which have no relation to each other. However, there may be some problems with complex or compound sentences.

8. FURTHER PREDICTION PROCEDURE

From the above discussion it becomes possible to predict noun phrases even in the case of left to right (forward) processing. That is, if one or two noun phrases have been recognized, some predicates are predicted. Even if the predicates have not been recognized, using case structures will make possible the prediction of noun phrases placed between already recognized noun phrases and the predicted predicates.

This procedure becomes straightforward by preparing four-items of the case structure: "a semantic attribute of a noun + a case particle (PC in table 1) + a kind of case + a predicate having such a case structure ". Using these four-items of the case structure, case particles following a noun, noun phrases following a noun phrase, predicates relating to a noun phrase or noun phrases relating to a predicate are predictable.

9. ASSOCIATIVE FUNCTION

Besides the various kind of information described above, it may be supposed that we have an associative function with which we can associate already recognized words or sentences with certain concepts or words and help ourselves to understand sentences. So we have examined words associated with words in the dictionary and have constructed an associative-words dictionary. Moreover, we have defined a distance between associative words based on associative information, and composed a distance-between-associative-words matrix [9].

10. EXPERIMENTAL RESULTS

Six sentences from the fairy tale "The three little pigs" uttered by 6 adult males having different accent patterns, have been used to adjust both the rules and three-layer neural networks for phoneme identification, and these thereafter are used speaker independently. The sentence recognition rates(including the top three) for a total of 72 sentences(including the above 36 sentences) uttered freely by the same 6 males are shown in table 6 for each speaker. In the table, system (1) does not use any associative function, system (2) uses the associative-word dictionary, and system (3) uses both the associative-word dictionary and the distance-between-associative-words matrix[10].

For 9 new sentences of another fairy tale, containing 3-9 phrases, uttered by two new speakers, the sentence recognition rates were (1) 38.9%, (2) 38.9% and (3) 50.0%, respectively. The very low 9%recognition rate with system (1) is mainly due to low phoneme identification, which is caused by the few training samples.

Table 6 Sentence recognition rates[%]

System	Speaker						Average
	NK	IT	MA	FU	KO	SE	
(1)	16.6	33.3	16.6	50.0	50.0	33.3	33.3
(2)	16.6	58.3	25.0	75.0	75.0	75.0	54.2
(3)	33.3	83.3	58.3	83.3	91.7	75.0	70.8

11. CONCLUSION

The strategy of semantic processing will be used task independently and also adapted to most languages. On the other hand, for a specific task, by using information particular to the task, the predictability will increase and the performance of the system may be much improved.

Now the recognition results are not sufficient. This is mainly due to the poor phoneme identification scores. For further development it is essential to identify each phoneme more accurately, and implement more various semantic knowledge.

The authors wish to hearty thank the students in our laboratory for their help.

References

1. K. F. Lee, H. W. Hon and R. Reddy, "An over view of SPHINX speech recognition system," IEEE Trans. on ASSP, 38,1,pp.35-45, (1990)

2. M. Shigenaga, Y. Sekiguchi and C. H. Lai, "Speech recognition system for Japanese sentences," Proc. COLING-80, pp.472-479, (1980)

3. Y. Sekiguchi, C. H. Lai and M. Shigenaga, "On syntactic informatin in a speech recognition system for Japanese sentences" Trans. IECE Japan, J65-D,6,pp.782-789, (1982)

4. Y. Sekiguchi and M. Shigenaga, "On semantic informatin in a speech recognition system for Japanese sentences" Trans. IECE Japan, J66-D,6,pp.629-636, (1983)

5. M. Shigenaga, Y. Sekiguchi, T. Yagisawa and K. Kato, "Speech recognition system for continuously spoken Japanese sentences - SPEECH YAMANASHI -," Trans. IECE Japan, E69, pp.675-683, (1986)

6. Y. Sekiguchi, T. Hanagata, Y. Suzuki and M. Shigenaga, "Prediction of predicates by using their case structures for speech recognition of Japanese sentences," Trans. IEE Japan, 108-C, pp.818-825, (1988)

7. M. Shigenaga, Y. Sekiguchi, T. Hanagata, M. Taki and T.Yamaguchi, "On prediction possibility of predicates and noun phrases for continuous speech recognition," Trans. IECE Japan, J72-DII, pp.1307-1312, (1989)

8. M. Shigenaga, Y. Sekiguchi, M. Taki, S. Furuta and R. Masuda, "A speech recognition system for large vocabulary and speaker adaptation function," IEICE Technical Report, SP89-88, (1989)

9. T. Yamaguchi, R. Masuda, M. Kawasaki, Y. Sekiguchi and M. Shigenaga, "Use of associated words in a speech recognition system with large vocabulary," IEICE Tecnical Report, SP90-76, (1990)

10. M. Shigenaga, Y. Sekiguchi, T. Yamaguchi and R. Masuda, "A large vocabulary continuous speech recognition system with high predictability," Trans. IEICE Japan, E74, 6, pp.1817-1825, (1991)

11. Y. Suzuki, Y. Sekiguchi and M. Shigenaga, "Detection of phrase boundaries using prosodics for Japanese speech recognition," Trans. IEICE Japan, J72-DII, pp.1609-1617, (1989)

12. S. Furuta, R. Masuda, Y. Sekiguchi and M. Shigenaga, "An acoustic processor for a speaker independent, large vocabulary continuous speech recognition system," IEICE Technical Report, SP89-73, (1989)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Syntax/Semantics-Oriented Spoken-Japanese Understanding System: SPOJUS-SYNO/SEMO

Seiichi Nakagawa, Yoshimitsu Hirata and Isao Murase

Department of Information and Computer Sciences, Toyohashi University of Technology
Tenpakucho, Toyohashi, 441 Japan

Abstract

This paper describes a syntax/semantics driven spoken-Japanese understanding system named "SPOJUS-SYNO/SEMO ". First, this system makes word-based hidden-Markov-models (HMM) automatically by concatenating syllable-based (trained) HMMs. Then a word lattice is hypothesized by using a word-spotting algorithm and word-based HMMs for an input utterance. In SPOJUS-SYNO, the time-synchronous left-to-right parsing algorithm is executed to find the best word sequence from the word lattice according to syntactic and semantic knowledge represented by a context-free semantic grammar. In SPOJUS-SEMO, the knowledge of syntax and semantics are represented by a dependency and case grammar. This system was implemented in the "UNIX-QA" task with a vocabulary size of 521 words. Experimental results show that the sentence recognition/understanding rate was about 87% for six male speakers for the SPOJUS-SYNO, but was very low for the SPOJUS-SEMO.

1. INTRODUCTION

In speech understanding systems, there are two basic control strategies for the syntactic analyses. One is a left-to-right parsing control strategy. This strategy has been used for a syntactic analysis of text inputs or speech inputs. The standard parsing algorithms are based on Earley's algorithm (top down), the CYK algorithm (bottom up), and the Augmented Transition Network grammar. The other control strategy is an island-driven strategy. This strategy is attractive for speech understanding systems, because candidate words obtained from speech input are not always correct. However, the latter consumes much computation time.

There are also the other basic choices for the parsing strategy. They are a backtracking search versus a parallel search. In speech understanding systems, the most optimal word sequence should be found in a word lattice, since the detected words are not perfect and have scores of reliability. In such a case, the parallel search is suitable and is usually implemented as a beam search [1,2]. The beam search is more efficient than a best-first (A*) search.

Before constructing a speech understanding system, we compared the left-to-right and top down parsing strategy with the island-driven and bottom up strategy by using a simulated phoneme recognizer [3]. Both strategies adopted the beam search. The syntactic constraint was represented by a context-free grammar. The word lattice for an utterance was generated by a word-spotting algorithm from an ambiguous phoneme sequence. The input of the parsers consists of a word lattice of candidate or spotted words, which are identified by their begin and end times, and the score of the acoustic phonetic match. Recently, Ward et al. have also studied a similar comparison [4]. They found that the island-driven parser produces parses with a higher percentage of correct words than the left-to-right parser in all cases considered. However, they did not use the grammatical constraints expressed in a context-free grammar, but trigram models of sentences with lexical and semantic labels. Their evaluation criterion was the rate of correctly recognized words. Our criterion was the rate of correctly recognized sentences. Therefore, our conclusion is not comparable with their results.

From the simulation experiments we found that (1) the left-to-right and top down parsing strategy was superior to the island-driven and bottom up strategy in terms of the processing time, (2) the recognition accuracy was almost the same for both strategies, and (3) when the initial part of an utterance was noisy, the island-driven strategy became superior to the left-to-right strategy. According to these comparison results, we developed a left-to-right parsing oriented spoken-Japanese understanding system named SPOJUS-SYNO. Many successful continuous speech recognition systems such as BYBLOS [5], SPHINX [6], and SPICOS [7] have adopted phoneme models based on HMM. The SPOJUS-SYNO used syllable-based HMMs as the basic unit of speech recognition. This system was implemented in the "UNIX-QA" task with a vocabulary size of 521 words. The experimental results show that the sentence understanding rate was 87% for six male speakers.

We also developed a semantic-oriented speech understanding system. The knowledge of syntax and semantics are represented by a dependency grammar (Kakari-Uke) between phrases and a case grammar among phrases. This system uses the same word lattice as SPOJUS- SYNO, obtained by the syllable-based HMM word-spotting algorithm. We show that the performance was very low in comparison with SPOJUS-SYNO.

2. SPOJUS-SYNO [8]

2.1. System Organization

Figure 1 illustrates the system organization. First, this system makes word HMMs automatically by concatenating syllable-based (trained) HMMs. Japanese comprises about 110 syllables, each of which is composed of a consonant and a vowel (CV), a syllabic nasal (N), a vowel (V), or a consonant, a semivowel and a vowel (CYV). We adopted a continuous output observation HMM with a discrete duration probability [8] [15]. This model consists of five states or four transitions. The four parameter set of duration (transition) and output probabilities (the mean vector and covariance matrix of feature vectors) were calculated using the Baum Welch estimation algorithm. Then a word lattice is hypothesized by a word-spotting algorithm and word-based HMMs. A hypothesized word

consists of a beginning frame, an ending frame, a matching score (probability) and a word name. Finally, a time-synchronous left-to-right parsing algorithm is executed to find the best word sequence from the word lattice according to syntactic and semantic knowledge represented by a context-free semantic grammar.

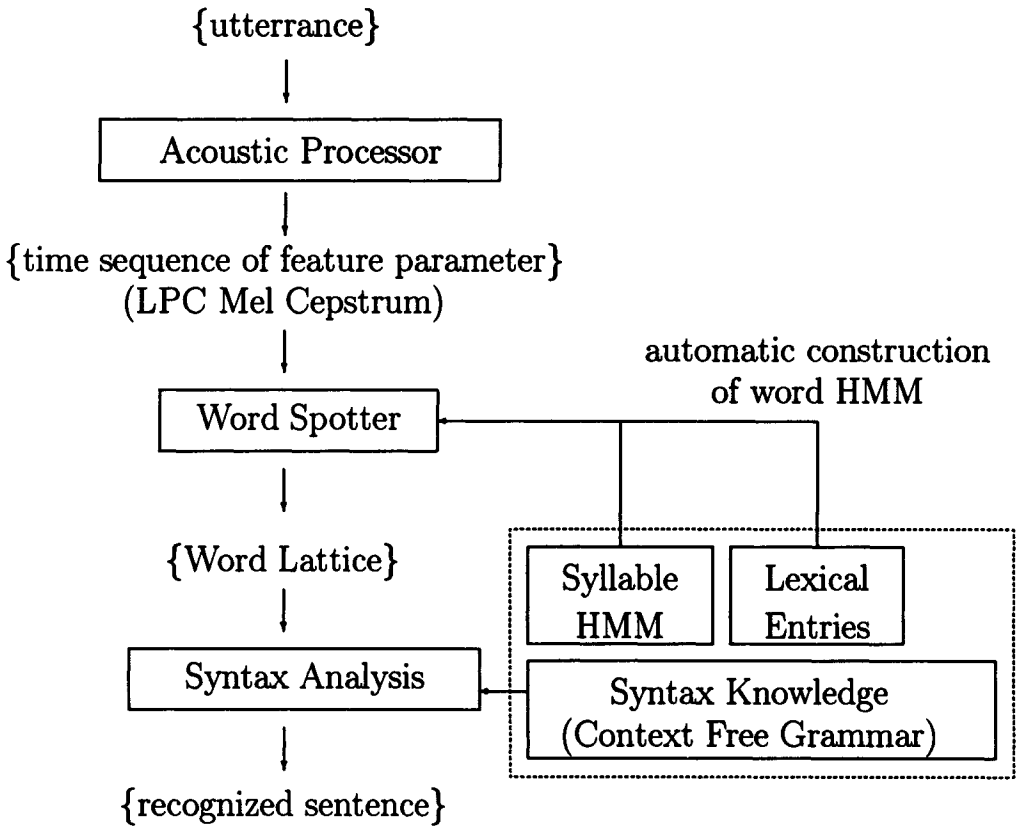


Fig.1. System Organization of SPOJUS-SYNO.

2.2. Sentence recognition algorithm from a word lattice [9]

(a)Representation of the grammar

The syntactic knowledge in terms of the "task" is given by the grammar. This grammar is represented by the context-free grammar as shown in Fig.2. A variable with the affix "@" is a non-terminal symbol, and a variable with "*" a word class (a kind of non-terminal symbols). A word class means the set of words with the same syntactical category. The numbers of the rows and columns define the position of a production rule. They will be

	0		1	2	3	4
8	@S	⟶	@NP	@VP		
16	@S	⟶	@NP	*AUX	@VP	
24	@S	⟶	*AUX	@VP		
32	@NP	⟶	*DET	@NP2		
40	@NP	⟶	@NP2			
48	@NP2	⟶	*ADJ	@NP2		
56	@NP2	⟶	@NP3			
64	@VP	⟶	*VERB			
72	@VP	⟶	*VERB	@PP		
80	@VP	⟶	*VERB	@NP		
88	@PP	⟶	*PREP	@NP		
96	@NP3	⟶	@NP3	@PP		
104	@NP3	⟶	*NOUN			

*NOUN	⟶	JOHN
*NOUN	⟶	MARY
*NOUN	⟶	MAN
*NOUN	⟶	I
*NOUN	⟶	TENNIS
*NOUN	⟶	GAME
*AUX	⟶	WILL
*AUX	⟶	CAN
*VERB	⟶	KNOW
*VERB	⟶	PLAY
*VERB	⟶	PLAYED
*DET	⟶	THE
*DET	⟶	A
*ADJ	⟶	BIG
*ADJ	⟶	YOUNG
*PREP	⟶	BY
*PREP	⟶	OF
*PREP	⟶	WITH

Fig.2 An Example of a Context-Free Grammar.

used in the parsing algorithm described in the next section.

(b)Time-synchronous context-free parsing algorithm

In the LITHAN speech understanding system [1], we proposed an efficient context-free parsing algorithm which is similar to the Earley algorithm [10]. We also use a modified algorithm.

Let the position of a production rule be represented by it's number. For example, the number "8" denotes @S, "9" @NP, and "19" @VP in the above grammar (see Fig.2). The basic problem is formalized as follows: Which words are predicted as the succeeding words when a partial sentence is given ? For example, when the partial sentence "MARY WILL PLAY" is given , which words could appear on the right-hand side ? In this case, the partial sentence is derived by the following production rules: @S→ @NP * AUX @VP→ @NP2 * AUX @VP→ @NP3 * AUX @VP→ * NOUN * AUX @VP→ MARY * AUX @VP→ MARY WILL @VP→ MARY WILL * VERB, MARY WILL * VERB @PP or MARY WILL * VERB @NP→ MARY WILL PLAY, MARY WILL PLAY @PP or MARY WILL PLAY @NP. Therefore the succeeding words could be predicted from @PP and @NP. Of course, "MARY WILL PLAY" may be regarded as a complete sentence in the first alternative derivation. "BY", "OF", "WITH", "THE", "A", "BIG", "YOUNG", "JOHN", "MARY", "MAN", "I", "TENNIS", and "GAME", are predicted.

We can memorize the application order of the production rules by the sequence of positions in the grammar. For the above example, "MARY WILL PLAY" is derived by the sequences "16"→ "17"→ "17 40"→ "17 41"→ "17 41 56"→ "17 41 57"→ "17 41 57 104"→ "17 41 57 105"→ prediction of *NOUN→ "18"→ prediction of *AUX→ "19"→ "19 64, "19 72" or "19 80"→ "19 65", "19 73" or "19 81"→ prediction of *VERB. For convenience sake, we call this sequence "grammar path". The recursive algorithm for parsing or prediction is given below: (PARSER)

1. Enter the given grammar path into the "path list".

2. If the path list is empty, stop. Otherwise, select a grammar path from the path list. Increase the number of the most right hand side in the grammar path by 1. This number indicates the next processed position in the grammar.

   If the variable on this position is a terminal symbol, predict the word (terminal symbol) and generate the grammar path. Then go to step 2.

   If the variable on this position is a word class with the affix "*", predict the set of words for the word class and generate the grammar path. Then go to step 2.

   If the variable on this position is a non-terminal with the affix "@", the production rules with the same non-terminal at the left hand side are predicted, that is, the head positions of these rules are concatenated at the most right hand side of the grammar path. Enter these paths into the path list. Then go to step 2.

If the variable on this position is empty, eliminate the number of the most right hand side in the grammar path and enter this path into the path list. Then go to step 2. In this procedure, we should pay attention to the representation of left recursion in a production rule, e.g.,@NP3→ @NP3 @PP→ @NP3 @NP3 @PP ···. Therefore we must restrict the application of such a rule. Although the times of the application are generally restricted , we restrict the length of the grammar path.

Next, we extend this parsing/prediction algorithm with the time-synchronous context-free parsing algorithm. We already proposed the basic idea, which combined the word spotting algorithm with a syntactical constrained connected word recognition algorithm[11]. We call it "Augmented Continuous DP Matching Algorithm". In the literature, the syntactic knowledge was represented by a finite state automaton or regular grammar. This algorithm was time-synchronous in terms of the ending frame of spotted words [Backward Algorithm W*]. We proposed another efficient time-synchronous parsing algorithm, S* [Forward Algorithm, [9]], which was time-synchronous in terms of the ending frame of generated partial sentences. This algorithm executes the prediction of words at the right-hand side of a partial sentence and the concatenation of a spotted word (candidate word) at the same time.

If all possible partial sentences are taken into consideration, the computation time or necessary memory space will become large. Therefore, we select a few best partial sentences and abandon the others. We proposed this pruning technique in the LITHAN speech understanding system [1]. In general, this search technique is well known as "beam search [2]". The number of partial sentences which cover the same range of the utterance, is restricted to less than a pre-set value, that is, the width or radius of the beam search. The sentence with the highest score which covers the whole of the utterance, is decided as the recognition result.

This forward algorithm S* is described in brief below.

[1] At the initial step, the partial sentence is set to "empty", and i=1.

[2] The algorithm predicts words for a partial sentence which covers the start to the i-th position of the input sentence.

[3] The algorithm expands the partial sentence (concatenation of the partial sentence and a predicted word which is found in the word lattice) and sorts the expanded partial sentences in terms of corresponding scores.

[4] If "i" is the last position in the input sentence, the best word sentence is regarded as the recognition result. Otherwise, i=i+1 and go to step [2]. Figure 3 illustrates this procedure.

3. EXPERIMENTAL RESULTS with SPOJUS-SYNO

3.1. Speech material and feature parameters

Six male speakers uttered 216 words, 80 loan (foreign) words, and 70 sentences in a soundproof room, respectively. These words and 20 sentences were segmented into syllable units by inspection and were used for training syllable-based HMMs. The other fifty sentences as test data were related to the content of "Question or Demand for Electronic Mail", which was a part of the task of UNIX-QA. The speed of utterances ranged from 8 to 9 morae per second (about 16 to 18 phonemes per second). This is moderately fast. The utterances were sampled/digitized with an accuracy of 12 bits / sampling by 12 kHz and analyzed using a 14 order LPC. We obtained 14 LPC cepstrum coefficients and signal power for every 5ms. These coefficients were transformed to 10 LPC mel-cepstrum coefficients. The vocabulary size of a part of the task is 521 words. The related sentences are generated by the context-free grammar which is represented by 534 rewriting rules, 259 non-terminal symbols, 268 word classes (a kind of non-terminal symbols), and 600 direct rewriting rules from word classes to terminal symbols. The average branching factor is about 26 (static) or 14 (dynamic). The perplexity is about 10.0 [12]. The number of plausible sentences in this subtask is about 10 to the power 37. Experiments were performed in multi-speaker mode and speaker-adaptation modes with speaker adaptation using isolated words and/or sentences.

3.2. Word spotting results

Table 1 shows the evaluation results of the word spotting performance after speaker adaptation using spoken sentences. The rate in the column of the n-th rank shows the percentage accuracy with which an input word is correctly identified as one of the best n spotted words in the neighborhood. The number of missing words denotes the total number of undetected input words in the total of 50 sentences. In comparison with the simulated of phoneme recognizer (see Table 2), our syllable HMM-based word spotting is superior to the performance of the word spotting of the simulator in the case of 80% phoneme recognition rate.

3.3. Sentence recognition results

The sentence recognition results are summarized in Table 3. For six male speakers, our system obtained an average sentence recognition rate of 60% in multi-speaker mode and 69% and 75% in speaker-adaptation mode using isolated words and sentences, respectively. One third of the errors were caused by confusion between semantically similar prepositions, such as "wa" and "ga", "ni" and "he" or "mo" and "o", and similar nouns, such as "messages" and "dengon". Therefore, a preposition word and an alternative preposition word combined with the preceding noun were spotted again, and a more reliable word detection was performed. Using this mechanism, the sentence recognition rate improved to 85%. This improvement suggests that a one-pass algorithm directed by a context-free semantic grammar may improve the sentence recognition accuracy. The sentence understanding rate was about 87% in speaker-adaptation mode using spoken sentences.

Table 1 Evaluation of word lattice by SPOJUS.

upper: multi-speaker mode,
middle: speaker-adaptation mode using isolated words,
lower: speaker-adaptation mode using sentences.

speaker	detection rate(%)				missing	average number of
	top1	top2	top5	top10	(total)	spotted words
SN	32.8	46.5	63.0	70.6	8	6962
	38.9	54.1	69.2	77.0	6	7763
	49.0	61.9	73.4	80.1	5	6962
TI	39.0	57.5	79.8	86.8	2	6395
	63.6	74.8	84.5	89.1	2	6077
	66.3	79.2	88.6	93.8	2	4818
HU	24.1	44.2	65.2	78.7	4	8278
	57.1	68.6	80.5	89.3	1	7724
	54.6	69.8	81.1	90.2	3	6987
KO	31.4	53.0	71.3	81.7	4	7087
	56.3	68.3	82.3	88.0	4	6314
	61.4	74.6	85.9	91.3	4	5762
MA	21.0	35.5	60.2	73.4	2	6436
	44.9	57.7	75.0	83.6	2	6049
	62.1	74.5	87.1	91.1	1	5323
SE	38.9	59.2	82.4	89.3	2	9086
	67.1	79.0	87.1	92.2	1	8588
	75.2	83.4	91.5	95.0	1	7645
average	31.0	49.0	70.0	79.7	3.7	7374
	54.1	66.7	79.5	86.3	2.6	7086
	61.4	73.9	84.6	90.3	2.7	6249

Table 2 Evalution of the lattice obtained by simulation
(phoneme recognition rate 80%, omission, insertion error rate 5%).

rank	detection rate(%)				missing	average number of
	top1	top2	top5	top10	(total)	spotted words
simul	53.5	66.7	78.6	85.5	4words	3930

Table 4 shows the sentence recognition rates for the word lattice obtained from the simulated phoneme recognizer (see Table 2). We find that our system is superior to the simulated recognizer.

Table 3 Sentence recognition / understanding rates by SPOJUS-SYNO (beam search width=45).

speaker	adaptation - mode			
	multi-mode	by words	by sentences	understanding
SN	50.0%	50.0%	64.6%	70.8%
TI	74.5	78.7	89.4	93.7
HU	63.0	71.7	76.1	91.3
KO	63.0	71.7	78.3	84.7
MA	65.3	79.6	85.7	93.8
SE	46.7	60.0	86.7	86.7
average	60.5	68.7	80.1	86.8

Table 4 Sentence recognition results by simulation (phoneme recognition rate=89%).

beam width	sentence recognition rate
40	52.0%
80	62.0%

4. SPOJUS-SEMO[13]

In this section, we describe our speech understanding system,which consists of three stages (see Fig.4). In Japanese there are four levels of hierarchy: syllables, words, phrase ("bunsetsu" in Japanese), and sentences. At the first stage, candidate words are recognized by concatenating HMM-based syllables and by checking a word dictionary. In this stage, a word lattice is made from words recognized with good scores. At the second stage, phrases are recognized using automaton-controlled Augmented Continuous DP matching [11]. This automaton represents the Japanese intra-phrase grammar. In this stage, a phrase lattice is made for the next stage. Finally, the best phrase sequence is selected by a backward KAKARI-UKE parsing algorithm that uses a dependency grammar between Japanese phrases and a case grammar among phrases.

The sentence recognizer gets the sentence recognition result by selecting phrases in the phrase lattice and concatenating them. Our system uses the backward KAKARI-UKE parsing algorithm that we proposed before[14]. In this algorithm, partial sentences are hypothesized and grown from right to left. In Japanese, verbs are located at the last phrase of the sentence. If we make partial sentences from left to right, we cannot use a dependency grammar for every phrase of the partial sentence because Japanese phrases have a dependency on other phrases on the right side (see Fig.5). For example, if the i-th phrase is the object of the j-th phrase (i<j), the dependency grammar cannot work until the analysis reaches the j-th phrase. This is the reason why our system makes partial sentences from right to left.

We define semantic features or markers to represent the meaning of the nouns. Every noun has one or more semantic features. For example, the word 'printer' has two semantic features, 'con' and 'sys', which represent a concrete thing and a part of a computer system. Every verb has some syntactic structure and semantic restrictions to objects. Although it is called valency grammar, it uses semantics. In a word dictionary this kind of information is described as follows:

```
* { hum } send { mail }[ to ]{ hum , loc }
* { hum } tell { hum }[ that ]{ s }
* { hum } write { mail }
```

In this example, the semantic features 'hum', 'mail', 'loc', and 's' represent human, mail, location, and sentence, respectively.

We define four key words related to the UNIX mail task. They are 'mail', 'message', 'file', and 'command'. Actually forty-four sentences of a total of fifty test sentences contain one or more key words. If a speech understanding system has detailed information about key words it is easy to prefer the sentences that contain those words. And this helps the system produce only meaningful sentences.

5. COMPARISON OF THE PERFORMANCES OF SPOJUS-SYNO AND SPOJUS-SEMO AND DISCUSSIONS

Table 5 summarizes the sentence recognition results with SPOJUS-SYNO and SPOJUS-SEMO using common word lattices for 300 sentences uttered by five male speakers. Three different kinds of context-free grammar for the SPOJUS-SYNO were used. From this table we can conclude that the recognition performance depends on the perplexity and the syntax and semantic grammar-driven parser is superior to the semantic-driven parser. Detailed descriptions and discussions are available in the literature [16,17]. Recently we also developed the SPOJUS-SYNO-X, which was based on a one-pass Viterbi algorithm directed by a context-free semantic grammar, and obtained a sentence recognition accuracy of 90%, as expected in section 3.3 [18].

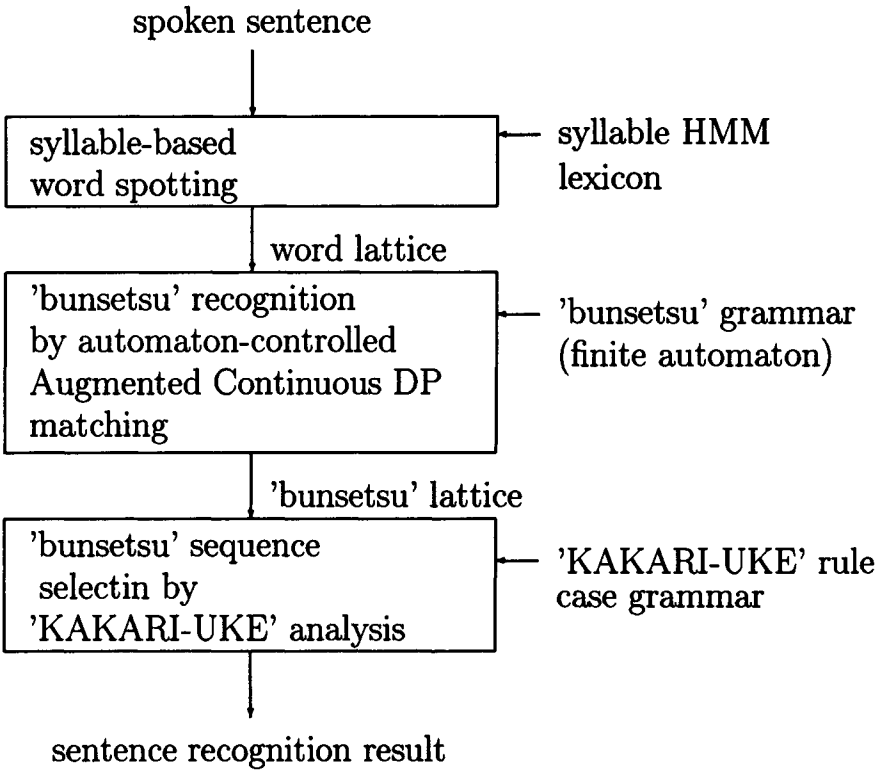


Fig.4. System Organization of SPOJUS-SEMO.

Bunsetu No.	1	2	3	· · ·	B-3	B-2	B-1	B

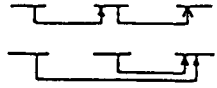
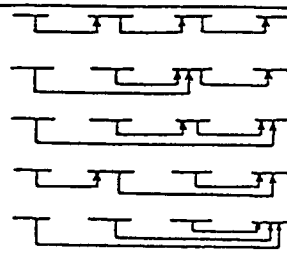
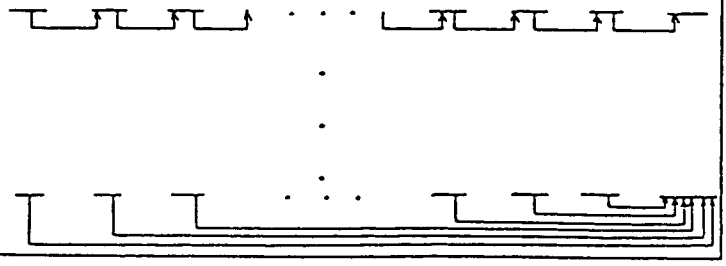
Step	Kakari-Uke Structure of Generated Partial Bunsetu Strings
1	
2	
3	
4	
· · ·	
B	

Fig.5. Generation of Kakari-Uke Structure by Backward Kakari-Uke Parsing.

Table 5 Sentence recognition results with SPOJUS-SYNO/SEMO

method	perplexity	sentence recognition rate	language model
SPOJUS-SYNO	10.0	68.7%	semantic grammar (CFG)
	25.4	41.8%	syntax and case grammar (CFG)
	50.7	27.4%	syntax only (CFG)
SPOJUS-SEMO	?	17.9%	dependency grammar and case grammar

Acknowledgement

We are indebted to Messrs. Y. Hashimoto and Y. Ohgruo for developing SPOJUS-SYNO and Messrs. T. Ito and A. Tanoue for SPOJUS-SEMO.

References

1. T. Sakai and S. Nakagawa, "Speech understanding system of simple Japanese sentences in a task domain," J. IECEJ, vol.60, No.1, pp.13-20 (1977)

2. B. T. Lowerre, "Harpy speech recognition system," PhD thesis Carnegie-Mellon University (1976)

3. S. Nakagawa and Y. Ohguro, "Comparison of parsing methods on continuous speech recognition," Proc. of Seech 88, 7-th FASE, pp.369-375 (1988)

4. W. H. Ward et al., "Parsing spoken phrases despite missing words," Proc. ICASSP, pp.275-278 (1988)

5. T. L. Chow, et al., "BYBLOS: the BBN continuous speech recognition system," Proc. ICASSP, pp.89-92 (1987)

6. K. F. Lee and H. W. Hon, "Large-vocablary speaker-independent continuous speech recognirtion using HMM," Proc. ICASSP, pp.123-126 (1988)

7. G.th. Niedermair, "Syntactic analysis in speech understanding," European Conf. Speech Technology, vol.1, pp.5-8 (1987)

8. S. Nakagawa, Y. Ohguro and Y. Hashimoto, "The syntax-oriented speech understanding system-SPOJUS-SYNO," Proc. Euro Speech, pp.224-227 (1989)

9. S. Nakagawa, "Speaker-independent continuous-speech recognition by phoneme-based word spotting and time-synchronous context-free parsing," Computer Speech and Language, vol.3, No.3, pp.277-299 (1989)

10. J. Earley, "An efficient context-free parsing algorithms," Comm. ACM. 13, 2, pp.9-10 (1977)

11. S. Nakagawa, "Connected spoken word recognition algorithm by constant time delay DP, O(n)DP and Augmented Continuous DP matching," Information Science, 33, pp.63-85 (1984)

12. S. Nakagawa, "An evaluation method for continuous speech recognition systems," Proc. ESCA workshop of speech I/O assessment and data-base, pp.4.9.1-4.9.4 (1989) or "Relationship among phoneme/word recognition rate, perplexity and sentence recognition, and comparison of language models," Proc. ICASSP, pp.I-589-592 (1992)

13. T. Ito and S. Nakagawa, "Sentence understanding of spoken Japanese using phrase spotting and dependency grammar," The second joint meeting of ASA and ASJ, pp.10 (1988)

14. S. Nakagawa and T. Ito, "Recognition of spoken Japanese sentence using monosyllable units and backword KAKARI-UKE parsing algorithm," Trans. Inst. Elect. Inf. Comm. Engrs, vol.70D, No.12, pp.2469-2478 (1987)

15. S.Nakagawa and Y.Hirata, "Comparison among time-delay neural networks, LVQ2, discrete parameter HMM and continuous parameter HMM," Proc. ICASSP, pp.509-512 (1990)

16. S.Nakagawa, Y.Hirata, I.Murase and T.Tanoue, "Comparison of syntax-oriented spoken Japanese understanding system with semantic-oriented system," Trans. Inst. Elect. Inf. Comm. Engrs, vol.E74, No.7, pp.1854-1862 (1991)

17. S.Nakagawa and I.Murase, "Comparsion of language models by context-free grammar, bigram and quasi/simplified-trigram," Trans. Inst. Elect. Inf. Comm. Engrs, vol.E74, No.7, pp.1897-1906 (1991)

18. A. Kai and S. Nakagawa, "A frame-synchronous continuous speech recognition algorithm using a top-down parsing of context-free grammar," Proc. Int. Conf. Spoken Language Processing, pp.257-260 (1992)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

An Application of Discourse Analysis to Speech Understanding

Yasuhisa Niimi and Yutaka Kobayashi

Department of Electronics and Information Science, Kyoto Institute of Technology
Matsugasaki, Sakyo-ku, Kyoto, Japan

Abstract

This paper describes a method for the discourse analysis performed in the speech dialogue system we are developing. The purpose of the analysis is to provide the system with top-down predictions. The predictions include words and syntactic rules likely to be used in the next utterance. Contextual information is analyzed in terms of topics and discourse goals. The transition of topics through a conversation is represented as an AND-OR tree of which the nodes correspond to topics. The prediction of topics is done by an expansion of the currently focused node. The structure of discourse goals is analyzed by a grammar as described in a context-free grammar. The terminal symbols of this grammar correspond to discourse goals of utterances. The top-down application of this discourse grammar hypothesizes discourse goals likely to appear in the utterance, each of which is translated into syntactic rules. The simulation of the dialogue system using typed input has proved that these top-down hypotheses reduce the vocabulary size effectively by about 60%.

1. INTRODUCTION

The recent advance in speech science and related technology has made it possible to build continuous speech recognition systems working in real time. Using such systems as an interface, we can construct man-machine dialogue systems [1,2]. In the speech dialogue system, discourse analysis, that is, the analysis of structures of dialogues, plays an important role in interpreting utterances.

This paper describes a method for the discourse analysis performed in the speech dialogue system we are developing [3]. The purpose of the discourse analysis is to provide the system with top-down hypotheses on words likely to appear in utterances of the partner in a dialogue.

The task performed by the man-machine dialogue is to make plans; for example, plans for seeing the sights of a city. The system is supposed to have a relational database about the sights of the city. A user (speaker) can access the system by voice and can collect information necessary to make plans. The speech dialogue system consists of three

components: a speech interface, a dialogue controller, and a planner. The speech interface recognizes utterances from users and passes their semantic interpretations to the dialogue controller. The dialogue controller analyzes the structure of the conversation and returns to the speech interface top-down expectations of what would be said next. The planner gives to the controller a guide line on how it should carry out a conversation with users as well as helps users to make plans by supplying information requested and suggesting some tour plans.

The discourse analysis is performed in terms of topics and discourse goals based on the work in ref.[4]. The transition of topics through a conversation is represented by an AND-OR tree. In this formulation the prediction of topics likely to be mentioned in the next utterance is equivalent to the determination of which node of the AND-OR tree is to be expanded next.

Each utterance in a dialogue has its own purpose (speaker's intention). The purpose of an utterance is called discourse goal. Since successive utterances are grouped to achieve a larger goal, discourse goals of a conversation can be described by a context-free grammar, which we call discourse grammar. The top-down analysis of a sequence of utterances (a partial dialogue) by this grammar predicts discourse goals the next utterances could express.

A simulation of the dialogue system using typed inputs has proved that these top-down hypotheses can reduce the search space for recognizing utterances by about 60%.

2. REPRESENTATION OF THE CONTEXT

Figure 1 illustrates an example of dialogues which the system could have with users. This example shows that the conversation is developing in two different modes: the

U101	I'd like to do a few day sights in Kyoto.
S101	Which are you interested in, temples, gardens, or traditional artifact?
U102	I've heard Kyoto is famous for gardens.
S102	I see. I introduce two day tour for visiting gardens. The gardens in what eras would you like to visit?
U103	Muromachi, Momoyama and Edo.
S103	O.K. Famous gardens in Muromachi era are the stone garden of Ryoanji temple, Kinkakuji temple and Ginkakuji temple.
U104	What is the stone garden?
S104	It is made of a few large stones and white sand only. It is believed that the simplicity of the garden is related to "Zen".
U105	That's great! Who built it?
S105	I'm sorry, it is not known.
U106	Who built Ryoanji?
S106	It's believed that Katsumoto Hosokawa did.

Figure 1. An example of a dialogue.

system-initiative and the user-initiative mode. In the system-initiative mode, the system, repeating questions to the user, elucidates the specification of his sightseeing tour, that is, the period of the tour, hotels to stay, places to visit, and so forth, and then offers some candidates for these items. In the user-initiative mode, the user, asking detailed information on these items, decides what is worth to involve in his plan. The questions issued by the user in this mode are interpreted as retrieval commands for the relational database. Responses of the system are generated based on the retrieved information.

2.1. Topic transition tree

It has been known [5] that topics in a goal-oriented dialogue move according to a task-dependent tree structure. In fact the topics in the illustrated example are specialized along the structure as shown in fig.2, which we call the topic tree. The nodes of the topic tree are entities related to the database such as names of relational tables, items included in the tables, and values of items. These correspond to topics the system can understand.

As we have reported in ref.[3], however, an AND-OR tree is more suited for representing movements of topics than a simple tree. In the AND-OR tree, which we call a topic transition tree, AND-nodes represent topics introduced by the user, and OR-nodes represent topics introduced by the system. If the user inquires about two or more sights (each assumed to be a topic), the system must offer information on all of them. On the other hand, even if the system proposes two or more candidates for a visit, the user is not interesting in all of them, and might move to the other topics. An AND-OR tree is suited to reflect this difference. The topic transition tree can be considered as a trace of a subtree of the topic tree.

2.2. Discourse goals

Each utterance in a conversation has its own purpose(speaker's intention). In the dialogue illustrated in fig.1, the utterance U101 inquires to make plans for sightseeing and

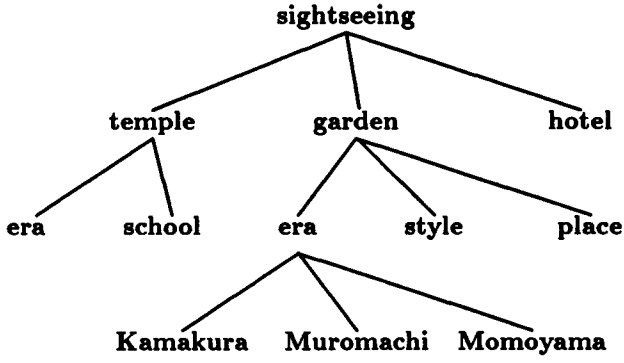


Figure 2. A part of the topic tree.

presents information on the period of the tour. The following four utterances, S102 to U103, have the larger purpose that the system tries to find out what places the user wants to visit. The next utterance, S104, proposes candidate sights to visit. In the utterances following S104, the user asks questions on those sights to judge which of them are worth visiting, and the system provides some information on them. Thus, the discourse goals in the dialogue form a hierarchical structure, like a tree.

This hierarchical discourse structure can be described by a context-free grammar in which the terminal symbols are discourse goals corresponding to a single utterance and the nonterminal symbols are larger discourse goals corresponding to a group of utterances. Figure 3 shows an example of the grammar for discourse analysis. Underlined strings indicate terminal symbols. For example, the terminal symbol 'prst-alt' (present alternatives) represents the discourse goal of an utterance, like S102 in fig.1, used to present multiple choices, and the terminal symbol 'slct-alt' (select alternatives) represents the discourse goal to select one of the alternatives presented. 'rqst-spec' (request for a part of specification) is a discourse goal of wh-questions issued by the system, and 'ans-spec' (answer a part of specification) works as an answer to 'rqst-spec'.

An utterance could have different discourse goals in different contexts, and a discourse goal could be expressed by various forms of utterances. Thus the relation between utterances and discourse goals is a many to many correspondence.

3. ANALYSIS OF THE DISCOURSE STRUCTURE

Figure 4 shows the flow of the discourse analysis. It involves bottom-up and top-down analyses. The bottom-up analysis performs the semantic analysis of an input utterance and then produces bottom-up hypotheses, that is, candidates for topics and discourse goals of the utterance. The top-down analysis predicts topics and discourse goals likely to appear in the current utterance referring to the context so far restored, which is represented by the AND-OR tree of topics and the parsing history of discourse goals. The

(1)	mk-plan	→	exm-spec, exm-plan
(2)	exm-spec	→	exm-spec, exm-spec | dcd-spec, exm-spec | dcd-spec
(3)	dcd-spec	→	prst-alt | prst-alt, slct-alt | prst-alt, chng-spec | rqst-spec, ans-spec | rqst-spec, chng-spec
(4)	exm-plan	→	exm-plan, exm-plan | {rqst-cand,} prst-cand, exm-plan-1
(5)	exm-plan-1	→	dcd-cand | exm-plan-1, exm-plan-1
(6)	dcd-cand	→	(apr-knwldg)*{, response}
(7)	aqr-knwldg	→	rqst-knwldg, ans-knwldg | recommend
(8)	response	→	accept | reject

Figure 3. A subset of the rules for the discourse analysis. { } indicates an optional term, and $(x)*$ indicates a null string or the repetition of x as many times as necessary. Rule (2-3), for example, refers to the second rule with the third alternative on the right-hand side.

bottom-up hypotheses are matched against the top-down predictions. The best match gives the interpretation of the utterance, which is preserved as contextual information.

3.1. Bottom-up analysis

The first stage of the bottom-up analysis is the semantic analysis of utterances, while the syntactic analysis of them is supposed to finish in the speech interface. It is performed based on the case grammar. Case frames associated with verbs are used to represent the meaning of sentences in the case grammar. They are described by a set of slots, each indicating one of the relations between the verb and a noun phrase, like an agent, object or instrument. The semantic analysis assigns noun phrases included in an utterance to some slot of the case frame of the main verb based on semantic markers of the noun phrases. The semantic interpretation of an utterance is represented by a list of four terms, a main verb, a case frame with slots filled, aspect information, and the style of a sentence.

The head nouns of the case slot fillers (noun phrases) are proposed as candidates for

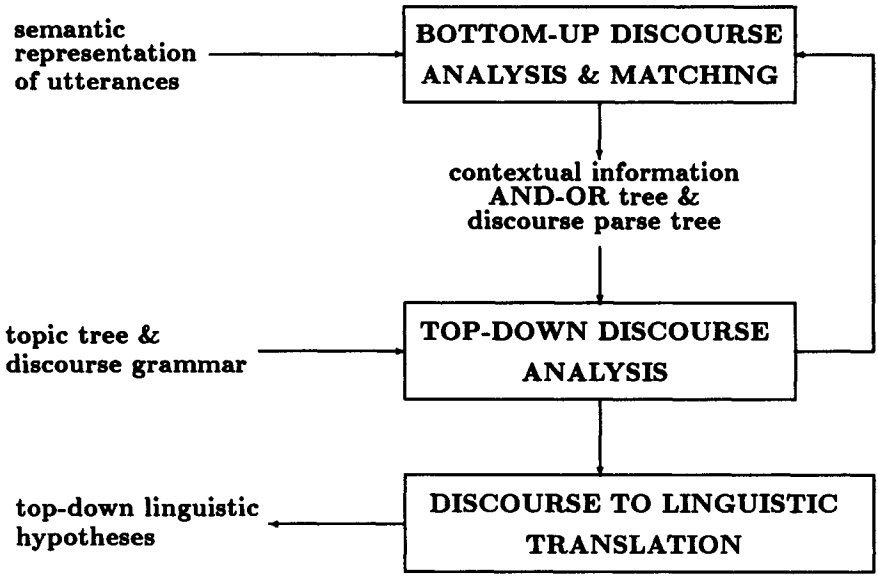


Figure 4. Flow of the discourse analysis

the topic of the utterance being analyzed. Those cases corresponding to topic, object and purpose are given a higher priority than others.

A lexicon is prepared to make bottom-up hypotheses on discourse goals. It contains the relation of a discourse goal with a verb, the aspect of the verb, and the style of a sentence including the verb. Bottom-up hypotheses on the discourse goal are built up by consulting this lexicon. It is generally difficult to uniquely determine the topic and discourse goal of an utterance only by the bottom-up analysis.

3.2. Top-down analysis

As mentioned in section 2.2, a grammar for the discourse analysis is formulated by a context-free grammar. Thus an analysis of the conversation so far carried out results in a tree structure. Leaves of a discourse parse tree correspond to utterances. Figure 5 shows an example of a top-down discourse analysis. It illustrates a discourse parse tree resulting from the utterances U101 and S102 shown in fig.1 and discourse goals possible to be expressed by an utterance following S102. The rules (1), (2-1), (2-3), (3-1), (2-2), {(3-2) or (3-3)} are used in the analysis of utterances U101 and S102. The last stage of the rule applications is ambiguous and incomplete. It is ambiguous in that there are two applicable rules, (3-2) and (3-3), and incomplete in that all terminal symbols of the right-hand sides of the applied rules are not used. Thus the next discourse goal would be 'slct-alt' if the rule (3-2) is applied and 'chng-spec' if the rule (3-3) is applied.

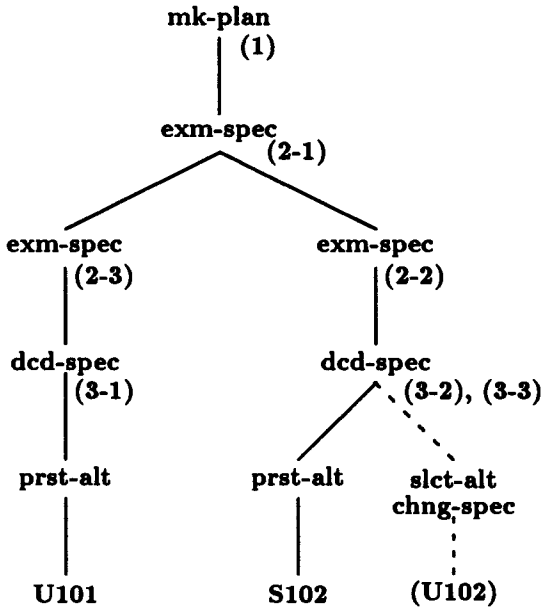


Figure 5. An analysis of a discourse structure and a prediction of discourse goals.

The movement of topics through a conversation is stored in the topic transition tree. Thus, for the top-down prediction of topics it is necessary to determine which node of this tree is to be selected or expanded next. New nodes which will result from the expansion can be found by referring to the topic tree because the topic transition tree is considered a trace of a subtree of the topic tree. In this connection, the concept of a focus is very important. The focus means the topic currently focused. We define it as the last of the topics which either the system or the user has uniquely mentioned.

Which node is to be selected or expanded depends on the discourse goal of the next utterance. Figure 6 shows the relation between discourse goals and topics in the top-down prediction. In order to shift the focus to nodes at higher levels than the currently focused node, it is necessary that all the topics under the new focus have terminated. The conditions for a topic to terminate can be stated as follows. (1) A node with AND successors can terminate only when all the successors have terminated. (2) A node with OR successors can terminate when one of the successors has terminated.

3.3. Translation of discourse hypotheses into linguistic hypotheses

A topic can be expressed by several words. For example, the topic 'garden' can be expressed by two Japanese words, 'niwa' (Japanese origin) and 'teien' (Chinese origin). So we have a table by which real words can be looked up from conceptual topic words. By using this table top-down hypotheses on the topic are translated into words likely to appear in utterances.

As mentioned in section 3.1, we have the lexicon describing the relation between a verb and discourse goals which the verb can express. This lexicon is also used to translate top-down hypotheses on the discourse goal into linguistic ones. First, a set of verbs capable of expressing a hypothesized discourse goal is found by consulting this lexicon, and then semantic categories of those nouns which can occur together with these verbs are obtained by looking up case frames of these verbs.

discourse goals	node to select or expand
slct-alt	select one of successors of the focus.
ans-spec	expand the focus.
chng-spec	expand a node at a higher level or select and expand one of successors of the focus.
rqst-cand	move the focus to a node at a higher level.
rqst-knwldg	expand the focus.

Figure 6. The relation between discourse goals and topics in the top-down prediction.

4. THE EFFECT OF THE DISCOURSE ANALYSIS

The speech dialogue system reported here has not been completed. So we simulated the dialogue system using typed sentences in order to measure the effect of the discourse analysis described in the previous sections. The dialogue controller accepted a dialogue consisting of typed sentences and generated top-down discourse hypotheses every time a sentence was input. Using the linguistic constraints translated from these hypotheses the linguistic processor of the speech interface analyzed the sentence following the input one, and predicted words possibly following each word of the analyzed sentence.

The average number of predicted words, a kind of branching factor, was computed as a measure of the effect of the discourse analysis on the speech recognition. Assuming the vocabulary consist of about 600 words of which about 360 are nouns, a conversation composed of 60 sentences was analyzed. The average number of predicted words was 240. This means that the discourse analysis has reduced the vocabulary size by 60%.

5. CONCLUSION

The method for the discourse analysis performed in the speech dialogue system we are developing, has been reported. The contextual information is analyzed in terms of topics and dialogue structures. The discourse analysis involves bottom-up and top-down analyses. The bottom-up analysis proposes candidates for topics and discourse goals based on the semantic representation of an utterance being analyzed. The top-down analysis makes hypotheses on topics and discourse goals referring to the contextual information. The bottom-up hypotheses are matched against the top-down hypotheses. The best match gives the interpretation of an utterance.

The top-down hypotheses are translated into hypotheses at the linguistic level, which are given to the speech recognition system. A simulation of the dialogue system using typed sentences has proved that these hypotheses can reduce the search space for recognizing utterances.

References

1. S. J. Young and C. E. Proctor, "Computer Speech & Language," 3, pp.329-353, (1989)

2. S. R. Young, et al., Com. ACM, 32, pp.183-194, (1989)

3. Y. Niimi and Y. Kobayashi, "Preprints of the Second Symposium on Advanced Man-Machine Interface Through Spoken Language," pp.33.1-33.8, (1988)

4. J. B. Grosz and C. Sidner, "Computational Linguistics," 12, pp.175-204, (1986)

5. J. B. Grosz, "Discourse knowledge," in: D. E. Walker (eds.), Understanding Spoken Language, pp.229-337, (North-Holland, New York, 1978)

Chapter 6

SPEECH SYNTHESIS

This Page Intentionally Left Blank

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Studies on Glottal Source and Formant Trajectory Models for the Synthesis of High Quality Speech

Satoshi Imaizumi and Shigeru Kiritani

Research Institute of Logopedics and Phoniatrics
Faculty of Medicine, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

Abstract

This paper describes a polynomial voice source model and a formant trajectory model for a Hi-Fi speech synthesizer. The voice source model represents time derivative of the glottal volume velocity waveform as a polynomial function. The formant model describes the formant trajectories as the summation of temporal functions: a second order delay function which represents vowel-to-vowel transitions, and two first order delay functions which represent the effects of surrounding consonants on the vowel formant trajectories. The models were tested through perceptual experiments for synthetic speech at slow and fast speaking rates. Results suggest that the models work well particularly at slow rates. Some additional strategies seem to be needed to improve the intelligibility of consonants at fast rates.

1. POLYNOMIAL GLOTTAL SOURCE MODEL

Although techniques for speech synthesis by rules have been significantly improved, synthesis of natural sounding speech with various voice qualities still remains a seemingly unattainable goal. Many researchers have been trying to reach this goal by developing voice source models by which intra- and inter-speaker variability in voice quality can be controlled [1-7].

For instance, Fant et al. [2,3] have introduced a four parameter model describing the time derivative of the glottal volume velocity waveform and have tried to synthesize female voice quality with high fidelity. Fujisaki and Ljungqvist [4] have proposed a seven parameter model, which might have more flexibility than other glottal source models. On the other hand, Klatt [5] and Hasegawa et al. [6], have insisted that an additive noise component must be included into the glottal source model to synthesize female voice quality with sufficient naturality. Although these studies provide a fruitful discussion on advanced techniques of Hi-Fi speech synthesis, there are few results reported on how naturally and how variously the voice quality can be reproduced by these glottal source models.

In this study, we examined how naturally and how variously a seven-parameter polynomial model can represent the voice quality of five male and two female speakers.

1.1. Method

1.1.1. Data recording

The following speech materials were recorded and analyzed.

(1) Sustained vowels and vowel sequences.

(2) Three sentences consisting only of vowels and semi-vowels.

These materials were recorded from 5 male speakers M_1, M_2, .. M_5, and 2 female speakers F_1 and F_2, who had no laryngeal pathology. Each speaker uttered each item three times at three loudness levels and at three pitch levels. The speech signal was recorded on a PCM Data Recorder using a high quality condenser microphone (B&K2234) whose frequency characteristic was flat (within 1dB) in the range 10Hz to 10kHz. An electroglottogram (EGG) [8] was also recorded simultaneously to indirectly observe the vocal cord vibration. The EGG signal yielded the glottal closure intervals which were used for a pitch-synchronous covariance LPC analysis [7,8].

The speech material reported here is vowel /a/ uttered at normal pitch and normal loudness for each speaker.

1.1.2. Inverse filtering

In order to estimate the glottal volume velocity waveform, formants were estimated based on a covariance LPC analysis with pitch synchronous frames corresponding to the glottal closure intervals derived from the EGG signal [7,8]. The glottal closure intervals were derived in the same way reported in other sources [7], that is, an interval was determined as to begin at a positive peak in the EGG time derivative and the end at the following negative peak, the length being $T_a(n)$ for the nth pitch period. The beginning of the actual analysis frame was shifted by a time dt later, according to the time delay it took for the sound wave to propagate from the glottis to the microphone, positioned 15cm away from the lips.

Because the formant trajectories obtained in this manner sometimes revealed cycle by cycle fluctuations, especially for the female voice, the formant frequencies and bandwidths were modified manually using an interactive program. This program displayed the speech waveform and its power spectrum, the inverse filtered waveform and its power spectrum, and the EGG time derivative, which indicated the glottal closure intervals. The optimal formant frequencies and bandwidths were searched manually so as to minimize ripples in the inverse filtered waveform during the glottal closure intervals and also the formant-like peaks in their power spectrum.

The time derivative of the glottal volume velocity waveform was estimated via inverse filtering in which only one set of the lower five formant frequencies and bandwidths selected from a steady portion of each utterance was used. In other words, cycle by cycle variation in formant trajectories was avoided.

1.1.3. The parameter estimation of the glottal source model

The inverse filtered waveform, or time derivative of the glottal volume velocity waveform, was approximated in each cycle by the following polynomial function,

$$\begin{aligned} g(t) &= a(t-t_1)^2 + b && 0 < t < t_1 \\ &= b && t_1 < t < t_2 \\ &= c(t-t_1)^3 + d(t-t_1)^2 + e(t-t_1) + b && t_2 < t < T \end{aligned} \tag{1}$$

where $t = 0$ is the negative peak in the inverse filtered waveform, and $t = T$ is the duration of one pitch. The parameters t_1, t_2, a, b, c, d, and e were determined based on the least-squares error criterion for the actual inverse filtered waveform $g_i(t)$ and the model $g(t)$. One example, from a female speaker, is shown in Fig. 1.

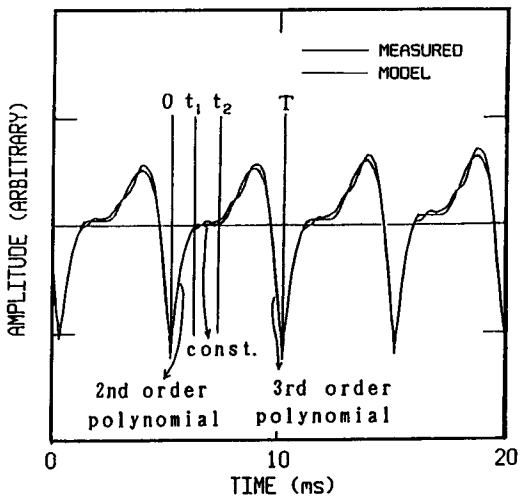


Figure 1. The polynomial model of the glottal source adapted to a measured glottal source waveform obtained by inverse filtering /a/ uttered by F_2.

1.1.4. Perceptual experiments

Three perception experiments were performed to examine how naturally and how variously the voice quality could be reproduced by the polynomial model of the glottal source. The subjects were 6 students with normal hearing capacity.

Experiment I was carried out to examine how closely the voice quality of the original vowel was reproduced by the polynomial model of the glottal source. The subjects rated the degree of resemblance between the original vowel and the vowel synthesized using the polynomial model. For the sake of comparison, they also rated the resemblance between the original vowel and the vowel synthesized using Rosenberg's Type B model of the glottal source [1]. The rating was performed in a paired comparison method using a scale

with 7 successive categories, 1:completely different, 2:very different, 3:different, 4:neutral, 5:similar, 6:very similar, 7:perfectly the same.

Experiment II was carried out to examine how variously the voice quality of vowels uttered by five male speakers M_1, .., M_5 were reproduced by the glottal model using a multi-dimensional scaling method [9,10]. Five vowel samples of 0.5 s in length, O_1, O_2, .. , O_5, corresponding to the five male speakers M_1, M_2, .. , M_5, were resynthesized using one pitch interval extracted from the inverse filtered waveform. Then, using one pitch interval from the polynomial model of the glottal source adapted to each vowel, five vowels G_1, G_2, .. , G_5 having a length of 0.5 s were synthesized. The pitch and its fluctuation were the same for all samples as those observed from /a/ uttered by M_1. The constant intervals corresponding to the glottal closure periods were lengthened or shortened to align the pitch for all samples.

The listening subjects rated the dissimilarity in voice quality for each of all the possible pairs of O_1, O_2, .., and G_1, G_2, .. , G_5. The ratings on dissimilarity were then analyzed using the multidimensional scaling method INDSCAL, included in the ALSCAL program [10], and the similarity among these 10 vowel samples was represented by the mutual distance in two-dimensional space.

Experiment III was carried out to examine the perception effects of fluctuations in the waveform (W), pitch (P) and amplitude (A) of the glottal source on the naturalness of the synthetic vowels. Five kinds of synthetic vowels — P_1, P_2, .., P_5 — were generated containing various fluctuations observed in the original vowel P_0. P_1 contained W+P+A; P_2:P+A; P_3:P; P_4:A; and P_5:no fluctuation. Here, waveform variation means the cycle to cycle variation in the modeled glottal voice source. The pitch fluctuation was the cycle to cycle variation in the intervals between negative peaks in the inverse filtered vowel waveform. The amplitude fluctuation was the cycle to cycle variation in the amplitude of the negative peaks in the inverse filtered vowel waveform.

All possible pairs of P_0, P_1, .. , and P_5 were made and presented to the listeners in random orders. Each listener selected the one member of each pair felt to be more natural than the other.

1.2. Results and Discussion

1.2.1. Experiment I

The results of the perception judgments on the degree of resemblance between the original vowels and the synthetic vowels are shown in Fig. 2. The samples used were /a/ uttered by five male speakers and 2 female speakers. The symbol G indicates the vowel synthesized using the polynomial model, and R represents the one synthesized with Rosenberg's glottal source model.

As shown in Fig. 2, for all speakers the ratings for the synthetic vowels synthesized with the polynomial model of the glottal source (G) are higher than those for the vowels synthesized with Rosenberg's model (R). This result shows that the polynomial model of the glottal source is better than Rosenberg's model at reproducing the voice quality of the vowels for which glottal source models are adapted.

For the vowels uttered by the male speakers, M_1, M_2, .., M_5, the medians of the ratings for the polynomial model scatter between 5:similar and 7:perfectly the same.

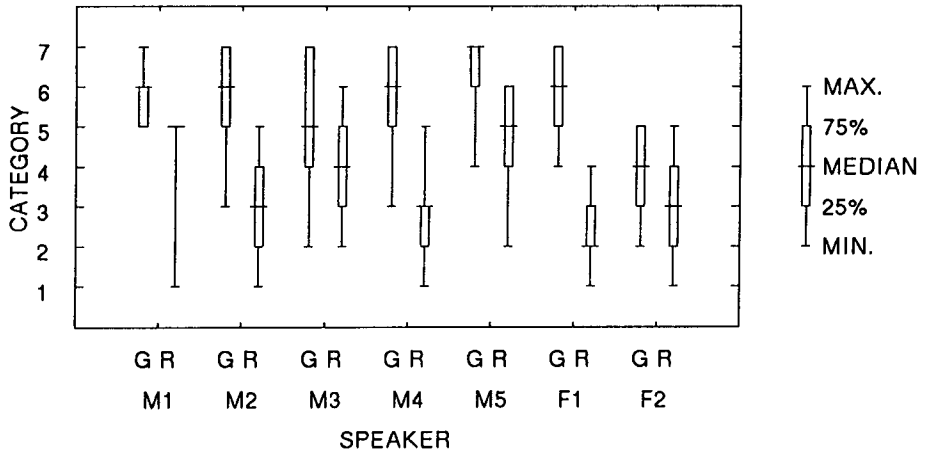


Figure 2. The results of perception judgments on the degree of resemblance between original vowels and synthetic vowels with the polynomial model of the glottal source (G), and that between original vowels and synthetic vowels with Rosenberg's voice source (R). Category 7 represents the greatest possible resemblance.

Those for Rosenberg's model lie between 3:different and 5:similar. This result indicates that the polynomial model of the glottal source can reproduce the voice quality of the male speakers analyzed here.

For the female speaker F_1 the median of the rating scores for the polynomial model is 6:very similar, while the median of the ratings for Rosenberg's model is 2:very different. On the other hand, for the female speaker F_2 the median of the ratings for the polynomial model is 4:neutral, and the median for Rosenberg's model is 3:different. These results indicate that the polynomial model of the glottal source can reproduce some female voice qualities. Figures 3(a) and 4(a) show the inverse filtered waveform and its model representation for F_1 and F_2, respectively. Figures 3(b) and 4(b) show their power spectra. The polynomial model of the glottal source for F_1 reproduces the voice quality of the original vowel very well, while that for F_2 does not.

In Fig. 4(a), the inverse filtered waveform or the measured glottal source have positive main lobes which skew right, and this characteristic is not represented well enough in the model. The intervals which are approximated by the constant b in the model contain waveform fluctuations in the measured glottal source. The negative peaks in the model source are too sharp compared to those of the measured glottal source. In Fig. 4(b), the harmonics higher than 2kHz in the power spectrum of the measured glottal source are not clear. On the other hand, the model shows clear harmonics for the higher range than 2 kHz. These discrepancies are not so large in F_1 as shown in Fig. 3, although the waveform fluctuation in the intervals which are approximated by constant b in the model are not approximated well.

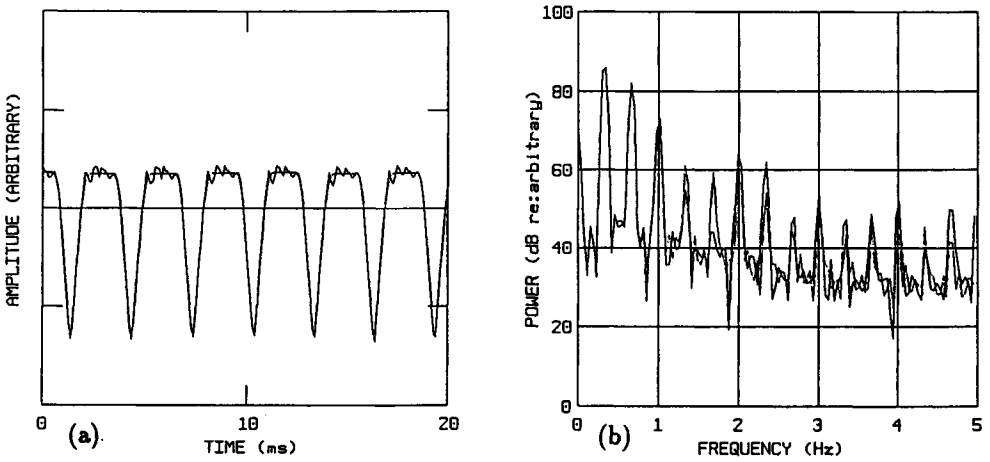


Figure 3. The measured glottal source waveform and its model representation (a), and their power spectra (b). Female speaker F_1.

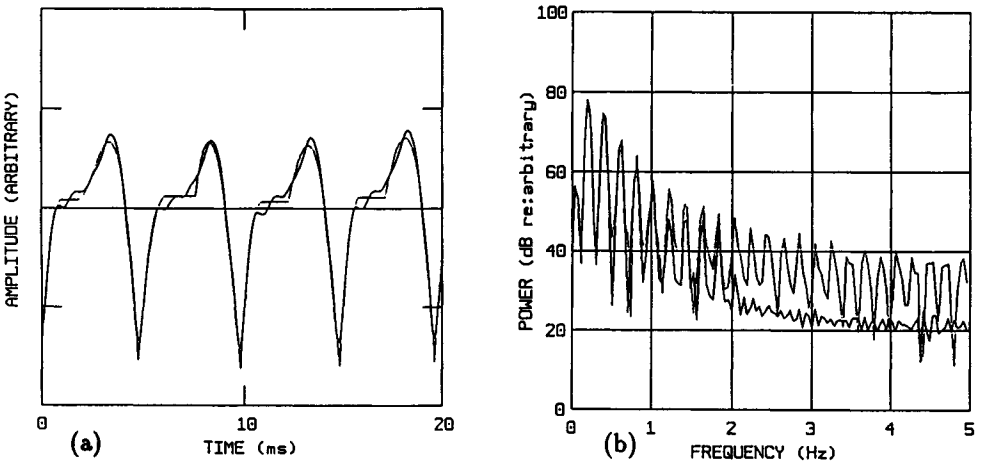


Figure 4. Same as in Fig. 3, but for female speaker F_2.

The skewing and waveform fluctuation observed in Fig. 4(a) might be effects of the source-tract interaction [11-13]. The disappearance of harmonics higher than 2 kHz might be due to the turbulence noise. These effects are not approximated in the polynomial model of the glottal source, thus the voice quality of F_2, which reveals these effects clearly, cannot be reproduced with high fidelity.

1.2.2. Experiment II

The result of Experiment II is shown in Fig. 5. In this figure, the similarity between the 10 synthetic samples were represented by their mutual distance in two-dimensional space.

Figure 5 shows that there are three types of similarity between O_n synthesized from the inverse filtered waveform and G_n synthesized from the model. Here, n indicates the speaker number. Type 1: for M_1, O and G are relatively close. Type 2: for M_2, M_4 and M_5, O and G are close in Dimension D_1, but distant in D_2. Type 3: for M_3, O and G are distant in D_1, but close on D_2. This result indicates that the voice quality of each speaker has various aspects, some of which can be reproduced by the polynomial model of the glottal source, and some of which cannot.

Figure 5 also shows that the voice samples O_n, resynthesized from the inverse filtered waveform, scatter in two-dimensional space, while G_n, resynthesized from the model, scatter in a one-dimensional manner on the line S_1 and separate into two groups G_2, G_3 and G_4 versus G_1 and G_5. In other words, the two-dimensional variability of the voice quality is maintained in O_n, but is reduced to one dimension in G_n.

These results must be interpreted through an examination of the acoustical and perceptual meanings of dimensions D_1 and D_2, or S_1 and S_2. According to our preliminary examination, S_1 may indicate the contrast between "strained" versus "asthenic" voice quality, or in another definition, a "hyper-fuctional/tense" versus "hypo-functional/lax" quality. G_1 and G_5 have stronger harmonics in the high frequency range than the others. On the other hand, S_2 may indicate a "breathy/noisy" versus "rough" quality. These results indicate that the polynomial model of the glottal source can reproduce the voice quality represented by S_1, but not that represented by S_2.

1.2.3. Experiment III

Figure 6 shows the results of Experiment III, which was carried out to examine the perception effects of fluctuations in the waveform (W), pitch (P) and amplitude (A) of the glottal source upon the naturalness of the synthetic vowel. In this experiment, five synthetic vowels — P_1, P_2, .., P_5 — were generated containing various fluctuations observed in the original vowel /a/, P_0, uttered by F_1. Then, all possible pairs of P_0, P_1, .. , P_5 were presented to four listeners in random orders. Each listener selected the one from each pair which was felt to be more natural than the other. The selection rate for the six samples is shown in Fig. 6.

As shown in Fig. 6, the original voice sample P_0 was selected as most natural. Although there were slight differences between the listening subjects, P_1, which contained fluctuation in waveform, pitch and amplitude, was selected as second in naturalness. P_2, which possessed fluctuation in pitch and amplitude, had almost the same selection rate

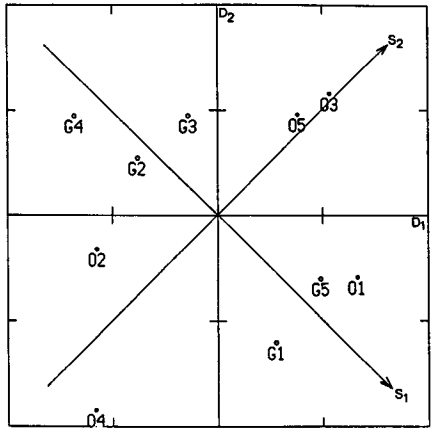


Figure 5. Two dimensional representation of the similarity between vowels resynthesized from the inverse filtered waveform O_n and those from the polynomial model G_n. Here, n indicates the speaker number M_n, $n = 1, 2, .., 5$. D_1 and D_2 are the dimensions extracted by the INDSCAL analysis, while S_1 and S_2 are their rotated version to interpret the configuration.

as P_1. Although P_3, containing only pitch fluctuation, had a lower selection rate than P_1 and P_2, it showed a higher rate than P_4, which possessed only amplitude fluctuation and P_5 which had no fluctuation.

This result indicates that fluctuation in pitch, amplitude and waveform affects the naturalness of synthetic vowels in this order. Proper modeling of the pitch fluctuation is quite important, because synthetic vowels without any pitch fluctuation here sound quite unnatural. On the other hand, waveform fluctuation in the glottal source did not largely affect the naturalness compared to pitch fluctuation in this study. However, the effect of waveform fluctuation on the naturalness might have been underestimated in this study, because a cycle by cycle estimation of the model parameters sometimes emphasizes waveform variation, which may generate a hoarse-like voice quality.

1.3. Conclusions

The present study gave the following results.

(1) For male voices, the polynomial model of the glottal source can reproduce to some extent the voice quality of the original vowels for which the model parameters are adapted. In a simple paired comparison based on a successive category method, Experiment I, the degree of resemblance between the original vowel and a synthetic one using the model was quite high. However, a detailed examination of the voice quality based on the multi-dimensional scaling method, Experiment II, showed that some aspects of the voice quality still remain unrepresented in the model.

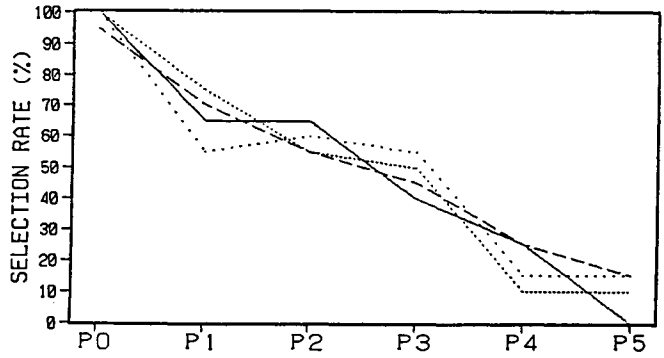


Figure 6. The rate of selection as the more natural vowel in paired comparisons by four subjects. The tested fluctuations were waveform variation (W), pitch fluctuation (P) and amplitude fluctuation (A). P_0: original vowel /a/ uttered by F_1; P_1:synthetic vowel which contained W+P+A; P_2:P+A; P_3:P; P_4:A; P_5:no fluctuation.

(2) For voices which contain turbulence noise in the high frequency range, and those which contain waveform perturbation and skewing possibly caused by source-tract interaction, the polynomial model fails to reproduce good voice quality.

(3) Proper modeling of pitch fluctuation is important for the naturalness of the synthetic voice.

2. FORMAT TRAJECTORY MODEL WITH VARIABLE SPEAKING RATE

In order to improve the quality of speech generated by a formant synthesizer, several models describing formant trajectories have been proposed. For instance, some studies have used smoothed step functions [14-17], where the step inputs represent putative targets of vowels [15-17] or even of consonants [16]. Some studies propose a linear summation model of the target formant frequencies of vowels and temporal functions representing the effects of adjacent consonants [18,19]. Although these models seem able to describe some phenomena in formant trajectories, for instance the undershoot at fast speaking rates, there have been very few assessment results showing the ability of these models to synthesize high quality speech with variable speaking rates.

On the other hand, as a basic issue in speech research, there are still numerous differences between the conclusions of studies on the effects of speaking rate [20-33]. Some studies [21,22] indicate that increased rates of speech result in systematic deviations in the obtained formant values from their putative targets, that is, "vowel reduction". Others [23-25] claim that such "vowel reduction" does not always occur at fast speaking rates. Still other studies claim that adjustments in speaking rate are achieved by strategies which differ between speakers [26,27] and by the carefulness of their articulation [28]. According

to electromyographic investigations [29,30], control of the speaking rate is achieved via a reorganization of motor commands.

One approach to this issue is to construct a model, by which we can test if undershoot or reorganization is necessary in generating high-quality speech at various speaking rates.

In this study, we proposed a functional model which describes formant transitions as the summation of two kinds of temporal functions: one represents vowel-to-vowel transitions, and the other represents consonant-to-vowel or vowel-to-consonant transitions. The model was assessed via an intelligibility test.

2.1. Method

2.1.1. Model of formant transition

The trajectory of the nth formant, $F_n(t)$, in a vowel segment is expressed as

$$F_n(t) = U_n(t) - C_{np}(t) - C_{nf}(t) \tag{2}$$

Here, $U_n(t)$ is the step response of a second order delay function which represents a vowel-to-vowel transition; $C_{np}(t)$ is a first order delay function which represents the effect of a preceding consonant; $C_{nf}(t)$ is a first order delay function which represents the effect of a following consonant.

To generate $U_n(t)$, the putative target frequency $R_{i,j}$ of each vowel in the sequence $V_1C_pV_2C_fV_3$, $(i = 1,2,3, j = 1,2,3)$ is assumed to be set at t_i as a step input. The suffix i represents vowel number, j indicates formant number. For the back vowels /a, u, o/, j represents the jth lower-formant frequency. For the front vowels /i, e/, $R_{i,1}$ is the lowest, $R_{i,2}$ the third and $R_{i,3}$ the second one. This numbering is adopted to take into account the continuity in formant trajectories [15,17].

Let $W_j(t)$ represent the step response of a second order delay function, expressed as

$$\begin{aligned} W_j(t) &= R_{1,j} + a_i(t)(R_{i,j} - R_{i-1,j}) \\ a_i(t) &= 1 - \{1 + b_j(t)\}exp(-b_j(t))u(t - t_i) \\ b_j(t) &= (t - t_i)/g_j \\ u(t - t_i) &= 1 \quad t > t_i, \qquad = 0 \quad t < t_i \end{aligned} \tag{3}$$

g_j:time constant representing the transition speed

For transitions from a back vowel to a front vowel, or vice versa, $W_2(t)$ and $W_3(t)$ intersect with each other. Such intersections never occur in actual speech, due to the coupling between the two resonance frequencies. Therefore, the resonance frequencies $W_j(t)$ are modified accounting for the coupling between $W_2(t)$ and $W_3(t)$ as follows [15,17].

$$\begin{aligned} U_1 &= W_1, \quad U_2 = h(W_2W_3)^{0.5}, \quad U_3 = (W_2W_3)^{0.5}/h \\ h &= s^{0.5}, \quad q = (W_2W_2 + W_3W_3)/W_2W_3 \\ s &= q - (q^2 - 4(1 - k^2))^{0.5}/(2(1 - k^2)), \quad k = 0.2 \end{aligned} \tag{4}$$

Two functions representing the effect of a preceding consonant $C_{np,i}(t)$ and the effect of a following consonant $C_{nf,i}(t)$ upon the formant trajectories in the segment V_i are assumed as follows.

$$C_{np,i}(t) = c_{np,i}exp\{-(t-t_{p,i})/g_p\}, \quad \text{for } t_{p,i} < t < t_{f,i} \tag{5}$$

$$C_{nf,i}(t) = c_{nf,i}exp\{-(t_{f,i}-t)/g_f\}, \quad \text{for } t_{p,i} < t < t_{f,i} \tag{6}$$

t_{pi} : initial time of vowel V_i

t_{fi} : final time of V_i

g_p, g_f : time constant representing the decay speed.

In this report, only the temporal parameters, t_i: onset time of the targets for vowel V_i, t_{pi}: initial time of V_i and t_{fi}: final time of V_i are variable depending on the speaking rate. This means that possible changes or "reorganization" in the vowel targets or other parameters such as g_p and g_f were not taken account of in this report.

2.1.2. Estimation of model parameters

The details of the recordings and the analyses of the speech material used for the modeling have been reported in other papers [31-34].

For the estimation of the model parameters, we assumed that the following is valid for vowels spoken clearly and slowly.

(1) The effects of surrounding consonants on the vowel formants decrease at the vowel midpoint, so we can set $C_{np,i}(t) = C_{nf,i}(t) = 0$.

(2) If the vowel segment is long, the formant frequencies $f_n(t)$ obtained by analysis are close to the putative targets, or $R_{i,j}$.

According to Assumption (2), $R_{i,j}$ is set to the formant frequencies $f_n(t)$ obtained by analysis at the midpoint t_i of vowel V_i. $U_n(t)$ is calculated using Eqs. (3) and (4), and then, the temporal function $X_n(t) = U_n(t) - f_n(t)$ can be calculated.

As shown in Fig. 7, $X_n(t)$ is large at the initial point of vowel V_i and decreases rapidly to zero. After the midpoint, it increases to its maximum at the endpoint of V_i. Thus, $X_n(t)$ can be approximated by the two first-order functions $C_{np,i}(t)$ and $C_{nf,i}(t)$. $C_{np,i}(t)$ is large at the beginning of V_i and then decreases exponentially, while $C_{nf,i}(t)$ is small at the midpoint and increases exponentially to its maximum at the end of V_i.

To determine $C_{np,i}(t)$ and $C_{nf,i}(t)$, $t_{p,i}$, the initial point of V_i is set at the point where the intraoral pressure starts to drop rapidly, or the release point of the consonant, and $t_{f,i}$, the final point of V_i, is set where the intraoral pressure starts to rise due to the vocal tract closure for the stop consonant. For $c_{np,i}$, g_p is adjusted so as to minimize the square error between $X_n(t)$ and $C_{np,i}(t)$ for the initial half of the segment V_i, and for $c_{nf,i}$, g_f is adjusted so as to minimize the square error between $X_n(t)$ and $C_{nf,i}(t)$ for the final half of the segment V_i.

Figure 7 (a) shows an example of the model functions estimated for /abiba/ spoken slowly and clearly. The uppermost curve is the original speech waveform, the second

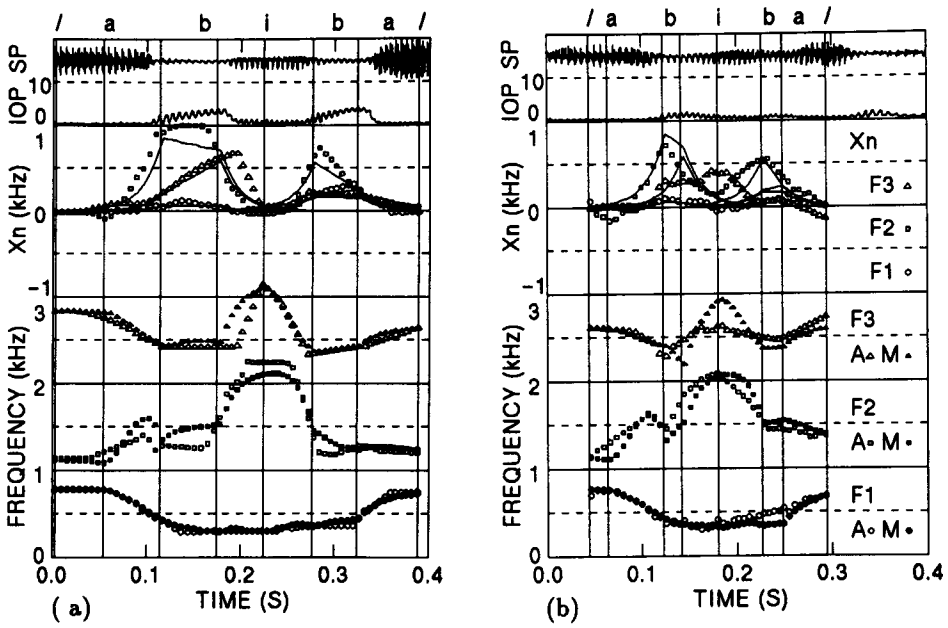


Figure 7. Model formant trajectories $F_n(t)$, and those obtained by analysis $f_n(t)$ for a slow (a) and a fast (b) utterances of /abiba/. "A" indicates $f_n(t)$, and M $F_n(t)$. See text for details.

curve is the intraoral pressure measured simultaneously [31-34]. $X_n(t)$ is shown by three kinds of dotted lines according to the formant number. In the lower section, the formant frequencies $f_n(t)$, obtained by analysis, and the model formant frequencies are shown together with $U_n(t)$. During the vocal tract closure, $F_n(t_{f,i-1})$ and $F_n(t_{p,i})$ are linearly interpolated. Figure 7 (a) shows that $F_n(t)$ fits well with $f_n(t)$.

2.1.3. Speech synthesis by rules at variable speaking rate

For the synthesis of speech at various speaking rates, rules for generating t_i, $t_{p,i}$, and $t_{f,i}$ should be constructed. We do not discuss such rules here. Instead, we discuss how well such a model predicts the formant trajectories observed in actual fast speech. For the assessment of the model proposed here, we compared speech samples actually uttered at a fast rate (FO), which was twice as fast as the slow rate examined, with synthetic speech generated based on a model where t_i, $t_{p,i}$, and $t_{f,i}$ were adapted to the fast speech FO. The other parameters were set to the same values, obtained from the slow utterances (SO), from which the model parameters were estimated.

Figure 7(b) shows one example of a fast /abiba/ uttered by the same speaker as in Fig. 7(a). Here, t_i, $t_{p,i}$, and $t_{f,i}$ are adapted to the actual utterance of /abiba/. $R_{i,j}$ for the vowel V_1 (initial /a/ in this case) are set to the actual average values of f_1,

f_2 and f_3 obtained by the analysis, because the vowel reduction for V_1 cannot fully be estimated without the preceding phonemes. Figure 7(b) shows that the model formant trajectories $F_n(t)$ for fast /abiba/ predict some of the gross characteristics in the f_1 and the f_2 transitions well, but fail to represent the large downward shift in f_3.

2.1.4. Intelligibility test

To assess how well the model could generate formant trajectories, an intelligibility test was carried out for two kinds of synthetic speech (G and M), and also for the original speech samples (O) from which model parameters were extracted. These speech samples were synthesized or recorded at two speaking rates, slow(S) and fast (F). The speech samples tested consisted of the following six groups.

SO: Original speech samples uttered slowly and clearly, from which the model parameters were extracted.

FO: Original speech samples uttered fast, from which the temporal parameters for the synthetic fast speech (FM, FG) were extracted.

SG: Synthetic slow speech, generated using the formant frequencies $f_n(t)$ obtained from SO by analysis and the glottal source obtained from the polynomial glottal source model.

FG: Synthetic fast speech generated in the same way as SG.

SM: Synthetic slow speech generated using the model formant trajectories $F_n(t)$ and the model glottal source.

FM: Synthetic fast speech generated in the same way as SM.

Each group of consisted of 48 V_1CV_2 samples, where V_1 and V_2 were one of /a, i, u,/, $V_1 = V_2$, and C was /b, d, g, r/. For SO and FO, $V_1C_pV_2$ and $V_2C_fV_3$ were extracted from the original utterances /korewa $V_1C_pV_2C_fV_3$ desu/. For the synthetic speech SG, FG, SM and FM, $V_1C_pV_2C_fV_3$ was synthesized to simulate the effects of articulatory undershoot, and then the segments $V_1C_pV_2$ and $V_2C_fV_3$ were extracted.

The subjects for the listening test were five adults with normal hearing who were not familiar with the purpose of this study, two phoneticians, one speech pathologist, one speech scientist and one graduate student majoring in speech science. The listening test was carried out only once to avoid possible adaptation to the synthetic speech. Each subject was instructed to transcribe each speech sample in phonetic symbols or in the Roman alphabet.

2.1.5. Four factors accounting for intelligibility

The following four factors were used to interpret the intelligibility of the six groups of speech, as shown in Fig. 8.

a_1: the decrement in percent due to the shortening of the segmental duration in fast speech

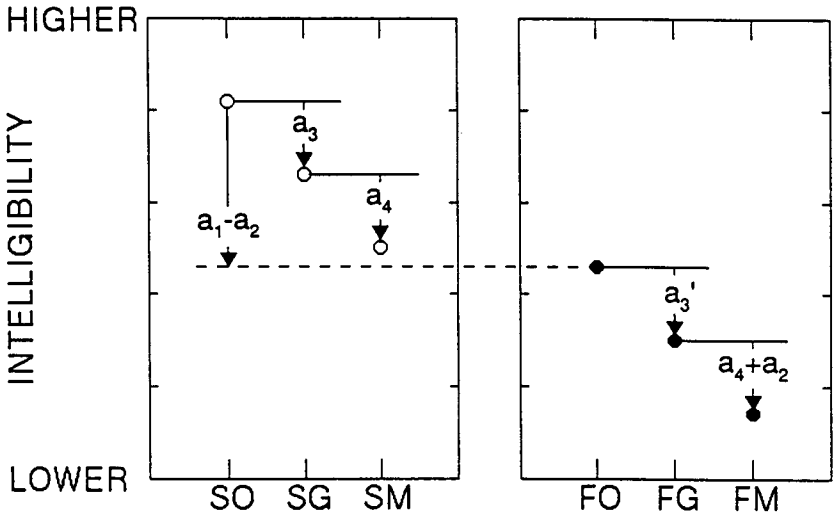


Figure 8. Four factors accounting for the intelligibility.

a_2: the decrement due to the omission of reorganization when the model is applied to fast speech ($-a_2$:the increment due to reorganization in fast speech)

a_3: the decrement due to the lack of plosive source for the stop consonants for the slow speech

a_3': same as a_3, but for fast speech

a_4: the decrement due to the formant model mismatch

The factor a_1 accounts for the decrement between the intelligibility of SO and that of FO. Because the speech samples FO have shorter duration, smaller formant transitions and also a larger undershoot or vowel reduction than SO, the intelligibility of FO may decrease largely compared to that of SO. However, if the speaker reorganizes the articulation for fast speech to increase intelligibility against the disadvantages mentioned above, this factor (reorganization:$-a_2$) might raise the intelligibility. As a result, the difference between FO and SO should be $a_1 - a_2$.

For SG, the intelligibility may be a_3 lower than that of SO, because SG is synthesized without a plosive source. And, for the same reason, the intelligibility of FG may be a_3' lower than that of FO. a_3' may be different from a_3 because the original fast speech may have only weak plosion or even no plosion.

The intelligibility of SM may be a_4 lower than that of SG due to the mismatch of the formant model. The intelligibility of FM is assumed to be $a_4 + a_2$ lower than FG, where a_4 represents the effect of the failure of the model to adapt to slow speech, and a_2 represents the effect of the failure of the model to predict the formant trajectories in fast speech

since it was devised based on slow speech. In other words, the factor a_2 represents the fact that the model does not take into account possible changes in articulation between two speaking rates, that is, reorganization.

2.2. Results and discussion

Figure 9(a) shows the average intelligibility of the three vowels /a,i,u/ for each subject in the six speech groups. The box-whisker graph in this figure shows the minimum, 25%-tile, median, 75%-tile and the maximum of the intelligibility scores, averaged for the three vowels with reference to each of the five subjects. Figure 9(b) shows the average intelligibility of the four consonants /b, d, g, r/ for the six speech groups. Table 1 shows the four factors estimated from the results shown in Fig. 9 based on the relationships shown in Fig. 8.

As shown in Fig. 9(a), the medians of the intelligibilities for the six speech groups are SO:100.0%, SG:100.0%, SM:100.0%, FO:92.7%, FG:91.7% and FM:93.8%. The intelligibility of FM is 93.8%, which is better than those of FO and FG.

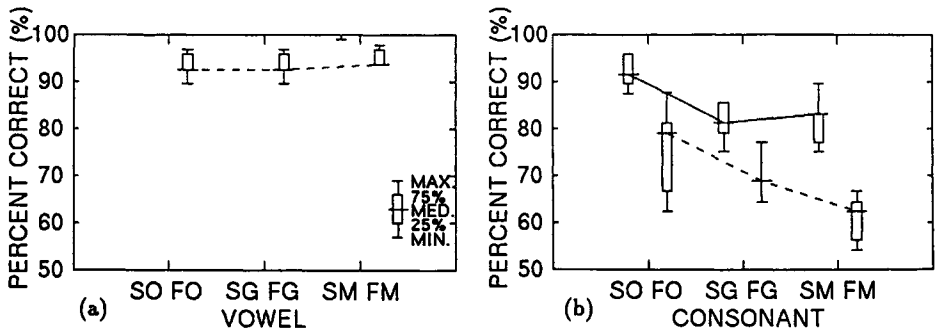


Figure 9. The average intelligibility of the three vowels (a), and of the four consonants (b).

This result indicates that the use of a formant model with a model voice source does not decrease the intelligibility (a_4, a_3, $a_3' = 0.0$, as shown in Table 1). The disregard of reorganization in the formant model slightly increases the intelligibility ($a_2 = -2.1$). Concerning the vowels, it can be suggested that the formant model maintains or even slightly improves the intelligibility compared to the original speech at slow and fast speaking rates.

On the other hand, as shown in Fig. 9(b), the medians for the consonants are SO:91.7%, SG:81.3%, SM:83.3%, FO:79.2%, FG:68.8% and FM:62.5%. The four factors accounting for the intelligibility are $a_1 = 20.8$, $a_2 = 8.3$, $a_3 = 10.4$, $a_3' = 10.4$ and $a_4 = -2.0$, as shown in Table 1.

For the consonants, the use of a model voice source without plosion decreases the intelligibility by about 10%. The use of the formant model slightly increases the intelligibility by about 2% ($a_4 = -2.0$). The disregard of reorganization may reduce the intelligibility

Table 1. Four factors estimated from the medians of the intelligibility. V:for 3 vowels, C:for 4 consonants.

Factor	V	C	/b/	/d/	/g/	/r/
a_1	5.2	20.8	16.6	8.3	50.0	16.6
a_2	-2.1	8.3	8.2	-8.3	-8.2	8.3
a_3	0.0	10.4	25.0	16.7	0.0	0.0
a_3'	0.0	10.4	0.0	33.4	0.0	8.4
a_4	0.0	-2.0	-16.6	8.3	16.6	8.4
a_1-a_2	7.3	12.5	8.4	0.0	42.8	8.3
a_4+a_2	-2.1	6.0	-8.4	0.0	8.4	16.7

by about 8%. Also, a_1 is estimated to be 20.8, which means that the decrement in the intelligibility due to speed or shortening in the fast speech is large. For the consonants in the slow speech, the formant model works well on average and even slightly improves the intelligibility compared to SG. However, for the fast speech the formant trajectories predicted by the model decrease the intelligibility by about 6%.

2.3. Conclusion

In this section, a model of formant trajectories at various speaking rates is proposed, and the intelligibility of VCV speech samples synthesized based on the model is reported. The intelligibility of vowels on average was 100% at a slow speaking rate and was about 93% at a fast rate, which

was about twice as fast as the slow rate. The intelligibility of the consonant, was 83% for the slow rate and about 63% for the fast rate. It was found that the formant model slightly improves the intelligibility of vowels at both speaking rates and that of consonants at the slow rate compared with the speech synthesized using formant trajectories obtained by analysis. However, for the consonants in the fast speech, the formant model decreases the intelligibility by about 6%. The use of a model voice source without plosion decreases the intelligibility by about 10%, and the disregard of reorganization was estimated to reduce the intelligibility by about 8% for the consonants.

References

1. A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am., 49 (2), pp.583-598, (1971)

2. G. Fant, J. Liljencrants and G. Lin, "A four-parameter model of glottal flow," STL-QPSR, 4/1985, pp.1-13, (1986)

3. G. Fant and Q. Lin, "Frequency domain interpretation and derivation of glottal flow parameters," STL-QPSR, 2-3/1988, pp.1-21, (1988)

4. H. Fujisaki and M. Ljungqvist, "A comparative study of glottal waveform models," IEICE Technical Report (EA85-58), pp.23-29, (1985)

5. D. H. Klatt, "Acoustic correlates of breathiness: First harmonic amplitude, turbulent noise, and tracheal coupling," J. Acoust. Soc. Am., 82(S1), p.S91, (1987)

6. K. Hasegawa, T. Sakamoto and H. Kasuya, "Effects of glottal noise on the quality of synthetic speech," Proceedings of ASJ Spring Meeting, pp.205-206, (1987)

7. S. Imaizumi, and S. Kiritani, "A study of formant trajectories and voice source characteristics based on the closed phase analysis," Preprints of the Second Symposium on Advanced Man-Machine Interface Through Spoken Language (1988)

8. D. Childers and J. Larar, "Electro-glottography for laryngeal function assessment and speech analysis," IEEE Trans. BME-31, 12, pp.807-817, (1984)

9. J. B. Kruskal, "Nonmetric multidimensional scaling: A numerical method," Psychometrika, 29, pp.115-219, (1964)

10. Y. Takane, F.W. Young and J. de Leeuw, "Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features," Psychometrika, 42, pp.7-67, (1977)

11. M. Rothenberg, "Acoustic interactions between the glottal source and vocal tract, in Vocal Fold Physiology," Ed. K. N. Stevens and M. Hirano, (Univ. Tokyo Press, Tokyo), pp.305-328, (1981)

12. T. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," STL-QPSR 1/1982, pp.1-30, (1982)

13. T. Koizumi, S. Taniguchi and S. Hiromitsu, "Two-mass models of the vocal cords for natural sounding voice synthesis," J. Acoust. Soc. America, 82(4), pp.1179-1192, (1987)

14. J. Liljencrants, "Speech synthesizer control by smoothed step functions," STL-QPSR 4/1969, pp.43-50, (1970)

15. H. Fujisaki, M. Yoshida, Y. Sato and Y. Tanabe, "Automatic recognition of connected vowels using a functional model of the coarticulatory process," J. Acoust. Soc. Jpn, 29, pp.636-638, (1974)

16. S. Yokoyama and S.Itahashi, "Approximation of formant trajectory by second order system with applications to consonants," Proc. Acoust. Soc. Japan, pp.89-90, (1975)

17. Y. Sato and H. Fujisaki, "Formulation of the process of coarticulation in terms of formant frequencies and its application to automatic speech recognition," J. Acoust. Soc. Jpn, 34,3, pp.177-185, (1978)

18. D. J. Broad and R.H. Fertig, "Formant-frequency trajectories in selected CVC utterances," J. Acoust. Soc. Am.,47, pp.1572-1582, (1970)

19. D. J. Broad and F. Clermont, "A methodology for modeling vowel formant contours in CVC context," J. Acoust. Soc. Am., 81(1), pp.155-165, (1987)

20. J. L. Miller, "Effects of speaking rate on segmental distinctions," (Perspectives on the study of speech. P. D. Eimas and J. L. Miller Eds., Lawrence Erlbaum Associates, New Jersey), pp.39-74, (1981)

21. B. Lindblom, "Spectrographic study of vowel reduction," J. Acoust. Soc. America, 35(11), pp.1773-1781, (1963)

22. T. Gay, "Effect of speaking rate on diphthong formant movements," J. Acoust. Soc. Am, 44, pp.1570-1573, (1968)

23. R. R. Rerbrugge and D. Shankweiler, "Prosodic information for vowel identity," J. Acoust. Soc. Am, 61, p.S39, (1977)

24. T. Gay, "Effect of speaking rate on vowel formant movements," J. Acoust. Soc. Am., 63(1), pp.223-230, (1978)

25. D. O'Shaughnessy, "The effects of speaking rate on formant transitions in French synthesis-by-rule," Proc. 1986 IEEE-IECEJ-ASJ, Tokyo, pp.2027-2039, (1986)

26. D. P. Kuehn and K. L. Moll, "A cineradiographic study of VC and CV articulatory velocities," J. Phonetics, 4, pp.303-320, (1976)

27. Y. Sonoda, "Effects of speaking rate on articulatory dynamics and motor event," J. Phonetics, 15, pp.145-156, (1987)

28. J. E. Flege, "Effects of speaking rate on tongue position and velocity of movement in vowel production," J. Acoust. Soc. Am., 84(3), pp.901-916, (1978)

29. K. Harris, "Mechanisms of duration change," (in Speech Communication 2, G. Fant Ed.,Almqvist & Wiksell), pp.299-305, (1974)

30. T. Gay, T. Ushijima, H. Hirose and F. Cooper, "Effect of speaking rate on labial consonant-vowel articulation," J. Phonetics, 2, pp.47-63, (1974)

31. S. Imaizumi, S. Kiritani, H. Hirose, S. Togami and K. Shirai, "Preliminary report on the effects of speaking rate upon formant trajectories," Ann. Bull. RILP, 21, pp.147-151, (1987)

32. S. Imaizumi and S. Kiritani, "Effects of speaking rate on formant trajectories and inter-speaker variations," Ann. Bull. RILP, 23, pp.27-37, (1987)

33. S. Imaizumi and S. Kiritani, "Perceptual evaluation of a glottal source model for voice quality control," Proc. 6th Vocal Fold Physiology Conference, Stockholm, pp.1-10, (1989)

34. S. Imaizumi and S. Kiritani, "A generation model of formant trajectories at variable rates," (in Talking Machines: Theories, Models, and Designs, C. Bailly, C. Benoit Eds., North-Holland, Amsterdam), pp.61-75, (1992)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

A System for Synthesis of High-Quality Speech from Japanese Text

Hiroya Fujisaki, Keikichi Hirose, Hisashi Kawai and Yasuharu Asano

Faculty of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113 Japan

Abstract

A text-to-speech conversion system for Japanese has been developed for the purpose of producing high-quality speech output. This system consists of four processing stages: (1) linguistic processing, (2) phonological processing, (3) control parameter generation, and (4) speech waveform generation. This paper focuses on the second and the fourth stages, especially on the generation of phonetic and prosodic symbols. The phonetic symbols are those having one-to-one direct correspondence with the pronunciation. These symbols are converted from the input orthographic text via phonemic symbols mainly using the word level linguistic information. The prosodic symbols, representing the prosodic structure of the text, are, on the other hand, generated using the linguistic information of a wider range, even to a paragraph. Rules for generating prosodic symbols were constructed based on the analysis of natural speech. As for the fourth stage, a new type of terminal analog speech synthesizer was designed. It consists of four paths in cascade of pole/zero filters which are, respectively, used for the synthesis of vowels and vowel-like sounds, the nasal murmur and the buzz bar, the frication, and plosion. The validity of the approach has been confirmed by the improvements both in the prosodic and in the segmental quality of synthesized speech.

1. INTRODUCTION

It is widely admitted that communication through speech can provide the most efficient and flexible means for the man-machine interface. The current technology for speech synthesis and speech recognition, however, is not sufficient to fully realize the inherent advantage of speech over the written language. As for the speech synthesis, in spite of many significant contributions in recent years, the quality of speech from conventional text-to-speech conversion system still needs to be improved. Among various factors, the following two are considered to be primarily responsible.

(1) Insufficient use of linguistic information. In conventional text-to-speech conversion systems, a full syntactic analysis of a sentence is often circumvented and is replaced

by the analysis of local dependencies within a phrase. Furthermore, few existing systems deal with linguistic information beyond the level of a sentence. Since the prosodic features of an utterance are deeply influenced by the syntactic structure of a sentence as well as by the structure of the discourse in which the sentence is produced, the lack of proper linguistic analysis will mainly lead to a poor quality of synthetic speech.

(2) Oversimplified configuration of the synthesizer. Speech synthesizers used in conventional text-to-speech conversion systems are based on a rather rude approximation of the process of speech production and hence fail to produce a close approximations to the actual speech signals. This leads mainly to the poor quality of synthetic speech.

These considerations have led us to design and construct a text-to-speech conversion system with special emphasis on the synthesis of natural prosody from high-level linguistic information and on the development of a terminal analog speech synthesizer capable of producing higher segmental quality than has been accomplished before.

2. PROSODIC FEATURES AND LINGUISTIC INFORMATION

In view of the importance of the time contour of the fundamental frequency (henceforth F_0 contour) among various prosodic features of Japanese, we have been intensively working on the analysis and synthesis of F_0 contours. A quantitative model has been presented which generates F_0 contours of sentences from a small number of discrete commands, viz., the phrase commands and the accent commands [1,2]. Detailed analysis was conducted on the F_0 contours of spoken sentences in Japanese, and the relationships between the F_0 contours and their underlying linguistic information were revealed [3–6]. Based on these studies, we have proposed a set of rules for generating prosodic symbols from the syntactic structure of the input text [7–13].

2.1. Model for F_0 contour generation

The F_0 contour of an utterance can be regarded as the response of the mechanism of vocal cord vibration to a set of commands which carry information concerning lexical accent, and the syntactic and discourse structure of the utterance. Two different kinds of command have been found to be necessary; one is an impulse-like phrase command for the onset of a prosodic phrase, while the other is a stepwise accent command for the accented mora or morae of a prosodic word. The consequences of these two types of commands have been shown to appear as the phrase components and the accent components, each being approximated by the response of a second-order linear system to the respective commands. If we represent an F_0 contour as the pattern of the logarithm of the fundamental frequency along the time axis, it can be approximated by the sum of these components. The entire process of generating an F_0 contour of a sentence can thus be modeled by the block diagram of Figure 1 [2].

The analysis and synthesis of F_0 contours were conducted using this model. For the analysis, an observed F_0 contour need to be decomposed into the phrase and the accent

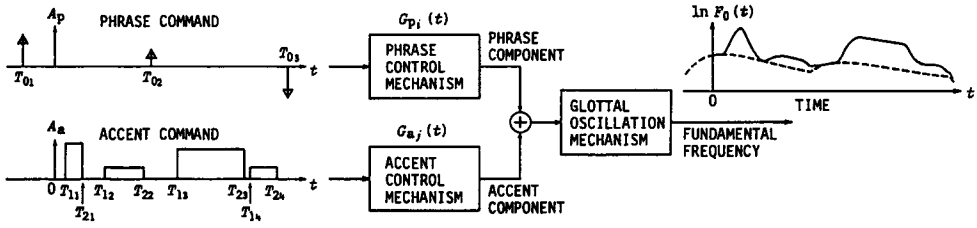


Figure 1. Block diagram of the functional model for the process of generating sentence F_0 contours.

components. This is accomplished by finding, by the method of analysis-by-synthesis, the optimum set of model parameters that gives the minimum mean-squared error between the observed F_0 contour and the model-generated F_0 contour. The phrase and the accent commands obtained as the result of such a decomposition are then used to examine the relationship between the linguistic information and the F_0 contour.

2.2. Prosodic units of spoken Japanese

As the minimal prosodic unit of spoken Japanese, we introduce the "prosodic word," which is defined as a part or the whole of an utterance that forms an accent type, and is usually composed of an independent word and the succeeding sequence of dependent words. As will be discussed later, a string of prosodic words, under certain conditions, can form a larger prosodic word due to "accent sandhi." On the other hand, a phrase component of the F_0 contour of an utterance defines a larger prosodic unit, i.e., a "prosodic phrase," which may contain one or more prosodic words. Generally, a prosodic word never extends over two prosodic phrases. Furthermore, in longer sentences, several prosodic phrases may form a section delimited by pauses. Such a section is defined as a "prosodic clause." As for the syntactic units, we adopt "bunsetsu," "ICRLB," clause and sentence, where "bunsetsu" is defined as an unit of utterance in Japanese, and, in most cases, consists of an independent word and succeeding dependent words. The word "ICRLB" is an abbreviation for "immediate constituent with a recursively left-branching structure," which is a syntactic phrase delimited by right-branching boundaries and contains only left-branching boundaries. Roughly speaking, the parallelism shown in Figure 2 exists between the hierarchy of syntactic units and the hierarchy of prosodic units[5,6].

2.3. Characteristics of Pause and Phrase Component

A pause is always accompanied by a phrase command, while a small phrase command usually occurs without a pause. There exists roughly a linear relationship between the duration of the pause and the magnitude of the command of the accompanied phrase component. Three ways are possible for starting a new phrase component:

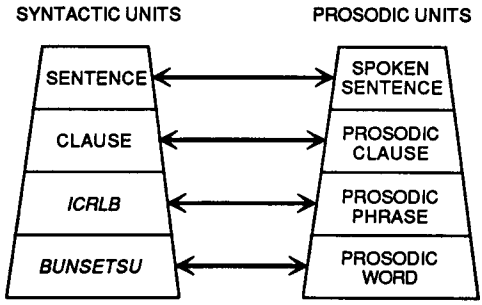


Figure 2. Hierarchy of the syntactic units and that of the prosodic units, and their relationship.

(1) preceded by a pause during which the immediately preceding phrase component is completely reset,

(2) preceded by a brief pause but the immediately preceding phrase component is not reset so that the new phrase component is superposed on the old one,

(3) simply added to the old one without pause.

Prosodic boundaries marked by these different ways are henceforth named TypeI, Type II, and Type III boundaries. Analysis of the F_0 contours was conducted on the spoken sentences of the news and weather forecast uttered by professional announcers. The results indicate that sentences and prosodic clauses are marked by prosodic boundaries of Type I and Type II, while prosodic phrases are marked by Type III boundaries.

The occurrence of these prosodic boundaries is primarily determined by the syntactic structure, and most probably coincides with syntactic boundaries. It is, however, also subject to other factors such as style of speaking, respiration, etc. As for the syntactic boundaries at which prosodic boundaries may occur, we may distinguish four different boundary types (1) between sentences, (2) between clauses, (3) between ICRLB's, and (4) within an ICRLB. The relationships between these syntactic boundaries and the three types of prosodic boundary are listed in Table 1 [5,11].

The correspondence between the syntactic boundaries and the prosodic boundaries is not one-to-one but is rather stochastic, and the probability that a prosodic boundary occurs at a given syntactic boundary is influenced also by the depth of the syntactic boundary [8,12]. Figure 3 shows an example of the result of the analysis along with the syntactic tree of the sentence. Each leaf of this syntactic tree is a prosodic word. The number on each leaf of the syntactic tree denotes the number of generations from the primal predicate of the sentence, i.e., the number of right-branches contained in the pass from the root to the leaf. Let us denote the "depth of a boundary" by $j-i+1(=k)$, where i and j, respectively, denote the numbers on the leaves at the left side and at the right side of the boundary. Using this notation, we can define the "left-branching boundary" as a boundary at which $k = 0$ and the "right-branching boundary" as a boundary at

Table 1. Classification of syntactic boundaries and their manifestations as prosodic boundaries.

PROSODIC BOUNDARIES	SYNTACTIC BOUNDARIES			
	BETWEEN SENTENCES	WITHIN A SENTENCE		
		BETWEEN CLAUSES	WITHIN A CLAUSE	
			BETWEEN ICRLB'S*	WITHIN AN ICRLB*
TYPE I (PHRASE RESETTING WITH PAUSE)	100%	20%		
TYPE II (PHRASE ADDITION WITH PAUSE)		30%	5%	
TYPE III (PHRASE ADDITION WITHOUT PAUSE)		50%	85%	15%

*ICRLB: AN IMMEDIATE CONSTITUENT WITH A RECURSIVERY LEFT-BRANCHING STRUCTURE.

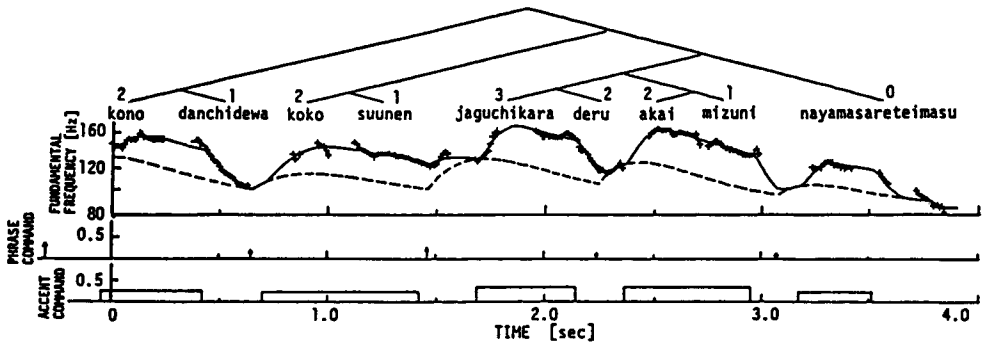


Figure 3. The F_0 contour and the syntactic tree of a Japanese sentence. The meaning of the sentence is "For the past several years people in this apartment house complex have been annoyed by the stained water coming out of the taps."

which $k > 0$. The results of the analysis indicate that longer pauses and larger phrase commands tend to occur at boundaries with larger k.

2.4. Characteristics of Accent Component

For the ease of representation, let us henceforth denote a prosodic word with a rapid downfall in the F_0 contour by "D-type prosodic word" or merely by "D" and denote one without any downfall by "F-type prosodic word" or "F". In Japanese with the Tokyo dialect, the manner of accentuation of the prosodic words can be classified into three groups:

(1) having a rapid downfall of the constituent independent word, such as in the case of "Fujisan-ga,"

(2) having a rapid downfall of the succeeding dependent word, such as in the case of "tomodachi-sae," and

(3) having the F-type accent, such as in the case of "otohto-no."

The grouping depends on (1) the accent type and the grammatical conditions of the independent word, and (2) the characteristics of accentuation of the dependent word. The rules for the determination of accent types of prosodic words have been studied and summarized [14].

When the prosodic words are uttered in isolation, there is a significant difference between the accent commands for D and F, being higher for D and lower for F. When more than two prosodic words are uttered in connected speech, however, they interact with each other and their accent commands change both in amplitude and in shape, depending on their accent types, the syntactic structure of the phrase in which they exist, and the focal conditions.

In order to clarify the rules underlying these changes, the F_0 contour analysis was conducted on utterances of sentences including noun phrases and verb phrases by a male speaker having the Tokyo dialect [4,5,9]. The phrases are composed of two or three prosodic words. Utterances for all the possible combinations for accent types, syntactic structures, and focal conditions were recorded and analyzed. The following characteristics were found for ICRLB phrases.

(1) In a phrase consisting only of D-type prosodic words, the accent command for the first word is essentially of the same amplitude as that when the word is uttered in isolation, while the accent commands for all the other words are suppressed. An example for three prosodic words is shown in panel (a) of Figure 4. When both the second and third prosodic words are of the F-type, they are concatenated to form one prosodic word with a low accent command.

(2) In a phrase of DFD, each of the two prosodic words of the D-type has a high command, while the word of the F-type has low one.

(3) In a phrase of FD or FFD, the F's have as high commands as the succeeding D. A new prosodic word is formed by concatenation of F and D.

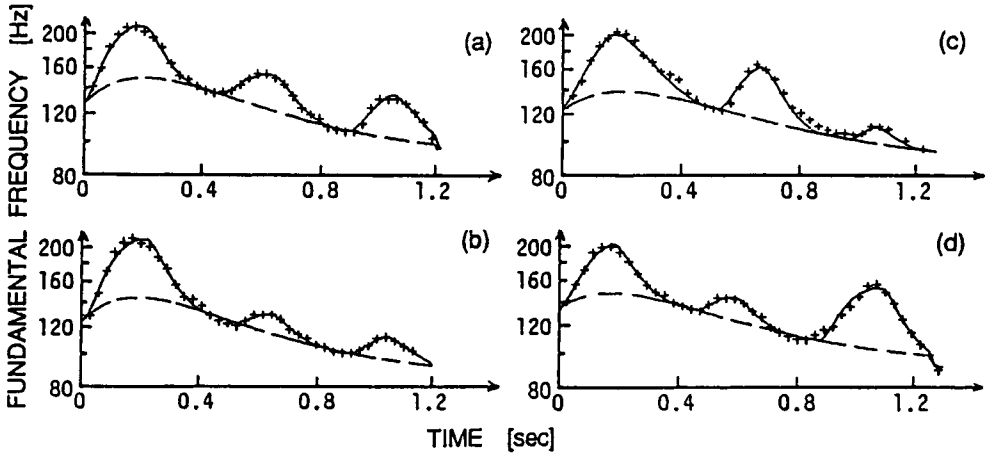


Figure 4. Results of the F_0 contour analysis for the noun phrase "aomorino (W_1) anino (W_2) amaguo (W_3)" uttered in four different manners: (a) without any obvious focus, (b) with focus on W_1, (c) with focus on W_2, (d) with focus on W_3. The meaning of the phrase is "a raincoat of my brother in Aomori."

(4) In a phrase consisting only of F's, they are usually concatenated to each other to form a new prosodic word.

(5) When a focus is placed on a prosodic word with a low accent command, its amplitude becomes higher (Panels (b), (c), (d) of Figure 4). If a phrase contains D-type prosodic words, the manner of concatenation is affected by the focus.

For phrases with right-branching boundaries, the characteristics of the accent component are somewhat different. For instance, a low phrase command often occurs at the right-branching boundary, and it usually prevents mutual interactions between the accent components of the prosodic words on both sides of the boundary.

3. RULES FOR GENERATING PROSODIC SYMBOLS

The term "prosodic symbols" is a generic name for the symbols necessary for the synthesis of the prosodic features of speech. In the present text-to-speech conversion system, they are pause, phrase and accent symbols. The pause symbols represent the duration of pauses and are necessary for the decision on the timing of the constituent syllable templates. On the other hand, the phrase symbols and the accent symbols, respectively, represent the magnitude of the phrase commands and the amplitude of the accent commands of the F_0 contour model, and are necessary for the decision on the timing and the magnitude/amplitude of the commands. Based on the results obtained in the previous section, prosodic symbols are selected as shown in Table 2. The symbols P0

and A0 are those for resetting the components to zero. The pause symbol S1 represents a pause between sentences. The pause symbols S2 and S3 occur at syntactic boundaries and are, respectively, followed by the phrase symbols P2 and P3. The phrase symbols P1, P2 and P3, respectively, correspond to the prosodic boundaries of Type I, Type II and Type III. The accent symbols can be classified into two groups, viz., those for D-type prosodic words and those for F-type prosodic words. The accent symbols DH and FM, respectively, correspond to the accent commands for D-type and F-type prosodic words uttered in isolation [15].

Table 2. Prosodic symbols and their values at the speech rate around 7 mora/sec.

PAUSE		PHRASE		ACCENT			
SYMBOL	VALUE	SYMBOL	VALUE	SYMBOL	VALUE	SYMBOL	VALUE
S1	700	P0	-0.50	DH	0.50	FH	0.50
S2	300	P1	0.35	DM	0.35	FM	0.25
S3	100	P2	0.25	DL	0.15	FL	0.10
		P3	0.15	(A0 = -DH, -DM, -DL, -FH, -FM, -FL)			

Based on the investigations on the pause duration and the F_0 contours, rules were constructed for the generation of the prosodic symbols [11,13]. The rules for the generation of the pause and the phrase symbols are as follows:

(1) Generate P1 at the beginning of the sentence. Generate P0 at the end of the sentence. Generate S1 between P1 and P0 (between sentences).

(2) Generate S2P2 at a clause boundary. If the distance from the preceding P1/P2 is shorter than $L1$ morae, generate P3 only.

(3) Generate P3 at an ICRLB boundary. If the distance from the preceding P1/P2/P3 is shorter than $L1$ morae, omit P3.

(4) Generate S3P3 at a boundary between parallel expressions. If the distance from the preceding P1/P2/P3 is shorter than $L1$ morae, omit S3.

(5) If an ICRLB is longer than $L2$ morae, insert P3's so that the length of all the prosodic phrases in the ICRLB are shorter than $L2$ morae. The insertion is conducted so that all the resultant prosodic phrases have similar lengths.

(6) If a clause is longer than $L3$ morae, insert S3's at ICRLB boundaries where insertions are allowed.

For a normal speech rate of 7 morae/s, $L1$, $L2$ and $L3$ are, respectively, set equal to 5, 15 and 40.

In the following rules for the generation of accent symbols, X denotes a prosodic word of any accent type, X^+ denotes a sequence of one or more X's, X^* denotes a sequence of zero or more X's, x denotes a sequence of any combinations of D's and F's, n denotes

the position of the prosodic word to be processed, and W_n denotes the prosodic word to be processed. The symbols +Emph, 0Emph, and -Emph, respectively, indicate the prosodic words to be emphasized, not to be emphasized, and to be suppressed due to focal conditions. The rules are applied within the scope of one ICRLB. The rules for F^+ sequences of prosodic words are as follows.

(1) For W1, generate FM if the focal condition is +Emph or 0Emph, and generate FL if the focal condition is -Emph. If W_1 is at the start of a sentence and the focal condition is -Emph, convert the P1 preceding the sentence to P2 and insert P3 just after W_1.

(2) For W_n $(n > 1)$, always generate FM. If the focal condition is +Emph, insert P3 just before W_n.

The rules for F*Dx sequences are as follows.

(1) For W_1 of F*, generate FH if the focal condition is +Emph or 0Emph, and generate FM if the focal condition is -Emph. If the focal condition is +Emph, convert all the 0Emph's for x into -Emph. If W_1 is at the start of a sentence and the focal condition is -Emph, convert the P1 preceding the sentence to P2 and insert P3 just after W_1.

(2) For $W_n(n > 1)$ of F*, generate FH if W_{n-1} is FH, and generate FM if Wn-1 is FM. If the focal condition is +Emph, convert all the 0Emph's for x into -Emph's.

(3) Always generate DH for a D following F^*. If the focal condition is +Emph, convert all the 0Emph's for x to -Emph. If W_n for D is at the start of a sentence and the focal condition is -Emph, change the P1 preceding the sentence to P3 and insert P2 just after W_n.

(4) Generate DH for D in an x sequence if the focal condition is +Emph, generate DM if the focal condition is 0Emph, and generate DL if the focal condition is -Emph. If the focal condition is +Emph for W_n, convert 0Emph's for prosodic words following W_n to -Emph.

(5) Generate FM for F in an x sequence if the focal condition is +Emph or 0Emph, and generate FL if the focal condition is -Emph. If the focal condition is +Emph, insert P3 just before W_n.

4. TERMINAL ANALOG SYNTHESIZER

A speech synthesizer of the terminal analog type is based on the simulation of the acoustic process of speech production and thus is capable of producing high-quality speech. Therefore, we have adopted this type of synthesizer as the final stage of the text-to-speech conversion system.

Because of the simplification in the realization of both the excitation sources and the vocal tract transfer functions, certain limitations exist on the quality of speech synthesized by conventional terminal analog speech synthesizers. One of the most used terminal analog

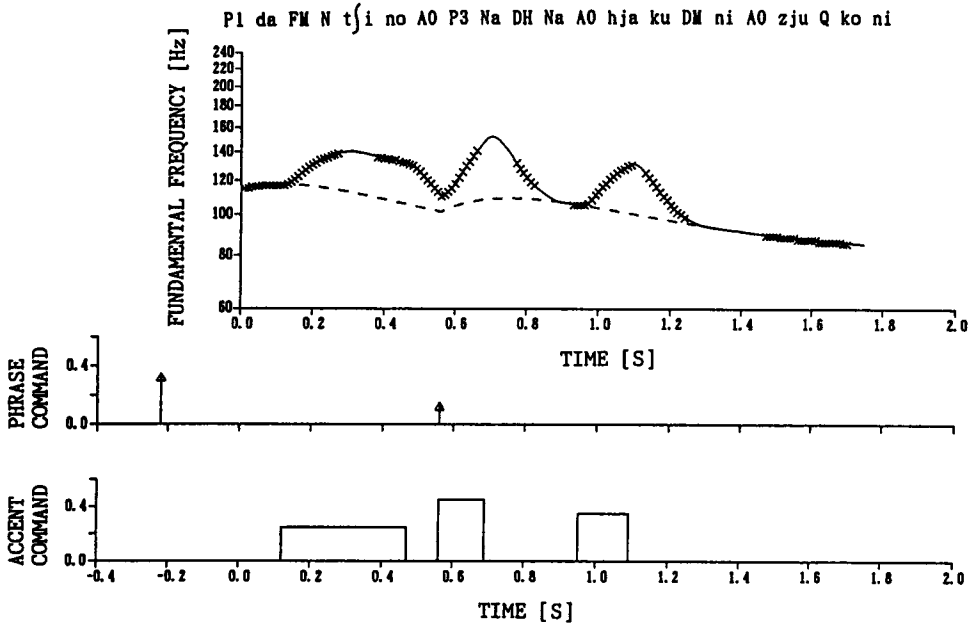


Figure 5. Examples for the prosodic symbols together with the phonetic symbols generated by the rules, and the synthesized F_0 contour.

synthesizers is the one developed by D. Klatt[13], where vowels and vowel-like sounds are generated by a circuit of in cascade connected pole filters, and fricatives are generated by a circuit with parallel connection. Although, the transfer function of the vocal tract can be represented either by cascade connection or parallel connection of pole/zero filters, the former has the advantage over the latter that a better simulation of the vocal-tract transfer function is available, since it represents the transfer function by the product of pole and zero filters and has characteristics close to that of the actual vocal tract. Moreover, relative gain control of each filter can be conducted automatically for the cascade connection.

Based on these considerations, we have developed a new terminal analog synthesizer for high-quality speech, as shown in Fig. 6[17–19]. This synthesizer consists of four paths of in cascade connected pole/zero filters and three source waveform generators. Nasal, vowel, fricative, and stop paths are, respectively, for the synthesis of the nasal murmur or the buzz bar, vowels or vowel-like sounds, the frication, and the plosion. The glottal waveform generator generates the voice source waveform (first derivative of the volume velocity) for the vocal path and the nasal path. The waveform is approximated by polynomials developed by one of the authors and controlled by three parameters: the fundamental frequency, skew and open quotient[20,21]. A random noise generator generates white Gaussian noise for the fricative path. This noise waveform is also supplied to the vowel path to produce the /h/-sound and devoiced vowels. An impulse from the impulse generator is fed to the stop path producing plosion for stop consonants.

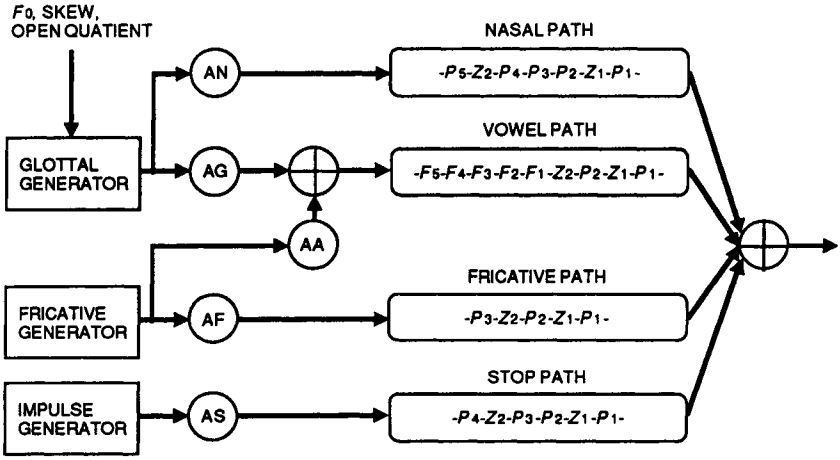


Figure 6. Configuration of a terminal analog speech synthesizer with four paths of in cascade connected pole/zero filters and with three source waveform generators. Symbols $F_1 - F_5$ and $P_1 - P_5$ indicate pole filters, while symbols Z_1, Z_2 indicate zero filters.

It is well known that Japanese vowels, typically back vowels, may be nasalized when attached to nasal murmur. Figure 7 shows the spectrum and its envelope for the vowel part of the utterance /ma/ extracted by the analysis-by-synthesis method. The result indicates that pole-zero pairs exist below the first formant and between the second and third formants[22]. Spectral analyses were also conducted for other utterances of various CV and VCV combinations using the LPC method and the analysis-by-synthesis method. Based on the results, the combination of pole and zero filters was decided on for each path of the synthesizer, as shown in Figure 6.

Precise simulation of the generation process of speech is possible using the proposed synthesizer. Figure 8 shows the waveforms of (a) natural speech and of (b) synthetic speech for /ka/. Stop, fricative and vowel paths are necessary for this synthesis. Various CV utterances were synthesized and the result indicated the advantages of the proposed synthesizer over the conventional ones.

5. SYSTEM CONFIGURATION

Based on the results obtained by the investigations above, a system has been constructed for the synthesis of high-quality speech from Japanese text[12,17,19,23]. As shown in Fig. 9, this system is based on the rule-synthesis of F_0 contours and on the concatenation of stored syllable templates. Roughly speaking, it consists of four stages, viz., stages for linguistic processing, phonological processing, control parameter generation, and speech waveform generation. The major functions of the system are as follows:

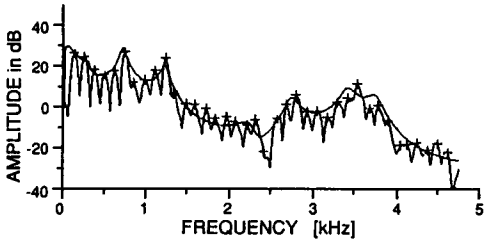


Figure 7. Spectrum and its envelope for the vowel part of the utterance /ma/ extracted by the analysis-by-synthesis method. Frequencies of five formants $F_1 - F_5$ are, respectively, 720 Hz, 1230 Hz, 2760 Hz, 3440 Hz, 3780 Hz, while those of two pole-zero pairs P_1, Z_1, P_2, Z_2 are, respectively, 283 Hz, 300 Hz, 2251 Hz, 2370 Hz.

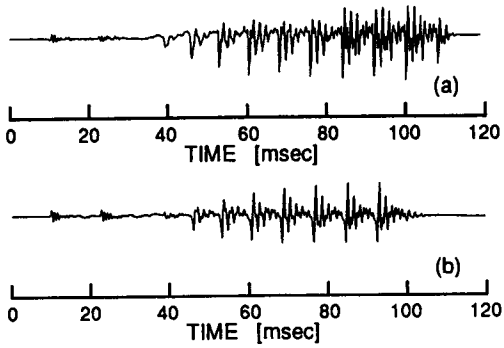


Figure 8. Waveform of (a) natural speech and of (b) synthetic speech for /ka/.

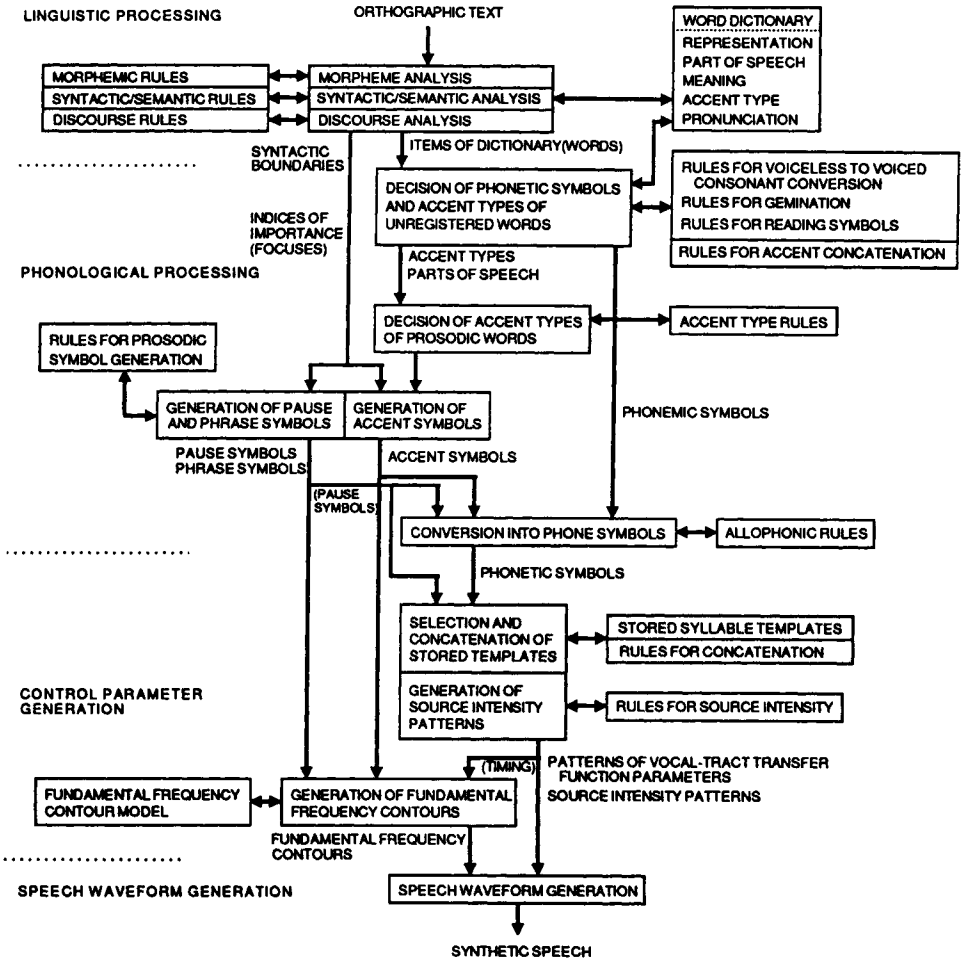


Figure 9. Block diagram of the system for synthesizing speech from Japanese text.

(1) Detection of word boundaries and identification of each lexical item. This is inevitable, since Japanese orthographic texts do not explicitly provide information of the word boundary. Ambiguities in word boundaries are minimized using rules for grammatical conjunction.

(2) Syntactic, semantic and discourse analyses to derive linguistic information, such as syntactic boundaries and focus locations.

(3) Determination of phonemic representations for words not registered in the lexicon, such as new compound words and new symbols.

(4) Context-dependent conversion of voiceless-to-voiced consonants a 2nd consonant gemination.

(5) Conversion of phonemic symbols to phonetic symbols using allophonic rules.

(6) Determination of accent types for unregistered words (mostly compound words) and prosodic words. This is done by applying accent sandhi rules and default accentuation rules.

(7) Generation of prosodic symbols using the syntactic structure and the indices of word importance.

(8) Selection of standard syllable templates as specified by the phonetic symbols. Time-patterns for the vocal-tract transfer function parameters and source intensities are generated by concatenating the selected templates.

(9) Generation of F_0 contours using the F_0 model and the prosodic symbols.

(10) Generation of the speech waveform by controlling the terminal analog speech synthesizer as indicated by the parameters for the vocal-tract transfer function and for the glottal source waveform.

The above functions (1) and (2), (3) to (7), (8) and (9), and (10), respectively, belong to the stages for linguistic processing, phonological processing, control parameter generation, and speech waveform generation.

6. CONCLUSIONS

In view of the reasons why the quality of synthetic speech is limited in the current technology, an intensive study was conducted on the synthesis of prosodic features using high-level linguistic information and on the improvement of the terminal analog synthesizer. Based on a quantitative analysis of the relationship between linguistic information and the F_0 contour of natural speech, rules were constructed for generating the prosodic features of speech. A novel terminal analog synthesizer was also proposed and constructed, which has an individual path for each type of speech with a different generation process. A total system for text-to-speech conversion was constructed for weather forecast sentences

in Japanese and the quality of the synthetic speech indicated the validity of the prosodic rules and of the proposed synthesizer.

This work was supported by a Grant-in-Aid for Scientific Research (No.01608003) from the Ministry of Education.

References

1. H. Fujisaki and S. Nagashima, "A Model for Synthesis of Pitch Contours of Connected Speech," Annu. Rep. Eng. Res. Inst., Univ. Tokyo, 28, pp.53-60, (1969)

2. H. Fujisaki and K. Hirose, "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese," J. Acoust. Soc. Jpn. (E), vol.5, No.4, pp.233-242, (1984)

3. K. Hirose and H. Fujisaki, "Analysis and Synthesis of voice Fundamental Frequency Contours of Complex Sentences," Proc. of 11th International Congress on Acoustics, Paris, pp.84-87, (1983)

4. H. Fujisaki, K. Hirose, N. Takahashi and M. Yoko'o, "Realization of Accent Components in Connected Speech," Trans. of the Committee on Speech Research, Acoust. Soc. Jpn., S84-36, (1984)

5. H. Fujisaki and H. Kawai, "Realization of Linguistic Information in the Voice Fundamental Frequency Contour of the Spoken Japanese," Proc. IEEE ICASSP 88, S14.3, pp.663-666, (1988)

6. H. Fujisaki, K. Hirose and N. Takahashi, "Manifestation of Linguistic and Paralinguistic Information in the Voice Fundamental Frequency Contours of Spoken Japanese," Proc. International Conf. on Spoken Language Processing, Kobe, 12.1, (1990)

7. K. Hirose, H. Fujisaki and M. Yamaguchi, "Synthesis by Rule of Voice Fundamental Frequency Contours of Complex Sentences," Proc. IEEE ICASSP 84, San Diego, 2.13, (1984)

8. K. Hirose, H. Fujisaki and H. Kawai, "Generation of Prosodic Symbols for Rule-synthesis of Connected Speech of Japanese ," Proc. IEEE ICASSP 86, Tokyo, 45.4, pp.2415-2418, (1986)

9. K. Hirose and H. Fujisaki, "Accent and Intonation in Speech Synthesis," Journal of IEICE, vol.70, No.4, pp.378-385, (1987)

10. H. Kawai, K. Hirose and H. Fujisaki, "Linguistic Processing in Text-to-Speech Synthesis," IEICE Tech. Report, SP88-10, pp.73-80, (1988)

11. K. Hirose, H. Kawai and H. Fujisaki, "Synthesis of Prosodic Features of Japanese Sentences," Preprints of the Second Symposium on Advanced Man-Machine Interface Through Spoken Language, Hawaii, pp.3.1-13, (1988)

12. K. Hirose, H. Fujisaki, H. Kawai and M. Yamaguchi, "Speech Synthesis of Sentences Based on a Model of Fundamental Frequency Contour Generation," Trans. IEICE, vol.J72-A, No.1, pp.32-40, (1989)

13. H. Kawai, K. Hirose and H. Fujisaki, "Rules for Generating Prosodic Features for the Synthesis of Japanese Speech," IEICE Tech. Report, SP88-129, pp.57-64, (1989)

14. Y. Sagisaka and H. Sato, "Accentuation Rules for Japanese Word Concatenation," Trans. IECE Jpn., vol.J66-D, pp.849-856, (1983)

15. H. Kawai, K. Hirose and H. Fujisaki, "Quantization of Accent Command Amplitude for Rule Synthesis of Fundamental Frequency Contours," Trans. of the Committee on Speech Research, Acoust. Soc. Jpn., SP86-93, (1987)

16. D. Klatt, "Software for a Cascade/Parallel Formant Synthesizer," J. Acoust. Soc. Am., Vol. 67, No. 3, pp.971-995, (1980)

17. H. Fujisaki, K. Hirose, H. Kawai and Y. Asano, "A system for Synthesis of High-quality Speech from Japanese text," Preprints of the Third Symposium on Advanced Man-Machine Interface Through Spoken Language, pp.12.1-16, (1989)

18. H. Fujisaki, K. Hirose and Y. Asano, "Terminal Analog Speech Synthesizer for High Quality Speech Synthesis," IEICE Tech. Report, SP90-1, pp.1-8, (1990)

19. H. Fujisaki, K. Hirose, H. Kawai and Y. Asano, "A System for Synthesizing Japanese speech from Orthographic Text," Proc. IEEE ICASSP 90, Albuquerque, S6a.5, pp.617-620, (1990)

20. H. Fujisaki and M. Ljungqvist, "Proposal and Evaluation of Models for the Glottal Source Waveform," Proc. IEEE ICASSP 86, Tokyo, 31.2, pp.1605-1608, (1986)

21. H. Fujisaki and M. Ljungqvist, "A New Model for the Glottal Source Wave form and Its Application to Speech Analysis," Trans. IEICE, vol.J72-D-II, No.8, pp.1109-1117, (1989)

22. E. Bognar and H. Fujisaki, "Analysis, Synthesis and Perception of the French Nasal Vowels," Proc. IEEE ICASSP 86, Tokyo, 31.1, pp.1601-1604, (1986)

23. K. Hirose, H. Fujisaki and Y. Asano, "A System for Speech Synthesis from Orthographic Text of Japanese," IEICE Tech. Report, SP90-42, pp.23-30, (1990)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

A Text-to-Speech System Having Several Prosody Options: GK-SS5

Ryunen Teranishi

Kyushu Institute of Design, Shiobaru, Minami-ku, Fukuoka, Japan

Abstract

This paper describes an original text-to-speech system, especially focusing on the structure and function of the recently finished prosody options. In order to synthesize supra-segmental features, an original parser based on the special bunsetsu (=phrase) grammar is used, and the prosody information is deduced automatically from the input text, which is written with kana letters and separated in each bunsetsu. For the pitch pattern composition from the above-mentioned prosody information, the "Fujisaki model" is applied. Recently, the system has been equipped with several options which make it possible to assign or to change the prosody style of the synthetic speech, systematically. These are the breath group option, the speed option, the rhythm option, the rising intonation option, and the pitch pattern expanding/flattening option. After many trials of the synthesis experiment with this system having these options, it is found that most of prosody styles and the variations which appear in the text reading mode speech are synthesized, but prosody styles in another speech mode, viz. conversation mode are hardly synthesized, and only acted conversation style can be simulated.

1. INTRODUCTION

Studies of by-rule-synthesis in Japanese have progressed remarkably, and several commercialized text-to-speech systems in Japanese have shown up one after another quite recently. Though they are not quite satisfactory for practical use in daily life, they seem to have reached a fair grade and they are expected to be improved and advanced more and more in the future. However, in the present state they have the following common faults: (1) they tend to misread *kanji* characters, (2) so far as concerning the intelligibility of the synthetic speech, they are not so nice yet; the syllables are not so articulate nor natural, comparing to those of real human speech, and (3) the supra-segmental features, viz. the prosody of the synthetic speech is still poor; they are fixed in a stereotyped pattern and cannot be changed at the user's will.

Because of these circumstances, a project concerning by-rule-synthesis in Japanese, focusing on the prosody problems, has been started at Kyushu Institute of Design in 1980 [1]. The structure of the system altered several times, seeking for a handy form suitable

for the purpose. We did not aim to make a system for practical use, but intended to construct a system with which we can easily conduct various experiments, through which we learn how to simulate the natural prosody in Japanese.

Therefore we designed it avoiding troublesome problems which are thought as secondary items from the viewpoint of the prosody research. So, in order to simplify the input system, the symbols to be input were restricted to only kana letters, which are similar to phonetic signs. The system is equipped with a word dictionary, which can be used not only for finding the accent of the word, but also for finding the class of the word for syntactic analysis, which is required for deducing the prosodic information automatically from the input sentences.

Concerning the segmental features of synthetic speech, we intended to utilize the established technology intensively, so we applied the LSP analysis and synthesis method. In our system, the principle of synthesis of the segmental features is based on the concatenation of the phonetic-unit data which are expressed in the LSP parameters. First we took phonemes as the unit, later this was changed to syllables, since then we were able to utilize the fine glide data included in each syllable, which were extracted from the spoken real human speech.Since then the system has settled using the LSP method and the CV unit system since 1985 [1]. Depending on the outcome of our research study done in recent years, we have finished equipping the system with prosody options with which the user can assign or change the prosody of the synthetic speech systematically. We call this newest version of the system GK-SS5 [2].

2. OUTLINE OF THE SYSTEM

Before describing the prosody options precisely, it would be better to give an outline of the whole system, because in this context it is easy to comprehend the role of the options. The outline of the system has been given already in other publications [1, 2], so the description is kept simple, here.

The hardware of the system can be any personal computer using "MS-DOS" and which is equipped with a DA converter. In my laboratory, a PC-9801VX of NEC with the high-speed processor 98XL-03 and extended with a 2 MB RAM board is utilized for the purpose. A "Sound Master" of Canopus Elec. Co. is installed in the computer as the DA converter. The whole system can be divided into four functional parts, as shown in Figure 1 (a).

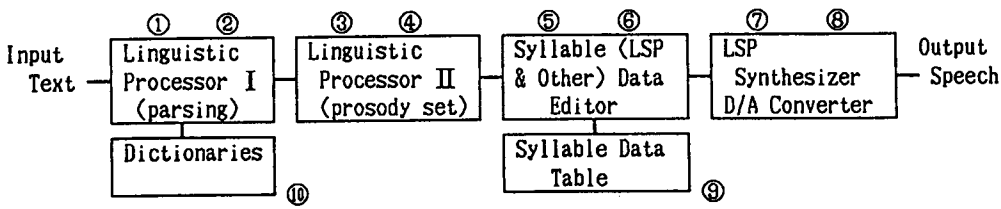


Figure 1. (a) Block diagram of the text-to-speech system GK-SS5

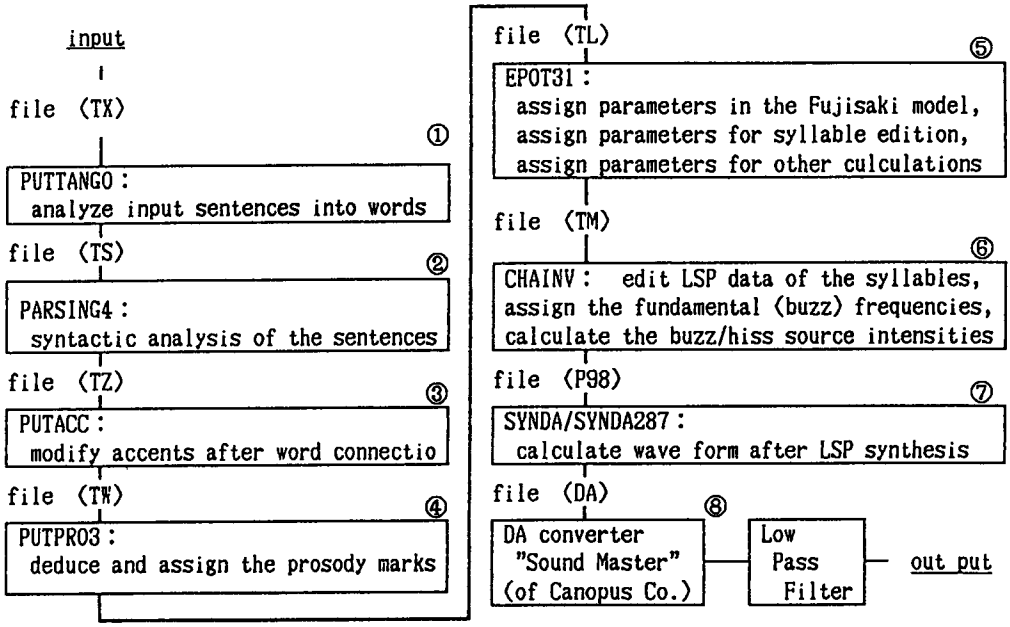


Figure 1. (b) Flow chart of the text-to-speech system GK-SS5

The software, which consists of 7 main procedure blocks and 2 tables (word dictionary and syllable data table), is stored on a 2HD type diskette. Figure 1 (b) shows the main flow of the synthesis process with the procedure blocks. A given text file at the input terminal is transformed to speech signals at the output terminal, as it is passed through each procedure block and the form and the file name converted from one to the next block. The upper half on the left side of the blocks in the picture deals with the linguistic process, and the lower half on the left side participates in the linguistic-phonetic process, while the right side upper conducts the phonetic-acoustic process and the right side lower the acoustic process.

At the first procedure block (PUTTANGO) in Figure 1 (b), each bunsetsu comprising the input text file is analyzed into an independent word and particles with the aid of a word dictionary. At the 2nd block (PARSING), each sentence in the text is analyzed into syntactic components, and the relations between them are defined using the bunsetsu grammar. At the 4th procedure (PUTPRO3), in accordance with the sentence structure and other factors, for example the rest breath volume, the sentence is divided into several breath groups, and to each of them a "phrase command" mark is attached for producing the pitch pattern. Before this work, each word is already marked "word accent" using the stored dictionary at the 3rd procedure block (PUTACC), and some of them are modified with a programmed rule, when the word is connected to the other. Thus the output file of the 4th block is composed of syllable code strings (including pauses) with (utterance-of-"phrase) command" marks and "accent (of word) command" marks.

The main work of the blocks in the right half in the picture is as follows: composing a pitch pattern using rules along the syllable code strings at the 5th block (EPOT31), editing or arranging the syllable data fetched from the stored table (NOBIT2), in which they are described using LSP parameters, and calculating the wave form of the synthetic speech at the last block (SYNDA).

3. PROSODY OPTIONS AND POSSIBLE STYLES

It seems that the important factors featuring the prosody style in text reading by a machine which simulates human recitation, are as follows: (1) breath group setting, after consideration of both of the sentence structure and the assumed speaker's breath condition, in other words, the decision rules when to put a pause and how long to continue the pause. (2) pitch (fundamental wave frequency) pattern, or the rules how to intone the voice pitch frequency during speech, (3) speed of the speaking, (4) rhythm rules for the recitation, (5) sound intensity pattern along the speech.

In the previous system, the above-mentioned factors were regarded as constant and the prosody control rules were tied to only one specific style, so that the user could not alter the prosody of the synthetic speech at his will. However, the system have been improved to one having prosody options, so the user can now choose a prosody style, or assign parameters relating to some prosody style. This is the most remarkable point of the system GK-SS5. The whole option system consists of several sub-options, viz. (1) breath capacity option, (2) speed option, (3) rhythm style option, (4) rising intonation option, (5) pitch pattern expanding or flattening option, and (6) intensity pattern modification option. Figure 2 shows a flow chart and the contents of the option system, which is installed in the 4th (PUTPRO3), the 5th (EPOT31), and the 6th (CHAINV) procedure blocks shown in the Figure 1 (b). Whenever the program execution comes to those blocks, the user can assign or change the parameter values of the items shown in Figure 2, if he likes to.

3.1. Maximum Breath Group Option

The prosody of text recitation is principally determined by two factors. One is the syntactic and/or semantic structure of the given sentence, and another is the speaker's breath condition, reflecting his physical and/or emotional state. Breath capacity is one of the most important points concerning the latter factor, and it can be expressed as the maximum duration of speech, or maximum number of morae (or syllables) within one expiration, when it is assumed that the speaking speed is kept constant.

With the first option in the Figure 2, the user can assign the reading style which seems suitable for the text. The style is controlled by two factors similar to those controlling human recitation, i.e. the given sentence and the breath capacity. In the previous system, there was a breath group limiter, of which the limit value was fixed assuming only about 2.5 s after some speech data. The prosody produced with this condition sounds quite nicely as far as the text is a simple story like a fairy tale, but it sounds unnatural when the text is an intricate article and explains some complicated subject, because the short breath group tend to split the sentence at unsuitable points. Then, the longer value (4

s) was added for reading an intricate text. Assigning the value long or short is left to the user's will. Table 1 shows the effect of the breath capacity option expressed as the speakable maximum number of morae within a breath, at the standard speech speed, i.e. 180 ms per mora. The long breath style is set as the default choice.

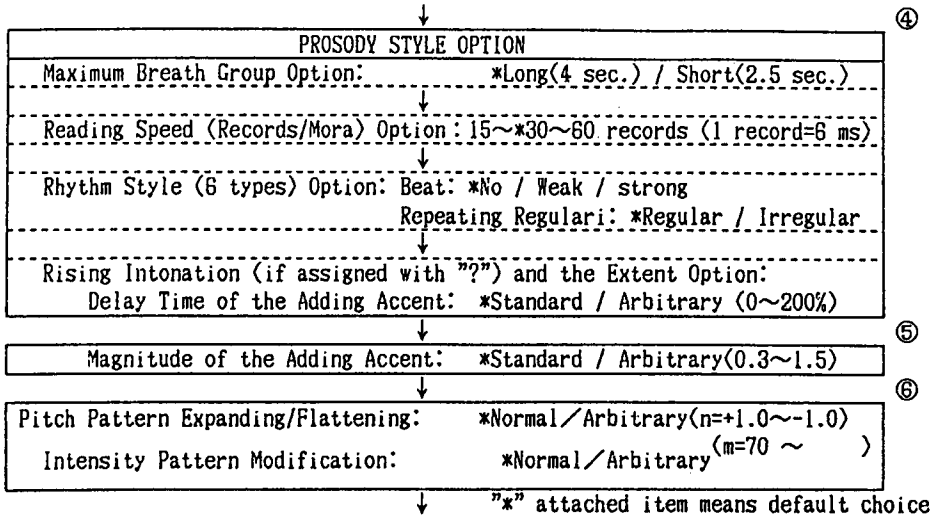


Figure 2. Structure and order of the options pertaining to the prosody style

The pitch pattern of the synthetic speech is produced in accordance with Fujisaki's model, and computed after the formula he has shown [3], using specific values of the utterance and accent component. In this system, each parameter value in the formula is set as the value shown in Table 2. As is shown there, in the short breath style the phrase component is 1.4 if it is an initial item in the breath group, otherwise 0.7. In the long breath style, those values are the same except for the case which occurs after a phrase size limiter (the newly set limiter, see Table 1); in that case the value becomes 1.1. The accent component value is 0.3 or 0.2 depending on the combined conditions, in both styles.

Figure 3 shows an example of the pitch pattern difference between the short breath style (upper picture) and the long breath style (lower). The resultant prosodies differ clearly, and the latter (lower) sounds more natural, because the sentence is a little intricate.

Table 1. Limit number of morae in the standard reading speed

	Short Breath Style	Long Breath Style
Maximum Morae in A Breath Group	15	25
Maximum Morae in A Phrase	-	15
Minimum Morae at A Sentencefinal Breath Group	5	5
Minimum Morae in A Breath-Group-Final Phrase	5	5

Table 2. The parameter values in the Fujisaki's model formula

		Short Breath Style	Long Breath Style
Fmin		70Hz	same to the left
Api	(when i=1)	1.4	same to the left
	(when i>1)	0.7	0.7 or 1.1
Aaj		0.3 or 0.2	same to the left
T_{θ}i	(when i=1)	1/αi-0.12	same to the left
T_1j		12 ms after *	same to the left
T_2j		12 ms after **	same to the left
αi	(when i=1)	2.5~8.3/s	same to the left
	(when i>1)	2.5/s	2.5~8.3/s
βj		20.0/s	same to the left
θ		0.9	same to the left

* means a syllable boundary changing low pitch to high
** means a syllable boundary changing high-pitch to low

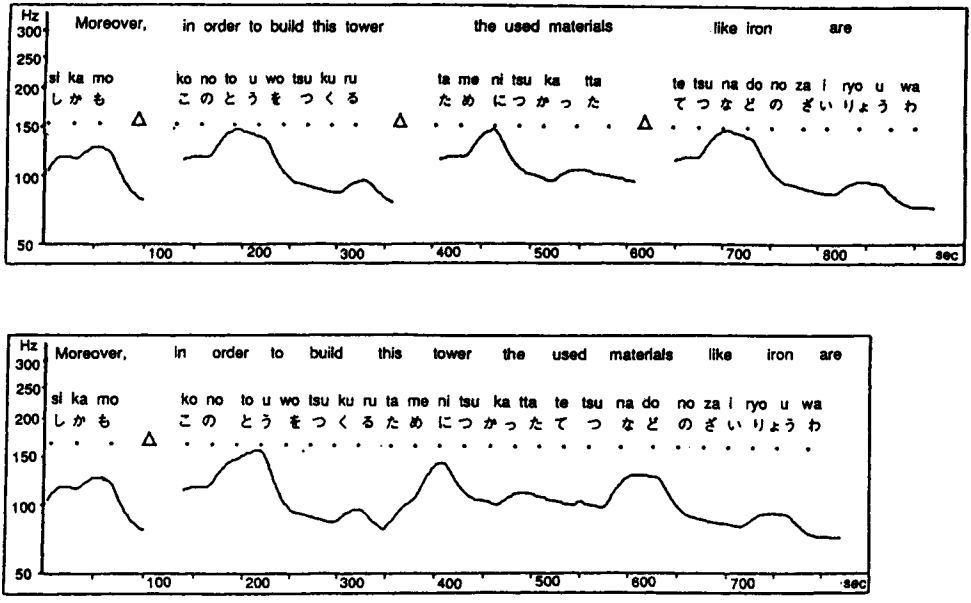


Figure 3. Example of the pitch pattern difference caused by the difference of maximum breath group length

3.2. Reading speed option

According the second option in the Figure 2, the user can choose and assign any speed from 11 morae per second to 2.8 morae per second. As default assignment, a standard speed (in this system 5.6 morae per second) is set. For speed increasing, the LSP data of each syllable are thinned homogeneously and compressed. The synthetic prosody is modulated automatically after speed assignment, so it sounds quite natural at each speed. This option has been set since 1985, improved since 1988, and already explained precisely [2].

3.3. Rhythm style option

The rhythm principle of spoken Japanese, especially for text recitation, is well-known as the "isochrony of mora". We will call this the first principle. However, another principle is proposed by Teranishi [4] and others.

Almost all the Japanese recite texts in this style whenever they are in one group and reading the text in chorus. We can hear a typical example in a common class room. Individual persons also uses this style commonly whenever he or she wants to speak very slowly. Then it seems that people tend to comply with the first principle at quick or moderate speaking speed, and with the second principle at moderate or slow speed. So at moderate speech speed, both the two types, based on the two principles, are possible to exist, and one of them is chosen by the speaker depending on various circumstances.

With the third option in the Figure 2, the user can choose and assign one of the 6 prepared rhythm styles, which are based on the first or second or the mixed principle. The prepared rhythm styles are defined as shown in the Table 3. Style I is set as the default choice.

Table 3. List of 6 kinds of rhythm

Beat	$D_1:D_2$	D_3	$D_1:D_2:D_3$ or $(D_1+D_2):D_3$	Rhythm type	No.
Without	1:1	$D_3=D_2$	1:1:1	Regular(in-tempo)	I
		$D_3=D_1+D_2$	1:1:2	Irregular	II
Weak	1:1.5	$D_3=D_2$	2.5:1.5	Irregular	III
		$D_3=D_1+D_2$	2.5:2.5	Regular(in-tempo)	IV
Strong	1:2	$D_3=D_2$	3:2	Irregular	V
		$D_3=D_1+D_2$	3:3	Regular(in-tempo)	VI

Let us assume that to each mora of input sentence a mora number in a given *bunsetsu*, the number 1 or 2 is attached alternately and repeatedly from the head of each *bunsetsu*, as shown in Figure 4. Each means the 1st or 2nd mora in the cluster. If the *bunsetsu* consists of an even number of morae, it is divisible by 2, so it does not leave an odd mora, but if it consists of an odd number of morae, it cannot be divided by 2, and an odd mora is left. This odd mora is numbered as the 3rd. In the Table 3, the duration of the 1st, 2nd, and 3rd mora are expressed as D1, D2, and D3.

Figure 4 shows the feeling of the resultant 6 rhythm style, using the same sentence. Rhythm style I is an extreme form of the "isochronal mora" and style VI is an extreme

Input Sentence	yamano kino hano u e ni yasun de i masitaga, kazeni yurio tosarete, o o zei やまの きの はの うえに やすんで いましたが、かぜに ゆりおとされて、おおぜい
Mora Number	1 2 3 1 2 1 2 1 2 3 1 2 1 2 1 2 3 1 2 3 1 2 1 2 1 2 3 1 2 1 2
Without Beat Regular I	yamanokinohano u eni yasun de imasitaga kazeniyuri otosarete. o o zei notomodati やまのきのはのうえに；やすんでいましたが；かぜにゆりおとされて；おおぜいのともだち •••••••••••△•••••••••△•••••••••△•••••••••
Without Beat Irregular II	yamano--kinohano u eni-- yasun de imasitaga-- kazeni--yurio tosarete-- o o zei やまのーきのはのうえにー；やすんでいましたがー；かぜにーゆりおとされてー；おおぜい ••• ••••••• △••••••••• △••• ••••••• △••••
With Beat Irregular III,V	yama- no- kino- hano- u e - ni- yasu- n de- i ma- sita- ga- kaze- ni- yuri- o やまーのーきのーはのーうえーにー；やすーんでーいまーしたーがー；かぜーにーゆりーお ◉• ⊙ ◉• ◉• ◉• ⊙ ◉• ◉• ◉• ◉• ⊙ ◉• ⊙ ◉• ◉
With Beat Regular IV,VI	yama- no-- kino- hano- u e - ni-- yasu- n de- i ma- sita- ga-- -- kaze- ni-- やまーのーーきのーはのーうえーにー；やすーんでーいまーしたーがーー； かぜーにー ◉• ◉ ◉• ◉• ◉• ◉ ◉• ◉• ◉• ◉• ◉ ◉ ◉• ◉

Figure 4. Example patterns of a recited sentence in 6 classes of rhythm style

form of the "isochronal beat" or "isochronal cluster". If we listen them after order of the rhythm number I, III, V, II, IV, VI, the feeling of the rhythm shifts quite smoothly.

3.4. Rising intonation and the extent option

According to this option, the user can assign various styles of the rising intonation. The rising style specification is assigned in three steps. At the first step, the rising intonation has to be assigned with the special mark "?" in the text, instead of the period mark ".". This means, unless the question mark shows up the ordinary falling intonation is set. After this assignment, the user is asked precise specification about the style he wants to make.

Principally the rising intonation style is actualized by adding an accent component to the sentence's final mora. Within this style, there are many variations, which convey various kinds of feeling, however, they can be systematically arranged with 2 parameters relating to the accent component to be added. So there are 4 possible types of pitch pattern for the sentence's final 2 morae, which are related to the accent to be added. Thus, the two parameter values, the value of the delay timing (this brings prolongation in the last mora) and the magnitude of the added accent component, with good consideration for the pitch pattern of the sentence's final pitch pattern type of the given sentence. In case of the default of this specification, the prepared standard values shown in the Table 4 are set automatically.

Figure 5 shows some examples of rising intonation patterns of the same sentence specified with different parameter values,which are as follows: (left to right) prolongation; 0%, 50%, 50%, magnitude (Aaj) 0.3+0.9, 0.3+0.9, 0.7+0.9.

Table 4. Standard values of accent component and delayed timing for generating the rising intonation

type of the ending 2 morae		magnitude of the adding accent component (Aaj)	timing delay of the adding accent in percentage
• • • • •	I	0.7	10 %
• • • • •	II	0.9	50 %
• • • • •	III	0.9	20 %
• • • • •	IV	0.9	30 %

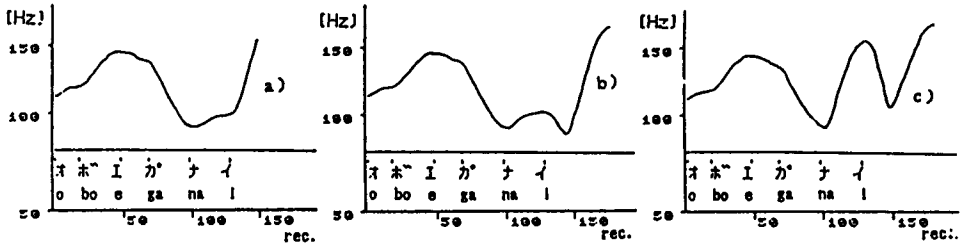


Figure 5. Example of the pitch pattern difference, within the rising intonation style, caused by the different parameter values of the added accent component. The literal meaning of this sentence is "You don't have the memory ?"

3.5. Pitch pattern expanding/flattening option

The user can modify the composed pitch pattern to expanding or flattening, by means of assigning two specific values of the continuous two variables "m" and "n". The variable "m" is the applicable lower limit frequency of the modification, and can be chosen as any value above 70 Hz. The other variable, "n", is the modification coefficient, of which the role is shown in the equation $f' = f(f/m)^n$. Here f is each frequency comprising the original pitch pattern, and f' is each frequency in the modified new pattern. The user can assign any value between -1 and +1 as "n". In case of the default of this assignment, $n = 0$ is set, and it is clear from the equation that no modification happens in this case.

4. CONCLUDING REMARKS

After many trials of the synthesis experiment with system, especially with the help of a combination of various optional parameter assignments, it is found that most of the prosody styles which appear in text reading mode speech can be synthesized, but the prosody styles in other speech modes, viz. conversation style can be simulated manual, not by rules from the input text. However, it is believed that to make it by rules may be possible by adding a few more paralinguistic marks to the input letters. It is also found that the rising intonation option and the pitch pattern expanding option are useful in simulating such a prosody style.

References

1. R.Teranishi, "A speech synthesis system by rule in Japanese; GK-SS4," Annual of the Information Processing Center Kyushu Inst. of Design, No.6, pp.17-29, (1987)

2. R.Teranishi, "A speech synthesis system by rule in Japanese at Kyushu Geikodai; GK-SS4/5," Preprints of the Second Symposium on Advanced Man-Machine Interface Through Spoken Language, pp.14,1-14,8, (1988)

3. H.Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in P.F. MacNeilage ed. The production of speech, Springer-Verlag, pp.39-55, (1983)

4. R.Teranishi, "Two-moras-clusters as a rhythm unit in spoken Japanese sentence or verse," J.A.S.A. vol.67 Suppl.1, pp.40, (1980)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

A Prolog-Based Automatic Text-to-Phoneme Conversion System for British English

J. Laver, J. McAllister, M. McAllister and M. Jack

Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK

Abstract

One of the challenges facing phonetics in the area of speech technology is the task of making speech knowledge explicit in a rule based form such that a computer can read text out loud in an intelligible fashion. The applications of a text-to-speech system of this sort are potentially very numerous. This paper describes a system, written in Prolog (extending to over 3,000 clauses) for automatic text-to-phoneme conversion for British English, offering details of modules for textual anomaly normalisation, word level pronunciation assignment, syntactic processing and word boundary phonology. Results of evaluation studies of the text-to-phoneme system are included.

1. INTRODUCTION

An automatic text-to-speech (TTS) system is a computer system which can take written text as input, and produce audible and intelligible speech as the output. There are two major components in such a system: a linguistic processor for converting the orthographic text to a phoneme string annotated for intonation, and a synthesiser for converting this string into speech sounds. In this paper, detailed consideration will be given only to the linguistic processor and comments on intonation and synthesis will be largely omitted. The internal strategy of the linguistic processor (text-to-phoneme) described here consists of a number of operations:

1. Preprocessing of the text to regularise anomalous forms such as acronyms.
2. Examining each word in the input text to establish its morphological make-up.
3. Stripping off any suffixes for separate treatment.
4. Looking up the pronunciation and grammatical category of the remaining morphological core of the word.
5. Assigning word-stress.

6. Making appropriate morpho-phonemic adjustments to the pronun- ciation of the whole recomposed word.

For those cases where the word is not morphologically complex or is not represented in the main dictionary of the system, then the pronunciation is generated by using a set of grapheme-to-phoneme spelling conversion rules. In addition, the overall intonation and rhythm of the utterance is generated on a phrasal basis partly with the help of a syntactic parsing mechanism. The structure of the system is illustrated overleaf.

Text in normal orthography is input to the system via a keyboard or is read from a file. Textual Anomaly Normalisation (Laver 1988) identifies any orthographic strings which do not conform to the system's limited notion of what constitutes a 'word' (e.g. digit sequences, abbreviations) and converts them into a form which is capable of being processed by the other modules of the system. The Syntax module performs an analysis of the sentence which will be used both by the Word Level Pronunciation modules and by the Intonation Assignment. The output of Textual Anomaly Normalisation is passed to a group of modules whose task is to generate the pronunciations of individual words. Morphological Analysis determines the morphological structure of words and assigns a pronunciation to each morph, either through dictionary lookup or (where this fails) by grapheme-to-phoneme conversion. In some instances, it is necessary to select between a pair of alternative pronunciations (for example, in the case of the noun and verb interpretations of the word **house**). This is one of the tasks of the Phonological Disambiguation module, which also resolves ambiguities in the suprasegmental properties of affixes. Morph Boundary Phonology introduces phonological modifications which arise when particular morphs are concatenated (for example, the change from /k/ to /s/ at the morph boundary when the suffix ity is added to the stem **electric**). In cases where lexical stress pattern is not marked in the dictionary, this is determined by rule in the Lexical Stress Assignment module. Vowel quality changes associated with stress and other phonological factors are introduced by the Vowel Reduction module. Finally, Intonation Accent Placement identifies those syllables in the sentence which should bear the major fundamental frequency movement. The resulting annotated phonemic string can then be passed to a speech synthesiser which produces audible output. We can now examine the structure of the main modules of the linguistic processor in more detail.

2. TEXTUAL ANOMALY NORMALISATION

The purpose of the textual anomaly normalisation process is to intercept and treat anomalous textual strings, i.e. strings which do not conform to the conditions of neutrality which characterise the input to the other modules of the system. This definition assumes preprocessing mechanisms for receiving textual data into the system, segmenting the character stream into strings, and separating the core textual strings from any meta-textual characters (such as punctuation) which may be appended to them. The Prolog interpreter environment provides facilities such that any program which expects input from the keyboard can have data directed to it from a file or the output of another operating system function. Therefore, the preprocessing and anomaly normalisation module is constructed as if input is always from the keyboard, and these system facilities are

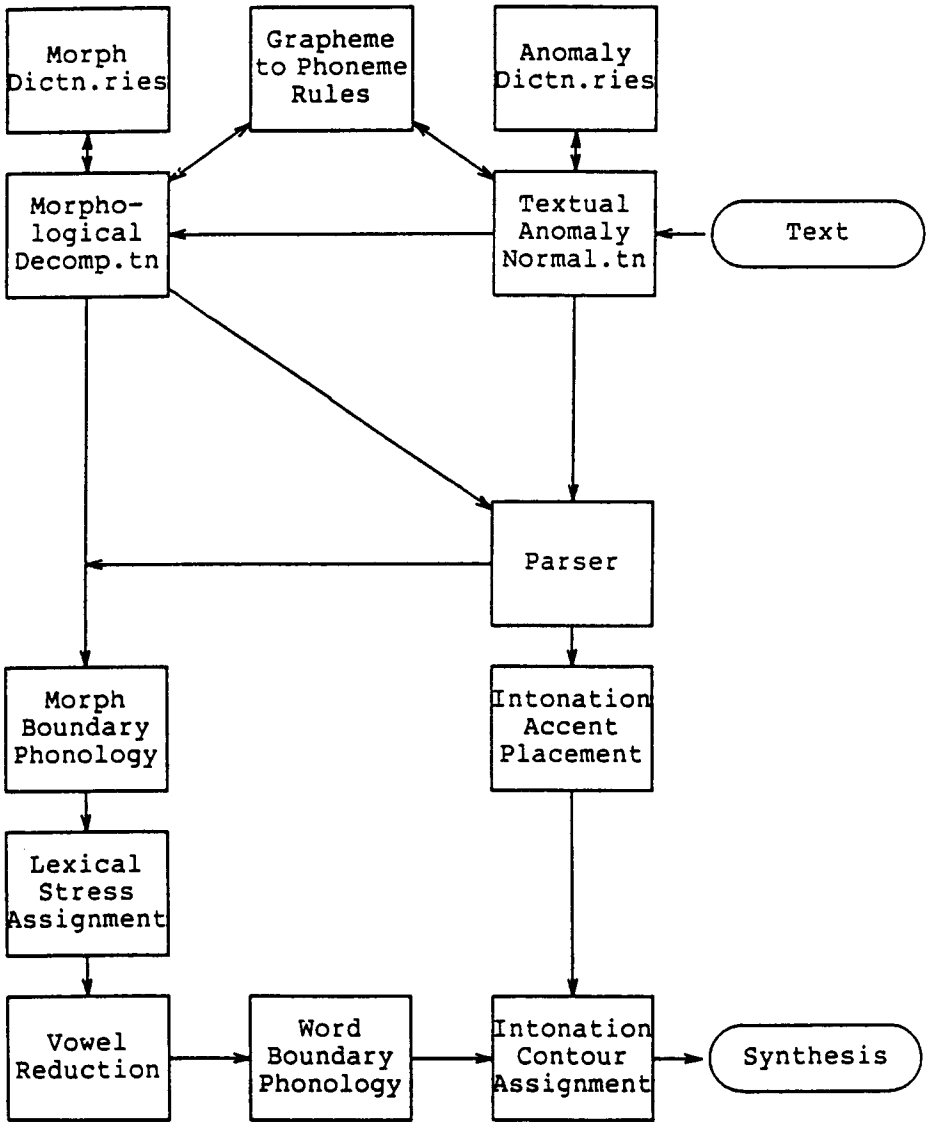


Figure 1. CSTR Text-to-Phoneme System.

invoked when this is not the case. The segmentation of the character stream into strings is primarily performed on the basis of 'orthographic islands', i.e. any strings delimited by space, tab or newline characters. (These characters form the set responsible for the production of white space in a printed version of a text). The only exception to this simple mechanism is the provision of a strategy to recombine words which have been hyphenated across a line break. The identification, removal and labelling of meta- textual characters from the extremes of strings has to take into account several interacting factors. Single quotes, for example, are sometimes used abutted to the beginning of otherwise neutral strings as an indication of an abbreviated form of a commonplace word (e.g. **'cause** for **because**). The anomaly dictionaries must be consulted prior to stripping non-alphanumeric characters in case those characters are a valid part of a known anomaly. Various abbreviations, acronyms, trade and technical names contain non- alphanumeric characters at their extremes, most commonly abbreviations terminated by a full stop. The closing item of paired punctuation characters (e.g. brackets) should only be stripped after a valid occurrence of the appropriate opening item. Non- alphanumeric characters which might form a valid part of a number string should not be removed where they abut digits in the text string. The precedence of these strategies is of particular importance as they can, at times, suggest contradictory actions.

Anomalies found in the anomaly dictionary are returned as a triple containing the dictionary entry and empty affix lists. Lower case strings are returned as themselves for decomposition. Special rules are invoked to deal with the pronoun I and the capitalised article **A** at the start of a sentence. Individual upper or lower case letters are given their letter name pronunciation, marked as proper nouns and returned as triples. Capital initial strings at the start of sentences are decapitalised if not found in the proper name dictionary. Capital initial strings elsewhere in sentences are marked as proper names whether found in the anomaly dictionary or not. In the case of any other anomalies, marked affixes are stripped, and the transcript of each alphabetic or numeric sub-string or each character name are appended into in the triple format. This mechanism allows the normal morph boundary phenomena to apply to the composite pronunciation.

Number strings may be divided into unit-based and unitless numbers. Integers, real numbers, fractions, ranges, factorials and comparators are all unitless. Of the unit-based number strings only quantity strings and currency strings have true, identifiable units. The remainder – sequence numbers, percentages, ordinals, times and dates – have what may be regarded as 'pseudo-units'. This distinction is made because, rather than the sparse set of combinations currently parsed by the system, it should be possible to combine every unitless type with the two true unit-based types. The omission of these combinations has been done on probabilistic grounds. The likelihood of say, a factorial expression of currency such as **$14!** is extremely low.

In order to evaluate the module, a set of anomaly-rich texts were taken from newspapers, technical journals, and other paper copy textual media. These texts were used as input data for the preprocessor and anomaly normalisation module, first in isolation then as part of the full system. In this way, separate judgements could be made on the accuracy of textual string separation and anomaly detection and the appropriacy of the

anomaly normalisation procedure selected. The results are shown below.

Anomaly Type	No. tested	No. correct	Accuracy
Meta-textuals	95	95	100%
Capital-initials	87	87	100%
Digit-bearing	35	34	97%
Hyphenation	15	15	100%
Other	24	24	100%
Proper names			72.5%
Strings in Dictn.ry			84.5%
Strings not in Dictn.ry			86.5%
Number strings			100%

3. WORD LEVEL PRONUNCIATION ASSIGNMENT

The philosophy adopted in the system has been to assign pronunciation by dictionary lookup wherever possible, and to resort to grapheme-to-phoneme rules only when the latter approach failed. It is well-known that, in English, morph boundaries frequently block the operation of grapheme-to-phoneme rules (Lee, 1969); for example, the rule which states that the sequence **sch** is pronounced /s k/ (as in **school**) does not operate in the word **mischance** because of the presence of the morph boundary between the s and the c. In order to determine the morphological structure of the word, it is analysed by the Morphological Decomposition module, which has access to a morph dictionary containing information about word stems, as well as to a list of the affixes used by the system. If no satisfactory analysis of the word can be found, all or part of it may be passed to the Grapheme-to-Phoneme module whose task is to arrive at a segmental transcription of the word, unmarked for stress and containing full vowels. Ambiguities in the pronunciations of words, as listed in the morph dictionary, or in the labels attached to affixes, are resolved by the Phonological Disambiguation module. Three final adjustments must be made to the segmental representations of words before they leave Word Level Pronunciation Assignment: modifications at morph pronunciation of the plural marker; lexical stress assignment, in cases where stress is unmarked in the dictionary or when the presence of certain affixes will result in stress shift; and reduction of certain vowels in unstressed syllables. The Morph Boundary Phonology, Lexical Stress Assignment and Vowel Reduction modules are responsible for these operations.

The morph dictionary consists of about 7,500 entries and was constructed by analysing the first 15,000 words in the American Heritage Word Frequency Book (a frequency-ordered list) into constituent morphs. In addition, some polymorphemic words which were exceptions by the criteria of other modules were included in the dictionary. For example, the word finish is considered by some writers to be polymorphemic (cf forms such as **final, infinite**). However, it is included in our dictionary because, as far as our system is concerned, the suffix **ish** signal an adjective (e.g. **peckish, loudish**). The dictionary contains almost 7,000 entries which are free forms of various kinds and over 500 bound stems. Each dictionary entry details (in the form of a Prolog list) the orthographic transcription of the word (a Prolog string), its phonological transcription (a list of Prolog atoms), grammatical category information (a Prolog list, which will be empty if the form

is a bound root), information about the morphological status of the form (i.e. whether it is free or bound) and a variable number of flags relating to aspects of the processing of the word. An example of a dictionary entry, for the word **crown**, is shown below:

["crown", [k, r, au, n], [noun([]), verb([main, bse])], free].

4. MORPHOLOGICAL DECOMPOSITION MODULE

The task of the morphological decomposition module is to provide a morphological analysis which will lead to correct segmental pronunciation. For example, the word **mishap** will be pronounced /" m i sh a p/ if the morph boundary is not identified. It also assists correct lexical stress assignment. For example, the word **engineer** will be stressed on the first instead of the third syllable if the suffix **-eer** is not identified. Similarly, it offers grammatical category information, certain affixes providing islands of reliability for the syntax module (e.g. the suffix **ity** unambiguously signals a noun).

The mechanism itself is the least important part of the module. The only important decision here has been to strip suffixes before prefixes. The morph dictionary provides the Morphological Decomposition module with information about any words that are considered to be exceptions for any reason. The dictionary has the highest precedence in decomposition; that is, if a word is found in the dictionary then no further decomposition will be attempted. The affix lists define the prefixes and suffixes which the system uses. These lists contain the orthographic and phonemic transcriptions of the affix, its word class properties, its stress properties and its weighting. The stem adjustment module ensures that stems are correctly adjusted when an affix has been removed. For example, the double orthographic consonant in **running** is due to the presence of the suffix **ing**.

Evaluation has been carried out using a composite list made from two machine-readable dictionaries plus a list of the word types in a large corpus of running text (over a million word tokens). The number of word types in this composite list was in excess of 85,000. From this list, 500 words were randomly selected. The 500 words were passed through the morphological decomposition module and the outputs were examined. The module performed at a 69.6% level of accuracy.

5. MORPH BOUNDARY PHONOLOGY

The Text-to-Phoneme conversion system is one whose linguistic processing is firmly centred on a morphological analysis of words input to it. This strategy yields a great deal of information about the stress pattern and possible word classes of a word and allows a more compact storage of segmental information. Pronunciation data are linked to the morphs that make up words and not the words themselves, thereby giving a generative expansion of the number words which can be handled against the number items stored. Component transcriptions, however, are generic across many instantiations of a morph and as such often need adjustment to take into account their phonological context.

For example, the past tense marking suffix **-ed** is stored with a transcription of /i d/, but it may take other forms in different contexts. Consider the word **fated, famed** and

faced. After morphological analysis, these give the stems **fate, fame** and **face**, respectively, followed by the suffix **ed**. Their pronunciations undergo transformations based on a pattern matching algorithm using a set of rules represented as facts in the Prolog database. The pattern matching algorithm (or 'engine') deals with the data structure of the decomposed word independently from the structure of the phonological rules. This allows both structures to be changed independently at any time without imposing constraints on the other. As well as comparing the segment sequences on either side of each morph boundary within the word with the patterns in the rules, the engine also allows word-final sequence to be matched against specifically marked rules.

6. LEXICAL STRESS ASSIGNMENT

The task of Lexical Stress Assignment (LSA) is to mark primary and secondary stress on the phonemic string which is passed to it by the Morph Boundary Phonology module. There is a strong theoretical and computational link between LSA and Phonological Disambiguation, on the one hand, and Vowel Reduction, on the other. The input to the Lexical Stress Assignment module is in the form of a Prolog list which contains both morphological and phonological information. Fudge (1984) defines the **stressable portion** (SP) of a word in terms of its morphological structure. The SP of a word is found by removing any affixes which bear the 'sn' accentual property label. For example, the word **inversion** is specified as follows:

```
[[[i, n], sr v]],
[[v, @@, sh]],
[[[@, n], psl, [[verb ([main, ]), noun([])],
[adj (bse), noun ([])]]],
[[z], sn, [[noun ([]), noun ([])],
[verb (main, bse]), verb ([main, gen])]]]]]
```

For the word **inversions**,, the SP is **inversion**. The 'get_sp' predicate in the LSA program removes all the information relating to -s from the suffix list, and primary stress is assigned to **inversion**.

A distinction is maintained in the LSA module between morphologically complex SPs, which contain suffixes and/or prefixes, and morphologically simple SPs, which do not. The latter, which the system can recognise by virtue of their empty prefix and suffix sub-lists, are assigned primary stress on the basis of phonemic information (specifically, the number and perhaps also the composition (or 'weight') of the syllables in the SP). In English, syllables consist of an obligatory peak (usually a vowel, although /m, n, r/ and /l/ may function as peaks) and optional preceding and/or following consonants. These optional consonant sequences are termed the onset and coda, respectively.

The basic rules for secondary stress placement use as the relevant criteria position of primary-stressed syllable relative to the beginning of the word and syllable weight, and are as follows:

1. If the second syllable receives primary stress, there is no s econdary stress (e.g. **lapel**/l @ "p e l/, **veranda** /v @ "r a n d @/, **America** /@ "m e r i k @/).

2. If the third syllable receives primary stress, the first syll able receives secondary stress, (e.g. **aluminium** /'a l y u "m i n i @ m/, **panorama**/'p a n @ "r aa m @/).

3. If the fourth (or later) syllable receives primary stress, th e weights of the preceding syllables must be considered as fo llows: call the syllable receiving primary stress syllable n . If syllable n-2 is strong, assign primary stress to syllab le n-2 (e.g. **encyclopaedia** /i n 's ai k l @ "p ii d i @/); otherwise, stress syllable n-3 (e.g. **pharmacopoeia** /'f aa m @ k @ "p ii @/).

The domain of the secondary stress rules is the whole word, not just the SP.

7. GRAPHEME-TO-PHONEME RULES

The final version of the grapheme-to-phoneme rules were tested on a total of 2409 free forms listed in the dictionary. When the outputs of the grapheme-to-phoneme module were compared with the dictionary transcription (on a segment by segment basis, ignoring markers) 75% were correctly transcribed. Some of the mismatches are due to the presence of reduced forms in the dictionary. Accordingly, the 611 entries which did not agree with the dictionary transcriptions were passed through the lexical stress assignment and vowel reduction modules and the total number of correctly transcribed items was then 2244 (93%).

8. WORD BOUNDARY PHONOLOGY

Word boundary phonology is a term used to denote those changes which words undergo when they are uttered in connected speech. For example, the citation form (i.e. the form produced when the word is spoken in isolation) of the word **ten** is /t e n/ whereas in connected speech this would change as in **ten miles** as ["t e m "m ai l z]. Two kinds of change may be expected depending on the quality of a following consonant:

1. If a following word begins with /k/ or /g/, the final segment of the word **ten** is realised as /ng/:

   ten cats ["t e ng "k a t s];
   ten geese ["t e ng "g ii s].

2. If the following word begins with /p/, /b/, or /m/, the final segment of the word ten is realised as /m/:

   ten people ["t e m "p ii p @ l];
   ten boys ["t e m "b oi z];
   ten men ["t e m "m e n].

Similar kinds of processes are found when word-final /n/ is followed by dental consonants (/th/, /dh/). A more general way of describing this kind of effect is to say that a word-final nasal segment assimilates to the place of articulation of the initial consonant of the following word. Assimilation is a common process in the production of connected speech,

affecting even carefully articulated, formal utterances such as those produced by television newsreaders (Brown 1977).

The rules were the result of a survey of several linguistics texts: e.g. Gimson (1980) and Brown (1977). Since the task of the rules is to convert the phonological input to its phonetic counterpart, in the form of a printed string representing many of the segmental characteristics of the eventual output of the system, a phonetic transcription is introduced in this module. Such a modification is necessary because elements such as the dental nasal, which do not function at the phonemic level in English, are frequently realised at the phonetic level.

9. VOWEL REDUCTION RULES

In English, many vowels in unstressed syllables (that is, in syllables not assigned primary or secondary stress) are realised at the phonetic level as the central vowel /@/. Consider, for example, the second vowel in each member of the following word- pairs:

invoke [i n "v ou k]	invocation ['i n v @ "k ei sh @ n]
maintain [m ei n "t ei n]	maintenance ["m ei n t @ n @ n s]

While the first member of each pair contains a full vowel , in the second syllable, the vowel in the second member is reduced. Similarly, some unstressed vowels are realised as [i]:

recite [r i "s ai t]	recitation ['r e s i "t ei sh @ n]
allege [@ "l e jh]	allegation ['a l i "g ei sh @ n]

This phenomenon is known as vowel reduction.

The corpus used in testing the vowel reduction module consisted of the 286 words correctly assigned stress in the LSA evaluation (see Section 6). These words were processed by the Vowel Reduction module and the results assessed for the correctness of vowel quality. It should be noted that correctness' here was interpreted in the light of the input passed to the program by higher-level modules such as Grapheme-to-Phoneme Conversion and Morph Boundary Phonology. In the test, 258 words (91%) were processed correctly.

10. SYNTAX

The nature of other modules in our system, particularly the morphological and intonation modules, places rather unusual constraints on the type of syntactic analysis which is required. In order to handle unrestricted text, which is not necessarily grammatical (in the sense of constituting valid English sentences), the syntactic analysis must provide extremely wide coverage of grammatical and quasi-grammatical sentences and sentential fragments. However, unlike most practical parsing systems, the parser must work without the benefit of fully specified lexical entries, since relevant syntactic information such as subcategorisation possibilities for verbs is not currently available from the TTS lexicon (and perhaps not from any morph- based lexicon). Moreover, for morphologically simple

words which are not members of one of the closed word classes, and for many morphologically complex items with syntactically ambiguous affixes, the parser must choose between a variety of different word class hypotheses which in the worst case might encompass noun, verb, adjective and adverb.

As the combination of all these factors would cause massively nondeterministic behaviour in the operation of many normally well-behaved sophisticated parsing formalisms, our syntactic analysis is based on a framework of simple phrase structure grammar. A highly sophisticated theory of grammar would also be largely redundant, since the syntactic analysis required by the intonation module need not be very detailed.

The grammar is written in an extended DCG notation which is easily understood by linguists, and is then translated before run time into a more efficient Prolog representation. The basic parsing strategy is a bottom-up, left-corner one, pursuing alternative parses serially which also has considerable predictive power, and is based on the BUP parser (Matsumoto et al 1983). The parser also keeps track of both failed and successful goals at run time by building a well-formed substring table (wfst), containing both types of information: this strategy significantly decreases the search space at any given point in a parse.

11. SYSTEM LEVEL EVALUATION

In addition to the individual evaluations of specific modules which are reported above, an evaluation of the system as a whole has been conducted. A corpus of 100 sentences was constructed by selecting 50 sentences from recent newspapers and journals and 50 Harvard sentences. The 'newspaper' sentences (mean length 20.48 words) were considerably more complex than the Harvard sentences (mean length 7.92 words).

The segmental outputs were then examined by hand. The number of words in each sentence which were assigned an inappropriate segmental pronunciation (taking into account phonemic context) was calculated, and the results are presented below.

Source	Total words	Words incorrect	%age correct
Newspaper sentences	1024	97	90.5
Harvard sentences	396	12	97.0
Total	1420	109	92.4

12. FUTURE WORK

The architecture of the text-to-phoneme conversion system described in this paper has been shown to be sufficiently robust to permit extension of the coverage of the system

to a "whole language" level. This does not include the highly specialised vocabularies of science or medicine nor the unusual phonology of a complete set of proper names. The eventual objective of the project will be for the system to be able to produce an intelligible version of any grammatical sentence capable of being written in ordinary English.

Acknowledgements

Many colleagues have contributed to the work of the project reported in this paper. They include R. Cann, H. Fraser, L. Friedman, D. Ladd, J. Lothian, D. McCluskey, J. Miller, A. Monaghan, M. Reape, J. Scobbie and L. Shockey. We gratefully acknowledge the support of Nippon Electric Company, C&C Laboratories, Tokyo, and the encouragement of Dr. Y. Kato.

References

1. G. Brown, "Listening to Spoken English," Longman, (1977)

2. E. Fudge, "English Word Stress," George, Allen and Unwin, (1984)

3. A. C. Gimson, "An introduction to the pronunciation of English," Edward Arnold, (1980)

4. J. Laver, M. McAllister and J. McAllister, "Preprocessing of anomalous text strings in an automatic text-to-speech system," In Ramsaran, S. (ed) "Studies in the Pronunciation of English: a commemorative volume in memory of A.C. Gimson," Croom Helm, (1988)

5. F. F. Lee, "Reading Machine: from Text to Speech," IEEE Transactions on Audio Electroacoustics 17, 275-282, (1969)

6. Y. Matsumoto, H. Tanaka, H. Hirakawa, H. Miyoshi and H. Yasukawa, "BUP: a bottom-up parser embedded in Prolog," New Generation Computing 2, (1983)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Data-Bank Analysis of Speech Prosody

Gunnar Fant, Anita Kruckenberg and Lennart Nord

Dept. of Speech Communication and Music Acoustics
Royal Institute of Technology (KTH), Stockholm, Sweden

Abstract

This is a brief summary of analysis procedures and results from studies of speech prosody and individual variations in text reading. Within a larger perspective we want to derive rules for good reading performance and rules related to voice types, sex, age, and reading style. We shall here discuss segmentation problems and provide data on objective and subjective studies of syllabic stress, speech rhythm, and the realization of phrase and sentence boundaries.

The general principles for the organization of our data base and of the search for durational rules have been developed by Carlson and Granström [1 - 4]. A first report on segmentation issues and on observations of individual speaker variations were given by Fant, Nord, and Kruckenberg [5, 6]. Results from studies of syllabic stress and phrase boundary marking have recently been reported [7 - 9]. A more comprehensive report on our work will appear in the STL-QPSR, 2/1989, Fant and Kruckenberg [10].

1. SEGMENTATION

We start out with a provisional phonemic transcription and segmentation, [2], which conforms with the letter-to-sound rules of the RULSYS synthesis. A combination of an initial automatic segmentation and a following careful manual correction and editing with oscillograms and spectrograms as a visual reference, see Figure 1, produces a temporally defined string of phoneme segments which is the basis for the durational studies.

This process is apparently not unproblematic. Phonemes may be so weakly manifested and be subjected to such extreme temporal spread that signal-driven segmentation strategies fail.

We often have to decide that a phoneme be given a zero duration but we retain its positional address. Irrespective of drop outs and fuzzy realizations, our system allows us to analyze and describe contextual and individual variations by rules. This seems preferable to attempting an initial narrow phonetic transcript thus avoiding ambiguities and inconsistencies of subjective notations.

A few examples follow. Boundaries are more clearly realized by changes in "manner" cues than in "place" cues. Thus, it is easy to find the boundary between a fricative and a vowel, but we have no consistent rules for defining boundaries between adjacent vowels or

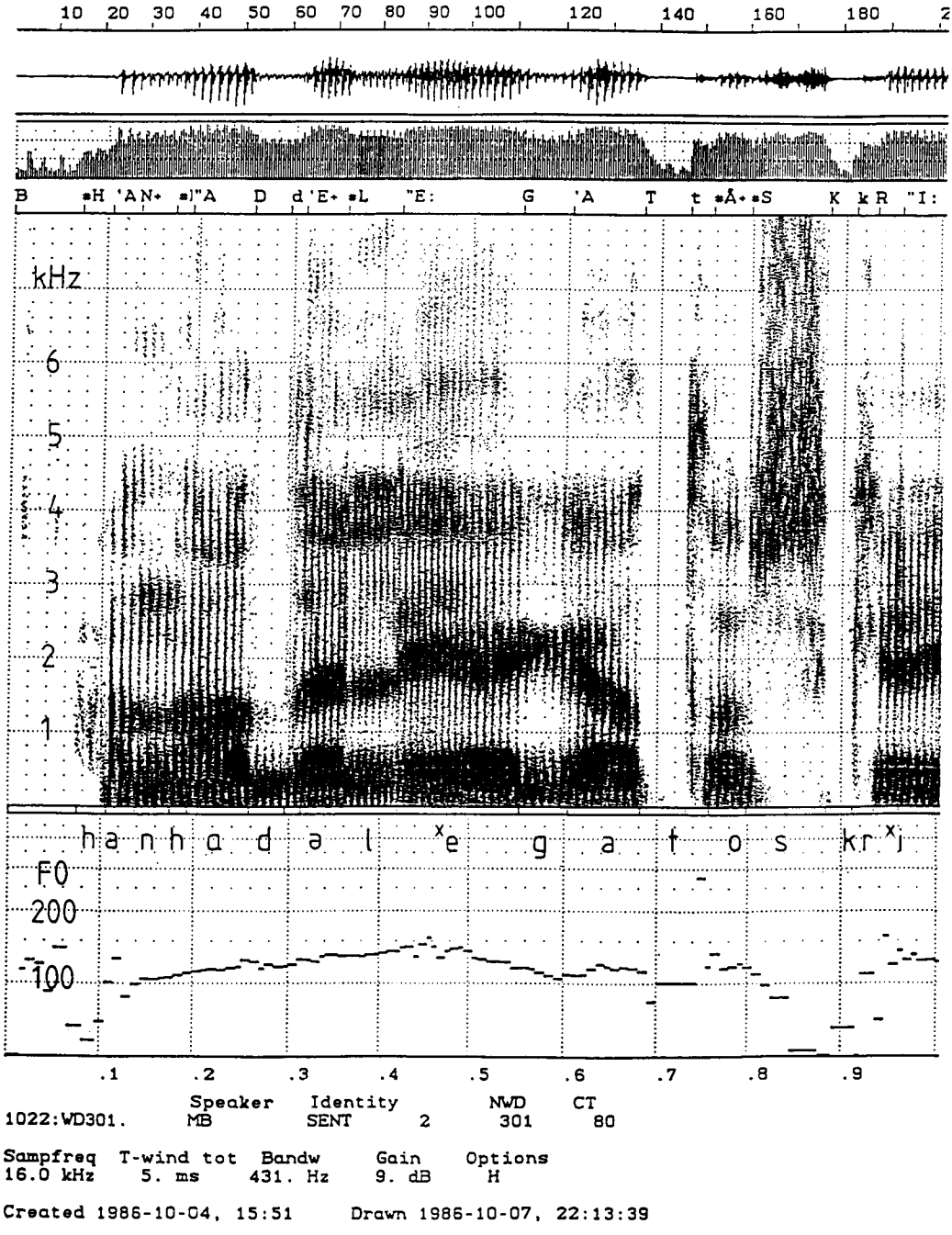


Figure 1. Sample of databank spectrogram print out.

between voiced consonants such as /v/, /j/, and /r/ and their combinations with vowels. A voiced intervocalic stop is not always associated with a stop gap, and phonemically unvoiced stops may attain voicing in unstressed positions. Lack of oral closure may affect nasals as well as stop sounds. An incomplete glottal abduction in an /h/ causes a continuation of voicing which obscures the segmentation. Here we are concerned with a class of articulatory reductions which tend to reduce temporal contrasts, especially between consonants and adjacent vowels. The notion of "articulatory contrast" and its consequence "dynamic contrast" is an important production parameter to consider.

Two adjacent unvoiced stops may share a single stop gap. As a rule we then assign one half each of the stop gap to the two phonemes. The boundary between a vowel and a following nasal may be hard or impossible to locate if the nasal element is realized by nasalization only. This happens frequently when the nasal is followed by an obstruent. In general, because of segmentation uncertainties, durational data on the sum of successive phonemes, e.g., a vowel plus a following consonant, are more reliable than data on the separate parts. However, in spite of all these difficulties, trained members of our group perform more consistent segmentations than one could anticipate, [2].

These phenomena are further illustration in [5], see also the incomplete closure of the voiced stop /g/ in the spectrogram of Figure 1.

2. SYLLABIC STRESS

Swedish is a stress-timed language with sequences of unstressed syllables alternating with stressed syllables. A stressed syllable carries a nucleus of a long vowel followed by one or two short consonants, or no consonant, or the vowel is short and is followed by a long consonant or a consonant cluster. A stressed syllable also carries one or two contrasting tones, accent 1 or accent 2. Duration appears to be the main correlate of stress in Swedish, at least it is more readily quantifiable than associated $F0$-measures. The zone of durational increase with increasing stress is the entire syllable but a larger part of the lengthening is confined to the vowel and the following consonant. This VC-nucleus will serve as our major objective reference. An unstressed VC has a typical duration of 110 ms and a stressed $V:C = 235$ ms when the vowel is long, and VC :=210 ms when the stressed vowel is short. In order to normalize durational measures, we first make statistics of the speaker's typical VC- durations for stressed and unstressed conditions, T_s and T_u. These are assigned normalized values of 2 and 1, respectively. Any duration measure T is then by linear interpolation converted to a normalized syllable index

$$S_i = 1 + (T - T_u)/(T_s - T_u) \tag{1}$$

A similar process was carried out for S_i values calculated on the basis of complete syllables and also for vowel-to-vowel units. These measures came out rather close to the VC-based S_i. Local F_0 rises and falls within a stress domain were also highly correlated with the S_i values.

From listening tests we derived subjective syllable stress ratings. These were linearly transformed into the same frame with subjective response $S_r = 2$ for typically stressed syllables. As a result, we could compare objective and subjective values within a common

frame. As seen in Figure 2, the overall correlation is very good. The temporal profile for a sentence is quite similar except that final lengthening did not induce the listeners to apply a higher score. An additional subjective test directed to quiet reading and introspection instead of listening also gave quite similar results. However, control experiments showed that listening performance was not essentially a top-down procedure. The listeners proved to be able to accurately follow individual variations of stress patterns.

3. INTERSTRESS INTERVALS AND RHYTHM

A substantial amount of work has been devoted to the study of interstress intervals, defined by the time from the onset of a stressed vowel to the onset of the next stressed vowel in the sequence. The durations of such stress feet were related to the number of phonemes and syllables in the foot. Excluding feet spanning syntactic boundaries, we found evidence for a linear regression

$$T_n = A + nB \tag{2}$$

where n is the number of phonemes, B is a constant durational increment of about 55 ms per extra added phoneme and A is a constant which includes all the durational increase within the stressed syllable. It is of the order of 50-200 ms dependent on speaker and speaking style. The incremental constant B is more stable and ranges from 50 to 65 ms. The foot length T_n ranges from 250 ms to 1000 ms with an average of about 550 ms, corresponding to $n = 7.4$ phonemes or close to three syllables.

A decomposition of the constant A showed that a trained reader prolonged the stressed vowel by about 60 ms, the following consonant by 65 ms, and the next consonant by 35 ms, whilst the consonant preceding the stressed vowel gained 15 ms only. A larger part of the lengthening associated with stress is that contributed by consonants. This also holds for individual variations. Vowel lengthening is usually less than the sum of consonant lengthenings. Similar conclusions arise from the Carlson and Granström study [2].

A study of stress intervals spanning pause gave interesting results. We found a covariation with negative correlation between pauses duration and final lengthening. Decomposing the interstress interval into sound and silence, we first calculated the amount of terminal lengthening as the amount that the sound duration exceeded the duration expected from the number of phonemes in non-spanning contexts, Eq. (2). The sum of the terminal lengthening and the duration of the pause is a net measure of the deviation of the interstress interval from that in a non-spanning context. In rhythmical reading it comes close to the average duration of the whole ensemble of non-spanning feet, about 550 ms. An example is shown in Figure 3.

Isochrony is thus physically manifested at phrase boundaries rather than in the uninterrupted sequence of stresses. We are thus in a position to substantiate and extend the conclusions of Lea [11] and Allen [12] as follows.

Speech rhythm is preserved by an inner clock which synchronizes on the average of non-spanning stress intervals. One such clock unit or silent foot is added at phrase boundaries. It is realized as a combination of pause and terminal lengthening. According to our findings, rhythmical reading preserves similar conditions at sentence boundaries, where

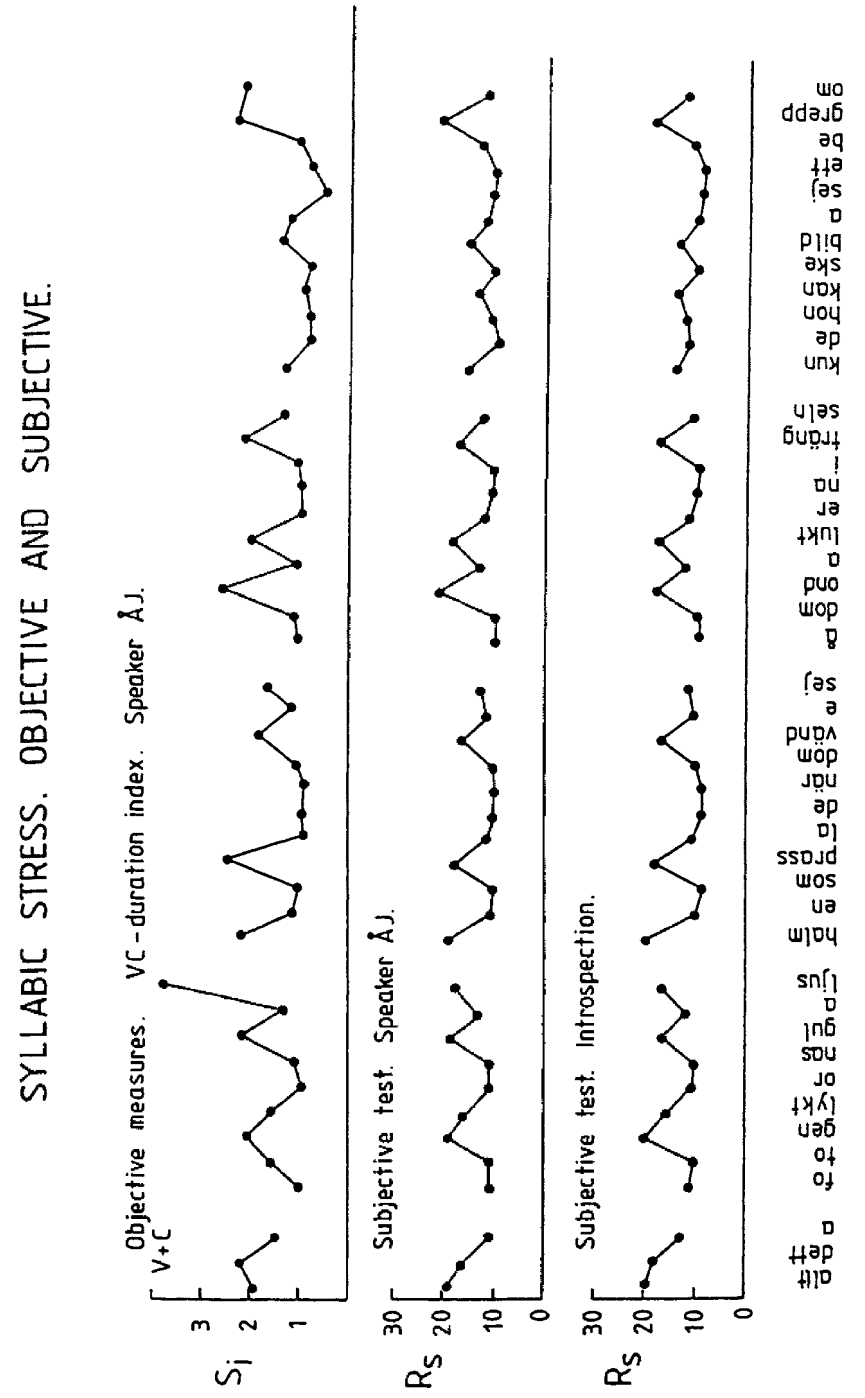


Figure 2. Objective and subjective measures of syllabic stress.

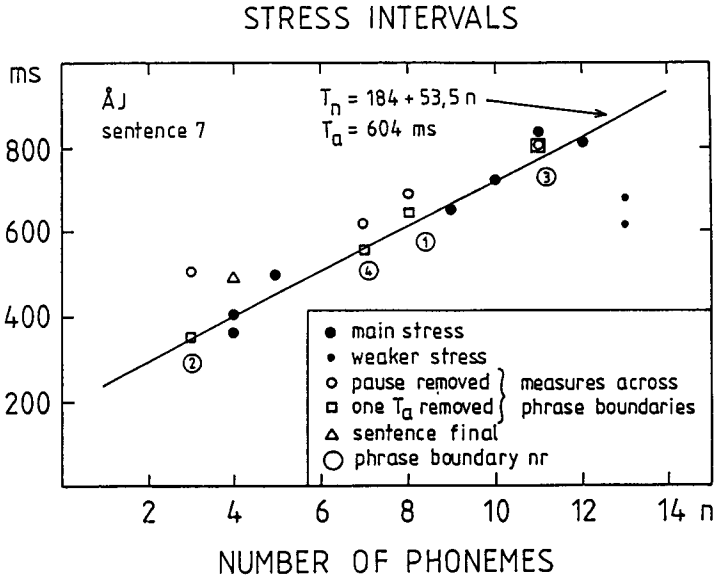


Figure 3. Duration of interstress intervals and related time intervals as a function of the number of phonemes involved.Filled circles refer to intervals not spanning boundaries.

the duration of pause and terminal lengthening tend to sum up to the integer multiples of the basic rhythm interval, usually two units.

However, deviations do exist. Phrase boundaries may be marked by terminal lengthening only, especially in case of vocalic juncture, where a local F_0-minimum also appears. The terminal lengthenings then are usually shorter than the clock unit. The presence of the terminal lengthening may conveniently be related to the interstress interval.

A study of 16 speakers' performance in realizing phrase boundaries in reading showed a large spread of the duration of the spanning interstress intervals. However, on the average, the sum of pause and terminal lengthening tended to come close to the ensemble mean of the subjects' average non-spanning foot.

Subjective scalings of each speaker's phrase boundary marking were performed. A linear increase of subjective rating versus the spanning interstress duration measure was observed, see Figure 4. The slope of the regression line tended to become less steep under conditions when both the left and the right syllable at the boundary were stressed.

The ratio of accumulated pause time to accumulated net reading time was found to be a characteristic individual feature ranging from 15 to 30%. One interesting observation from our study is that speakers who make relatively short pauses tended to have relatively longer net speaking time.

This trend is opposite to expectancy but could illustrate one more instance where parts of speech combine in a compensatory fashion.

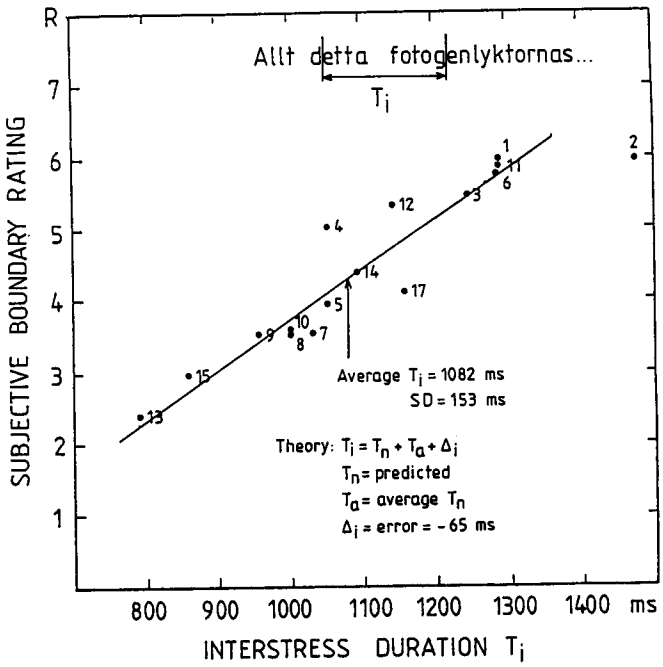


Figure 4. Subjective rating of syntactic boundary marking as a function of the duration of the boundary spanning interstress.

References

1. R. Carlson, and B. Granström, "Rule controlled database search," STL-QPSR, pp.29-42, No.4 (1985)

2. R. Carlson, and B. Granström, "A search for durational rules in a real-speech data base," Phonetica vol.43, pp.140-154, (1986)

3. R. Carlson and B. Granström, "Swedish durational rules derived from a sentence data base," STL-QPSR, pp.13-26, No.2-3 (1986)

4. R. Carlson and B. Granström, "Durational rules from a speech data base," paper FF6 presented at the 116th Meeting of the Acoustical Society of America, (1988)

5. G. Fant, L. Nord and A. Kruckenberg, "Individual variations in text reading. A data-bank pilot study," STL- QPSR, pp.1-7, No.4 (1986)

6. G. Fant, L. Nord and A. Kruckenberg, "Segmental and prosodic variabilities in connected speech. An applied data-bank study," Proc. XIth ICPhS, vol.6, Estonian Academy of Sciences, Tallinn, USSR, pp.102-105, (1987)

7. G. Fant and A. Kruckenberg, "Contributions to temporal analysis of read Swedish," paper presented at the Phonetic Symposium, Lund, to be published in Working Papers, University of Lund, Linguistics Department, (1988)

8. G. Fant and A. Kruckenberg, "Some durational correlates of Swedish prosody," paper presented at the VIIth FASE Symposium, SPEECH 88, (1988)

9. G. Fant and A. Kruckenberg, "Stress and interstress intervals in reading," paper FF10 presented at the 116th Meeting of the Acoustical Society of America, (1988)

10. G. Fant and A. Kruckenberg, "Preliminaries to the study of Swedish Prose Reading and Reading Style," STL-QPSR, pp.1-83, No.2 (1989)

11. A. Lea, "Prosodic aids to speech recognition," in Trends in Speech Recognition, ed. W. A. Lea, Prentice Hall Int., (1980)

12. G. Allen, "Speech rhythm: its relation to performance universals and articulatory timing," J. of Phonetics, vol.3, (1975)

Chapter 7

DIALOGUE SYSTEMS

This Page Intentionally Left Blank

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Parsing Grammatically Ill-Formed Utterances

Kuniaki Uehara and Jun'ichi Toyoda

The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, 567 Japan

Abstract

Any practical speech understanding system must be able to deal with a wide range of grammatically ill-formed utterances, both because people regularly form ungrammatical utterances and because there are a variety of forms that cannot be easily included in the current grammatical models. In this paper, we will describe some language phenomena commonly considered grammatically ill-formed in real language usage and present some relaxation mechanisms based on preference semantics proposed by Wilks.

1. INTRODUCTION

Among the components included in a Speech Understanding System (SUS) is a grammar, which specifies much of the linguistic structure of the utterances that can be expected. However, it is certain that a SUS will often receive ill-formed input, both because, unlike written text input, a SUS takes a multiple number of hypotheses as input for a particular voice input and because people regularly form ungrammatical utterances. Furthermore, there are a variety of forms that cannot be easily included in the current grammatical models (i.e. extra-grammatical). Therefore, a SUS requires, at the very least, some attempt to interpret, rather than merely reject, what seem to be grammatically ill-formed utterances.

In this paper, we will first describe some language phenomena commonly considered ungrammatical or extra-grammatical. We will then discuss some relaxation mechanisms directed at integrating them as much as possible into the conventional syntax-based framework of grammatical processing performed by a SUS. This parser is called FleP (FleP is an acronym for Flexible Parser). FleP has been intended to primarily focus on the relaxation mechanism and its use of syntactic knowledge.

In the remainder of this paper, we will consider an alternative model of parsing grammatically ill-formed utterances, a model which addresses many of the issues suggested from the development of FleP. This model is speculative, since it is currently not supported by an implementation. At this point, the model's major value is that it suggests the kind of computational model which may form the basis for the development of broader models of parsing and more general techniques for building a SUS.

2. SYNTAX BASED APPROACH: FleP

This section will introduce the types of ill-formed utterances we have studied, propose some relaxation mechanisms aimed at solving them, and discuss how these mechanisms are used. At the end, some limitations inherent to FleP will be discussed and extensions suggested.

2.1. Language Phenomena

There are a number of distinct types of grammatically ill-formedness, and not all types are found in all types of communication situations. The grammatically ill-formed utterances treated within FleP include syntactic constraint violations, especially, co-occurrence violations, extraneous forms, and some kinds of conjunctions.

(1) Co-Occurrence Violations

Our first type of grammatically ill-formedness is concerned with co-occurrence restrictions within an utterance. The most common form of co-occurrence violation is agreement failure between subject and verb, or determiner and head noun as in:

He want to have a cups of coffee.

Violations as the above involve coordination between the underlined words. Such phenomena do occur naturally. For example, Eastman and McLean [3] analyzed 693 English queries to a database system, although the queries are not spoken inputs but written text inputs. In this experiment, co-occurrence violations, including subject/verb disagreement, tense errors, apostrophe problems, and possessive/plural errors arose in 12.3% of the queries.

(2) Extraneous Forms

Another type of ungrammaticality occurs when a speaker puts unnecessary phrases in an utterance. In other words, people often repeat words, break off what they are saying and rephrase it, or put some interjected phrases in the utterance.

(a) Repeated words:
I would like to have two cups — two cups of coffee.
(b) Broken-off and restarted utterance:
Can I — Would you give me a cup of coffee?
(c) Interjected phrases:
I want the book titled, you know, Feigenbaums's the handbook of artificial intelligence.

In the experiment described above, extraneous forms occurred in as little as 1.6parse extraneous forms may not be a critical requirement for a SUS. However, in human communication with the aid of spoken input rather than written text input, it is not uncommon for people to form this kind of language phenomenon because of a change of intention in the middle of an utterance, an oversight, or simply for emphasis.

(3) Conjunctions

The use of conjunctions is not ungrammatical, but it is generally not included in grammars. This is because conjunctions can appear in so many places that their inclusion would dramatically increase the size of the grammar. Several types of conjunction we have considered are as follows:

(a) Simple form of conjunction:
John washes his face and reads a newspaper in the morning.
(b) Gapping:
He chooses a rose and she a poppy.
(c) Hacking:
John enjoyed and my friend liked the party.
(d) List form of conjunction:
John gets up, washes his face and reads a newspaper in the morning.

2.2. The Mechanisms and How They Apply

In this section, we will propose some relaxation mechanisms which can address the language phenomena discussed above. All of the mechanisms follow a general paradigm, wherein a 'normative' grammar is assumed. The normative grammar specifies a set of acceptable inputs to the parser. We have chosen the Definite Clause Grammar (DCG) model [8] as the tool in which to express our ideas. This framework may be as follows:

(1) Co-Occurrence Violations

First of all, FleP processes an input using a 'normative' grammar. During parsing, if FleP finds a co-occurrence violation in the input utterance, the grammatically applicable condition on the right-hand side of a generalized phrase structure rule is marked with a 'relaxable point'. At this time, the condition is not considered until after all possible analyses have been attempted, thereby insuring that the input sentence will be handled correctly. When all possible analyses have failed, the parser checks and sees the relaxable point and locally relaxes the condition allowing the acceptance of the violation.

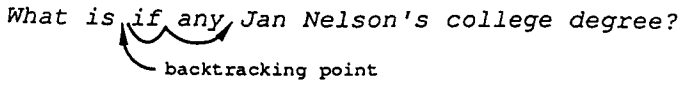
(2) Extraneous Forms

If no relaxable point was found in the 'normative' grammar, although FleP could not parse the utterance, it then assumes that the input utterance may include an extraneous form. Extraneous forms may be inserted at almost any place in an input utterance. We can not manually add rules for handling extraneous forms to the normative grammar. Extraneous forms, thus, must be dealt with as they arise by relaxation mechanisms.

(a) Repeated Words and Interjected Phrases

When the parser finds extraneous forms in the input sentence, which cannot be accepted by the normative grammar, it skips a word, creates a 'backtracking point', and analyzes the rest of the sentence. If the parser fails to analyze the remaining of the sentence, it returns to the backtracking point, skips one more word, and attempts to continue

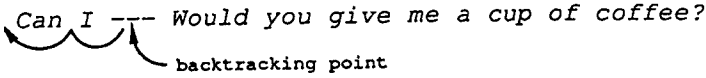
parsing. FleP repeats this step until an extraneous form is deleted from the input.



(b) Broken-off and Restarted Utterance

It is relatively straightforward for FleP to simply ignore interjected phrases, such as 'if any', 'please', or 'I think', and repeated words. More troublesome are broken-off and restarted utterances.

In the example described in section 2.1, the first fragment 'Can I' is not a complete utterance, so the same technique as proposed above is being attempted and fails due to the missing constituents. In this case, FleP returns to the backtracking point, 'unreduces' all the previously skipped words in the input, skips the word prior to the backtracking point, and attempts to continue parsing. This process is repeated until the utterance does not have an incomplete fragment.



(3) Conjunctions

In addition to these techniques, FleP can handle some types of conjunctions in a rather simple and reasonable way. As was described above, conjunctions, similar to extraneous forms, are not included in a 'normative' grammar. So, the mechanism for treating conjunctions is quite different from the one for treating co-occurrence violations or extraneous forms. At this point, we will consider the elided constituent in a conjunction, which is of

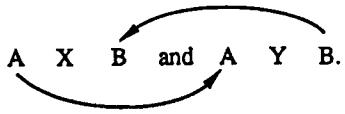
the form:

A X and Y B.

For example

John washes his face and reads a newspaper in the morning.
A X Y B

where the underlying deep structure is of the form:



Conjunctions are, thus, treated by FleP as follows: processing proceeds normally until a conjunction is encountered. If the conjunction appears, processing is suspended and a particular process (called a demon) is activated. The demon, at first, saves the history of the parse up to that time, and analyzes the string following the conjunction by use of its sub-grammar. The demon, then, reactivates the parser using the history of the parse which has already been saved. For instance, in the example above, processing proceeds until the conjunction 'and' is encountered. The action of the demon is, therefore, to unreduce the elided constituent complementing the information of A (in this example, 'John').

2.3. Limitations of the Syntax-Based Approach

Although we have attempted to incorporate sentential level relaxation mechanisms into a DCG model, we have found that this (syntax-based) parsing paradigm itself is not well-suited to the kinds of grammatically ill-formed utterances discussed above.
The reasons are as follows:

First, semantic information is very important in recovering from many types of ungrammatical input, and this information is not flexibly available in a purely syntactic DCG model. It is true that semantic information can be brought to bear on DCG based parsing, either through the semantic grammar based approach [9] in which joint semantic and syntactic categories are used directly in the DCG, or by allowing 'extra tests' on the right-hand side of a generalized phrase structure rule [8] depending on semantic criteria.

However, the natural way to use these techniques is to employ the semantic information only to confirm or disconfirm parses arrived at on syntactic grounds. So, the rigidity of the DCG formalism makes it very difficult to bring the available semantic information to bear effectively on extra-grammatical input.

Second, FleP cannot infer an alternative interpretation if its initial conjecture was incorrect. Furthermore, since the ability of variant syntax is limited, FleP can handle variations in only a subset of the total classes of syntax its parsing mechanism handles. Thus, FleP cannot understand utterances with arbitrary out-of-order words and missing words. The parser's problem in each case is to put together a group of recognizable sentence fragments without the normal syntactic glue of function words or position cues to indicate how the fragments should be combined.

Third, DCGs naturally operate in a top-down left-to-right mode, although a bottom-up capability is essential for many relaxation mechanisms, and directional flexibility often enables easier and more efficient operation of the mechanisms. Of course, the top-down left-to-right mode of operation is a characteristic of the DCG interpreter, not of the DCG formalism itself. In the bottom-up mode, all the fragments can be recognized independently, and purely semantic constraints can also be used to assemble them to a single interpretation on the basis of semantic considerations.

Finally, in the case of out-of-order words, a parser which relies heavily on a strict left-to-right scan will have much greater difficulty than one with more directional freedom. Thus, a parser which scans fragments, and subsequently attempts to assign them appropriate syntactic categories from the surrounding input, is more amenable to this type of recovery than one dependent upon rigid word order constraints.

3. SEMANTICS BASED APPROACH

Given the importance of both the relaxation mechanism and its use of semantic knowledge, the best way to address the issues described in the last section is through experimentation in the context of a real domain. That is, we must develop an alternative parsing model for grammatically ill-formed inputs and implement a system which tests this model. Based on such experiments, we can develop broader models and characterize the usefulness of different relaxation mechanisms.

3.1. Language Phenomena

Our second parser is intended to parse grammatically ill-formed utterances in a language with a relatively free word order, like Japanese, although FleP has been specialized for a language with a rigid word order, like English. In Japanese, an utterance begins with a certain number of noun phrases followed by a verb phrase. Almost all nouns and noun phrases have one or more postnominal suffixes marking case relationships. A verb phrase is always placed at the end of the utterance. The verb phrase consists of a verbal root and one or more ordered suffixes marking tense, aspect, modality, voice, negativity, politeness level, question, etc.

In this experimental system, we will consider the following grammatically ill-formed utterances: omission of a postnominal suffix, extraneous forms, and free word order.

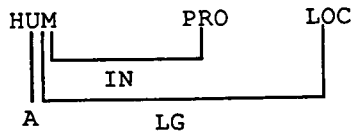
(1) Omission of a Postnominal Suffix

It is not uncommon for the native user of Japanese to omit a postnominal suffix from his utterance, either by mistake or in an attempt to be cryptic. Furthermore, the postnominal suffix is too short for a SUS to be recognized correctly. However, fortunately, a Japanese postnominal suffix, by itself, does not always provide all the necessary information for case assignment. Conversely speaking, the omission of postnominal suffixes is usually (though not always) recoverable by using semantic features of nouns and the case frame of each verb.

For example, in an ill-formed utterance

Watashi jitensha gakkou iku.
(I) (bicycle) (school) (go)

where the semantic features of the nouns 'watashi', 'jitensha', and 'gakkou' are 'HUMan', 'PROduct', and 'LOCation', respectively. Each verb then must have a case frame specifying which cases are required or allowed with it. The case frame of the verb 'iku' specifies that 'Agent' must be a 'HUMan'; 'Agent' moves to the 'Locational Goal' where the locational goal of this action is 'LOCation'; an 'INstrument' for the action 'iku' must be a 'PROduct'. These case relationships are represented as follows:



How is the case frame of the verb to be matched with each semantic feature of the noun in the utterance? In this example, we can pattern match 'watashi' with 'Agent', 'jitensha' with 'INstrument', and 'gakkou' with 'Location', respectively. Furthermore, in order to specify the case relationships of the verb 'iku', 'Agent' requires the postnominal suffix '-wa', 'INstrument' usually requires '-de', and 'Location Goal' must be marked by '- ni'. Finally, we can infer the following complete utterance:

Watashi-wa jitensha-de gakkou-ni iku.

(2) Extraneous Forms

Since the language phenomena of an extraneous form in Japanese are similar to those in English described above, we will not go into any more detail.

(3) Free Word Order

The relatively free word order of Japanese further complicates the situation, as in the six sentences listed below, which are all grammatical and all mean "I go to school by bicycle", but each with different noun phrases given prominence. In this paper, we will not deal with the differences in meaning of these utterances.

a. Watashi-wa jitensha-de gakkou-ni iku.

b. Watashi-wa gakkou-ni jitensha-de iku.

c. Gakkou-ni watashi-wa jitensha-de iku.

d. Gakkou-ni jitensha-de watashi-wa iku.

e. Jitensha-de watashi-wa gakkou-ni iku.

f. Jitensha-de gakkou-ni watashi-wa iku.

3.2. Preference Semantics

Now we will introduce a method of parsing grammatically ill-formed utterance by use of preference semantics, proposed by Wilks [7]. The preference semantics approach is different from a conventional selection restriction approach. The selection restriction approach embodies a binary principle of well-formedness, that is, a semantic marker either fits a selection restriction or it does not. Whereas the preference semantics approach adopts a different, unary principle of well-formedness. That is, even if a preference in a sentence is violated, an interpretation is still produced for the utterance as if it is well-formed. The decision whether to accept the interpretation or not depends on whether there are other possible interpretations for that utterance. So, the difference between the selection restriction approach and the preference semantics approach is the criterion for ill-formedness. In the former approach, an interpretation can be treated as ill-formed by examining it alone. Whereas in the latter approach, the interpretation can only be considered as ill-formed after comparing it with the other interpretations.

3.3. Blackboard Model

The blackboard model is a popular problem solving vehicle for expert systems. We have adopted its concept and utilized it for parsing grammatically ill-formed utterances. The blackboard model is usually described as consisting of three major components:

(a) Knowledge Sources

The semantic knowledge needed to parse an utterance is partitioned into knowledge sources, which are kept separate and independent.

(b) The Blackboard Data Structure
The states of parsing the data are kept in a global database, called the blackboard. Knowledge sources produce changes in the blackboard that lead incrementally to a semantic interpretation of the utterance. Communication and interaction among the knowledge sources take place solely through the blackboard.

(c) Control

The knowledge sources respond opportunistically to changes in the blackboard.

3.4. A Model of Semantics Based Language Analysis

Analysis of a Japanese utterance by our model proceeds in several stages. First, the input is broken into fragments (at conjunctions, postnominal suffixes, etc.). To each fragment (i.e. phrase) are assigned its own semantic features, one semantic feature for each head word in the fragment. The head word is the most important word in the fragment, and its semantic feature expresses the most general semantic category under which the word sense in question falls. Now, the dictionary may contain several semantic features for each word, representing its different senses, but we will focus on a restricted type of communication situation in a limited domain. Thus, we assume that each noun has only a single semantic feature.

The semantic feature of a noun can be used to correlate the meaning of different words in an utterance. These case relationships can be inferred from the semantic features of the nouns. Hereafter, we will call them semantic roles. For example, if the semantic feature of the noun is 'HUMan', such as 'watashi' (I), the noun may be 'Agent', 'Object1' (the direct object of a stative verb), 'Object2' (the direct object of a nonstative verb), etc. If the semantic role of the noun is 'ACTion', such as 'benkyou' (study), it may be used as 'INstrument', 'MoTive', 'Object1', 'Object2', 'Non-locational Goal', etc. Note that, in our notation, the semantic role correlating with the subject is marked by the suffix '1', whereas the same semantic role modifying the object has the suffix '2'.

For each noun in an utterance, our model produces an ordered list of semantic role candidates along with the semantic feature on which they depend. They are arranged in the order of frequency in use. This structure is called a semantic role lattice. Figare 1

shows a sample utterance with the corresponding semantic role lattice.

fragment	*watashi* *(I)*	*radio* *(radio)*	*ongaku* *(music)*	*kiku* *(listen)*
semantic feature	HUM	PRO	ABS	
candidate semantic roles	A O1 O2 : :	O1 O2 IN A :	O2 O1 NG NL :	

Fig.1 Semantic role lattice

The next step is to build from the semantic role lattice a complete semantic interpretation of an entire utterance. To do this, we have defined some interesting constraints for semantic role lattices. These constraints are used to reject inappropriate semantic roles in the lattice. Examples of these constraints are as follows:

(1) Each semantic role cannot appear more than once in the same utterance.

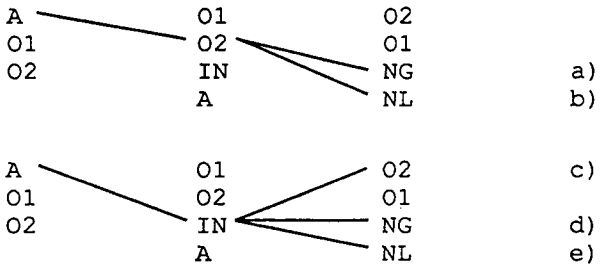
(2) The direct object of a stative verb ('O1') and the agent of a nonstative verb ('A') cannot appear in the same utterance.

(3) Either 'A' or 'O1' must appear in the utterance.

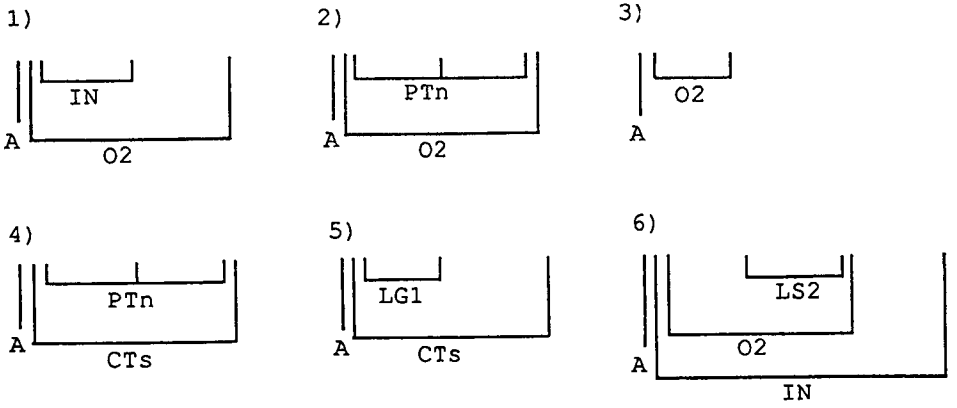
(4) Noun phrases with the same semantic role, although their suffixes are different, cannot appear in the same utterance.

Applying these constraints to the semantic role candidates associated with each word, we can get the following combinations of semantic roles, that is, case frame candidates of

the verb:

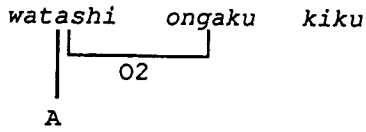


With these semantic role candidates, we must determine the 'preferable' case assignment among them. The verb 'kiku' has the following possible case frames:

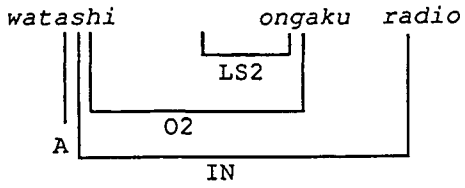


Case frame (1) can match with candidate (c) completely, whereas case frame (3) can partially match with candidates (a), (b), or (c), respectively. For example, the result of matching (a) with (3) turns out to be the following case assignment of the utterance. In

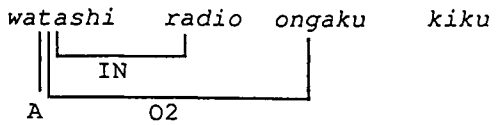
this case, we can consider that the noun 'radio' is an extraneous form of the utterance.



Case assignment (6) can also partially match with candidate (c). In this case, we consider that the noun corresponding to the semantic role 'LS2' (i.e. Location Source) is missing.



The following case assignment has a larger number of satisfied preferences, or greater 'semantic density', so it is preferred. In the case of the assignments shown above, all of them contain some failed preferences, but the following assignment is accepted because there are no other competing assignments.



4. RELATED WORKS

Kakigahara at ATR Interpreting Telephony Research Laboratories has already discussed a method of generating a Japanese sentence by inferring postnominal suffixes from the valency pattern of a verb in a utterance [5]. The valency pattern for each verb specifies the relationship between the semantic features of nouns and their possible postnominal suffixes. For example, the valency pattern of the verb 'nagetukeru' (throw at) has the valency pattern:

N[HUM] '-ga' + N[CON] '-wo' + N[CON] '-ni' + V.

where 'N' is the noun, 'V' is the verb, and 'HUM' and 'CON' are semantic features of the associated nouns.

Kakigahara's method of selecting the correct postnominal suffix proceeds in two steps: (1) generate a meaningful utterance by inferring suitable postnominal suffixes for a given sequence of nouns and a verb by use of the valency pattern, and (2) compare each inferred postnominal suffix with the corresponding candidates in the phrase lattice output by the SUS to select the most appropriate one.

In his study, he assumed that the task domain is rather restricted, so as not to cause ambiguity in the utterance. Furthermore, his selection mechanism of valency patterns is binary, that is, a semantic feature either fits a valency pattern or it does not. With the binary principle, there is an absolute criterion for ill-formedness: a semantic interpretation can be labelled ill-formed by examining this interpretation alone, without applying any other relaxation mechanism. Consider the following utterance, which means 'The dog throws a ball at the wall'.

Inu-ga ball-wo kabe-ni nagetukeru.
('dog') ('ball') ('wall') ('throw at')

The best reading of this utterance shows a conflict between the valency pattern of the verb 'nagetukeru', expecting a 'HUMan' agent as subject, and the actual data, because the subject ('inu') is an 'ANImal'. Although this utterance is semantically ill-formed in the real world, whether that (semantically violated) reading is acceptable or not depends on what the parser believes the state of the world to be, and how far it can be extended by rules with the aid of the knowledge structures available. In other words, if we apply Kakigahara's method to a different task domain, we must set the system up with other types of semantic features, depending on the state of the task domain.

Whereas we think much of semantic roles rather than semantic features in our model, the utterance described above may be accepted, if either there are no other interpretations of the utterance or if all the other interpretations of the utterance have more semantic violations.

5. CONCLUDING REMARKS

Any practical speech understanding system must be able to deal with a wide range of grammatically ill-formed utterances. This paper proposed a taxonomy of the usual forms of grammatically ill-formedness in real language usage and presented some relaxation mechanisms for them. We also discussed that the preference semantic approach provided the best framework among the commonly used relaxation mechanisms. It is our hope that by pursuing the approaches described above we can obtain a parser that has robustness in more general language settings.

References

1. K. Jensen, G. E. Heidorn, L. A. Miller and Y. Ravin: "Parse Fitting and Prose Fixing: Getting a Hold on Ill-Formedness," *AJCL*, vol.9, No.3-4, pp.147-160, 1983.

2. S. C. Kwansny and N. K. Sondheimer: "Relaxation Techniques for Parsing Grammatically Ill-Formed Input in Natural Language Understanding Systems," *AJCL*, vol.7, No.2, pp.99-108, 1981.

3. C. M. Eastman and D. S. McLean: "On the Need for Parsing Ill-Formed Input," *AJCL*, vol.7, No.4, pp.257, 1981.

4. D. S. McLean and C. M. Eastman: "A Query Corpus Containing Ill-Formed Input," *Technical Report 84-CSE-9*, Southern Method-ist University, 1984.

5. K. Kakigahara and T. Aizawa: "Completion of Japanese Sentences by Inferring Function Words from Content Words," *Proc. COLING 88*, pp.291-296, 1988.

6. D. Fass and Y. Wilks: "Preference Semantics, Ill-Formedness, and Metaphor," *AJCL*, vol.9, No.3-4, pp.178-187, 1983.

7. Y. Wilks:A Preferential, "Pattern-Seeking, Semantics for Natural Language Interface," *Artificial Intelligence*, vol. 6, pp.53-74, 1975.

8. F. C. N. Pereira and D. H. D. Warren: "Definite Clause Grammars for Language Analysis — A Survey of the Formalism and a Comparison with Augmented Transition Networks," *Artificial Intelligence*, vol.13, pp.231-278, 1980.

9. G. G. Hendrix: "Human Engineering for Applied Natural Language Processing," *Proc. IJCAI-5th*, pp.183-191, 1977.

Recent Research Towards Advanced Man-Machine Interface
Through Spoken Language, H. Fujisaki (Editor)

A Dialogue Analyzing Method Using a Dialogue Model

Atsuko Takano, Hideki Kashioka, Makoto Hirai and Tadahiro Kitahashi

The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, 567 Japan

Abstract

This paper describes a dialogue analysis system focusing on conversational coherence. Although conversational coherence has diverse components, we deal with it from three points of view: structural coherence, cohesion, and coherence of the dialogue contents. The system produces a structure which characterizes the input dialogue. The analysis scheme employs a dialogue model on the basis of utterance pairs and utterance groups, which are recognized in terms of the planning by the participants and changes in topic. The relationships between utterances are recognized by integrating the three components of "coherence", formalized as the dialogue model.During the process, when needed, omissions, references, implications, etc. are resolved so as to maintain coherence of the utterance contents.

The reasoning mechanism for understanding a dialogue realized in this scheme, is similar to the human's. It generates hypotheses about the problems, then provides evidence for the hypotheses. During the process of structural determination, the dialogue structure is partially reconstructed if the process fails to maintain coherence, where the reconsidered parts are as small as possible.

1. INTRODUCTION

Analyzing utterances in a dialogue can be viewed as recognizing the conversational coherence in it. Along this direction, we have developed a system for analyzing dialogue in Japanese. The system understands the dialogue by recognizing its structure so as to maintain conversational coherence.

There are two major approaches to understanding a dialogue or discourse in general. One is to understand the intention and the semantic and pragmatic contents of individual utterances or groups of utterances in a specific domain [2][3]. This requires various knowledge, such as "common sense" and "domain knowledge". The other is to recognize structural relationships among groups of utterances in order to explain linguistic phenomena [4]. The former approach discusses the semantical aspects of conversational coherence, while the latter one discusses the structural aspects.

A comprehensive dialogue system, however, should include a mechanism for integrating both these approaches. Grosz [5] studied relationships between discourse structures and intentions, but failed to construct a precise mechanism for analysis. Although the work presented here is similar to Grosz's in basic approach, it investigates a theory of analysis which stresses structural coherence and cohesion through an integration of the coherence involved in the contents of the dialogue.

We are also developing a reasoning mechanism to realize the theory. Considering the importance of the coherence of structure and cohesion, we provide a model of the dialogue structure on the basis of the relationships between paired utterances (e.g. question and answer) and utterance groups consisting of paired utterances. These considerations delineate dialogue regularity as structural rules and heuristic principles. The heuristic principles can specify the integration of structural coherence and cohesion in terms of the linguistic concepts "topic" and "focus".

The process of determining the structure often necessitates supplementation of information. For example, determination of omitted words (this linguistic phenomenon frequently appears in Japanese), reference resolution, and utterance implication resolution are needed. The information is derived from several pragmatic rules based on the cooperation principles proposed by Grice [1].

The effectiveness of the process is strengthened by introducing a dialogue model and constraint propagation and satisfaction [9] , as will be discussed in sections 5.2 and 5.3. Thus, we have established an elaborate mechanism in which the two kinds of approaches to dialogue understanding are integrated, that is, a mechanism capable of dealing with the two different aspects of coherence in a natural way, as will be discussed later.

The features of the reasoning mechanism described in this paper can be summarized as follows.

- First this mechanism reasons out the relationships between utterances and the preceding dialogue using abduction, and then verifies the reasoning in the succeeding process. This is similar to the process by which people understand dialogues.
- Most of the processes are invoked on demand in order to minimize the amount of computation.

2. BASIC CONCEPTS OF THE SYSTEM

2.1. Conceptual Structure of the System

A block diagram of the analyzing system is shown in Figure 1. The semantical representation of the utterance is obtained in terms of a network produced by the "Sentence Analyzer". In our research, a network representation is used for both the input and the inner representations. The "Dialogue Structure Builder" analyzes the dialogue structure, and the "Information Supplementer" is equipped with heuristics for supplying missing information.

The conversational coherence is transformed to the "Dialogue Model", expressed in the form of rules, procedures, and heuristics. It is interpreted in terms of style, cohesion, and meaning of utterances.

2.2. Basic Concept of the Reasoning Mechanism

The framework of the reasoning scheme employed here, which discerns the utterance's position in the dialogue, is shown in Figure 2. The details will be described later, along with some examples. Candidates for the next utterance's position are produced from the preceding dialogue. The most appropriate hypothesis is selected using structural rules which express the relationships between the utterance's positions and the superficial character itself of the utterances. The rules have default expressions that mean "if A then is possible B" in the form of "A -M→ B". Assuming that a hypothesis holds, the structure is extended. In this process the hypothesis is confirmed if it is verified by constraints and procedures. On the other hand, if this process fails to maintain coherence, another hypothesis will be selected.

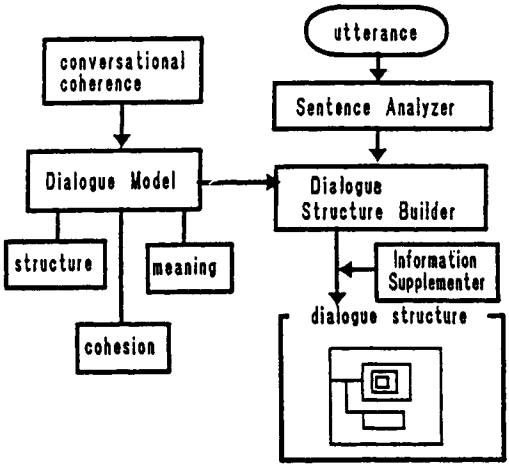


Fig.1 Block diagram of the analyzing system

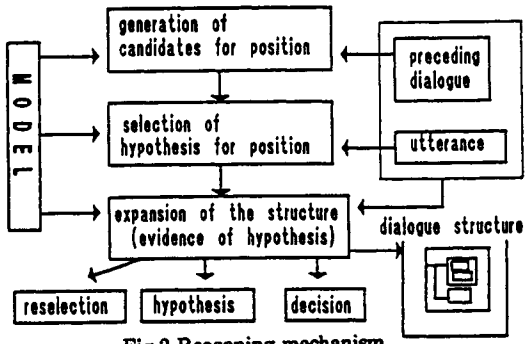


Fig.2 Reasoning mechanism

2.3. Dialogue Example

To illustrate the approach to the analysis, an example will be given. The example conversation has two participants C (clerk) and T (tourist) and it concerns a registration for a sightseeing tour. The utterances enclosed in parentheses are in Japanese.
T1: I would like to register for a sightseeing tour.
(kankou tua- no moushikomi wo shitainodesuga.)
C2: Which tour do you prefer?
(dono tua- ni nasaimasuka.)
C3: We have half-day tours and full-day tours.
(hanniti tua- to itiniti tua- ga arimasuga.)
T4: Which one goes to the best-known places?
(dono tua- ga yu-meina basyo he ikimasuka.)
C5: Are you interested in the fine arts?
(bijyutu ni kyoumi ga oaridesuka)
T6: Yes, I love European paintings.
(hai, yo-roppa no kaiga ga sukidesu.)
C7: Well, a half-day tour goes to some famous art museums.
(soredeshitara, hanniti tua- ga yu-meina bijyutukan ni ikimasu.)
T8: Then I'd like the half-day tour.
(deha hanniti tua- ni shimasu.)
C9: Could you write the telephone number of your office on this form please?
(kaisha no denwa bangou wo kono youshi ni okakikudasai.)
T10: Do you have a pen?
(pen ha arimasuka)
C11: Here you are.
(korewo douzo.)
T12: At what time does it start?
(nanji ni hajimarimasuka)
C13: It starts at nine in the morning.
(asa no 9-ji desu.)

3. DIALOGUE MODEL

Utterances of two participants often form pairs, such as a question and an answer, a request and a consent. These pairs of utterances are the structural elements of the dialogue. Some of the pairs form an utterance group, sharing a topic. These groups serve as semantic elements of a dialogue and are referred to as "dialogue units". The nature of the pairs and pragmatics of the relationships within and between units of utterances provide the basis of the analysis of the dialogue structure in this paper.

3.1. Utterance Pair and Dialogue Unit

Our dialogue analyzing system works on the two levels of dialogue structure, the "utterance pair" and the "dialogue unit". A dialogue unit normally consists of a few paired utterances.

To identify these levels of hierarchical structure we employ two sorts of relationships observed in a dialogue. One is a correspondence between a pair of utterances (e.g. a question and an answer), which identifies the "utterance pair". The other is a semantic identity among a group of utterances, which identifies the "dialogue unit".

The utterance types which stand for the former sort of relations (referred to as *ptype*(s)), are classified into the following five types:

- information transmission (INFOTRAN),
- response (RESPONSE),
- Yes/No-information request (YN-INFREQ),
- WH-information request (WH-INFREQ) and
- act request (ACTREQ).

These types are discerned by syntactic cues in utterances. Syntactic types of utterances, such as declarative sentence, interrogative sentence, imperative sentence, or other idiomatic expressions are examples of syntactic cues. these syntactic types are referred to as *stype* in this paper. Other cues, such as expressions closing a sentence and modal terms (e.g. "possible" or "desirable), are also useful in identifying these types. Combinations of utterances which are expected to form basic pair relations are listed below:

YN-INFREQ	RESPONSE
	INFOTRAN
WH-INFREQ	INFOTRAN
ACTREQ	RESPONSE
INFOTRAN	RESPONSE

The utterance types which stand for the latter sort of relations (referred to as *mtype*) are classified into three types: GOALPRE, TOPICPRE and CONTI. These types concern pragmatics and planning of the dialogue and are defined as follows. The utterances whose mtypes are either the goal presentation (GOALPRE) or the topic presentation (TOPICPRE) open dialogue units. Succeeding utterances which are chained by a pair relation, are included in the same dialogue unit. The type of these utterances is "continuation" (CONTI). They discern the "dialogue unit", which is structured by chaining or embedding paired utterances as follows:

(Dialogue unit made by a link of two utterance pairs:)
WH-INFREQ T12: At what time does it start?
⇕pair
INFOTRAN C13: It starts at nine in the morning.

⇕pair
RESPONCE T14: Is that so?

(One paired utterance is embedded in another:)
WH-INFREQ T4: Which one goes to the best-known places?
YN-INFREQ C5: Are you interested in the fine arts? ⇑
⇕pair ‖pair
INFOTRAN T6: Yes,I love European paintings. ⇓
INFOTRAN C7: Well, the half-day tour goes to some famous art museums.

3.2. Focus and Topic

The linguistic concepts "focus" and "topic" are useful in analyzing a dialogue structure. The definitions of these concepts so far proposed seem to be unclear. In this paper we will determine FOCUS partly from syntactic information. The notion of topic involves continuation of the speaker's attention. FOCUS, as used here, refers to the linguistic unit to which attention is directed. We basically relied on the method proposed by Kameyama [11] to identify FOCUS. Roughly speaking, the following items are checked to see whether or not they could be FOCUS in turn:

(1) items in question in an interrogative sentence;

(2) items which represent a subject of conversation;

(3) items which arouse sympathy in the listener;

(4) grammatical subjects;

(5) grammatical objects.

According to this, we will give a more precise idea of the relation between FOCUS and TOPIC. When a FOCUS is referred to by a pronoun or an omitted word in succeeding utterances, it turns into a TOPIC. In other words, we assert that a FOCUS involves local and explicit attention whereas a TOPIC involves global and underlying attention.

3.3. Relationships Between Dialogue Units

The set of relationships that can hold between dialogue units is considered to represent the underlying structure of a dialogue. The relationships between the dialogue units that form the dialogue structure are classified as follows.

Subissue Relation: when a succeeding dialogue unit decomposes the issue of the preceding one, the succeeding unit has a *subissue relation* to the preceding one. More specifically, if a succeeding utterance opens a new unit when the pair started by the preceding utterance has not yet been completed and if it has the same topic, the succeeding unit is a *subissue* of the preceding one.

Joining Relations: two units which contain the same topic have a *joining relation.*
Shift Relation: when an object in a preceding unit becomes the topic of a succeeding one, the succeeding unit has a *shift relation* to the preceding one.

The dialogue structure of the example dialogue discussed above is depicted in Figure 3 in terms of the paired utterances and dialogue units.

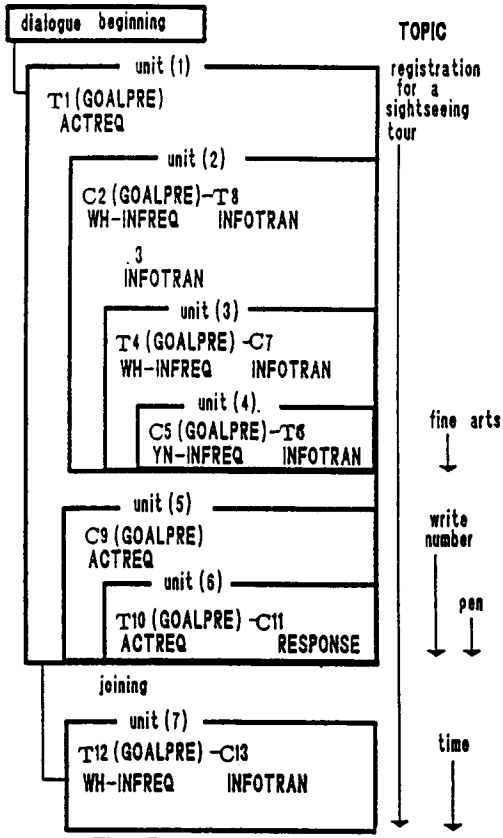


Fig.3 Dialogue structure

3.4. Prediction of Succeeding Utterances and Structural Rules

As described in section 2, some preceding utterances require subsequent utterances of specific types in order to complete pairs. For instance, a question requires an answer. These requirements can be represented as candidates for the structural position of the succeeding utterance. In order to determine the most appropriate position, structural rules based on a certain regularity of speech in a dialogue are applied to the syntactic information of the utterance.

The table in Figure 4 shows expected pairings between the preceding utterances and the succeeding ones which are needed for forming paired utterance or dialogue units.

Candidates for the position of the succeeding utterance are arranged in order of probability or priority for the hypotheses. Among the candidates the most appropriate requirement imposed by the preceding utterances would be a feasible hypothesis of the structural position of the succeeding utterance. The requirement of forming a pair will be removed when the pair is completed. Likewise all the requirements within a dialogue unit will be removed when the unit is closed.

structural level (U: unit P: pair)

preceding utterance		succeeding utterance			
mtype	ptype		mtype	ptype	side effect
beginning			TOPICPRE	INFOTRAN	
GOALPRE	*-INFREQ	P	CONTI	INFOTRAN	make pair, fix
		U	GOALPRE	*-INFREQ	topic hold, fix (shift)
			GOALPRE	ACTREQ	make subissue unit
			GOALPRE	*-INFREQ	topic indirectly hold (shift)
					make subissue unit
	ACTREQ	P	CONTI	RESPONSE	make pair, fix
		U	GOALPRE	*-INFREQ	make subissue unit
			GOALPRE	ACTREQ	make subissue unit

note: "fix" shows the hypothesis has been finalized
* means YN or WH

Fig.4 Utterance position expectation(requirement)

On the other hand, the same requirement posed by the succeeding utterance may be applied to the preceding utterance. For instance, a RESPONS utterance requires the preceding utterance to be of either YN-INFREQ or ACTREQ type.

In order to reason about the position in the dialogue structure and to single out a proper candidate, certain structural default rules have been formalized. Some of them are as follows:

SR1 INFOTRAN(U) -M→ stype(U, declarative)
SR2 WH-INFREQ (U, X) -M→ stype(U, WH-interrogative, X)
SR3 YN-INFREQ (U) -M→ stype(U, YN-interrogative)
SR4 ACTREQ(U, Action) -M→ stype(U, imperative), verb(U, Action)
SR5 ACTREQ(U, Action) -M→ stype(U, declarative), modal(U, desirable), end(U, hesitation), verb(U, Action)
SR6 ACTREQ(U, Action) -M→ stype(U, YN-interrogative), modal(possible), agent(U, listener), verb(U, Action)

3.5. Heuristic Principles

The relations between topic, focus and dialogue structure, and other dialogue regularities are formalized as heuristic principles, as follows:
HP1. The priority of requirements is as follows:

- requirement concerning pair $\succ$ requirement concerning unit,
- requirement occurred later $\succ$ requirement occurs earlier.

HP2. When an utterance whose mtype is either GOALPRE or TOPICPRE opens a unit which is not a *subissue* of any unit, its FOCUS becomes a TOPIC.
HP3. A word which is referred to by a pronoun or zero pronoun must appear in one of the following dialogue units:

- the current unit,
- the unit which is related to the current unit by a *subissue*, *joining* or *shift* relation.

HP4. The structural position of a unit is confirmed when another unit which has a *subissue relation* to that unit is identified.
HP5. *Joining* and *shift* relations (described in section 3.3) are established between two units which have the same level (i.e. degree of nesting by *subissue relation* and are adjacent to each other.
HP6. Requirements imposed on the preceding utterance by the succeeding one should be satisfied.

4. INFORMATION SUPPLEMENT

As is well known, utterances tend to provide insufficient information and are ambiguous. Nevertheless people can understand dialogues. This is because they use knowledge, whether it is "common sense" or domain specific. Our system concentrates on lexical knowledge, including information concerning the usage and meaning of words.

4.1. Coherence in the Participant's Knowledge

Hirai and Kitahashi [8] have formulated several pragmatic rules representing knowledge shared by communicating persons in order to determine omitted words and resolve references. Their formulation is used in our system.

4.2. Lexical Knowledge

Lexical knowledge is considered necessary both to recognize answers to questions and to capture implications of utterances. We represent the "concrete meaning" of nouns and verbs from a functional point of view in order to describe presuppositions and effects of

actions. The process of implication resolution using this lexical knowledge is illustrated in section 5.3.

As has often been pointed out, the intention of a speaker can frequently be different from the superficial meaning of an utterance. The utterance T10 in the example suggests an indirect intention. What does participant T want to imply by the utterance T10? In order to determine the implication of the utterance T10, the following information is necessary to supplement the typical lexical definition of "pen" and "have".

≪pen≫
func (pen, write ((agent, X), (locate, Y), (object, Z), (instrument, pen)))
≪give≫
presuppose (give ((agent, K), (goal, L), (object, M)), have ((agent, K), (object, M)))
effect (give ((agent, K), (goal, L), (object, M)), have ((agent, L), (object, M)))

5. ANALYSIS MECHANISM

A dialogue structure and utterance contents are represented by networks. A network consists of nodes and links which represent concepts and the relations between nodes, respectively [12] . Inference is performed by attaching or removing the nodes and links and matching subnetworks. They are classified into two groups. One is a set of general processes activated in any case. The other is a set of processes which are executed only when the general process fails to maintain coherence. The processes are attached to the node at which the processes are involved.

5.1. Representation of the Utterance Contents and Dialogue Structure

The network representing a dialogue structure and utterance contents consists of some of the following five kinds of nodes.

(1) mtype,

(2) ptype,

(3) stype,

(4) additional information node: modal concepts, closing expressions, tense, idioms, social relationships,

(5) node containing propositional contents of utterances: predicate, case elements.

Every node is expressed by a triple consisting of the following three elements:

(1) A concept name,

(2) Names of an object and other information: "*" represents undecided, which means that it may be bound to any object which meets conditions concerning the concept names of "*".

(3) A performable procedure: the procedure discussed here is performed when reconsideration is required. For example, the procedure "supplementary lexical knowledge" might derive more information from lexical knowledge.

The network of the utterance T1 is represented in Figure 5.

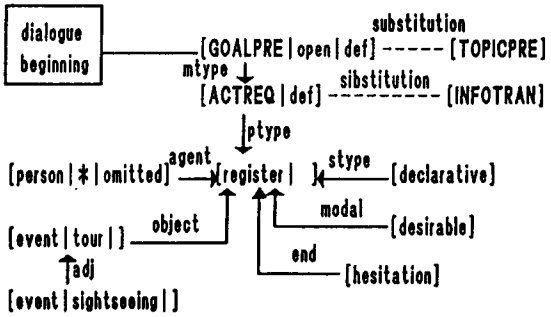


Fig.5 Representation of T1

Matching between subnetworks is a basic procedure in this analyzing process. In order for two concepts or two names to match, they are not necessarily identical to each other as far as one has 'similar' or 'isa' relations to the other.

(a) [person|*]→[be interested|*]←[art|fine arts]
agent object

(b) [person|*]→[love|*]←[fine arts|painting]←[place|European]
agent object adj

The networks a) and b) match since they are linked by the relations similar(be interested, love) and isa(painting, fine arts) in the semantic thesaurus.

5.2. General Process

When the Dialogue Structure Builder receives an utterance represented by a network from the Sentence Analyzer, it tries to reason out the structural position of the utterance. To put it concrete, the reasoning out of the structural type (mtype and ptype) is achieved by applying the structural rules and heuristic principles to the candidates for the assumption of the prior order.

We will show the analyzing process of the following utterances C5 and T6 as an example.

C5: Are you interested in the fine arts?

T6: Yes, I love European paintings.

First, the process tries to find the candidates for the structural position, which are obtained from the table in Figure 4 by referring to the Pair column, imposing the priority conditions in HP1. The first prior requirement for C5 is INFOTRAN to the utterance T4, and the second priority is RESPONSE to the utterance T1. In this case, there is no structural explanation for these candidates. In detail, there is no structural rule (described in section 3.4) that explains any of the two positions. Consequently, the nextction prior requirement (SUB) GOALPRE to the utterance T4 is selected because the structural rule SR3 explains that the ptype of the utterance C5 is YN-INFREQ. This hypothesis for the

structural position is confirmed by the side effect which is attached, as shown in the table in Figure 4. In this case, the topic 'sightseeing tour' is realized to be held indirectly. When the indirect relation between the topic and the utterance is recognized in the succeeding process, this hypothesis will be verified. The requirements of the unit and of the pair to the participant T produced by C5 are mtype: GOALPRE and ptype: INFOTRAN, respectively. In the same manner, the utterance T6 is assumed to be INFOTRAN and form pair with C5, using the rule SR1. In order to evidence this position of T6, matching is tried between the subnetworks C5 and T6. Since the matching succeeds, as described in section 5.1, by the side effect column in the table, this structural position is concluded to be definite.

5.3. Implication Resolution

Many researchers [10] have been studying the issues on indirect requests. Indirect requests stemming from the speaker's respect for the listener usually produce an utterance posing the listener's intention or ability instead of asking the request itself. Therefore it might be a question about preconditions for the request contents. This pragmatic pattern of request is formalized as a structural rule as follows:

SR7 ACTREQ (U, Action) -M→ stype (U, YN-interrogative),
agent (U, listener), presuppose (Action(X), V(Y)), verb (U, V)

We will show how to discern the implication of utterances which request indirectly with utterances T10 and C11.

T10: Do you have a pen?

C11: Here you are.

Although T10 is recognized to be YN-INFREQ in the general process, the succeeding utterance C11 is RESPONSE representing an action of "giving". In this case, the utterance of the RESPONSE type expects the preceding utterance to be of the ACTREQ type. However, this requirement cannot be satisfied, and coherence is not maintained by HP6. As a result, the system starts the procedure involved in the node for the utterance whose position has not yet been decided. In detail, since the performable procedure attached to the node for the predicate "give" is "supplementary lexical knowledge", lexical information of "give" which is described in section 4.2 is derived and the structural position of the utterance T10 is considered again. Then the ptype of the utterance T10 is assumed to be ACTREQ by the structural rule SR7.

6. CONCLUSION

We have presented a theory of dialogue analysis using conversational coherence from three points of view: structural coherence, cohesion, and coherence of dialogue contents. It differs from the current dialogue theories dealing with conversational coherence in that the three different aspects of coherence are integrated and treated by a similar reasoning mechanism. This mechanism generates hypotheses using default rules and incomplete information, and supports or reconsiders them with constraints. This work, in a general

sense, could be viewed as an approximate theory of disambiguation, which encompasses several kinds of ambiguities encountered in natural language analysis.

This framework for a dialogue analysis system is required to investigate the reconstruction of dialogue structures. If the analyzing process fails to maintain coherence, it runs some processes to reconstruct partial dialogue structures to the minimal extent.

Acknowledgments

The authors wish to express their gratitude to Dr. Daniel G. Hays, who is a visiting Professor at Osaka University, for his continuing efforts to improving English and valuable suggestions and comments, and to Professor Abe and Dr. Dan for frequent, stimulating, and helpful discussions. It is a pleasure to acknowledge the hospitality and encouragement of the member of the natural language group in Kitahashi laboratory.

References

1. H. P. Grice: "Logic and conversation," In P. Cole and J. Morgan (Eds.), *Syntax and semantics.* Academic Press. pp.41-58, 1975.

2. Philip R. Cohen and C. Raymond Perrault: "Elements of a Plan-Based Theory of Speech Acts," *Cognitive Science,* 3, pp.177-212, 1979.

3. K. Dohsita: "Identifying Zero-Pronouns Referring to Persons in Japanese Dialogue," *Proc. SDUMA*, 1989.

4. R. Reichman: "Conversational coherency," *Cognitive Science,* 1, 421-441, 1978.

5. J. Grosz: "Attention, intentions, and the structure of discourse," *Computational Linguistic,* vol. 12, No. 3, 1986.

6. Jerry R. Hobbs: "The use of abduction in natural language," *Proc. Nagoya Inter. Symp. on Knowledge Information and Intelligent Communication* 1989.

7. H. Kashioka, A. Doi, M. Hirai and T. Kitahashi: "Analysis of dialogue structure based on the informed knowledge," *Proc. Conference of Natural Language Processing Interest Group of Japan Information Processing Society,* 1990.

8. M. Hirai and T. Kitahashi: "Determination of Omitted Words in Japanese Sentence by Pragmatic Rules," *Proc. 1st Ann. Conf. JASI,* pp.380-388, 1987.

9. A. Doi, H. Kashioka, S. Dan, M. Hirai, N. Abe and T. Kitahashi: "Dialogue analysis by constraint-learning," *Proc. the Third Symposium on Advanced Man-Machine Interface Through Spoken Language*, 1989.

10. Geoffrey N. Leech: "Principles of Pragmatics", 1983.

11. M. Kameyama: Japanese zero pronominal binding: "Where syntax and discourse meet." In William J.Poser, editor, *Second International Workshop on JAPANESE SYNTAX. CSLI*, 1988.

12. H. Maruyama: "Semantic Analysis Using Graph Matching," *Proc. Conference of Natural Language Processing Interest Group of Japan Information Processing Society*, 1986.

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Discourse Management System for Communication Through Spoken Language

Yoichi Yamashita*, Tetsuya Yamamoto** and Riichiro Mizoguchi*

* The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, 567 Japan
** Faculty of Engineering, Kansai University, 3-3 Yamate-cho, Suita, 564 Japan

Abstract

The authors have been involved in the development of a speech understanding system called SPURT-I, which accepts utterances describing simple scenes. In order to realize communication through spoken language, it has to accept discourse. To this end, the authors are currently developing a speech understanding system with a discourse management system. This paper describes the current status of the development of a new version of the system. The discourse management system takes care of the identification of the correspondence between a requirement and a response based on the SR plan and two kinds of stacks. The interaction between the discourse management system and two other subsystems is also discussed.

1. INTRODUCTION

We have been developing a speech understanding system SPURT-I (Speech Understanding system with Rule-based and Topic-directed architecture [1, 2]. The basic assumption of our approach is that acoustically close phoneme sequences rarely correspond to semantically similar words. Although this assumption makes acoustic processing easier, language processing becomes more difficult.

SPURT-I accepts utterances describing simple scenes. In order to realize communication through spoken language, it has to accept discourse. To this end, the authors are currently developing a speech understanding system with a discourse management system. One of the characteristics of discourse is that the user does not always respond to a question posed by the system. When the system asks a question, for example, the user may ask another question instead of answering if he/she does not know what to answer or may answer incorrectly. Furthermore, the representations of the answers take various forms. Therefore, the system has to identify whether the utterance is an answer or a question and which question it corresponds to when it is an answer.

Sentences appearing in a discourse have several particular properties in addition to those presented above. They sometimes do not follow legal syntax. They often include anaphora or omission of words. Although these are important issues in discourse analysis,

Utterances of the user
(a) Requirement: stimulus to the system.
(b) Response: answer to the question or requirement of the system.
(c) Confirmation: acknowledgement to the response of the system.
(d) Statement: auxiliary information.

Utterances of the system
(a) Requirement: stimulus to the user.
(b) Response: answer to the question or requirement of the user.
(c) Evaluation: evaluation of the response of the user.
(d) Statement: auxiliary information.

Figure 1. Categories of utterances.

this paper deals only with the issue discussed earlier, since other research groups take care of them.

The discourse management system has two other functions. One is to predict the type of response of the user in order to give useful information to the language processing subsystem, and the other is to standardize the information obtained from the user before giving it to the problem solver. In order to perform these three functions, we introduce the SR plan, which represents fundamental relationships between requirements and responses and has a two-stack architecture for managing them. This paper describes the details of the management system [3, 4].

2. BASIC CONCEPTS

2.1. Characteristics of a Discourse

This paper deals only with discourses with definite purposes, such as those appearing in consultation, information retrieval, CAI and so on. In such discourses, utterances are classified as shown in Figure 1.

A sentence in an utterance can be divided into one of the above categories; however, an utterance usually contains more than one sentence. So, let us investigate what types of sentences are contained in an utterance. A discourse with a definite purpose contains either a stimulus (requirement) to the opponent or a response to the stimulus given by the opponent or both explicitly or implicitly. Communication is thus performed by giving stimuli or responses to each other, and some sentences effecting this interaction, such as confirmation, statement, or evaluation, are given in an utterance if necessary.

2.2. Overview of the Discourse Understanding System

A block diagram of the total system is shown in Figure 2. The discourse management system is located between the speech understanding system and the problem solver. Two problem solving systems, i.e. ITS (Intelligent Tutoring System) and registration desk system, are considered in our research, but this paper deals only with the former task.

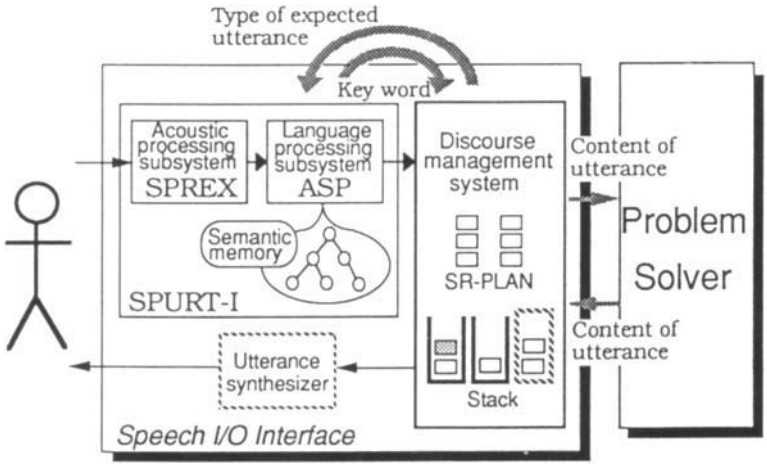


Figure 2. Block diagram of our total systems.

System SR plan	User SR plan
Requirement of the system	Requirement of the user
Response of the user	Response of the system
(Evaluation of the system)	(Confirmation of the user)

Figure 3. The structure of SR plans.

Discourse processing is tightly connected to the ASP (ASsociation-based Parser) [5, 6], which recognizes utterances of the user. The discourse management system has various knowledge for discourse processing, some of which is also useful for the ASP, so the two subsystems share parts of their knowledge bases.

An utterance of a user who uses the problem solving system, is used as input for the acoustic processing subsystem, named SPREX (SPeech Recognition EXpert) [7, 8], and it is converted to a sequence of phonemes. Then, the phonemes are used as input for the ASP to identify the utterance. The discourse management system accepts the utterance and generates appropriate responses by interacting with the problem solver.

3. DISCOURSE MANAGEMENT SYSTEM

3.1. Outline

A discourse management system has a knowledge base of the possible interaction between the user and the problem solving system. We call the interaction pattern the SR (Stimulus and Response) plan, which will be discussed later. We have two kinds of SR plans: one is for the user and the other is for the system. Figure 3 shows the structure of the SR plan.

The discourse management system employs two stacks for manipulating SR plans. Requirements or questions (SR plans) are pushed onto the stack. When a response is

S-QUIZ	<Questions>
1. SP-ASK-COMPONENT	to ask objects or locations of an action
2. SP-ASK-REASON	to ask reason of an action
3. SP-ASK-FACT	to ask whether a fact is true or not
S-COMMAND	<Order>
4. SP-COMMAND	to order the student to do some tasks
S-EXPLAIN	<Explanation and requirement>
5. SP-AFTER-EXPLAIN	to confirm whether the student understands correctly or not
S-DEMAND-U.SPEECH	<Requirement to the user's utterance>
6. SP-CONFIRM	to confirm the correctness of system's understanding
7. SP-SUPPLEMENT	to ask missing information
8. SP-DISAMBIGUATE	to ask for disambiguation

Figure 4. System's SR plan.

given, the topmost element is usually popped and stored in the history data base. In ordinary cases, a response corresponds to the latest requirement independent of the history of requirements and responses. This shows the adequacy of a stack structure. However, there may be some cases where a response corresponds to a requirement in a position lower than the topmost element. The process in such a case will be discussed in section 4.2. The system has a stack for system SR plans and one for user SR plans. These two stacks help to simplify the management process of the interaction.

3.2. SR Plan

The SR plan represents some procedures useful for identifying interactions based on stimulus and response. Figures 4 and 5 show examples of SR plans for the system and user, respectively. The following is a common procedure for the system SR plan.

Step 1: Identify the SR plan corresponding to the requirement of the problem solver.

Step 2: Send the content of the utterance to the utterance synthesizer.

Step 3: Predict the possible types of responses to the utterance, consulting the SR plan, and send the information to the ASP.

Step 4: When the result of the analysis of the user's utterance is given by the ASP and it is a response, standardize the result and send it to the problem solver, asking its evaluation. When it is not a response, this plan is suspended.

Step 5: When the response turns out to be correct from the result of the evaluation, this SR plan is popped from the system stack. Otherwise, this plan is suspended. Some SR plans may terminate at step 4.

The procedure for the user SR plan will not be given.

U-DEMAND-K.BASE <Requirement concerning knowledge base>
1. UP-ASK-DEFINITION to ask definition of terms
2. UP-ASK-COMPONENT to ask objects or locations of an action
3. UP-ASK-REASON to ask reason of an action
4. UP-ASK-FACT to ask whether a fact is true or not
5. UP-ASK-WAY to ask how to do

U-COMMAND <Order>
6. UP-COMMAND to order the system to do some tasks

U-DEMAND-S.SPEECH<Requirement to the system's utterance>
7. UP-ASK-UTTERANCE to ask what the system said
8. UP-ASK-CONFIRM-UTTERANCE to confirm what the system said
9. UP-ASK-MEANING to ask the meaning of the system's utterance

Figure 5. User's SR plan.

After the requirement of the system	After the response of the system
(1) Response to it	(1) Its confirmation
(2) New requirement	(2) Response to the suspended requirement of the system
	(3) New requirement

Figure 6. Expectation of user's utterances.

4. BEHAVIOR OF THE DISCOURSE MANAGEMENT SYSTEM

4.1. Interaction with the ASP

In SPURT-I, the ASP plays an important role in identifying utterances correctly. The ASP employs various knowledge, such as syntax, semantics, and association relations between topic and vocabulary. In the previous implementation, the ASP accepted utterances describing simple scenes, where expectations were done using the association relation mentioned above. Our new discourse understanding system can often expect the user's utterances, since major parts of them are responses to the requirements of the system. Figure 6 shows expected utterances of users.

The discourse management system predicts the user's utterances according to the information shown in Figure 6. Then, it sends to the ASP templates of the expected responses of the user, which are written in the system SR plan, or templates of possible requirements of the user, which are written in the user SR plan. The ASP can perform top-down analysis of the input utterance utilizing this information.

An example of a system SR plan, SP-ASK-FACT, is shown in Figure 7. As shown

System: *Ikou suru toki, sono kou no fugou wo kaemasuka?*
(Do you change the sign of the term when you move it?)

type-1 [affirmative, negative]	*hai, iie* (yes, no), *etc.*
type-2 [repetition]	*kaemasu* (I change), *etc.*
type-3 act:[guess] obj:([repetition])	*kaeru to omoimasu* (I am not sure but I change the sign), *etc.*

Figure 7. Description of SP-ASK-FACT.

in the figure, an SR plan has some templates of expected responses, which are used for identifying the correspondence between requirement and response. "Repetition" in the figure indicates that the response consists of a repetition of the words contained in the utterance of the system. So, it is instantiated when the request of the system is given.

Top-down analysis is thus introduced in the ASP. However, there exist at least two shortcomings in this formulation.

(1) There is no ordering of multiple templates.
(2) There is no expectation which requirement (user SR plan) comes next.

In order to overcome these difficulties, we introduce bottom-up analysis. When a word lattice is given, the ASP scans all the candidate words to find some keywords which remind of some templates. For example, the template "UP-ASK-REASON" is expected when the keyword "why" is found in the word lattice. When a template is expected in the bottom-up analysis, top-down analysis can be done as described above. If many templates are expected, they are ordered according to some heuristics. The success of the analysis using the template expected indicates that the utterance is of a type of directly inferred from the template. When the top-down analysis fails, the next probable template is sent and the process is repeated. If all the expectations fail, then complete bottom-up processing using the dependencies between the words, is performed.

4.2. Extension of the Stack Manipulation

The stack manipulation discussed thus far makes the latest SR plan active, in other words, the user's response is considered to correspond to the latest requirement, which is the topmost element of the stack. However, this manipulation is no longer valid in the following cases:

(1) When the user responds to a previous requirement of the system.
(2) When the system asks another question without answering the question posed by the user after he/she answered incorrectly to the previous question of the system.

Figure 8 shows an example of type (1).

In this example, U5 corresponds not to S4, which is on top of the system stack, but to S1 located under S4. To cope with such a case, the management system does not pop the requirements having incorrect answers. If and only if the utterance of the user

S1:	*Houteishiki X-2=7 wo tokinasai.* (Solve the equation X - 2 = 7.)	*(requirement)*
U2:	*5 dato omoimasu.* (I think it is 5.)	*(response)*
S3:	*Chigaimasu.* (You are incorrect.)	*(evaluation)*
S4:	*-2 wo ikousuruto X=7-2 desu ka?* (Do you have X = 7 - 2 when you move the term -2 to the other side?)	*(requirement)*
U5:	*Wakatta! X=9 desu ne.* (I got it! X = 9.)	*(response)*

Figure 8. An example requiring irregular manipulation of the stack.

cannot be recognized, whether it is a response to the topmost SR plan of the system stack or a new requirement to the system, the management system goes down the system stack to see if there exists a requirement corresponding to the utterance. When such a requirement is found and the utterance is a correct answer to it, all the elements higher than this requirement are popped including itself. The second case is concerned with the user stack. It can be detected as a case where the system neglects the requirement of the user. When the requirement of the user is confirmation of his/her answer, the user SR plan is popped, considering the system giving an implicit (negative) answer to it. This irregularity is rather easy to deal with, since it is caused by the system intentionally.

4.3. Detailed Description of the Behavior of the System

Let us take an example of a discourse in which an ITS is teaching a user to solve linear equations, as shown in Figure 9. In the rest of this subsection, the behavior of the discourse management system is described.

Suppose that the management system is requested by the problem solver, ITS in this case, to pose the question if X = 2 is the solution of the equation 6 - 2X = 3X - 4. This is a requirement asking a fact, so the system SR plan SP-ASK-FACT is activated and pushed onto the system stack. Then, the content of the utterance is sent to the utterance synthesizer and S1 is uttered. Next, templates of the expected response are generated according to the information written in SP-ASK-FACT. In this case, templates of affirmative, negative, assertion of a fact (noun phrase), or guess of a fact are expected. Then, they are instantiated according to the situation, especially using the words appearing in the utterance of the system. For example, "assertion of a solution" and others are obtained in this case and they are sent to the ASP.

The next utterance, U2, is successfully analyzed and is recognized as a response to S1. Then it is sent to the problem solver to see if it is correct or not. Since it is an incorrect answer, the SR plan is not popped, though a corresponding response is obtained. The problem solver sends the content of the next utterance to the discourse management system, the type is recognized as SP-COMMAND SR plan by the system and it is pushed onto the system stack. The content is further sent to the synthesizer and S3 is uttered.

S1: *Dewa, tsugi no toi ni kotaenasai. X=2 wa,houteishiki 6-2X=3X-4 no kai desuka?*
(Now, answer the following question.
Is X=2 a solution of the equation 6-2X=3X-4?)

U2: *Kai dewa nai.*
(It isn't a solution.)

S3: *Houteishiki ga yoku wakattenai youdesune. Tatoeba, X=3 ga 2X+1=5 no kai ka douka shirabemasu. X=3 wo sahen no 2X+1 ni dainyuusite kudasai.*
(It seems to me you don't understand linear equations well. For example, let's examine whether X=3 is the solution of 2X+1=5: Substitute X with 3 of the left hand side 2X+1.)

U4: *Mazu, 3 wo 2 ni kakeru no desune?*
(Well, I multiply 2 by 3, don't I?)

S5: *Sou desu.*
(Yes.)

U6: *7 desu.*
(It is 7.)

Figure 9. Example of a simple discourse.

The template "assertion of numerals" and others are expected according to the information in SP-COMMAND, but the analysis fails, since U4 is not a response but a new requirement. Then, templates of possible requirements are generated according to bottom-up processing. In this example, UP-ASK-FACT is identified, since the utterance is of the form "fact + interrogative" and is pushed onto the user stack, suspending the system stack. At this moment, therefore, UP-ASK-FACT is active and SP-COMMAND is inactive. The result of the analysis of U4 is sent to problem solver, it evaluates U4, and S5 is uttered. Then, the system expects utterance of confirmation, such as "*wakarimashita* (I see)", "*wakatta* (I've got it)" and so on, but it fails, since U6 is not a confirmation. Then it tries to identify whether it is a response to the suspended requirement. In this case, SP-COMMAND is the candidate, so the templates expected according to the plan, such as "assertion of numerals", are sent to the ASP again. The analysis in the ASP succeeds this time.

5. CONCLUDING REMARKS

We have discussed a speech understanding system with discourse management capability. The basic mechanism of the discourse management subsystem is a two-stack architecture based on SR plans. Currently, the whole system is being implemented in Common Lisp and Flavors on Symbolics 3620.

References

1. M. Hori, K. Tsujino, R. Mizoguchi and O. Kakusho: "A speech understanding system SPURT-I – Dynamic clustering method and performance evaluation –," *Trans.IEICE*

Japan, J72-D-II, 8, PP.1291-1298, 1989.

2. M. Hori, K. Tsujino, R. Mizoguchi and O. Kakusho: "A speech understanding system: SPURT-I," *Proc. WESTPAC-III*, pp.779-782, 1988.

3. R. Mizoguchi, T. Yamamoto and Y. Yamashita: "Dialog management system based on stack structure," Research Report No. PASL 1-6-4, 1989.

4. T. Yamamoto, H. Ozaki, M. Hori and R. Mizoguchi: "Discourse management system in speech understanding system SPURT-I," *National Conference JSAI*, pp.499-502, 1989.

5. M. Hori, R. Mizoguchi, M. Kawachi, K. Uehara, J. Toyoda, O. Kakusho: "Association-based parser for speech understanding system – Framework design based on Cognitive exploration," *Trans. IEICE Japan*, J71-D, 5, pp.774-781, 1988.

6. M. Hori, M. Ohsumi, H. Ozaki, R. Mizoguchi and O. Kakusho: "Association-based parser for speech understanding system – Details knowledge and performance evaluation," *Trans. IEICE Japan*, J71-D, 5, pp.782-789, 1988.

7. R. Mizoguchi, K. Tsujino and O. Kakusho: "A continuous speech recognition system based on knowledge engineering techniques," *Proc. ICASSP*, pp.1221-1224, 1986.

8. K. Tsujino, T. Sakurai, Y. Nomura, S. Chigusa, R. Mizoguchi and O. Kakusho: "A continuous speech recognition system with a powerful environment for the knowledge base construction: SPREX-II," *Trans. IEICE Japan*, 3, pp.531-542, 1988.

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Towards Habitable Systems: Use of World Knowledge to Dynamically Constrain Speech Recognition

Sheryl R. Young and Wayne H. Ward

Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract

Current state-of-the-art speaker-independent continuous speech recognizers are able to achieve word recognition rates well above 90 percent with lexicons of 1000 words or less using grammars with perplexity 60 or less. Performance of these systems decreases rapidly as the perplexity of the grammar increases. As we allow users more flexibility in interacting with recognition systems, the size of the lexicons and perplexity of the grammars increase greatly. Allowing spontaneous speech instead of read speech compounds the problems even more. Other sources of knowledge may be available to help constrain the ever more complex search spaces in such systems. When recognition systems are used in performing problem solving tasks, predictable features of the user's behaviour can be used to aid recognition. We describe a system (MINDS) which uses additional constraints based on dialog interactions. The constraints are applied in a manner that allows optimum performance when users behave predictably, and degrades gracefully when they do not. We also present an evaluation of the system's performance to show the utility of the additional knowledge sources.

1. OVERVIEW

One of the biggest problems in computer speech recognition is coping with large search spaces. The search space for speech recognition contains all the patterns associated with words in the lexicon as well as all the legal word sequences. The most widely used recognition systems are hidden Markov model (HMM) based. In these systems, typically, each word is represented as a sequence of phonemes, and each phoneme is associated with a sequence of Markov states. As search space size decreases, recognition performance increases. Knowledge can be used to constrain the exponential growth of a search space and hence increase processing speed and recognition accuracy [6, 11, 17]. Currently, the most common approach to constraining search space is to use a grammar. The grammars used for speech recognition dictate legal word sequences. Normally they are used in a strict left to right fashion and embody syntactic and semantic constraints on individual

sentences. These constraints are represented in some form of probabilistic or semantic network which does not change from utterance to utterance [3, 11, 12].

As we move toward habitable systems and spontaneous speech, the search space problem is greatly magnified. Habitable systems permit system users to speak naturally. Grammars which try to cover even naturally elicited syntactically accurate sentences have perplexities that are an order of magnitude larger than the perplexities of grammars typically used by speech recognizers. Spontaneous speech grammars will have even larger perplexities. Spontaneous speech is ungrammatical and contains much editing. These edits can occur anywhere within a sentence and are often preceded by interjections. Additionally, spontaneous speech exhibits human noise in the form of filled pauses. Finally, the above phenomena are compounded by the presence of multiple sentences uttered without pausing at sentence boundaries and silent pauses within incomplete phrases. Hence, the notion of a well formed sentence exhibiting typical syntactic regularities is not applicable when processing spontaneous speech. These problems point to the need of using knowledge sources beyond typical syntax and semantics to constrain the pattern matching process in speech recognition.

There are many other knowledge sources besides syntax and semantics. Typically, these are clustered into the category of pragmatic knowledge. Pragmatic knowledge minimally includes inferring plans, using context across clausal and sentence boundaries, determining local and global constraints on utterances and dealing with definite and pronominal reference. Work in the natural language community has shown that pragmatic knowledge sources are important for understanding language. People communicate to accomplish goals, and the structure of the plans to accomplish them are well understood [7, 18-21] [1, 4, 5, 9, 16]. When speech is used in a structured task such as problem solving, pragmatic knowledge sources are available for constraining search spaces.

In the past, pragmatic, dialog level knowledge sources were used in speech to either correct speech recognition errors [2, 8] or to disambiguate spoken input and perform inferences required for understanding [12, 14, 15]. In these systems, pragmatic knowledge was applied to the output of the recognizer.

In this manuscript we describe an approach for flexibly using contextual constraints to dynamically circumscribe the search space for words which can be matched against a speech signal. We use pragmatic knowledge to derive constraints about what the user is likely to say next. Then we loosen the constraints in a principled manner. Hence, we generate sets of predictions which range from very specific to very general ("layered predictions"). To enable the speech system to give priority to recognizing what a user is most likely to say, each prediction set dynamically generates a grammar which is used by the speech recognizer. The prediction sets are tried in order of most specific first, until an acceptable parse is found. This allows optimum performance when users behave predictably, and displays graceful degradation when they do not. The implemented system (MINDS) uses these layered constraints to guide the search for words in our speech recognizer. For our recognizer, we use a modified version of SPHINX (Lee, 1988) large vocabulary, speaker independent, continuous speech recognition system.

The following section places the research described in the context of the overall MINDS system architecture. The following two sections describe the methods used to generate predictions and use them to guide recognition. We then present the results of two studies

which illustrate both perplexity reduction and performance improvements resulting from the use of predictions. Finally, we describe our current work in progress.

2 MINDS SYSTEM ARCHITECTURE

The MINDS system uses pragmatic knowledge sources predictively to circumscribe the search space for words in a speech signal [10, 22]. The pragmatic knowledge sources are embodied in an elaborate dialog model. The dialog model infers plans, performs plan tracking, deals with clarification subdialogs and dynamically computes constraints using local and global focus, or contextual information propagated from prior information seeking stages. To allow for diverse user behavior, MINDS uses a principled, general algorithm for relaxing constraints. Constraints are organized into sets that are successively more general, called "layers". When some constraints are violated, we use the non-violated constraints to reduce search space. Additionally, the flexible use of constraints allows the use of knowledge sources that are less certain to be true. Users that behave consistently can benefit greatly from enhanced recognition and the system will show a graceful degradation on those who do not.

To enable the MINDS system to generate predictions and use them to guide the speech recognizer, we have partitioned the system into interacting modules, as seen in Figure 1. The bf speech module is composed of a modified version of SPHINX speaker-independent, continuous, large vocabulary speech recognizer. This version of SPHINX uses finite state grammars to constrain search. The grammars are dynamically generated after each utterance by the dialog module and sent to the completion module. Hence, the speech module receives input from the completion module and the speaker. It sends its output to both the completion module and the display module.

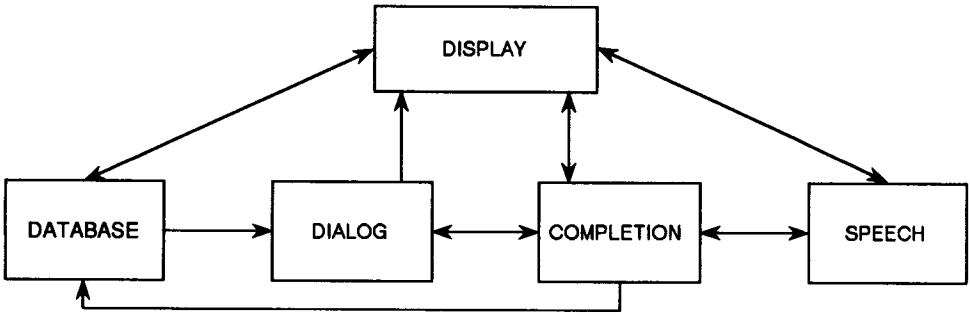


Figure 1: MINDS System Modules.

The **completion module** is composed of a semantic parser, representations of the domain, the database, and the finite state grammar. The completion module communicates with the speech module, the database, and the dialog module. It takes the speech output, parses it and performs any necessary disambiguation. Then it takes its semantic representation and communicates it to the dialog module and generates a database query.

Once the predictions are generated, the completion module indexes them into precompiled portions of the finite state, semantic grammar and places restrictions on the expansions of the rewrite rules embedded in the finite state nets. The nets are then merged.

The **dialog module** is composed of a domain knowledge base, a hierarchical representation of possible domain problem solving plans, and a set of heuristics for propagating constraints, inferring plans and tracking plans. The dialog module receives input from both the completion module and the database module so that it can track all information communicated. The dialog module is responsible for generating layered sets of predictions. It communicates these to the completion module so they can be expanded into potential surface forms.

The **database module** is composed of the Informix^{TM} relational database management system filled with a domain database, an "expert" interface to the database, and a natural language generator. The database module receives input queries from the completion module. If these are either ambiguous or computationally expensive, the "expert" interface has the option of querying other system modules or the user for clarification / further specification. The "expert" interface translates query inputs into a form necessary for the database. Additionally, it translates the output into a semantically meaningful form. The output is then sent to the dialog module while the natural language generator produces the sentential output which is then communicated to the display module.

The **display module** is composed of four displays. Two displays are maps which display the current version of the world and can zoom into areas of interest to the system user. The third display depicts detailed information previously communicated in the dialog, while the fourth display is devoted to communicating with the user and all the system modules. The fourth display contains a type in window which also displays the generated natural language database response as well as spoken utterance. Additionally, it contains windows for displaying clarifications requested by other system modules, and a window for displaying the test set perplexity of the just parsed utterance. This module communicates with all other modules and knows the complete system state.

When spoken information is input to the system, it is first processed by the speech module using the predictions generated earlier. Its output is sent to both the display module (where it can be corrected if necessary) and the completion module. The completion module performs a semantic parse on the information and generates a database query. The semantic parse is sent to the dialog module and the database query is sent to the database module. The dialog module then determines which possible plan steps were activated by the input and uses the database response to gather further context. It then generates a new set of layered predictions and passes these back to the completion module for expansion and use by the speech module. Each of the system modules described above run in a distributed environment.

In the next section we describe the use of plans to limit search space and the algorithms which enable the MINDS system to generate layered sets of predictions.

3. PLAN BASED CONSTRAINTS: PREDICTION GENERATION

The idea underlying the MINDS system is that tracking all information communicated (user questions and database answers) enables a system to infer a set of possible problem solving plans and to track progress through these plans. In the convention of Newell and Simon (1972) these plans are represented as hierarchically organized goal states. *For example, in the domain of dealing with disabled ships, a goal state would be finding a replacement ship.* As each new input sentence is spoken, the system analyzes the utterance to determine the concepts expressed, and uses these concepts to activate goal states. To derive plan based constraints on future utterances, active goal states are assessed to determine legal next states. *For example, when finding a replacement ship, some of the legal next states which follow a question about the ships in some region are questions about more ships in the region, questions about availability of these ships, and questions about the ships' equipment.* Because speech systems use grammars to guide word transactions, we associated a list of required and optional concepts with each goal state (*e.g. concepts associated with a goal state for ship equipment include equipment, weapons, aircraft, electronics, etc.*). The list of possible next states is used to generate a set of possible concepts which could be spoken in the next utterance. This set is then limited by local and global focus which takes into account prior context, rules about reference, etc. The speech recognizer only searches for surface forms expressing concepts in this set.

3.1. Layered Predictions

Plan based constraints are quite effective in reducing search space by delimiting the types of information likely to be communicated [10, 22]. But plan based constraints are based upon inferring user plans. Usually it is not possible to either definitively select a single plan step given an input utterance. Similarly, users may exhibit unexpected behavior by either violating the hierarchical nature of a plan or leaving plan steps incomplete. As both domain size increases and spontaneously generated speech is used for generating queries, these problems become magnified.

To overcome the problem of multiple active plans and unexpected user behavior, we instituted three procedures. First, we designed an algorithm to select "the best" plan step or goal state from the list of possible goal states activated by the preceding utterance and database response. Here we preferred goal states that were both complete and most likely to follow given previous goal states activated. Second, we maintained a list of all other active goal states, including those which were not hierarchically embedded. These activated states were used to generate some alternate predictions about what the user could say. Third, we generated sets of layered predictions about the content of the following utterance. The predictions ranged from very specific to very general. These layered predictions were rank ordered to reflect both amount of constraint provided as well as the reliability of the knowledge sources used to generate them. It should be noted that the least constraining prediction layer allowed all domain concepts. This means that the system could cope with any statement the user might say even if it's not included in the system grammar. However, the system cannot cope with words which are not included

in the system lexicon.

By layering predictions, we allow the system to reparse a speech signal with a different grammar until such time as a good parse is received. The ability to reparse an utterance also enables us to use less reliable knowledge sources to further constrain our predictions. Hence, we added two additional knowledge sources to the system: user domain expertise models and preference orderings for conjunctive goals.

Observing that system users with significant domain expertise solved problems using very different plans than novice users, we attempted to model the effects of expertise by constructing domain knowledge models of novice, intermediate and expert system users. Our user models were represented as subsets of the domain knowledge base. The knowledge differed primarily by the existence of relations between domain objects. *For example, an expert user would know that each class of ships has a set of default equipment and is suited for particular types of tasks, while a novice user might not be aware that shiptypes are divided into ship classes.* The user models were then used to construct schemas which specified which goal states were exclusive. *To further the last example, a control schema for an expert user would show that if the user asked about a shipclass they would not ask about default equipment.* These models were hand coded from the training set data.

Similarly, we used the training data to derive probabilistic orderings on conjunctive subgoals. These orderings told us which conjunctive goals would be executed first, second, etc.. The orderings were computed across individuals (although our training data only came from two people). However, there is no reason why these could not be automatically obtained for individual system users in future systems.

Thus, the MINDS system used the following knowledge sources to derive predictions about the content of a user's next utterance:

1. knowledge of problem solving plans represented as a hierarchical goals,
2. semantic knowledge about the application domain's objects, attributes and their interrelations (a domain knowledge base),
3. domain independent knowledge about methods of speaking, appropriateness of references and partial utterances (local and global focus)
4. dialog history knowledge about information previously communicated,
5. discrete models of user domain expertise as described above, and
6. information about user preferences for ordering conjunctive subgoals

These knowledge sources were used by the prediction module to perform iterative analyses of the dialog after each input/database response pair and generate sets of restrictions on the next utterance. The predictions generated. Each successive layer is less constraining than the prior layer.

The most constraining prediction set is generated using all knowledge sources listed above. The next set does not use user models and uses a larger non-overlapping set of goal states. Further sets are generated by moving upward in the goal hierarchy, allowing

more plans to be executed. The prediction sets become successively more general, hence the term "layered". Ultimately, the entire system grammar will be used. If this fails, an "allword" recognition is attempted where any word sequences are allowed (providing of course that the words are in the system lexicon).

3.2 Derivation of Predictions

To illustrate how the information contained in a goal state or plan step is used to generate predictions, we present simplified, although prototypical representations of both a plan step and the control information which encodes our "less reliable" information about users. These are depicted in Figures 2 and 3, respectively.

```
[Shipclass
 :Concepts-Requited ((Shipclass single-use Chile-restrictions*
                                        (knoxclass perryclass)))
                    Concept Times-used Restriction-pointers
 :Optional-Concepts ((Region single-use Chile-restrictions*
                                        (persian-gulf)))
 :Optional          (Not for expert-user)
                    True/Nil/User-consideration
 :Next-states       (Find-Replacement) goal-state
 :Parent            (Find-Replacement)
 :Children          (none)
 :Control           (none)
]
* = Computed by local and global context
```

Figure 2: Example Goal State Schema.

```
[Control00030 - for Find-replacement
 :Exclusive        ((Shipclass Equipment))
 :Omit             (Shiptype)
 :order            ((.90 Shipclass .10 Mission-Info)
                    (.90 Mission-Info .10 Shipclass))
]
```

Figure 3: Example Control Schema.

The concepts-required and **optional-concepts** slot values are used to specify the concepts relevant when a user transits to the goal state. The number of concepts per goal state and the number of goal states a user could progress to next determine the size of the lexicon the speech recognition system must analyze. The **control** slot contains a pointer to a **control schema** whenever the child slot is not empty.

Control schemas predict whether any child states are likely to be omitted and any preferred orderings on the states for a specific system user. They are used to generate the most constraining prediction layers.

As seen in Figure 3, there are three slots in a control schema. The **order** slot stores information about preferred orderings among non-optional, conjunctive subgoal states. The **exclusive** slot stores pairs of goal states which are exclusive because the information in the first allows the user to infer the information in the second. The **omit** slot store a list of goal states the user omits because they are unaware of the domain concepts.

Control schemas are attached to parent goal states to predict which child states will be visited. Hence, they are also used to dynamically compute the value of the *optional* slot for each child schema. When a state is predicted to be omitted, the optional slot value becomes true for that cycle of input and database response. The **optional** and **concepts-required** slots are important for determining when a goal state is complete.

3.2.1 Algorithm for Prediction Generation

These structures are used to derive predictions in the following process. When an incoming utterance and database response are processed, we select the most likely plan steps executed. If a plan step is not complete, then our most constraining prediction set reflects the assumption that the user will complete the plan step. The other prediction layers do not change. If one or more plan steps are complete, we identify the next goal states to which a system user could transit. Identifying possible next states is the basis for each layer of predictions. Next, we take all of the possible next plan step which would follow from the just completed step and store them. Following this, we apply our less reliable knowledge sources to further prune the set of next, most likely steps. To do this, we first use any and all knowledge of user ordering preferences and states which could be omitted. Then we back off first on the ordering information and then on the states which could be omitted. Then, since all goal states and possible problem solving plans are represented in a hierarchical manner, we progressively move up a layer in the hierarchy of incomplete, yet active plan steps or goal states to determine state to which the user could next progress. Once we have determined the next states, we can take the concepts associated with the states and compute restrictions on their expansions, restrictions on references given the state and the context, and restrictions on partial utterances.

Once the predictions are generated, they are expanded into potential surface forms and used by the speech recognition module to guide the pattern matching process, as described below.

4 USE OF PREDICTIONS TO GUIDE RECOGNITION

The idea behind the MINDS system is to use pragmatic knowledge to reduce the amount of search performed by the speech recognizer thereby reducing the recognition errors caused by ambiguity and word confusion. Hence, pragmatic knowledge is used predictively. These predictions take the form of semantic concepts with restrictions on their children and restrictions on methods of referencing the concepts. The underlying motivation for using a semantic representation was that speech recognizers

The algorithm is somewhat simplified for purposes of this discussion. A forthcoming paper will discuss predictions re: clarification subdialogs and non-hierarchical open focus spaces. can be guided by using a semantic grammar. Furthermore, a semantic grammar

can be represented with non-terminal rewrite rules which group semantically related surface forms. Thus, once the layered predictions are generated, appropriate portions of the semantic grammar are "activated" and restrictions are placed on the expansion of rewrite rules as dictated by the predictions. The expanded "active" grammar is then used to guide the speech recognizer.

4.1. Expanding Prediction into Potential Surface Forms

To expand the prediction sets, we must relate the abstract concepts to words sequences which represent the conceptual meaning of the concepts.

For each concept, we have a partially precompiled set of possible surface forms which can be used in actual utterances. These individual concepts usually expand into noun phrases.

In addition to the individual concepts, we have a complete semantic network grammar which is indexed according to the combinations of semantic concepts expressed. The semantic network grammar is partitioned into subnets. A subnet defines allowable syntactic surface forms to express a particular combination of semantic concepts. *For example, all the ways for asking about a ship's mission are grouped into subnets.* The subnets are also partitioned along syntactic lines, such as ellipsis (a partial utterance), single anaphora (we, he) plural anaphora (they, them, those) and definite reference (the). This multidimensional indexing allows predictions about syntactic forms as well as concepts. Thus, the surface forms associated with each combination of semantic concepts are segmented into a number of subnets.

The grammar is precompiled into finite-state networks. The nodes of the nets represent non-terminal categories which expand into words. This also allows us to add additional words to the system lexicon without modifying the grammar.

4.1.1. Algorithm

As illustrated above, the grammar is multidimensionally segmented into subnets. Our algorithm for using this information to translate each set of layered predictions into a form usable by the speech recognition module is as follows.

First we find the set of subnets which contain one or more of the predicted semantic concepts. Forms that violate predictions on ellipsis or anaphora are pruned from this set. This set defines the nets to be active for the next utterance. Once the set of subnets is defined, we look for all the semantic concept categories, and check if their membership has been reduced by the predictions from the dialog module. This step represents a restriction on concept words that are active. The module then forms an active lexicon list and grammar based on the resulting subnets and restrictions derived from this algorithm.

The final expansion of predictions brings together the partitioned semantic networks that are currently predicted and the concepts in their surface forms. Through an extensive set of indexing, we intersect all predicted concept expressions with all the predicted semantic networks. This operation dynamically generates one combined semantic network grammar which embodies all the dialog level and sentence level constraints on the sentences which can be matched.

This operation is repeated for each set of predictions and results in a set of layered semantic networks. These networks are used by the recognizer to guide the pattern matching process.

To illustrate this point, let us assume that the frigate "Spark" has a disabled sps-48 radar. One layer of our predictions expects the user to ask when it will be repaired. The dialog tracking module predicts the "shipname" concept restricted to the value "Spark", the estimated time of repair concept and the "ship-capabilities" concept, restricted to radar and SPS-48. Single anaphoric reference to the ship is also expected, but ellipsis is not meaningful at this point. The current damage assessment dialog phase allows queries about features of a single ship.

During the expansion of the concepts, we find the word nets such as "the ship", "this ship", "the ship's", "this ship's", "it", "its", "Spark" and "Spark's". We also find the word nets for the radar capabilities such as "surface search radar", "sps-48", "radar", etc., and word nets for repair questions.

We then intersect these with the sentential forms allowed during this dialog phase. Thus we obtain the nets for phrases like "Display / list etr / estimated time to repair / estimated repair time / projected time for repair on / for surface search radar / sps-48 / radar / sps-48 surface search radar", and "Display / what is / its / Spark's / this ship's / the ship's etr / projected repair time /", and many more. This semantic network now represents a maximally constrained grammar which reflects the constraints embodied in this layer of predictions.

4.2 Recognizing Speech Using Dynamic Networks

As explained above, predictions are used to define an active set of subnets and an active set of words to be used in processing the next utterance. We use the SPHINX system as the basis for our recognizer. It has been modified to use finite state nets to control word transitions instead as opposed to word-pairs or bigrams. SPHINX creates word models by concatenating Hidden Markov Models of phonemes. These word networks are precompiled.

During recognition, the speech module performs a time-synchronous beam search. The search traces through the active nodes of our nets to control word transitions. As the search exits a word it forms a set of words to transit to form successor states in the nets. Only the active finite state nets and active words are used to compute the successor word set. The search then transits to the words in this set. Paths falling below a threshold score are pruned. The network is used to allow only "legal" transitions. It does not affect the score of a path but simply restrict words which can continue the path.

The recognizer is given several sets of predictions which are successively more general (less constraining). The most constraining set is used first. If no string is found which exceeds a threshold score, the input is reprocessed using the next more general set of predictions. If an acceptable recognition is not found using the most general set of predictions, the entire set of nets is used.

After input has been processed, the word string with the best score is passed back to the system for parsing. In addition to the word string, the subnet matched, the overall score and individual word scores are passed back.

5 RESULTS

The above described use of plans in speech recognition is currently embodied in the MINDS system (Young and Ward, 1988; Young, Hauptmann and Ward, 1988; Hauptmann and Young, 1988). MINDS is a multimedia interactive dialog system where users solve problems by interacting with a database. Users can speak, type or point to input information and both the system and the user can initiate clarification dialogs when appropriate. It uses an adapted version of the SPHINX (Lee, 1988) speech recognition system with a 1000 word vocabulary. Its task domain is naval resource management. Here users must query a relational database to determine whether a disabled vessel should be replaced with another vessel, scheduled for a later repair, or whether the mission should be delayed.

To test the ability of our layered predictions to both reduce search space and to improve speech recognition performance, we performed two experiments. The first experiment assessed perplexity reduction enabled by the predictive use of pragmatic knowledge. The second experiment measured improvement in recognition rates resulting from the use of layered predictions. Both studies used an independent test set. This means that the utterances processed by the system to obtain the experimental results had not been previously seen by the system. Furthermore, the test set did not include any clarification dialogs.

5.1 Test and Training Sets

Our test data (10 scenarios) were adapted versions of three problem solving sessions taken from the TONE database. The TONE database is a set of transcripts from NAVAL personnel solving problems about what to do with a disabled vessel. The personnel must determine whether to delay a mission, find a replacement vessel or schedule a repair for a later date. They use a database to find necessary problem solving information. In addition to the three scenarios from the TONE database, we created seven additional sessions by paraphrasing the original three. These scenarios were not used to train upon.

Our training data were five different problem solving scenarios from the TONE database. The training scenarios were used for writing grammars and developing user models. Problem solving plans were derived from an abstract description of the stages and options available to a problem solver. The abstract plan descriptions were provided by the Navy.

Our database was different from the one used in gathering the TONE transcripts. While it contained the same fields, the information about particular ships differed across the two databases. To enable testing with the TONE transcripts, we had to adapt the test scenarios. Our adaptations consisted of the following:

- Shipnames were changed to correspond to those in our database.
- Lexical entries not in our lexicon (such as 'employment schedule') were replaced with equivalent concepts in our lexicon (such as 'mission' and 'mission importance').
- Database inconsistencies were resolved in favor of the CMU database. For example, if in the naval database, ship X required capability Y for its mission but in the CMU

database ship, X required mission capability Z, all references in a scenario to Y were replaced with references to Z.

These adaptations have minimal impact on the integrity of the data.

5.2 Reduction in Search Space

Our first experiment was designed to test the search space reduction resulting from applying pragmatic constraints. Thus, we used all 10 of our test scenarios. The scenarios contained an average of 9 sentences.

To measure the constraint imposed by the knowledge sources, we use an index called perplexity. This is an information theoretic measure that is widely used in speech systems to characterize the constraint provided by a grammar. Perplexity represents the geometric mean of the number of alternative words at any point. Search space size for a given test sentence is computed by raising perplexity to the number of words in a sentence.

To measure the reduction in perplexity and search space it was necessary to collect test set perplexity measurements for each of the parsed sentences in two conditions:

- Total domain grammar alone

- Using predictions

Test set perplexity is the perplexity of the actual sentence parsed. It is different than total grammar perplexity because it takes into account only those alternatives which are legal next words given the grammar.

To measure the perplexity of all the sentences in each of the test scenarios using the entire system grammar is relatively straight forward. However, measuring the test set perplexity of sentences which are parsed with layered predictions is not. Since prediction layers fail, we must report the perplexity of the layers which were successful. However, since some layers are non-overlapping, the number we report is the perplexity of the successful prediction layer merged with all the unsuccessful layers attempted.

As seen in Table 1, test set perplexity was reduced in excess of an order of magnitude, from 279.2 to 17.8.

Put differently, the knowledge sources reduced the search space for lexical entries by 9 orders of magnitude on the average 8 word sentence when the predictions were expanded into potential surface expression forms for future utterances.

Table 1:

Reduction in Branching Factor and Search Space		
Constraints used:	grammar	layered predictions
Test Set Perplexity	279.2	17.8
Search Space	3.81×10^{19}	1.01×10^{9}

5.3 Recognition Performance

To evaluate the effects of using layered predictions on recognition performance we used 10 speakers (8 male, 2 female) who had not been used to train the recognizer. Each speaker read 20 sentences from the adapted test set provided by the Navy. Each of these utterances was recorded. The speech recordings were then run through the SPHINX recognition system in two conditions:

- using the system grammar (all legal sentences)

- using the successful prediction layer merged with all unsuccessful layers

The results can be seen in Table 2.

As can be seen, the system performed significantly better with the predictions. Error rate decreased by a factor of five. Perhaps more important, however, is the nature of the errors. In the "with predictions" condition, 89 percent of the insertions and deletions were the word "the". Additionally, 67 percent of the substitutions were "his" for "its". Furthermore, none of the errors in the "with predictions" condition resulted in incorrect database query. Hence, semantic accuracy was 100%.

Table 2:

Recognition Performance		
Constraints used:	grammar	layered predictions
Test Set Perplexity	242.4	18.3
Word Accuracy	82.1	96.5
Semantic Accuracy	85%	100%
Insertions	0.0%	0.5%
Deletions	8.5%	1.6%
Substitutions	9.4%	1.4%

6 SUMMARY

In summary, by identifying and using knowledge sources which can intelligently reduce search space, we progress toward developing robust, interactive problem solving environments where speech is the primary mode of communication. One such knowledge source is pragmatics. The use of layered predictions derived from pragmatic knowledge sources appears to be a powerful technique for improving speech recognition and reducing search space. Layered predictions allow the recognition system to capitalize upon pragmatic knowledge sources without impairing the system's ability to recognize less likely utterances. The more consistent the users' behavior, the better the recognition. As user behavior deviates, recognition accuracy degrades gracefully and the system is capable

of recovering and generating further pragmatic predictions based upon both the users' expected and less expected behavior. However, as domains continue to scale up and we begin to process spontaneously generated speech, additional knowledge sources will become increasingly important.

References

1. J. F. Allen and C. R. Perrault: "Analyzing Intention in Utterances," *Artificial Intelligence 15*, pp.143-178, 1980.

2. A. Biermann, R. Rodman, B. Ballard, T. Betancourt, G. Bilbro, H. Deas, L. Fineman, P. Fink, K. Gilbert, D. Gregory and F. Heidlage: "Interactive natural language problem solving: A pragmatic approach," Conference on Applied Natural Language Processing, pp.180-191, 1983.

3. L. Borghesi and C. Favareto: "Flexible Parsing of Discretely Uttered sentences," COLING-82, Association for Computational Linguistics, Prague, pp.37-48, 1982.

4. J. G. Carbonell: "POLITICS: Automated Ideological Reasoning," *Cognitive Science 2*, pp.27-51, 1978.

5. P. R. Cohen and C. R. Perrault: "Elements of a Plan-Based Theory of Speech Acts, " Cognitive Science, pp.177-212, 1979.

6. L. D. Erman and V. R. Lesser: "The Hearsay-Speech Understanding System: A Tutorial, " In Lea, W.A., Ed., *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, pp.340-360, 1980.

7. R. E. Fikes and N. J. Nilsson: "STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving," *Artificial Intelligence 2*, pp.189-208, 1971.

8. P. K. Fink and A. W. Biermann: "The Correction of Ill-Formed o73 Input Using History-Based Expectation With Applications to Speech Understanding," *Computational Linguistics 12*, pp.13-36, 1986.

9. B. J. Grosz and C. L. Sidner: "Attention, Intentions and the Structure of Discourse," *Computational Linguistics 12*, pp.175-204, 1986.

10. A. G. Hauptmann, S. R. Young and W. H. Ward: "Using Dialog Level Knowledge sources to Improve Speech Recognition," *Proc. 7th NCAI*, 1988.

11. O. Kimball, P. Price, S. Roucos, R. Schwartz, F. Kubala, Y. L. Chow, A. Haas, M. Krasner and J. Makhoul: "Recognition Performance and Grammatical Constraints, " *Proc. DARPA Speech Recognition Workshop*, Science Applications International Corporations Report No. SAIC-86 / 1546, pp.53-59, 1986.

12. W. A. Lea, (Ed.): "*Trends in Speech Recognition*," Prentice-Hall, Englewood Cliffs, NJ, 1980.

13. K. Lee: "*SPHINX: Large Vocabulary, Speaker-Independent Speech Recognition*," Ph. D. Th., Carnegie-Mellon University, 1988.

14. S. E. Levinson and K. L. Shipley: "A Conventional-Mode Airline Information and Reservations System Using Speech Input and Output," *The Bell Systems Technical Journal 59*, pp.119-137, 1980.

15. S. E. Levinson and L. R. Rabiner: "A Task-Orientated Conversational Mode Speech Understanding System," *Bibliotheca Phonetica 12*, pp.149-196, 1985.

16. D. J. Litman and J. F. Allen: "A Plan Recognition Model for Subdialogues in Conversation," *Cognitive Science 11*, pp.163-200, 1987.

17. B. Lowerre and R. Reddy: "The Harpy Speech Understanding System," In Lea, W.A., Ed., *Trends in Speech Recognition*, Prentice-Hall, Englewoo Cliffs, NJ, pp.340-360, 1980.

18. A. Newell and H. A. Simon: "*Human Problem Solving*," New Jersey.: Prentice-Hall, 1972.

19. E. D. Sacerdoti: "Planning in a Hierarchy of Abstraction Spaces," *Artificial Intelligence 5*, pp.115-135, 1974.

20. R. Wilensky: "*Understanding Goal-Based Stories*," Ph. D. Th., Yale University, 1978.

21. R. Wilensky: "*Planning and Understanding*," Addison Wesley, Reading, MA, 1983.

22. S. R. Young, A. G. Hauptmann and W. H. Ward: "An Integrated Speech and Natural Language Dialog System: Using Dialog Knowledge in Speech Recognition," Tech. Rept. CMU-CS-88-128, Carnegie Mellon University Computer Science Technical Report 1988, also submitted.

Chapter 8

SPEECH ENHANCEMENT

This Page Intentionally Left Blank

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Noise Elimination of Speech by Vector Quantization and Neural Networks

Kazuo Nakata and Akihiko Sugiura

Faculty of Technology, Tokyo University of Agriculture and Technology
2-24-16 Nakamachi, Koganei, Tokyo, 184 Japan

Abstract

Noise elimination/reduction or speech enhancement is a very necessary pre-processing technique in every practical application of speech recognition. For this purpose, a method is proposed which combines the data clustering function of vector quantization with the pattern classification function of neural networks.

The method is effective for a noise elimination, for example, in the range of an average SNR from 9 to 2dB with the learning at an SNR of 5dB. The features of the method are:

(1) the use of different effective parameters for clustering and classification, respectively, and

(2) the rejection of non-speech segments both in learning and classification, to avoid confusion. Applications for recognition and coding are discussed. The large-vocabulary case is beyond the scope of the present paper and now under study.

1. OBJECTIVES OF RESEARCH

Noise elimination/reduction or speech enhancement is a very necessary pre-processing technique in every practical application of speech recognition. "Speech enhancement " is defined here as the reduction or elimination of noise directly from speech waveforms, and the reproduction of speech waveforms which sound noise free. "Noise reduction/elimination" means here to derive any parameters effective for speech recognition or synthesis, which are robust against the noise mixed with the speech.

Previous work on noise reduction or speech enhancement is mostly based on iterative adaptations of noise canceling filters using two separate microphones [1 - 3]. Our method studied here is using only one microphone, and a quite different and new one.

2. OUTLINES OF THE METHOD

Our method of noise elimination is designed for the use of one microphone and based on the combination of two functions of new technologies: clustering of continuous input

speech data into a finite number of code vectors by "vector quantization" (VQ) and classification of noisy patterns into clustered categories defined by the VQ by "neural networks" (NN).

Speech data flow continuously in time and consecutive segmental analysis of 20-30 ms. duration is assumed. The noise elimination process is carried out segment by segment. The noise elimination process is prepaired and performed using the following four steps, shown in Figure 1.

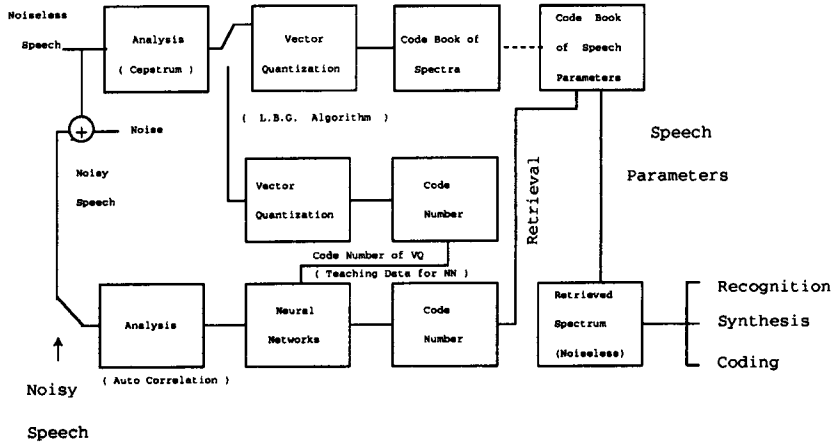


Figure 1. Noise elimination process by vector quantization and neural networks.

(1) Vector quantization of noiseless speech.

   (a) Generation of code vectors by the regular LBG algorithm with speech samples for learning [4, 5].

   (b) Quantization of the speech samples using the codebook vectors, and generation of the desired output for NN training.

(2) Learning/training of the neural networks. For classification (VQ) of noisy speech,neural networks are trained at a typical SNR condition with the back propagation algorithm and the desired outputs [6, 7].

(3) Vector quantizing of noisy speech. The above-mentioned trained neural networks classify the noisy input speech, segment by segment, and output a sequence of code numbers.

(4) Retrieve any desired pre-analyzed speech parameters. The desired speech parameters can be retrieved using the output code number of the NN as an index. The retrieved speech parameters have been pre-extracted from each segment of noiseless speech from which each code vector is derived, respectively.

Then, if the neural networks can classify each noisy segment correctly, the final speech parameters are noise free in quantized accuracy, and noise cannot disturb the results at all.

The point is how well neural networks can classify noisy speech into correct categories, segment by segment, after the learning, and a series of experiments are carried out to study this point.

3. PRELIMINARY STUDIES ON VOWELS CLASSIFICATION BY NEURAL NETWORKS

Preliminary experiments are carried out to specify the configurations of the neural networks, the input parameters, and to confirm the idea of noise elimination by five Japanese vowels classification.

3.1. Configuration of the neural networks

Concerning classification of five Japanese vowels, the number of categories of the patterns to be classified is relatively small, i.e. five.

It is concluded after the experiments that the following simple Perceptron type of two-layer NN, as shown in Figure 2, is good enough for the classification both in noiseless and noisy conditions.

(1) all connections are feed forward and only to the next layer,

(2) the number of neural units in the hidden layer is relatively small, presumably greater than the number of output categories and less than twice that number.

3.2. Input parameters for the VQ and NN

(1) VQ: LPC Cepstrum coefficients.

Several parameters are compared with each other as input for VQ clustering. Our criterion for the choice is simple: we select the parameters which show the best correspondence to the phonemic segmentations which are made manually, by inspections of the various types of analyses.

As shown in Figure 3, the LPC Cepstrum coefficients show the best correspondence and are chosen as input parameters for the vector quantization.

(2) NN: Quasi-normalized auto-correlations.

In all experiments,random noise is assumed and generated in a computer and added to digitized speech data with controlled amplitude in order to get a given average SNR.

LPC based parameters cannot be used at a very low SNR as inputs for noisy speech classification because they are seriously distorted by the existance of heavy noise. Input parameters for noisy speech classification must themselves be as robust as possible. Concerning auto-correlation,as far as the noise is independent of the signal and additive, the effect of noise is simply a linear addition of ther auto-correlation of the noise itself to that of the signal [8].

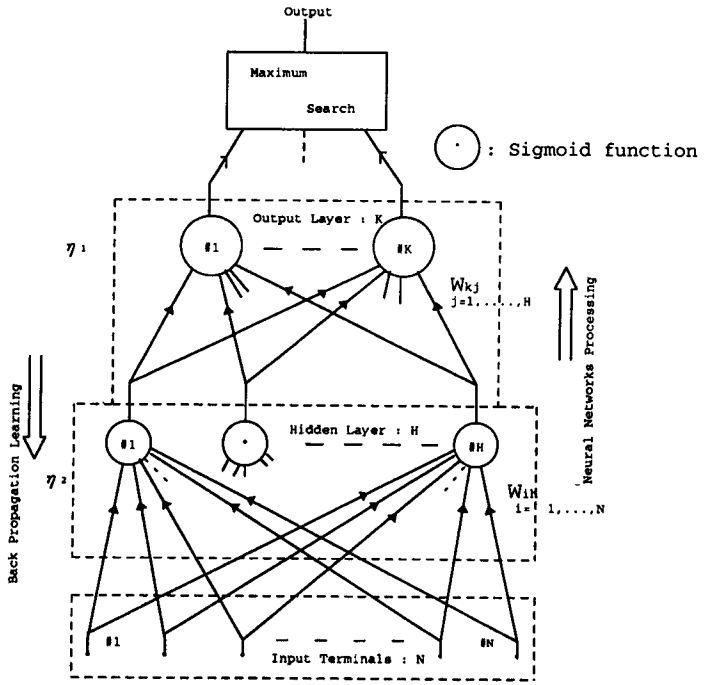


Figure 2. Configuration of two-layer Perceptron-type neural networks.

Furthermore, auto-correlation of random noise is mostly concentrated in the first, zero-time delay term, as shown in Figure 4. Quasi-normalized auto-correlation coefficients are used. Quasi-normalization means the normalization using the maximum value in the range of consideration excluding the first (zero-time delay) term.

3.3. Classification and noise elimination

Errorless classification by neural networks is possible for a single speaker's five Japanese vowels under noiseless conditions.

Concerning the classification of noisy vowels,the range of good performance is rather limited around the learning condition, but the results are far better than those of the usual classification using pattern matching by the Euclidean distance. Therefore, our idea of noise elimination is confirmed in principle.

4. MAIN EXPERIMENTS AND THEIR RESULTS

4.1. Conditions of the experiments

(1) Digital data: 15kHz sampling and 12 bit linear A/D conversion

(2) Analysis: 20ms (300 samples) duration and 10 ms (150 samples) shift.

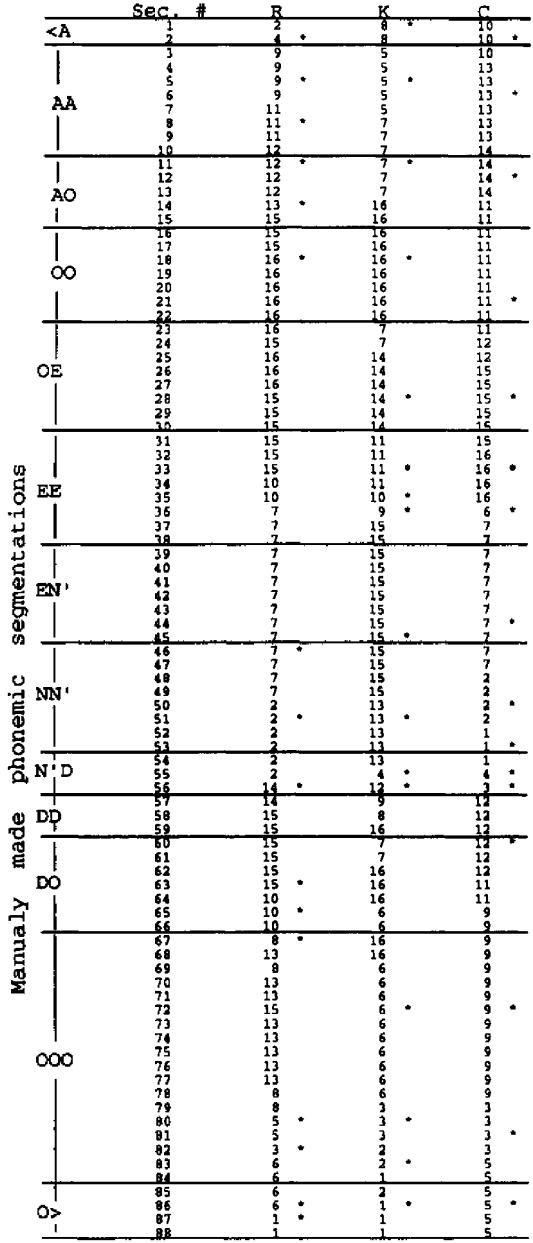


Figure 3. The correspondences between code vectors and manual phonemic segmentations of the word "AOENDO" by the vector quantization of various parameters. Marked * are code vector segments. R: auto-correlation, K: PARCOR, and C: LPC Cepstrum.

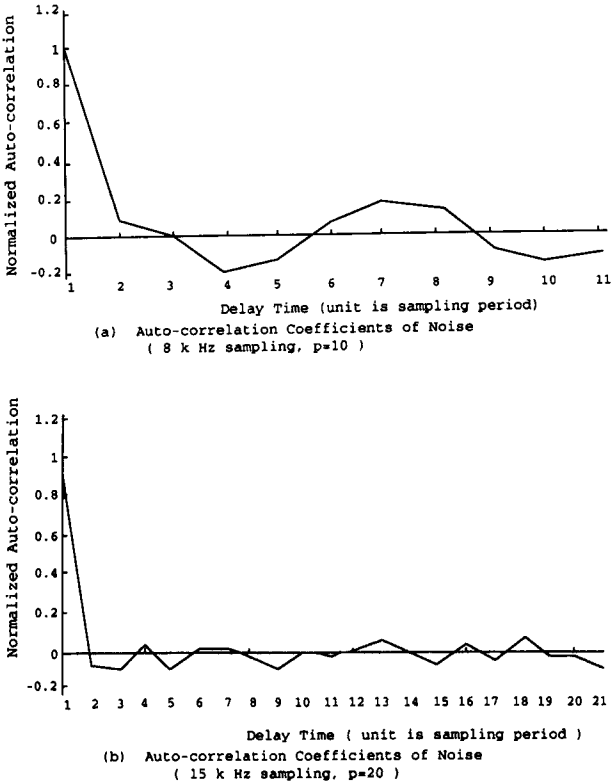


Figure 4. Examples of the auto-correlation of random noise.

(3) Vector Quantization: Inputs are the LPC Cepstrum coefficients of the 1st to 19th order.

(4) Neural Networks: Inputs are quasi-normalized auto-correlations of the 1st to 19th order.

4.2. The small vocabulary case

(1) Subject: The word "AOENDO" (green peas) spoken by a male speaker. The number of analyzed frames is 88, and the number of code vectors is 16.

(2) Neural networks: Two-layer Perceptron with 20 inputs and 16 outputs. The number of neural units in the hidden layer is 18.

(3) Noiseless speech: Perfect classification is obtained after enough learning.

4.3. Classification of noisy speech by trained NNs

(1) Learning: NNs have been trained at an average SNR of 5dB in 5000 to 10000 times.

(2) Testing speech: The average SNR ranged from 10 to 1dB with 1dB steps.

(3) The results: The rate of correct classification into VQed categories reaches higher than 98% on the average, in the range of SNR of 9 to 2dB,as shown in Figure 5.

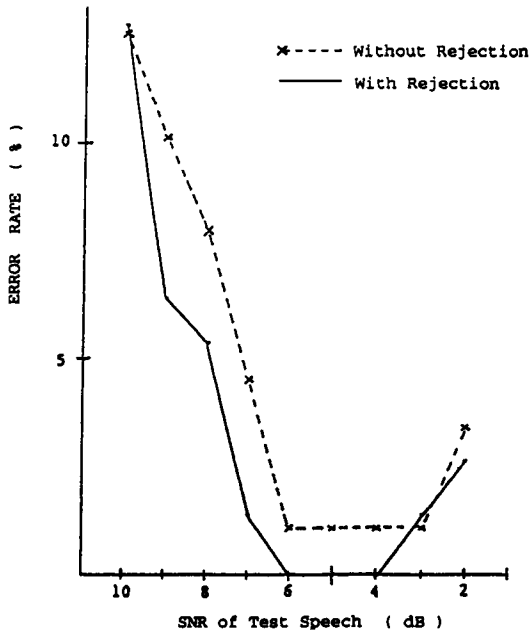


Figure 5. Error rate and its reduction by the rejection of non-speech segments in the neural networks classification. Error rate = No. of error frame / No. of speech frame.

First, a speech/non-speech decision is made in each segment, and non-speech segments are rejected in learning and classification. This process can reduce errors in classification, as shown in Figure 5, due to the elimination of confusing learning of a non-speech segment as a speech segment.

4.4. Results

Noise is adequately eliminated in a relatively wide range at low SNR, i.e. below 10dB, around the learning condition. A typical example of the original noiseless speech,noisy speech and noise eliminated speech is shown in Figure 6. This is only for the purpose of demonstration of the noise elimination,the residuals of the original noiseless speech are used as input for noise eliminated, synthesized speech.

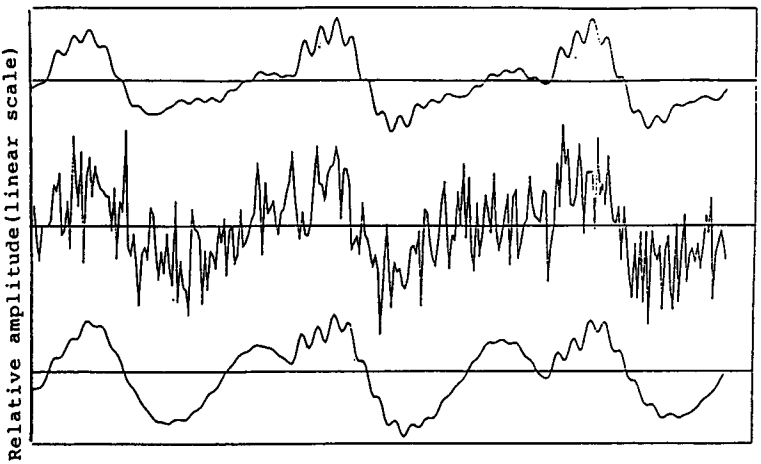


Figure 6. A typical example of noisy speech and noise eliminated speech at SNR = 2dB.
Top: original noiseless speech.
Middle: noisy speech with average SNR = 2dB.
Bottom: noise eliminated speech (synthesized).

4.5. The large vocabulary case

(1) Vocabulary: Subject of the analysis are several decades of words of one male speaker. The number of analyzed segments is about 1000 in total, and the number of code vectors is 128.

(2) Neural Networks: Two-layer Perceptron with 20 inputs and 128 outputs. The number of units in the hidden layer is 128 or 256.

(3) Inputs: The parameters used in VQ clustering and NN classification are the same as in the previous small-vocabulary case.

(4) Results: No perfect classification can be obtained even in the noiseless case after enough learning. A three-layer NN or a multi-staged of two-layer NN are now under consideration and a series of learning and classification experiments are in progress.

5. CONCEIVABLE APPLICATIONS

The following are examples of conceivable applications of the noise elimination process described above.

5.1. Recognition — Input to an HMM

The direct application of the noise elimination process for speech recognition is shown in Figure 7. The application is direct and the concept is easy to under- stand. But experiments are not carried out yet.

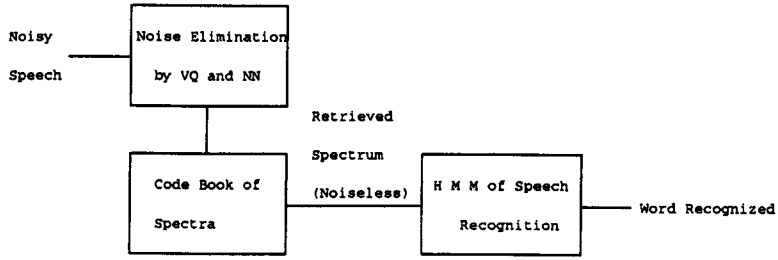


Figure 7. Application of the noise elimination process to speech recognition

5.2. Coding — Coding of noisy input speech

An other interesting application of noise elimination is, as shown in Figure 8, coding of noisy speech. The noise elimination process can extract noiseless spectrum data from noisy speech in quantized form by retrieving of the LPC Cepstrum, PARCOR, LPS or any desired parameters.

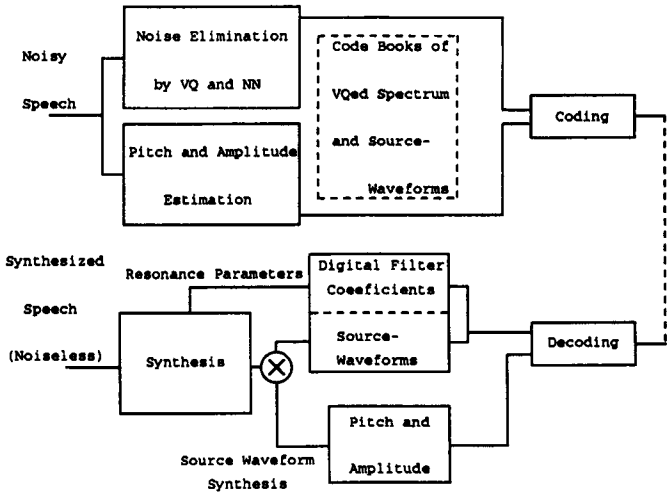


Figure 8. Application of the noise elimination process to noisy speech coding.

Then, if the data of the excitation source can be extracted from noisy speech properly, noiseless speech could be reproduced from noisy speech by synthesis.

One possibility of the method is as follows: waveforms of the excitation source can be described by three data; the unit waveform of the excitation source, the pitch period or exciting timing and the source amplitude. As for the first one, the unit waveform, a pulse (voiced) or a random noise (unvoiced) is conventionally used, and this is one of the main reasons why the synthesized speech does not sound natural or human-like.

One possible alternative for the unit waveform of excitation is extracted from residuals of each segment from which each code vector is derived, that is,to pick out a typical pitch

period of the residuals derived from the speech of that segment, and they are written in the codebook of the excitation waveform with the same code number of the spectrum.

The vector quantization of the spectrum quantizes not only the spectrum but also the unit waveform of excitation. Then the difficult problems remained are how to estimate the excitation timing and the source amplitude from the noisy speech. One feasible solution is to take a simple running average of the input noisy speech and detect the position of its peaks and measure the amplitude of each peak. The method is simple but very effective for noisy speech. Needless to say, some detailed adjustments or modifications are necessary in the practical use of the method.

An experimental example is shown in Figure 9, and the overall synthesized (simulation of coded-and-decoded) speech is shown in Figure 10, with the original noiseless speech. As expected from the figures, the synthesized speech sounds very natural and noise free. This type of application is quite new and also interesting from a practical point of view.

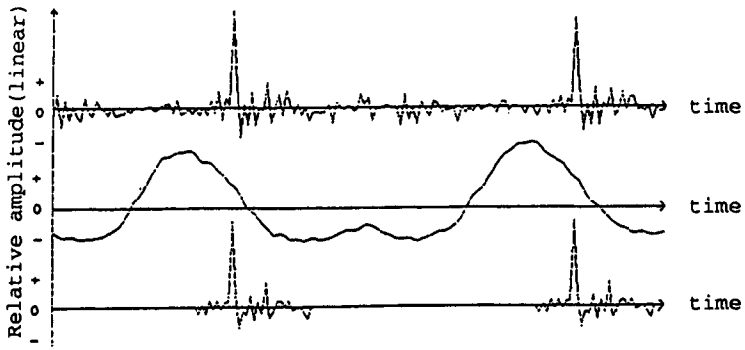


Figure 9. An example of synthesized excitation waveforms.
Top: residuals of the original speech.
Middle: running average of noisy speech.
Bottom: synthesized excitation waveforms.

6. CONCLUSIONS AND FURTHER PROBLEMS

Random noise mixed with speech is adequately eliminated by the combined use of vector quantization for data clustering and neural networks for pattern classification. The typical range of noise elimination is a SNR of 9dB to 2dB when learning at 5dB. The method can be applied not only for speech recognition but also for noisy speech coding. A problem remained unsolved the expansion of these results from the small vocabulary case to the large-vocabulary case and this is now under study.

7. Acknowledgements

The research work was supported by a Grant-in-Aid for Scientific Research on Priority Area, Advanced Man-Machine Interface Through Spoken Language, Ministery of Education, Science and Culture, 1987-1989.

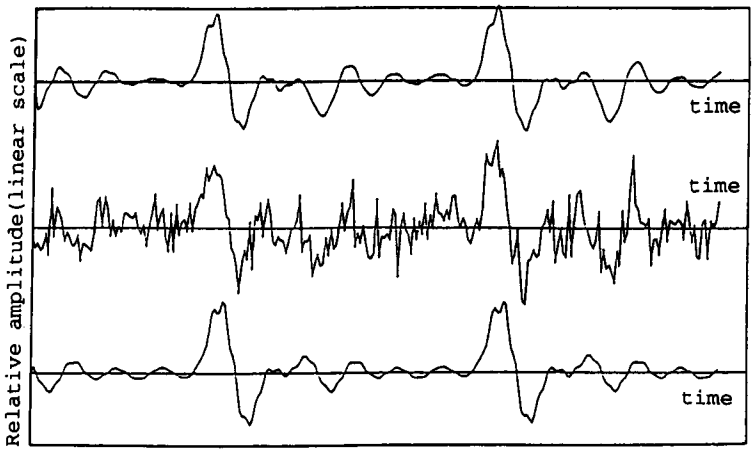


Figure 10. An example of noisy speech and its coded-decoded synthesized speech at SNR = 2dB.
Top: original noiseless speech.
Middle: noisy speech with SNR = 2dB
Bottom: synthesized (coded-decoded) speech.

References

1. R. L. Zinser Jr. and J. B. Evans, "Some experimrntal and theoretical results using a new adaptive filter," Proc.ICASSP-85, pp.1253-1256.

2. W. A. Harison, J. B. Lim and E. Singer, "A new application of adaptive noise cancellation," Trans. IEEE-ASSP-34, No.1, pp.21- 27, (1986)

3. J. J. Rodoriguez, J. S. Lim and E. Singer, "Adaptive noise reduction in aircraft communication system," Proc. ICASSP-87, pp.169-172, (1987)

4. Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantization design," Trans. IEEE, COM-8, pp.84-95, (1980)

5. R. M. Gray, "Vector Quantization," IEEE ASSP Magz., (1985)

6. R. P. Lippman, "An introduction to computing with neural nets," IEEE ASSP Magz., (1987)

7. D. E. Lumerhar, G. E. Hinton and R. J. Wiliams, "Learning internal representations by error propagation, Parallel Distributed Processing," MIT Press., (1986)

8. D. Mansour and B. H. Juang, "The short-time modified coherence representation and noisy speech recognition," Trans. IEEE-ASSP-37, No.4, pp.796-804, (1989)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Speech/Nonspeech Discrimination Under Nonstationary Noise Environments

Hidefumi Kobatake and Akira Ishida

Faculty of Technology, Tokyo University of Agriculture & Technology
2-4-16 Nakamachi, Koganei, Tokyo, 184 Japan

Abstract

The progress in the speech processing technique makes it more and more complicated and it tends to lose robustness against noise distortion. A speech recognition system is usually used under noisy environments and effective noise processing techniques are desired to be developed. This paper presents a study on speech/nonspeech discrimination under real life noise environments. This paper propose several acoustic parameters as features effective for speech/nonspeech discrimination. Experiments to test the performance of the speech/nonspeech discrimination system which are based on the proposed feature parameters are also discussed.

1. INTRODUCTION

The speech recognition technique has made rapid progress and it is now put to practical use. However, the performance of speech recognition systems degrades severely by additive noise. This is because the processing in the speech recognition system becomes more and more complicated with the progress in speech technology and, as a result, its robustness against noise distortions tends to be lost. It should be robust against background noise, and, for this purpose, noise processing techniques are desired to be developed. This paper presents an important part of a noise processing system. To keep the performance of speech recognition systems high in a real life noise environment, precise detection of speech segments and speech enhancement (or restoration of the speech spectrum) are very important. This paper deals with the former problem. Considerations on feature parameters for speech/nonspeech discrimination are given. Experiments to test the performance of the proposed speech/nonspeech discrimination system are also demonstrated.

2. NOISE PROCESSING SYSTEM

In many cases, background noise in a read life environment can be modeled as a sum of stationary continuous noise and isolated noises such as door clicks, footsteps, barking, and so forth. Speech is also nonstationary and usually isolated. If we assume that there

is no overlapping between speech and isolated noise, an observed signal $x(n)$ is simply modeled as the sum of a stationary background noise $d(n)$ and an isolated signal $s(n)$, which may be isolated speech or isolated noise. That is,

$$x(n) = s(n) + d(n). \tag{1}$$

With thie assumption, the signal $s(n)$ can be identified as a nonstationary region in the stationary background noise $d(n)$. Therefore, the noise processing problem becomes to detect nonstationary regions in $x(n)$ and to restore speech signal by reducing the disturbances by the background noise if the detected region contains speech signal. Figure 1 shows the block diagram of the noise processing system under development. Its processing steps consist of three parts. The first processing step is the precise segmentation of the nonstationary signal. A new segmentation method has been developed by which nonstationary segments can be detected precisely. The next step is the speech/nonspeech discrimination for the nonstationary segment detected by the first processing step. If the segment includes speech, the third processing step begins to work. It is the speech enhancement stage, and stationary noise $d(n)$ is removed to restore the speech waveform. This paper treats only the second processing step.

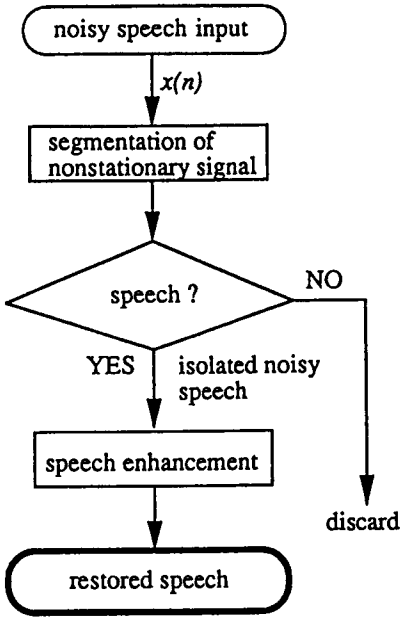


Figure 1. Noise processing system.

3. FEATURE PARAMETERS FOR SPEECH/NONSPEECH DISCRIMINATION

A detected nonstationary segment is disturbed by a stationary background noise. The degree of disturbance on the nonstationary signal characteristics depends on the relative power levels of the two signals. It is natural that the acoustic characteristics of the nonstationary signal are well preserved if the signal-to-noise ratio (SNR) is high. A vowel has generally much a larger power than any consonant. This means that characteristics of vowels are preserved much better than those of consonants. It is hard to specify acoustic characteristics of all sorts of isolated noises in a real life environment. Therefore, we can say that whether the nonstationary signal has vowel-like characteristics or not gives the primary clue to the speech/nonspeech discrimination. The speech/nonspeech discrimination subsystem processes a subregion of the detected nonstationary segment where the power level of $x(n)$ is above a threshold L. The threshold level is determined so that the power level of most vowel segments exceeds it.

The fundamental feature parameters for speech/nonspeech discrimination adopted in our system are:

(1) periodicity

(2) pitch frequency (f_0),

(3) optimum order of the prediction model (p_0),

(4) Q of the first formant (Q_1) and

(5) minimum LPC cepstrum distance between the 5 vowels (d_{min}).

These feature parameters reflect the similarity to vowel characteristics. Kamiya and Tanaka have proposed the sum of the absolute values of the Cepstrum coefficients as a feature parameter for speech/nonspeech discrimination. In our system, not only the spectral information but also other acoustic features are adopted, so that the system works well under various noise environments.

The speech/nonspeech decision is made in two steps. At the first step, a frame by frame decision is performed. If the feature parameters extracted from an analysis frame satisfy the following conditions at the same time:

(1) periodicity,

(2) $f_L < f_0 < f_H$,

(3) $p_L < p_0 < P_H$,

(4) $Q_L < Q_1$, and

(5) $d_{min} < d_H$, the subsystem assigns the frame to the category speech. That is

if (1) $\cap$(2) $\cap$ (3) $\cap$ (4) $\cap$ (5) is true, then the analysis frame is of speech
The following items are examined for overall decision:

(6) smoothness of the change of pitch frequency,

(7) smoothness of the change of the optimum order of the linear prediction model, and

(8) percentage of the frames assigned to the category of speech (r).

If the changes of f_0 and p_0 are both smooth and the percentage r is lager than a predetermined threshold, the whole of the detected nonstationary segment is classified as speech. The smoothness measures of f_0 and p_0 are given as follows.

$$S_f = \frac{\sum_{n=n1}^{n_2} df_0(n)}{n2 - n1 + 1},$$

and

$$S_p = \frac{\sum_{n=n1}^{n_2} dp_0(n)}{n2 - n1 + 1},$$

where,

$$df_0(n) = |f_0(n) - f_0(n-1)|,$$

and

$$dp_0(n) = |p_0(n) - p_0(n-1)|.$$

4. EXPERIMENTS

4.1. Experimental Conditions

The purpose of this paper is to show the effectiveness of the adopted acoustic parameters for speech/nonspeech discrimination. Nonspeech signals used in the experiments include various office room noises, computer room noises, sounds of musical instruments, barking, chirping of insects, and so on. Total number of nonspeech signal is 120. Test signal is sampled at 10kHz with 12bits accuracy. The length of the analysis frame and its interval are 25.6ms and 12.8ms, respectively. The autocorrelation function of the prediction residual is used for periodicity judgment. The threshold level for the decision is 0.25. The pitch frequency is obtained for every analysis frame whether the maximum peak value of its autocorrelation functions greater than the threshold or not. The lower and the higher boundaries of the pitch frequency of typical speech are assumed to be 80 and 350Hz, respectively. The optimum order of the prediction model is defined in the experiments as the order which gives the minimum for the AIC criterion. From a preliminary experiment, it was shown that the optimum order of the linear prediction model is typically 12 and it lies usually between 8 and 15, which are adopted as the lower and the higher boundaries of the optimum order of the speech model. The Q of the first formant is obtained from the pole of the linear prediction model.

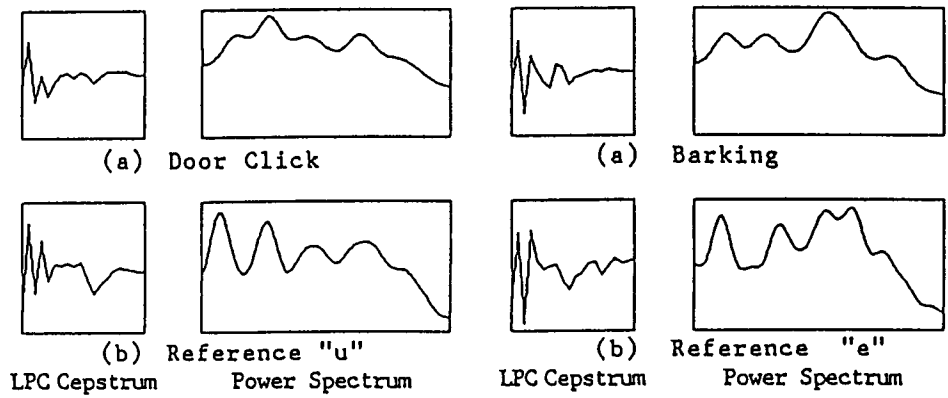


Figure 2. Comparison of spectral patterns between speech and nonspeech. Left: door click and /u/, right: barking and /e/.

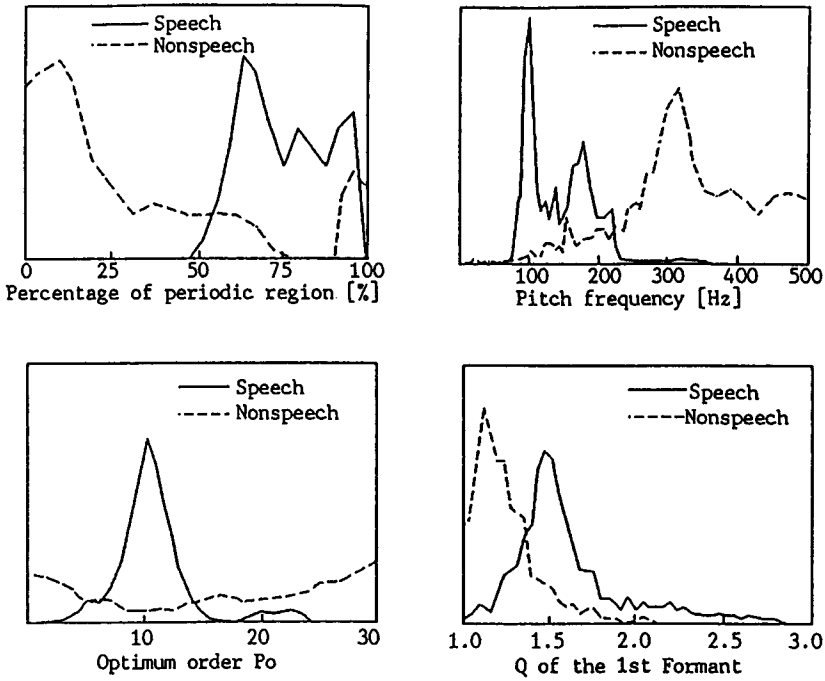


Figure 3. Distributions of each feature parameter. Upper left: percentage r, upper right: pitch frequency, lower left: optimum order, lower right: Q.

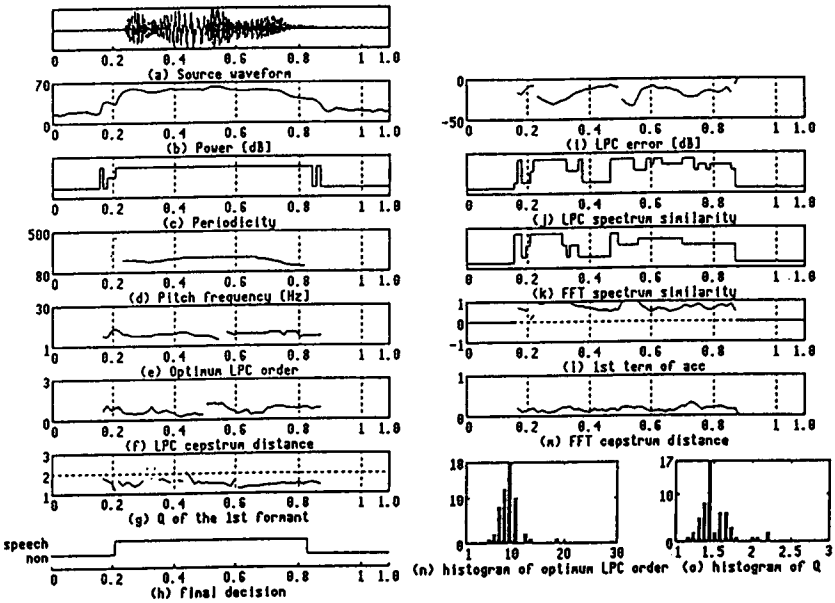


Figure 4. An example of system output for female speech /koganei/.

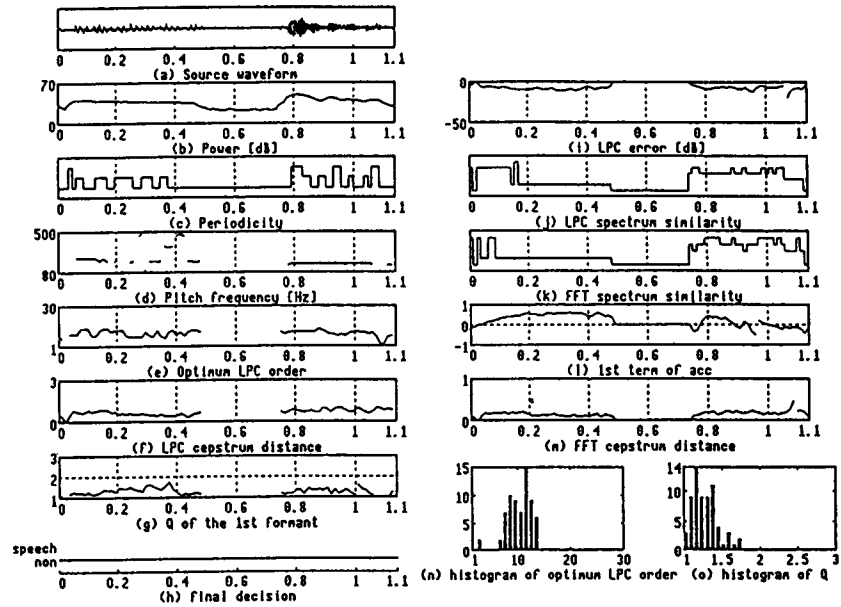


Figure 5. An example of system output for musical sound of a guitar.

4.2. Distribution of each feature parameter

There are many kinds of noises in real life environments whose instantaneous spectra are similar to those of vowels. Two examples are shown in Figure 2. A door click and barking show vowel-like spectra in their transient states. The feature parameters (a) - (d) are selected so that they make up for the defectiveness of the LPC Cepstrum distance measure. Distributions of the adopted feature parameters are shown in Figure 3. Solid lines and dotted ones correspond to speech and nonspeech, respectively. These figures show that speech and nonspeech are well separated to each other by these feature parameters. Musical sound is periodic and its pitch frequency is usually in the pitch frequency range of typical speech. And, moreover, the spectral patterns of some musical instruments are very similar to those of speech. However, they are identified correctly by evaluating the Q of the first formant. The Q of music sound is usually smaller than that of speech. The prameter value Q_L is experimentally determined as 1.4.

4.3. Speech/nonspeech discrimination

Figures 4 and 5 show examples of outputs of the proposed system. Figure 4 shows an example for female speech /koganei/. In this example, every feature parameter satisfies the condition of speech. On the other hand, Figure 5 shows an example of sound of a guitar. From the figure, it is known that the LPC Cepstrum distance remains small throughout the signal segment and the pitch frequency and the optimum order of the prediction model also remain in the reasonable range of speech. We can see, however, that the distribution of the Q of the first formant is different from that of speech. Mean value of the Q is apparently smaller than that of speech, by which it was classified into the category of nonspeech.

Table 1 shows the result of speech/nonspeech discrimination. The results shown in the upper half of Table 1 are correct classification rates when one feature parameter assigned by a small circle was used. These results show that almost all speech is classified correctly and about half of nonspeech signals are categorized into speech. However, the adopted feature prameters are shown to be compensative to each other. Correct classification rates when four or five feature prameters were combined are shown in the lower part of Table 1.

From experimental results, we can say;

1) Male voice is easy to recognize as speech.

2) On the contrary, female voice is difficult to classify into speech. It is because the distributions of the proposed feature prameters of the female voice are wider than those of male voice. Especially, the optimum order of the linear prediction model changes its value widely in the case of female speech.

4.4. Speech/nonspeech discrimination under stationary noise

The background stationary noise is not considered in the experiments described in the section 4.3. However, such background noise must be considered in a real life environment.

Table 1. Correct classification rate in percent. Small circle (o) show the adopted feature parameters for speech/nonspeech discrimination.

Adopted feature parameters					Correct classification rate [%]			
(a)	(b)	(c)	(d)	(e)	Male speech	Female speech	Non-speech	Mean
o					100	95	57	76
	o				100	95	43	67
		o			100	98	44	69
			o		100	98	41	67
				o	100	95	63	79
o	o	o	o		100	88	88	91
o	o	o		o	100	88	90	92
o	o		o	o	100	90	91	93
o		o	o	o	100	88	91	93
	o	o	o	o	100	90	87	91
o	o	o	o	o	100	85	94	94

(a) percentage of periodic region, (b) pitch frequency, (c) optimum order of the linear prediction model, (d) LPC cepstrum distance, (e) Q of the first formant.

Table 2. Correct classification rates when speech and nonspeech are disturbed by white noise.

S / N ratio	Classification rate [%]		
Signal	10dB	5dB	0dB
Male speech	100	100	88
female speech	84	80	71
nonspeech	94	94	96
Total	94	92	87

Experiments to test the usefulness of the proposed features under stationary background noise contamination were performed. A white noise was used as a contaminating noise and feature parameters were extracted from the non-stationary signal disturbed by the white noise. Table 2 shows the results. The experiments show that the proposed system keeps its performance well if the signal-to-noise ratio is equal to or above 5dB. It means that the proposed features are robust against noise disturbances.

5. CONCLUSION

Feature parameters for discrimination between speech and nonspeech signal have been studied. Periodicity, pitch frequency, optimum order of the linear prediction model and the Q of the first formant have been investigated as acoustic feature parameters effective for the discrimination of speech/nonspeech.

Experiments to test the performance of the system have been made, whose results show that the proposed system works well and especially male speech is classified correctly. However, classification of female speech is not accurate. The reason has been given.

References

1. H. Takahashi, Y. Matumoto and H. Kobatake, "Studies on noisy word recognition," Systems and Computers in Japan, 17, 1-7, (1986)

2. K. Nakata, "Advanced techniques for speech processing in the presence of noise and/or interference," Preprints of the First Symp. on Advanced Man-Machine Interface through Spoken Language, 63-67, (1988)

3. J. S. Lim ed., "Speech enhancement," Prentice-Hall Inc., (1986)

4. J. H. Hansen and M. A. Clements, "Iterative speech enhancement with spectral constraints," Proc. ICASSP, 189-192, (1987)

5. H. Kobatake and K. Tawa, "Precise detection of isolated words in noise by linear prediction method," Proc. Jpn-U. S. Joint Meeting on Acoustics, (1988)

6. H. Kobatake and A. Ishida, "Speech/nonspeech discrimination for speech recognition system under real life noise environments," Proc.ICASSP, pp.365-368, (1989)

7. S. Kamiya and A. Tanaka, "Voicepass filter," Preprints of the Autumn Meeting of Acoust. Soc. Jpn., 1-4-1, (1985)

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Spatially Selective Multi-Microphone System

Hikaru Date and Tomio Watanabe

Department of Information Engineering, Yamagata University, Yonezawa, 992 Japan

1. INTRODUCTION

The present paper introduces "spatial selectivity", a new concept in sound reception, the principle behind its realization, and several results of computer simulation which manifest the foundation of the design.

Spatial selectivity of a microphone system is defined as a function which rejects all sound signals coming from the external region specified by the system.

Its practical applications include the suppression of howling in public address systems and conference telephony, the improvement of the signal-to-noise ratio of microphone output to automatic speech recognition machines, and the improvement in separability of the signal picked-up from a particular musical instrument from that of other instruments played simultaneously in the same studio.

The reason is quite simple; most unwanted signals such as noise, reflected sound, radiated sound from loudspeakers in public address systems, etc., come from distant sources or images, therefore these can be effectively eliminated by a microphone system with appropriate spatial selectivity.

2. PRINCIPLE

The principle of spatial selectivity is a direct consequence of the representation of the sound field by an integral equation. Figure 1 shows a closed region with its boundary surface and related vectors. The sound pressure, $\psi(\vec{r},t)$, is given by the well-known Kirchhoff equation:

$$\begin{aligned}\psi(\vec{r},t) &= \int_{-\infty}^{+\infty} dt_0 \iiint_{\Omega} dv_0 \left[\frac{1}{R}\delta(R/c-(t-t_0))\rho(\vec{r}_0,t_0)\right] \\ &\quad +\frac{1}{4\pi}\oint_{\Gamma} dS_0 \left[\frac{1}{R}\frac{\partial\psi}{\partial n_0}-\frac{(\vec{R}\cdot\vec{n_0})}{R^3}\psi-\frac{(\vec{R}\cdot\vec{n_0})}{R^2 c}\frac{\partial\psi}{\partial t_0}\right]_{t_0=t-R/c}\end{aligned} \tag{1}$$

where $\vec{r}$ in the Green's function is an observation point inside the closed region, $\vec{r}_0$ in the Green's function refers a source point, $\vec{R}=\vec{r}-\vec{r}_0$, and $R=|\vec{R}|$.

Transposing the left side term to the right side and the volume integral term on the right side to the left side, we get:

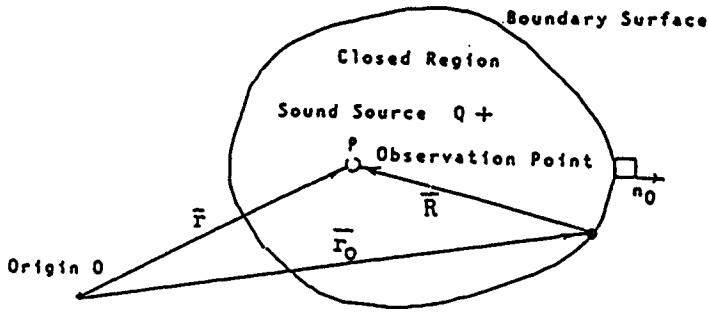


Figure 1. Vectors, boundary surface,and closed region for integral representation of the sound field.

$$\begin{aligned} &\int_{-\infty}^{+\infty} dt_0 \iiint_{\Omega} dv_0 \left[\frac{1}{R}\delta(R/c-(t-t_0))\rho(\vec{r}_0,t_0)\right] \\ &= \psi(\vec{r},t) - \frac{1}{4\pi}\oint_{\Gamma} dS_0 \left[\frac{1}{R}\frac{\partial\psi}{\partial n_0} - \frac{(\vec{R}\cdot\vec{n}_0)}{R^3}\psi - \frac{(\vec{R}\cdot\vec{n}_0)}{R^2 c}\frac{\partial\psi}{\partial t_0}\right]_{t_0=t-R/c} \end{aligned} \tag{2}$$

The value of the left side of this equation is zero if there is no source in the closed region. Therefore, we can receive sound signals coming from sources inside the closed region only, if we measure the sound pressure at an arbitrary point $\vec{r}$ inside the closed region, and at the same time measure the pressure, the pressure gradient normal to the boundary surface and the time derivative of the pressure at the boundary by appropriate means, and then process these signals according to Eq.(2). Other signals coming from the outer region of the boundary surface are eliminated at the output of the system. In other words, we can get a spatially selective sound receiving system.

2.1. Discrete Arrangement of Microphones

In order to realize such a system, we use a spherical surface as the boundary and divide it ideally into M domains of the same shape and with the same area. And we let $\vec{r}$ be the center, 0, of the sphere. Then, the right side of Eq.(2) can be approximated by $\alpha(r_q,t)$, defined as

$$\alpha(r_q,t) = \psi(o,t) - (1/M)\sum_{i=1}^{M}[a(\partial\psi_i/\partial n_{oi}) + \psi_i + (a/c)(\partial\psi_i/\partial t_0)]_{t_0=t-a/c}, \tag{3}$$

where a is the radius of the sphere and c is the sound velocity.

Figure 2 shows as example the microphone arrangement when a regular dodecahedron is used for equal division of the sphere: boundary microphone pairs C_i or D_i, where $i = 0$ to 5, are located at the center of each regular pentagon on the boundary in order to pick

up the sound pressure and the pressure gradient normal to the boundary: a microphone 0 is placed at the center of the regular dodecahedron.

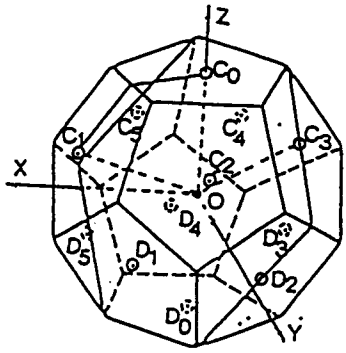


Figure 2. Location of microphones for the case of a dodecahedron.

2.2. Signal Processing

Figure 3 shows a block diagram of the signal processing system, whose function is the same as Eq.(3). It consists of adders, constant multipliers, differentiators, and a single delay. Therefore, this block diagram can be easily realized in either analog or digital form, according to the precision required.

3. SIMULATION

3.1. Relative Gain

Figure 4 shows the center and boundary of the sphere, the sound source Q, and related vectors used in the simulation. The output of the system for a spherical wave incident from a sound source which is at finite distance r_q can generally be expressed as

$$\alpha(r_q, t) = (1/r_q) \cdot \exp\{j(\omega t - kr_q)\} \cdot \{RelativeGain\} \tag{4}$$

This means that the "Relative Gain" is the system output normalized by the output of the center microphone proportional to the sound pressure and it can serve as a suitable criterion for representing the spatial characteristics of the system.

3.2. Spatial Cut-off Characteristics

Figure 5 shows the simulation results for the case of the dodecahedron-type division of the sphere, that is, $M = 12$. The ordinate is the relative gain and the abscissa is the normalized distance of the sound source, KQ, i.e., the distance r_q multiplied by the

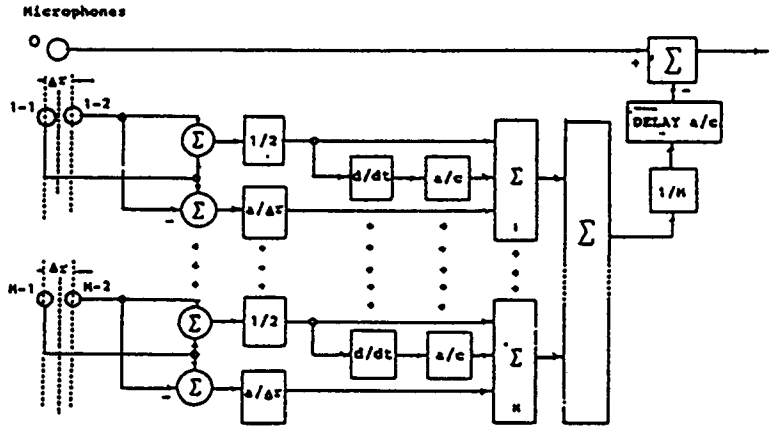


Figure 3. Block diagram for signal processing in the multi-microphone system with range-dependent sensitivity.

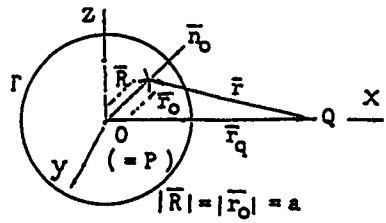


Figure 4. Relationship of vectors for analyzing the range-dependent sensitivity.

wave number. The parameter K is the normalized radius of the sphere, that is, the sphere radius a multiplied by the wave number. Very sharp spatial cut off characteristics are obtained. Because the cut off position of the so-called "passband" lies on the points $KQ = K$, the cut off positions coincide with the sphere boundary. The slope of the cut off characteristics is about 35dB per octave distance, which is sharp enough for eliminating external noise in most cases. It was also found, in another simulation, that the value of the slope increases as M increases, for example, 11.5dB per octave distance for $M = 4$, and 37.5dB per octave distance for $M = 20$. These two properties of the cut off position and slope are essentially independent of frequency.

On the other hand, however, the minimum value of the so called "stopband", which is obtained when KQ approaches infinity, increases as K increases, that is, as the frequency or the radius increases. This means that the spatial selectivity is more significant at lower frequencies than higher frequencies. This property is very convenient for suppression of acoustic noise, which usually has a greater power spectrum in the lower frequency band than in the higher frequency band.

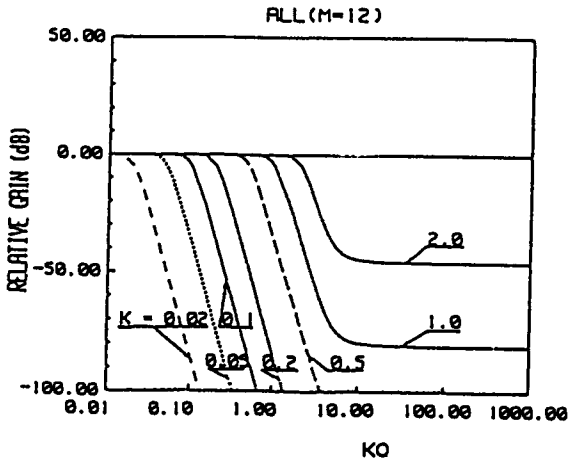


Figure 5. Spatial cut off characteristics of the system with twelve boundary microphone pairs.

3.3. Response for Plane Wave Incidence

Figure 6 shows the relative gain as a function of the normalized radius K for all types of regular polyhedrons when a plane wave is incident, that is, when r_q is equal to infinity. This corresponds to the relative gain of the stopband in the last figure. The more M increases, the more attenuation is obtained. Furthermore, we find these curves can be divided into three groups, i.e., the first is $M = 4$, the second is $M = 6$ and 8, and the third is $M = 12$ and 20. This is due to similarities of the geometrical properties within each group and is interesting for the design of this multi-microphone system.

3.4. Modification to Close-talking Microphones

We can give a directional property to the spatial selectivity of the system if we let the sensitivity of one or several boundary microphones increase or decrease. Figure 7 shows an example of this property. In this case, one boundary microphone pair is removed. The upper curve corresponds to the direction of the removed microphone pair and the lower curve to the other directions. This suggests one possibility for realizing a new type of close-talking microphone with high suppression characteristics for noise incident from distant sources.

Figure 8 shows equicontours for the spatial distribution of the relative gain when one of the twenty boundary microphone pairs is removed, that is, with $M = 19$. A gradual change in the relative gain near the multi-microphone system promises us the realization of close-talking microphones which are easy to use.

Figure 9 shows an example of the directional pattern of the relative gain, where KQ/K, that is the distance from the microphone system normalized by the sphere radius, is the

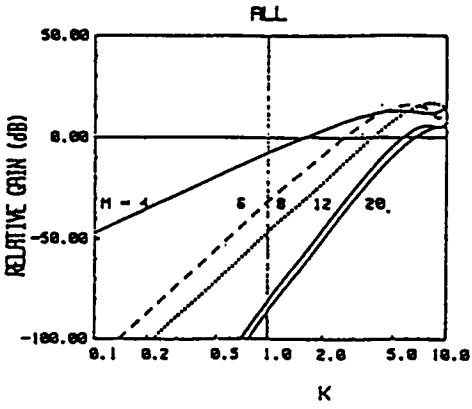


Figure 6. Relative gain for plane wave incidence as a function of sphere radius multiplied by wave number.

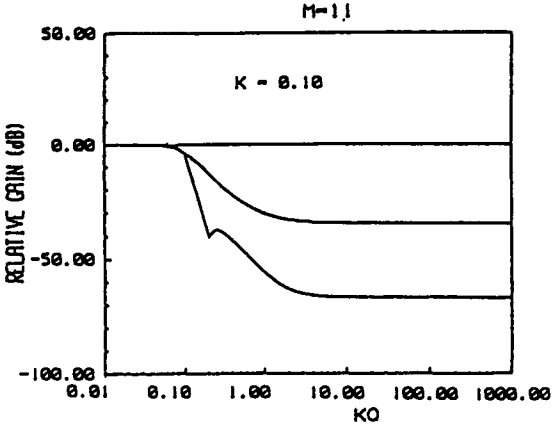


Figure 7. Range-dependent sensitivity characteristics when one boundary microphone pair is removed.

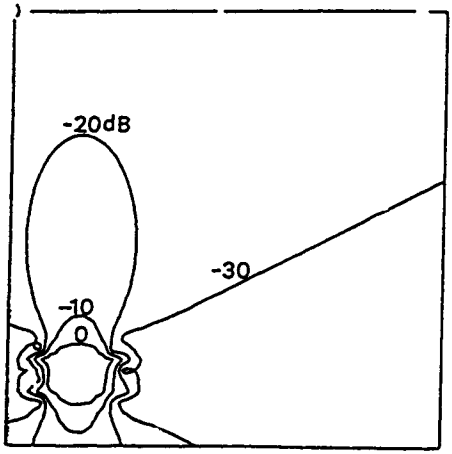


Figure 8. Equicontours of the relative gain spatial distribution for $M = 19$.

parameter. A clear dependency of the relative gain on both the normalized distance and the direction of the sound source can be seen.

Figure 10 illustrates the maximum relative gain in the directional patterns with various normalized distances as a function of K. Obviously, these values correspond to the excess values of the howling margin, compared with a non-directional microphone located at the center of the multi-microphone system when the other conditions are kept equal. We can expect a very high howling margin value when the distance is over ten times greater than the sphere radius.

4. Conclusion

The concept of spatial selectivity, the principle behind its realization, and the results of computer simulation have been discussed. The possibility of developing a new type of close talking microphone was also discussed.

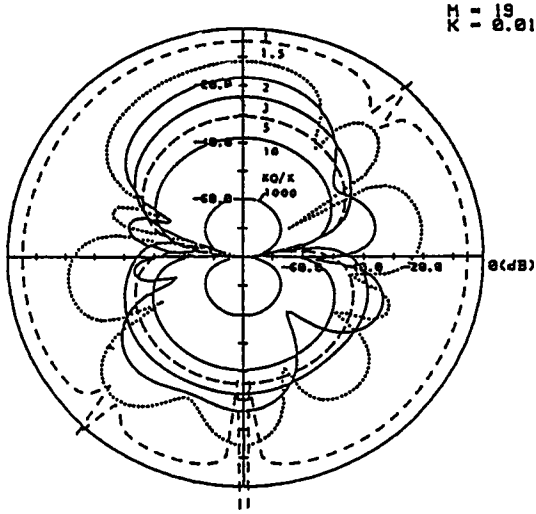


Figure 9. An example of the directional characteristics of the relative gain.

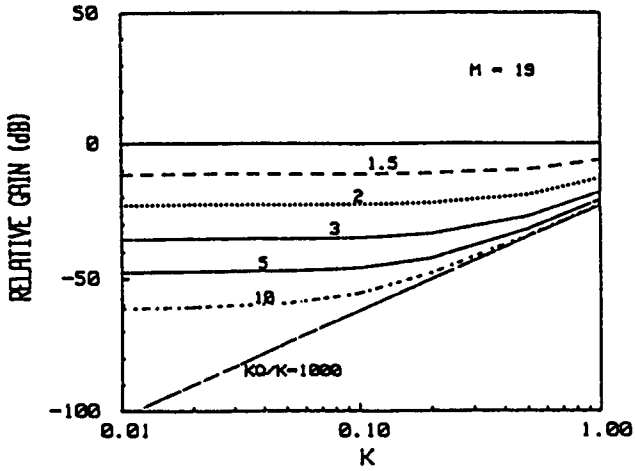


Figure 10. Reducation of noise output of the multi-microphone system for isotropic incidence of a noisy wave from various distances. 0 dB refers to the output of a non-directional microphone when the other conditions are kept equal.

Chapter 9

EVALUATION

This Page Intentionally Left Blank

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Classification of Japanese Syllables Including Speech Sounds Found in Loanwords

Shizuo Hiki

School of Human Sciences, Waseda University, Tokorozawa, Saitama, 359 Japan

Abstract

Methods of classifying Japanese syllables are discussed with regard to their use in evaluating the performance of speech processing techniques for both synthesis/recognition systems and user's speaking/listening characteristics. Firstly, by taking into account of the initial/intervocalic contrast of utterance, sixteen intervocalic syllables are added to the traditional list of one-hundred Japanese syllables which has been used in articulation test for sppech transmission channels. Then, by rearranging these syllables based on place/manner of articulation and by supplementing missing kinds of combinations of the preceding consonant and following vowel, more than sixty new syllables which can be pronounced in the loanwords are derived. A revised version of the text ("The North Wind and the Sun") as Japanese specimen for illustrating the International Phonetic Alphabet as an application of the classification of Japanese syllables by the International Phonetic Association, which contains a greater variety of syllables, is proposed.

1 LEVELS OF CLASSIFYING JAPANESE SYLLABLES

Purpose of this trial of classifying Japanese syllables is two-fold. The first is to rearrange systematically the traditional Japanese syllables by combining with new ones found in loanwords, which have become indispensable in speaking/listening and even to speech synthesis in modern Japanese 1), 2). The second is to provide various kind of lists of Japanese syllables according to the different rules of adding new syllables to the traditional ones, so that the most suitable one can be chosen for each purpose of evaluation of speech processing techniques.

The rules used in each level of adding syllables are as follows:

Level 0: This is the basic list of the traditional list of one-hundred Japanese syllables which has been widely used in articulation test for speech transmission channel and also for speech synthesis/recognition systems. They have been composed by utilizing the fifty Japanese Kana chart which correspond to a set of syllables consisting of <unvoiced consonant + vowel> (except [p]) and <semi-vowel [w, j] + vowel>, and by adding their voiced counterparts and <consonant + semi-vowel [j] + vowel> counterparts.

Level +1: For the consonants [g], [dz] and [dʒ] which are found in word (or phrase) initial utterances, the intervocalic counterpart [ŋ], [z] and [ʒ] are introduced, and 16 new syllables are added. (They are underlined in Table 1.)

Level +2: By separating [ʃ] found when followed by the vowel [i] from [s] in the /s/ row of the fifty Japanese Kana chart, and by supplementing the missing kinds with following vowels [i] for [s] and [e] for [ʃ], new syllables are derived. The same rules are applied to the initial and intervocalic voiced counterparts, adding six new syllables in this level.

Level +3: By separating the consonants in the /t/ row into [t], [ts] (when followed by vowel [u]) and [ts] (when followed by [i]), and supplementing missing kind of vowels, nine new syllables are introduced.

Level +4: The /h/ row is decomposed into [h], [ç] and [Φ] rows, and nine new syllables are derived. All the syllables adding in these levels are often found in the speech sounds of loanwords.

Level +5: New sounds [kw], [gw], and [v] which are found only in the utterance of loanwords are derived, and 13 syllables are added.

Level +6: For the syllables consisting of <consonant + vowel> with [t], [d], [Φ] and [v], 16 possible counterparts of <consonant + semi-vowel [j] + vowel> are added.

Level +7: For the syllables consisting of <consonant + semi-vowel [j] + vowel> with eight kinds consonant, namely [p], [b], [k], [g], [r], [m], [n] and [ŋ], a missing kinds with following vowel [e] are supplemented.

The number of syllables introduced in Level +2 through +7 for the speech sounds of loanwords are 61. (They are underlined in Table 2.) The total number of Japanese syllables becomes 177 when they are added to Level 0 and +1. These levels are introduced in the order of familiarity in the modern Japanese speech, but any combination of the levels can be adopted according to the purpose of its use.

2 SELECTING TEXT FOR TEST MATERIAL

In order to derive a revised version of the text ("The North Wind and the Sun") as Japanese specimen for illustrating the International Phonetic Alphabet by the International Phonetic Association 3), firstly, the text of the English specimen was translated into modern Japanese. Then, some of the words in the text were replaced by alternative ones so that the text contained a greater variety of syllables of the Level +1 list of Japanese syllables. The text was intended to be kept in the style of narration which is natural to both young and old adults. (The English text and Japanese text, in both Japanese letters and phonetic symbols, are shown in Table 3.)

The revised text has following features regarding the distribution of vowels, consonants and syllables 4):

Table 1. List of Japanese syllables in Level 0 and Level +1.

NOTATION: [PHONETIC SYMBOL] KANA
/PHONEMIC SYMBOL/ ROMAN REPRESENTATION

UNVOICED CONSONANT + VOWEL, SEMI-VOWEL + VOWEL (44 SYLLABLES)

[a] あ / 'a / a	[i] い / 'i / i	[u] う / 'u / u	[e] え / 'e / e	[o] お / 'o / o	/ '/ :5
[ka] か / ka / ka	[ki] き / ki / ki	[ku] く / ku / ku	[ke] け / ke / ke	[ko] こ / ko / ko	[k] :5 / k/ :5
[sa] さ / sa / sa	[ʃi] し /sji / (si) shi	[su] す / su / su	[se] せ / se / se	[so] そ / so / so	[s] :4 [ʃ] :1 / s/ :5 / j/ :1
[ta] た / ta / ta	[ʧi] ち / ci / (ti) chi	[tsu] つ / cu / (tu) tsu	[te] て / te / te	[to] と / to / to	[t] :3 [ʧ] :1 [ʦ] :1 / t/ :3 / c/ 2
[na] な / na / na	[ni] に / ni / ni	[nu] ぬ / nu / nu	[ne] ね / ne / ne	[no] の / no / no	[n] :5 / n/ :5
[ha] は / ha / ha	[ç i] ひ / hi / hi	[ɸu] ふ / hu / hu fu	[he] へ / he / he	[ho] ほ / ho / ho	[h] :3 [ç] :1 [ɸ] :1 / h/ :5
[ma] ま / ma / ma	[mi] み / mi / mi	[mu] む / mu / mu	[me] め / me / me	[mo] も / mo / mo	[m] :5 / m/ :5
[ja] や /'ja / ya		[ju] ゆ /'ju / yu		[jo] よ /'jo / yo	[j] :3 / '/ :3 / j/ :3
[ra] ら / ra / ra	[ri] り / ri / ri	[ru] る / ru / ru	[re] れ / re / re	[ro] ろ / ro / ro	[r] :5 / r/ :5
[wa] わ / wa / wa					[w] :1 / '/ :1 / w/ :1

VOICED CONSONANT + VOWEL, /p/ + VOWEL (33 SYLLABLES)

[ga] が / ga / ga	[gi] ぎ / gi / gi	[gu] ぐ / gu / gu	[ge] げ / ge / ge	[go] ご / go / go	$[g] :5 / g/ :5
[ŋa] が / ga / ga	[ŋi] ぎ / gi / gi	[ŋu] ぐ / gu / gu	[ŋe] げ / ge / ge	[ŋo] ご / go / go	*[ŋ] :5 / g/ :5
[ʣa] ざ / za / za	[ʥi] じ ぢ /zji / (zi) (di) ji	[ʣu] ず づ / zu / zu (du)	[ʣe] ぜ / ze / ze	[ʣo] ぞ / zo / zo	$[ʣ] :4 $[ɕ] :1 / z/ :5 / j/ :1
[za] ざ / za / za	[ʒa] じ ぢ /zji / (zi) ji (di)	[zu] ず づ / zu / zu	[ze] ぜ / ze / ze	[zo] ぞ / zo / zo	*[z] :4 *[ʒ] :1 / z/ :5 / j/ :1
[da] だ / da / da			[de] で / de / de	[do] ど / do / do	[d] :3 / d/ :3
[pa] ぱ / pa / pa	[pi] ぴ / pi / pi	[pu] ぷ / pu / pu	[pe] ぺ / pe / pe	[po] ぽ / po / po	[p] :5 / p/ :5
[ba] ば / ba / ba	[bi] び / bi / bi	[bu] ぶ / bu / bu	[be] べ / be / be	[bo] ぼ / bo / bo	[b] :5 / b/ :5
[a] /a /	[i] /i /	[u] /u /	[e] /e /	[o] /o /	$: INITIAL *: INTERVOCALIC
17	14	15	15	16	77
NUMBER OF FOLLOWING VOWELS					NUMBER OF SYLLABLES (CONSONANT + VOWEL, SEMI-VOWEL + VOWEL)

NUMBER OF [PHONETIC SYMBOLS]
/PHONEMIC SYMBOLS/
IN PRECEDING CONSONANT

UNVOICED CONSONANT + SEMI-VOWEL + VOWEL (21 SYLLABLES)

[kja] きゃ /kja / kya		[kju] きゅ /kju / kyu		[kjo] きょ /kjo / kyo	[kj] :3 / k/ / j/ :3
[ʃa] しゃ /sja / sya sha		[ʃu] しゅ /sju / syu shu		[ʃo] しょ /sjo / syo sho	[ʃ] :3 / s/ / j/ :3
[ʧa] ちゃ /cja / tya cha		[ʧu] ちゅ /cju / tyu chu		[ʧo] ちょ /cjo / tyo cho	[ʧ] :3 / c/ / j/ :3
[ɲa] にゃ /nja / nya		[ɲu] にゅ /nju / nyu		[ɲo] にょ /njo / nyo	[ɲ] :3 / n/ / j/ :3
[ça] ひゃ /hja / hya		[çu] ひゅ /hju / hyu		[ço] ひょ /hjo / hyo	[ç] :3 / h/ / j/ :3
[mja] みゃ /mja / mya		[mju] みゅ /mju / myu		[mjo] みょ /mjo / myo	[mj] :3 / m/ / j/ :3
[rja] りゃ /rja / rya		[rju] りゅ /rju / ryu		[ryo] りょ /ryo / ryo	[rj] :3 / r/ / j/ :3

VOICED CONSONANT + SEMI-VOWEL + VOWEL, /p/ + SEMI-VOWEL + VOWEL (18 SYLLABLES)

[gja] ぎゃ /gja / gya		[gju] ぎゅ /gju / gyu		[gjo] ぎょ /gjo / gyo	$[gj] :3 / k/ / j/ :3
[ŋja] ぎゃ /gja / gya		[ŋju] ぎゅ /gju / gyu		[ŋjo] ぎょ /gjo / gyo	*[ŋj] :3 / k/ / j/ :3
[ʤa] じゃ ぢゃ /zja / zya (dya) ja		[ʤu] じゅ ぢゅ /zju / zyu (du) ju		[ʤo] じょ ぢょ /zjo / zyo (dyo) jo	$[ʤ] :3 / z/ / j/ :3
[ʒa] じゃ ぢゃ /zja / zya ja (dya)		[ʒu] じゅ ぢゅ /zju / zyu ju (dyu)		[ʒo] じょ ぢょ /zjo / zyo jo (dyo)	*[ʒ] :3 / z/ / j/ :3
[pja] ぴゃ /pja / pya		[pju] ぴゅ /pju / pyu		[pjo] ぴょ /pjo / pyo	[pj] :3 / p/ / j/ :3
[bja] びゃ /bja / bya		[bju] びゅ /bju / byu		[bjo] びょ /bjo / byo	[bj] :3 / b/ / j/ :3

[a] /a /	[i] /i /	[u] /u /	[e] /e /	[o] /o /	$: INITIAL *: INTER VOCALIC
13	0	13	0	13	39
NUMBER OF FOLLOWING VOWELS					NUMBER OF SYLLABLES (CONSONANT + SEMI-VOWEL + VOWEL)
[a] /a /	[i] /i /	[u] /u /	[e] /e /	[o] /o /	
30	14	28	15	29	116
NUMBER OF FOLLOWING VOWELS					NUMBER OF SYLLABLES (TOTAL)

Table 2. List of Japanese syllables in Level 0 through +7.

NOTATION: [PHONETIC SYMBOL] KANA (KATAKANA: LOAN WORD) NUMBER OF [PHONETIC SYMBOLS]
/PHONEMIC SYMBOL/ ROMAN REPRESENTATION /PHONEMIC SYMBOLS/
IN PRECEDING CONSONANT

[pa] ぱ / pa / pa	[pi] ぴ / pi / pi	[pu] ぷ / pu / pu	[pe] ぺ / pe / pe	[po] ぽ / po / po	[p] :5 / p/ :5
[ba] ば / ba / ba	[bi] び / bi / bi	[bu] ぶ / bu / bu	[be] べ / be / be	[bo] ぼ / bo / bo	[b] :5 / b/ :5
[pja] ぴゃ /pja / pya		[pju] ぴゅ /pju / pyu	[pje] ぴぇ /pje / pye	[pjo] ぴょ /pjo / pyo	[pj] :3+1 / p/ / j/ :3+1
[bja] びゃ /bja / bya		[bju] びゅ /bju / byu	[bje] びぇ /bje / bye	[bjo] びょ /bjo / byo	[bj] :3+1 / b/ / j/ :3+1
[ta] た / ta / ta	[ti] てぃ / ti / ti	[tu] とぅ / tu / tu	[te] て / te / te	[to] と / to / to	[t] :3+2 / t/ :3+2
[da] だ / da / da	[di] でぃ / di / di	[du] どぅ / du / du	[de] で / de / de	[do] ど / do / do	[d] :3+2 / d/ :3+2
[tja] てぃゃ /tja / tya		[tju] てぃゅ /tju / tyu	[tje] てぃぇ /tje / tye	[tjo] てぃょ /tjo / tyo	[tj] :4 / t/ :4
[dja] でぃゃ /dja / dya		[dju] でぃゅ /dju / dyu	[dje] でぃぇ /dje / dye	[djo] でぃょ /djo / dyo	[dj] :4 / d/ :4
[tsa] つぁ / ca / tsa	[tsi] つぃ / ci / tsi	[tsu] つ / cu / (tu) tsu	[tse] つぇ / ce / tse	[tso] つぉ / co / tso	[ts] :1+4 / c/ :1+4
[dza] ざ / za / za	[dzi] ずぃ / zi / zi	[dzu] ず ず / zu / zu (du)	[dze] ぜ / ze / ze	[dzo] ぞ / zo / zo	$[dz] :4+1 / z/ :4+1
[tʃa] ちゃ /cja / tya cha	[tʃi] ち / ci / (ti) chi	[tʃu] ちゅ /cju / tyu chu	[tʃe] ちぇ /cje / tye	[tʃo] ちょ /cjo / tyo cho	[tʃ] :4+1 / c/ / j/ :4+1
[dʒa] ぢゃ じゃ /zja / zya (dya) ja	[dʒi] ぢ じ / zi / (zi) (di) ji	[dʒu] ぢゅ じゅ /zju / zyu (du) ju	[dʒe] ぢぇ じぇ /zje / zye	[dʒo] ぢょ じょ /zjo / zyo (dyo) jo	$[dʒ] :4+1 / z/ / j/ :4+1
[sa] さ / sa / sa	[si] すぃ / si / si	[su] す / su / su	[se] せ / se / se	[so] そ / so / so	[s] :4+1 / s/ :4+1
[za] ざ / za / za	[zi] ずぃ / zi / zi	[zu] ず / zu / zu	[ze] ぜ / ze / ze	[zo] ぞ / zo / zo	*[z] :4+1 / z/ :4+1
[ʃa] しゃ /sja / sya sha	[ʃi] し /sji / (si) shi	[ʃu] しゅ /sju / syu shu	[ʃe] しぇ /sje / sye she	[ʃo] しょ /sjo / syo sho	[ʃ] :4+1 / s/ / j/ :4+1
[ʒa] じゃ ぢゃ /zja / zya ja (dya)	[ʒi] じ ぢ /zji / (zi) ji (di)	[ʒu] じゅ ぢゅ /zju / zyu ju (dyu)	[ʒe] じぇ ぢぇ /zje / zye je	[ʒo] じょ ぢょ /zjo / zyo jo (dyo)	*[ʒ] :4+1 / z/ / j/ :4+1
[ka] か / ka / ka	[ki] き / ki / ki	[ku] く / ku / ku	[ke] け / ke / ke	[ko] こ / ko / ko	[k] :5 / k/ :5
[ga] が / ga /. ga	[gi] ぎ / gi / gi	[gu] ぐ / gu / gu	[ge] げ / ge / ge	[go] ご / go / go	$[g] :5 / g/ :5
[kja] きゃ /kja / kya		[kju] きゅ /kju / kyu	[kje] きぇ /kje / kye	[kjo] きょ /kjo / kyo	[kj] :3+1 / k/ / j/ :3+1
[gja] ぎゃ /gja / gya		[gju] ぎゅ /gju / gyu	[gje] ぎぇ /gje / gye	[gjo] ぎょ /gjo / gyo	$[gj] :3+1 / k/ / j/ :3+1

[a] /a/	[i] /i/	[u] /u/	[e] /e/	[o] /o/	
[ha] は / ha / ha [a] あ / a / a			[he] へ / he / he [e] え / e / e	[ho] ほ / ho / ho [o] お / o / o	[h] :3 / h/ :3 / '/ :3
[ça] ひゃ / ça / hya [ja] や / ja / ya	[çi] ひ / çi / hi [i] い / i / i	[çu] ひゅ / çu / hyu [ju] ゆ / ju / yu	[çe] ひぇ / çe / hye [je] イェ / je / ye	[ço] ひょ / ço / hyo [jo] よ / jo / yo	[ç] :4+1 / ç/ :4+1 [j] :3+1 / '/ :4+1 / j/ :3+1
[ɸa] ファ / ɸa / fa [wa] わ / wa / wa	[ɸi] フィ / ɸi / fi [vi] ウィ ヰ / vi / vi	[ɸu] ふ / ɸu / hu fu [u] う / u / u	[ɸe] フェ / ɸe / fe [we] ウェ ヱ / we / we	[ɸo] フォ / ɸo / fo [wo] ウォ ヲ / wo / wo wo	[ɸ] :1+4 / h/ :1 / f/ 4 [w] :1+3 / w/ :1+3 / '/ :1
[ɸja] フィャ /fja / fya		[ɸju] フィュ /fju / fyu	[ɸje] フィェ /fje / fye	[ɸjo] フィョ /fjo / fyo	[ɸj] :4 / f/ / j/ :4
[va] ヴァ / va / va	[vi] ヴィ / vi / vi	[vu] ヴ / vu / vu	[ve] ヴェ / ve / ve	[vo] ヴォ / vo / vo	[v] :5 / v/ :5
[vja] ヴィャ /vja / vya		[vju] ヴィュ /vju / vyu	[vje] ヴィェ /vje / vye	[vjo] ヴィョ /vjo / vyo	[vj] :4 / v/ / j/ :4
[kwa] クァ /kwa / kwa (kwa)	[kwi] クィ /kwi /kwi		[kwe] クェ /kwe / kwe	[kwo] クォ /kwo / kwo	[kw] :4 / k/ / w/ :4
[gwa] グァ /gwa / gwa gwa	[gwi] グィ /gwi / gwi		[gwe] グェ /gwe / gwe	[gwo] グォ /gwo / gwo	[gw] :4 / g/ / w/ :4
[ra] ら / ra / ra	[ri] り / ri / ri	[ru] る / ru / ru	[re] れ / re / re	[ro] ろ / ro / ro	[r] :5 / r/ :5
[rja] りゃ /rja / rya		[rju] りゅ /rju / ryu	[rje] りぇ /rje / rye	[ryo] りょ /ryo / ryo	[rj] :3+1 / r/ / j/ :3+1
[ma] ま / ma / ma	[mi] み / mi / mi	[mu] む / mu / mu	[me] め / me / me	[mo] も / mo / mo	[m] :5 / m/ :5
[mja] みゃ /mja / mya		[mju] みゅ /mju / myu	[mje] みぇ /mje / mye	[mjo] みょ /mjo / myo	[mj] :3+1 / m/ / j/ :3+1
[na] な / na / na	[ni] に / ni / ni	[nu] ぬ / nu / nu	[ne] ね / ne / ne	[no] の / no / no	[n] :5 / n/ :5
[ɲa] にゃ / ɲa / nya		[ɲu] にゅ / ɲu / nyu	[ɲje] にぇ /ɲje / nye	[ɲo] にょ / ɲo / nyo	[ɲ] :3+1 / n/ / j/ :3+1
[ŋa] が / ŋa / ga	[ŋi] ぎ / ŋi / gi	[ŋu] ぐ / ŋu / gu	[ŋe] げ / ŋe / ge	[ŋo] ご / ŋo / go	*[ŋ] :5 / g/ :5
[ŋja] ぎゃ /ŋja / gya		[ŋju] ぎゅ /ŋju / gyu	[ŋje] ぎぇ /ŋje / gye	[ŋjo] ぎょ /ŋjo / gyo	*[ŋj] :3+1 / g/ / j/ :3+1
30 +9 =39	14+11 =25	28 +7 =35	15+24 =39	29+10 =39	116 +61 =177

NUMBER OF FOLLOWING VOWELS

$: INITIAL
*: INTERVOCALIC

NUMBER OF SYLLABLES (LOAN WORD)

Table 3. The English text and the revised Japanese text.

(The Principles of the International Phonetic Assosiation)
The north wind and the sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be considerd stronger than the other. Then the north wind blew as hard as he could, but the more he blew the more closely did the traveller fold his cloak around him, and at last the north wind gave up the attempt. Then the sun shone out warmly, and immediately the traveller took off his cloak. And so the north wind was obliged to confess that the sun was the stronger of the two.

北風と太陽が、どちらが強いかで言い争っているところへ、旅人が暖かそうな外套にくるまってやってきました。そこで、その旅人の外套をさきに脱がせることができた方が、強いのだとゆうことにしようときめました。まず北風が、旅人に向かってせいいっぱい烈しく吹きつけました。しかし、吹けば吹くほど、旅人は外套をますますしっかりと体にまとい付けるので、北風はへとへとにくたびれて、とうとうあきらめました。つぎに太陽が、旅人の上からじわじわと暖かい光を注ぐと、たちまち旅人は外套を脱ぎました。それで北風も残念ながら、太陽の方が強いと認めなければなりませんでした。

kʲitakazeto táijo:ŋa dóʃiraŋa tsujóikade i:arasótteiru tokoróe, tabibitoŋa atatakasó:na gaito:ni krumátte jattekimáʃʲita. sokode, sono tabibitono gaito:o sakini nuŋáseru kotóŋa dékʲitaho:ŋa, tsʷjoinodato ju:kotóni ʃijó:to kimemaʃʲita. mázu kʲitakazeŋa, tabibitoni mukatte se:ippai haŋeʃʲikʷ ɸʷkitsʷkemáʃʲita. ʃíkaʃʲi, ɸʷkebaɸʷkuhodo, tabibitowa gaito:o masúmasʷ ʃʲikkárito karadani matoitsʷkérunode, kʲitakazewa hetohetoni kʷtabírete, tó:to: akʲiramemáʃʲita. tsuŋíni táijo:ŋa, tabibitono uekara ɕiwaʒiwato atatakái hʲikarío sosóguto, taʧímaʧʲi tabibitowa gaito:o nuŋimáʃʲita. sorede, kʲitakazemo ɕannennáŋara, táijo:no hó:ŋa tsʷjóito mitomenakéreba narimaséndeʃʲita.

Consonants, several samples of each of the 21 kinds of Japanese consonants, as well as choked sound and syllabic nasal, are found among the 202 consonants in the text.

Vowels, vowels in the text, 282 in total, show a distribution of the five Japanese vowels similar to ordinary Japanese speech. More than ten examples are found for each of the elongated and diphthongized vowels. Examples of devoicing of [i] and [u] are also found to be more than ten for each.

Syllables, total number of syllables in the sentence is 212. At least one example is found for all 44 kinds of syllables consisting of <semi-vowel [w, j] + five vowels> and <unvoiced consonant + five vowels> (except [p]). Examples are found for the 17 kinds of syllables all 23 kind of syllables consisting of <voiced consonant + vowel> (plus [p]). On the other hand, syllables consisting of <consonant + semi-vowel [j] + vowel> are not included in the text.

In addition to these phonetic or phonemic features, the text as test material has to be designed as a typical example of Japanese sentences with regard to words, phrases, closes and sentences. The different words among the 136 words in total, in this text are 71. They can be grouped into 53 phrases each having a down skip of word accent. The text consisted of 6 sentences, five of them having two clouses. The Average number of words in a sentence is nine. Examples of all types of word accent are found. These features reflect the statistic characteristics of Japanese speech 5), 6).

The use of the text will be examined in detail through the acoustical analysis of the utterances by speakers of standard Japanese.

References

1. S. Hiki: "Advanced methods of evaluating techniques for speech processing," *Preprint, The First Symposium on Advanced Man-Machine Interface Through Spoken Language*, pp.69-76, 1988.

2. S. Hiki: "Classification of Japanese syllables including speech sound found in loanwords," Research Report PASL No.01-8-1, pp.1-16, 1989.

3. The Principles of the International Phonetic Association, being a description of the International Phonetic Alphabet and the manner of using it, illustrated by text in 51 languages, International Phonetic Association, 1949.

4. S. Hiki: "Selecting sentence text for test material taking "The North Wind and the Sun" as an example," Research Report PASL No.01-8-3, pp.1-8, 1989.

5. H. Sato: "Statistical analysis, of Japanese phoneme concatenations for speech synthesis," *Trans. IEICEJ and ASJ*, SP82-77, pp.609-616, 1983.

6. K. Itoh and H. Sato: "Phoneme occurrence characteristics in Japanese conversational speech," *Proc. Fall Meeting of the ASJ*, pp.151, 1988.

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

A Study of the Suitability of Synthetic Speech for Proof-Reading in Relation to the Voice Quality

Hideki Kasuya

Faculty of Engineering, Utsunomiya University
2753 Ishii-machi, Utsunomiya, 321 Japan

Abstract

Even if synthetic speech is intelligible and natural-sounding, its voice quality must still be evaluated in terms of the suitability for individual applications; particular the voice quality of synthetic speech may be a severe burden for the users. We first summarize several human factors pertaining to the voice quality of synthetic speech which we thought important from our interviews with professional users of synthetic speech, working at the proof-reading department in a newspaper company, where synthetic speech has extensively been used as an aid for the proof-reading of manuscripts in a computer. This paper describes new findings obtained from perceptual experiments on the subjects' preference for voice quality of synthetic speech, primarily focusing on the suitability of pitch characteristics, speaker's sex, and speaking rate, in the task where subjects were asked to proof-read a printed text while listening to the speech.

1. INTRODUCTION

Most of the research on the evaluation of synthetic speech produced by rules has primarily focused on intelligibility test at the levels of phonemes, syllables, words, and sentences. It was not only concerned with the measurement of the extent to which the linguistic information has been correctly transmitted to the subjects, but also with improvement of the intelligibility resulting from a priori knowledge about linguistic structures and pragmatics [1-8].

There are some reports on the evaluation of the naturalness of synthetic speech, most of which are related to preference test performed on multiple synthetic speech samples and to the diagnostic evaluation in terms of the degree of manifestation of segmental and prosodic features in synthetic speech [5, 6].

Little research, on the other hand, has been reported on the suitability of synthetic speech for a specific application. Even if synthetic speech is intelligible and natural-sounding, its voice quality must still be evaluated in terms of the suitability for a specific application; particular the voice quality of synthetic speech may be a severe burden for the users.

In this paper, we will first summarize several human factors involved in the use of synthetic speech which we thought important from our interviews with professional users of synthetic speech, working at the proof-reading department in a newspaper company, where synthetic speech is extensively used as an aid for the proof-reading of manuscripts in a computer. This paper describes new findings obtained from perception experiments on the subjects' preference for voice quality of synthetic speech, primarily focusing on the suitability of pitch characteristics, speaker's sex, and speaking rate, in the task where subjects were asked to proof-read a printed text while listening to the speech.

2. HUMAN FACTORS INVOLVED IN THE PROOF- READING TASK

There are many applications of Japanese synthetic speech produced by rules in real use. One typical example can be seen in a newspaper company, that has made full use of a Japanese text-to-speech synthesis system as an aid for proof-reading of a text. The major specifications of the synthesis system employed at this newspaper company are: (1) a diphone-type synthesizer with CV/VC/VV sound segments as the unit for the synthesis, (2) LPC parameters for spectral representation, (3) eight variable speaking rates, (4) option of a male or a female voice, and (5) a loudspeaker or a headset output. Operators at the proof-reading department are engaged in the proof-reading of manuscripts at computer voice terminals about four hours a day with adequate intermissions. We interviewed several operators and a head of the department being responsible for the whole system. Human factors extracted from the interview are summarized as follows:

(1) Male voices are exclusively used by all the operators except when interference with other voice terminals may take place because of the use of all the terminals, although the naturalness of the synthetic female voice is nearly the same as that of the male voice.

(2) A speech synthesizer should be able to produce a wide variety of voice qualities from which the users will benefit for a change of pace.

(3) The operators do not care about unnaturalness of synthetic speech, as long as its intelligibility is assured, since they can adapt themselves to a computer voice.

(4) Controllability of the speaking rate is very helpful.

(5) Operators are forced to considerably slow down the work when errors occur resulting from inadequate linguistic analysis of a text, e.g. errors in syntactic boundary assignment, phonetic transformation, accentuation, etc.

(6) It is helpful to have additional pauses at "proper places" in a sentence.

(7) Use of a headset hastens fatigue. A loudspeaker at the terminal is exclusively used unless interference with other terminals is expected.

(8) Four hours a day suffices.

The first through third items are all related to the voice quality of synthetic speech and the fourth to the prosodic properties. From these we felt that the pitch characteristics as a constituent of both voice quality and prosodic properties play important roles in promoting usability of synthetic speech for a specific task, such as the proof-reading in the newspaper company. We tried to discover from perception experiments in a laboratory (1) an adequate range of the average pitch frequency for each of the male and female voices, (2) preference for a male or a female voice, and (3) a suitable dynamic range of pitch variations, assuming that the synthetic speech is used at proof-reading under the same working conditions as those in the newspaper company mentioned above. The adequate speaking rate was also studied.

3. PITCH FREQUENCY CHARACTERISTICS SUITABLE FOR PROOF-READING

3.1. Analysis-Synthesis System

Since none of the Japanese speech synthesizer products currently available provides the ability of flexible pitch control, a PARCOR analysis-synthesis system was employed to generate speech material for the experiment. The major specifications of the system used are: analysis window lengths of 20 ms for males and 10 ms for females, an inverse filter of the 12th order, a frame rate of 10 ms, a sampling frequency of 10 kHz, and a quantization accuracy of 12 bits. The pitch frequency was measured by a flexible pitch extractor based on the autocorrelation method with nonlinear preprocessing [9]). Pitch frequency errors and inadequate PARCOR parameters were corrected by hand. The pitch contours thus obtained were further modified according to the experimental conditions for the synthesis. In the synthesis routine, the spectral and pitch parameters were updated synchronously with the pitch using a linear interpolation method. The excitation source was an impulse train for voiced and white noise for unvoiced.

3.2. Speech Material

Two male (K and I) and two female (U and Y) public radio announcers read a written text of about 12 s at their comfortable pitch, loudness, and speaking rate. While the speakers K and U had relatively low fundamental frequencies, I and Y showed rather high fundamental frequencies in each of the sex groups. The text was selected from the radio news.

3.3. Method

Three perception experiments were carried out: (1) subjective judgments on the suitable average pitch frequency of the individual announcers for the proof-reading task (Experiment I), (2) preference test for a male or a female voice (Experiment II), and (3) measurements of the preferable dynamic range of pitch frequency variations (Experiment III). The effect of the environmental noise level on the preference was also examined.

Experiment I

The speech samples consisted of synthesized speech utterances with five different pitch contours for the individual speakers. The pitch contours included the original and four modified contours, being raised or lowered by 0.2 octave/step. The average pitch frequency of the samples ranged from about 90 to 180 Hz for the males and from 160 to 300 Hz for the females, as shown in Table 1. It is confirmed that the intelligibility of the synthesized speech with modified pitch contours is well preserved within the used pitch range.

32 male and 22 female subjects participated in the perception experiments. They proof-read a printed text while listening to the synthesized speech samples in an office with a background noise level of 50 dBA and in a sound-proof room at 20 dBA. Each of the 20 texts (5 pitch contours $\times$ 4 announcers) included different errors that were intentionally added considering types of errors commonly encountered in many Japanese word processor products. The subjects were asked to rate the speech samples on preference scale, giving 1 to the worst and 3 to the best of the five different pitch patterns of the individual speakers. They were instructed to imagine that they were engaged in the proof-reading while listening to each of the speech samples about four hours every day as a professional operator. They were also asked to pay much attention to the difference in pitch.

Experiment II

On the basis of the judgments made in the first experiment, the subjects were required to state their preference for a male or a female voice in the same task.

Experiment III

The pitch contour of a sentence utterance was decomposed into its phrase and accentual components, assuming that the phrase component could be approximated by a line on a logarithmic frequency scale [10]. The frequency range of the accentual components, that were obtained by subtracting the phrase component from original pitch contour, was decreased by 24 or 43%. These modified accentual components were again added to the phrase component, thereby producing two different pitch contours. Speech samples were synthesized with the three pitch contours, i.e. the original and the two compressed contours. Eight female subjects gave their preference values for the three pitch patterns under the same instructions as for the other two experiments.

3.4. Results and Discussion

The results of Experiment I are illustrated in Figure 1, where the preference score is normalized to 10 points as the maximum for the individual speakers and the two dashed vertical lines are the average pitch frequencies of the original male and female utterances. Irrespective of the speakers' sex, the score rapidly decreases as the average pitch frequency increases. The reasons for their taking a dislike to higher pitch frequencies were related to fatigue, irritation, annoyance, and so on. Very low average pitch frequencies also gained a low preference score because of what was felt a depressed quality. For the two male speech samples the judgments are nearly the same for male and female listeners and the

Table 1. Relationship between the amount of pitch shift and average pitch frequencies

Speaker	Amount of pitch shift (oct.)	Average pitch frequency (Hz)
	-0.2	96.6
	0.0	110.4
K	0.2	127.7
	0.4	146.8
	0.6	168.8
	-0.6	97.4
	-0.4	112.0
I	-0.2	128.7
	0.0	147.0
	0.2	170.2
	-0.4	169.2
	-0.2	194.6
U	0.0	221.3
	0.2	257.6
	0.4	296.0
	-0.6	164.3
	-0.4	189.0
Y	-0.2	217.5
	0.0	247.1
	0.2	288.1

preferred average pitch frequency is between 110 and 130 Hz. This is also the case for the two female speech samples, where a pitch range between 190 and 220 Hz is preferred. It should be noted that the subjects gave higher scores to lower average pitch frequencies than the original one for one male and two female announcers. Little difference was observed in the judgments for quiet and somewhat noisy environments.

In Experiment II, 22 out of 32 (69%) male subjects and 19 out of 22 (86%) female subjects preferred male voices in the proof-reading task.

The preference rate for male voices was the same for the sound-proof room and the office environment. Most of the subjects answered that the male voice was more suitable for the task because of the same reasons as mentioned at Experiment I. The results are quite similar to those founed at the interviews with the operators in the newspaper company.

In Experiment III, seven out of eight subjects voted for the pitch contour with the accentual component compressed by 24% and the other for the one compressed by 43%. The original pitch contour gave an impression of somewhat exaggerated quality, resulting in the least preference. Almost all the subjects felt sleepy or bored by the overly flattened pitch pattern.

We may conclude from the experiments that a male voice quality with a low average

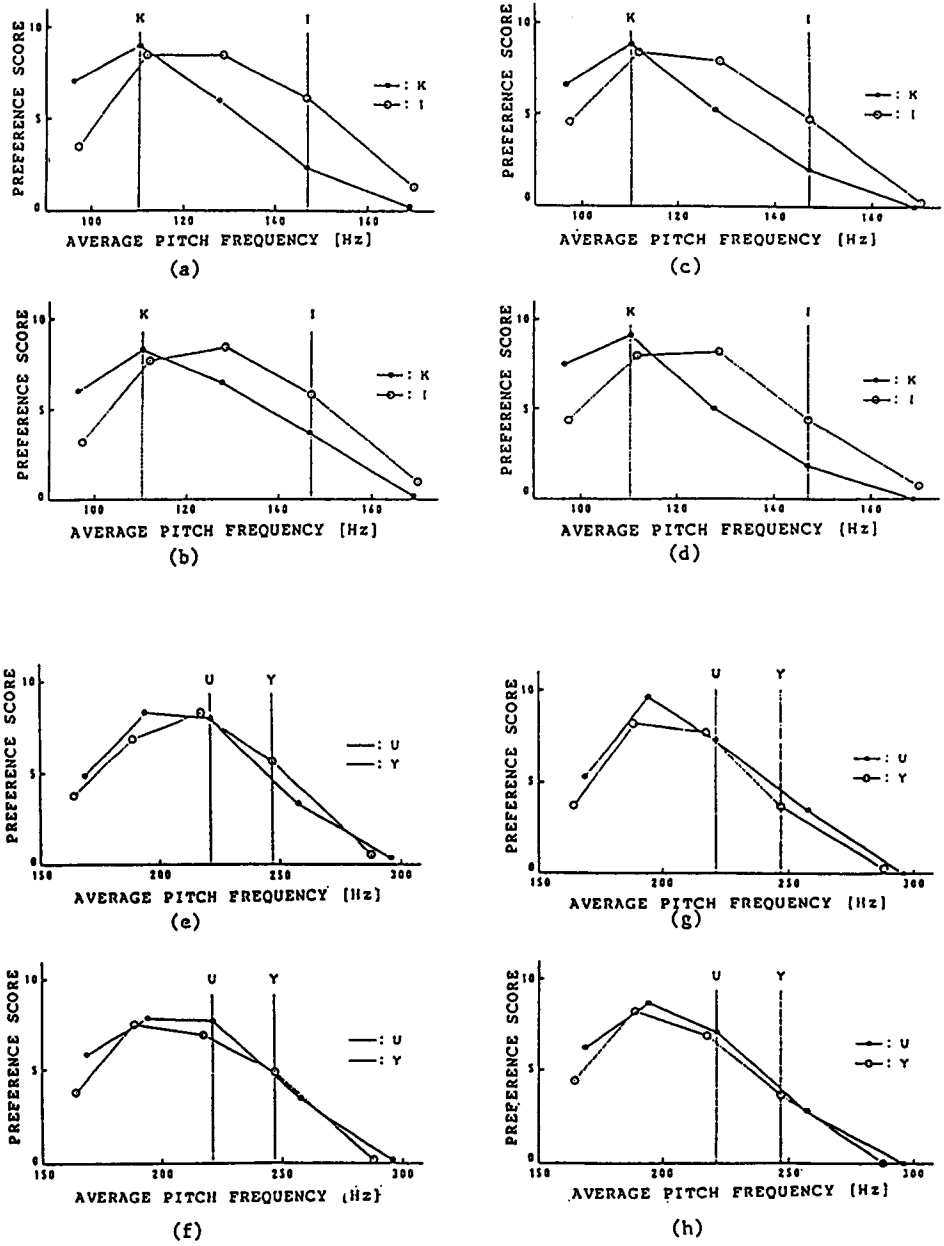


Figure 1. The normalized preference score as a function of the average pitch frequency: (a) the scores of the male subjects for the male speech samples in the office and (b) in the quiet room, (c) the scores of the female subjects for the male speech samples in the office and (d) in the quiet room, (e) the scores of the male subjects for female speech samples in the office and (f) in the quiet room, and (g) the scores of the female subjects for the female speech samples in the office and (h) in the quiet room.

pitch frequency with a moderate extent of pitch variations is suitable for the proof-reading task. This preference is primarily related to the task conditions, which the subjects were asked to judge in terms of every day use of the synthetic speech as a professional operator.

4. SPEAKING RATE

4.1. Experimental Method

A sequence of randomly ordered four digit numbers were synthesized at five different speaking rates with a commercially available speech synthesizer. Five subjects, who had no experience in listening to synthetic speechs, were asked to find wrong numbers in a written text while listening to the sequences of four numbers read by a synthesizer. Each of the texts included wrong numbers at error rates of 5 and 15%. After the listening sessions, they rated the speaking rates in terms of their preference, giving 1 to the worst and 3 to the best. The experiments were performed over six days.

4.2. Results and Discussion

The results are illustrated in Figure 2, where the preference score is again normalized to 10 as a maximum, and no. 1 is the slowest (5.7 morae/s) and no. 5 the fastest rate (11.3 morae/s). To see the overall trends of the judgments, the ones made on two consecutive days were summed over all the subjects. As the error rate of the text increases, slower speaking rates are preferred. Although we expected a shift of the preferred speaking rate to a higher rate as the subjects get accustomed to the task and synthetic speech, little change was observed over the repetitions. This is again due to the task condition given to the subjects that they should imagine to be engaged in the same task every day. Higher speaking rates certainly improved the efficiency of the task but required much concentration and produced fatigue.

The results of the judgments of the five subjects could be classified into two groups, as shown in Figure 3; group 1 preferred speaking rates of no. 2 and no. 3, but group 2 did No. 1 as well. Since the preferred speaking rates depend on the listeners, the ability of a variable speaking rate is indispensable for speech synthesizer products.

5. SUMMARY

People being professionally engaged in the proof-reading of manuscripts with synthetic speech produced by rules, provided significant human factors in terms of a suitability assessment of a synthesizer. Many human factors were related to the control of prosodic properties and voice quality of synthetic speech.

From the subjective judgments on the preference for three prosodic features and the voice quality, i.e. average pitch frequency, dynamic range of pitch variations, speaking rate, and sex of a speaker, it was concluded that a male voice with a relatively low average pitch frequency with rather small pitch changes was well accepted by users for the proof-reading task. This preference was considered to be related to the every day use of synthetic speech for the proof-reading task.

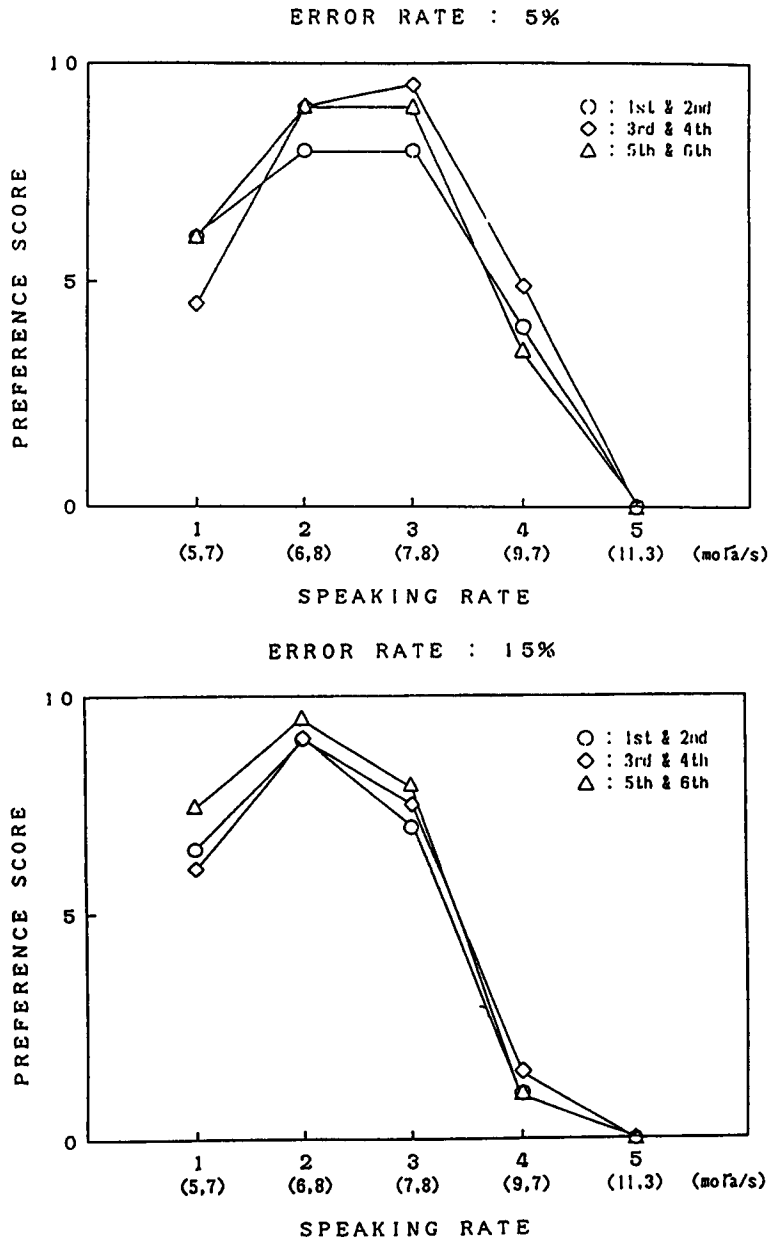


Figure 2. The normalized preference score as a function of the speaking rate for the two error rates in the digit strings, (a) 5 % and (b) 15 %. The judgments made on two consecutive days are added together.

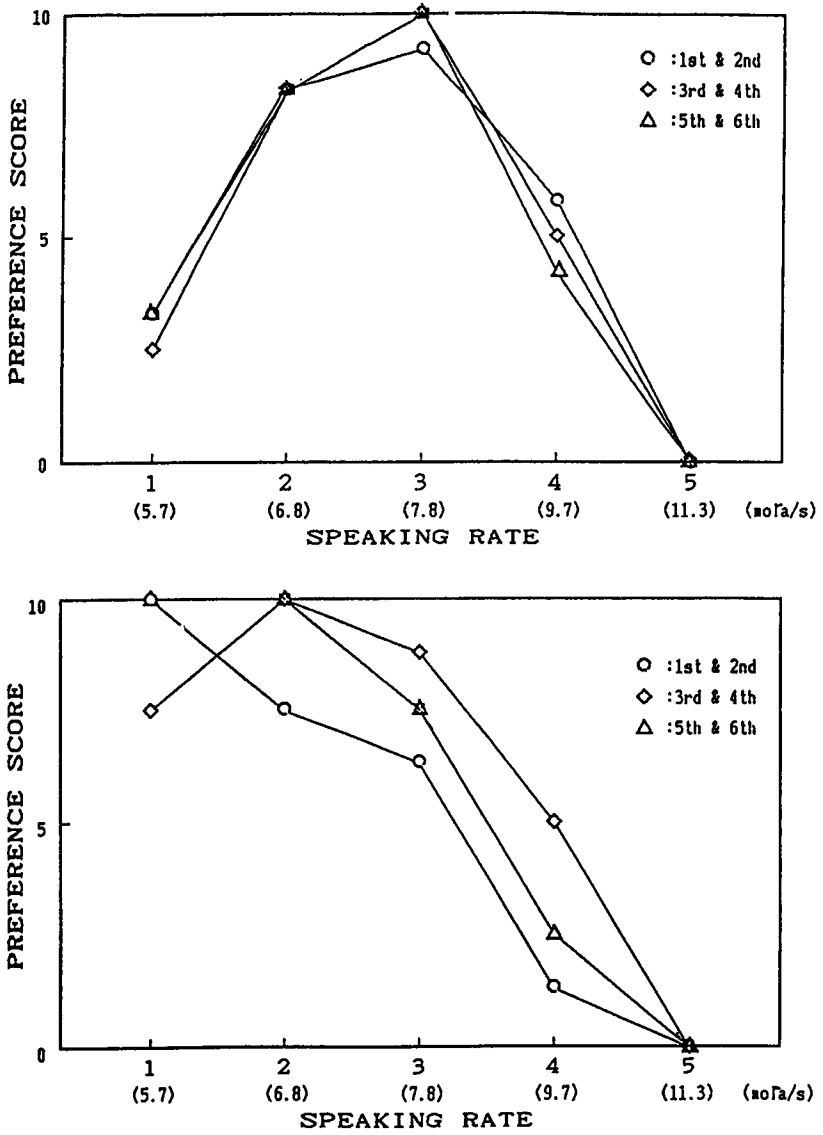


Figure 3. The normalized preference score as a function of the speaking rate for the digit strings with an error rate of 5 %. The type of the score distributions was divided into two groups.

Acknowledgements

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, "Advanced Methods of Evaluating Techniques for Speech Processing," from the Ministry of Education, Science and Culture, Japan. The author would like to thank Dr. S. Hiki, School of Human Sciences, Waseda University, for his valuable comments. He also thanks Messrs. K. Tajima, K. Morita and S. Kasuya, Utsunomiya University, for their very able assistance with experiments.

References

1. P. W. Nye and J. Gaitenby: "The intelligibility of synthetic monosyllable words in short, syntactically normal sentences," *Haskins Lab. Status Report*, SR-37/38, pp.169-190, 1974.

2. D. B. Pisoni and S. Hunnicut: "Perceptual evaluation of MITalk: the MIT unrestricted text-to-speech system," *Proc. ICASSP*, pp.572-575, 1980.

3. D. B. Pisoni, H. C. Nusbaum and B. G. Green: "Perception of synthetic speech generated by rule," *Proc. IEEE*, 73, pp.1665-1676, 1985.

4. B. G. Green, J. S. Logan and D. B. Pisoni: "Perception of synthetic speech produced automatically by rule: intelligibi lity of eight text-to-speech systems," *Behavior Res. Methods, Instruments & Computers*, 18, pp.100-107, 1986.

5. J. S. Logan and D. B. Pisoni: "Preference judgments comaring different synthetic voices," *J. Acoust. Soc. Am.*, 79, s24, 1986.

6. N. Higuchi, S. Yamamoto and T. Shimizu: "Evaluation of Intelligibility and naturalness of the synthetic speech synthe sized with a Japanese speech synthesizer by rule," *Trans. IECE Jpn.*, SP87-138, pp.61-68, 1988.

7. T. Watanabe and S. Hayashi: "Influence of listening conditions on intelligibility of synthesized speech by rule," *Proc. Spring Meeting of Acoust. Soc. Jpn*, pp.165-166, 1988.

8. M. Spiegel, M. J. Altom, M. Macchi and K. Wallace: "Using a monosyllabic test corpus to evaluate the intelligibility of synthesized and natural speech," *Proc. American AVIOS Conf.*, 1988.

9. X. Gao, Y. Kikuchi and H. Kasuya: "An improved algorithm of autocorrelation pitch detection," *Trans. IECE Jpn.*, E67, pp.291-292, 1984.

10. K. Hakoda and H. Sato: "Prosodic rules in connected speech synthesis," *Trans. IECE Jpn.*, J63-d, pp.715-722, 1980.

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Improving Synthetic Speech Quality by Systematic Evaluation

Louis C.W. Pols

Institute of Phonetic Sciences, University of Amsterdam
Herengracht 338, 1016 CG Amsterdam, The Netherlands

Abstract

In the joint Dutch research program for developing a high-quality text-to-speech synthesis system, much emphasis is put on systematic speech quality evaluation. This is not just done to produce performance figures, but even more so to support the developers of the various linguistic and acoustic synthesis modules by indicating to them ways for improvement. This ap proach compares favourably with most other projects in which no diagonistic testing is done at all, or once in the final phase in order to produce (incomparable) performance figures which do not lead to further improvements. The joint project is sponsered by SPIN (Dutch National Program for the Advancement of Information Technology).

1. INTRODUCTION

A complete text-to-speech synthesis-by-rule system consists of many different components originating from such diverse areas as text processing, language processing, and signal processing. By improving the performance of single components one hopes to improve the performance of the total system. However, more often than not, different experts in the various fields develop single components and leave the remaining problems to others. For instance the acoustic front end presupposes a correct phonetic input, whereas grapheme-to-phoneme conversion can easily introduce errors here. The intonation module requires correct stress markers, whereas rules to define the position and the character of those markers are not yet fully developed. The morphological decomposition requires error-free word sequences, and the text expander requires knowledge about how to interpret the text. Should, for instance, the digit sequence 14.18 be pronounced as a number, as a money value, or as a time indicator?

Whenever performance figures are given at all, they mainly represent the results of one final test. Such results specify in an absolute or relative way (in comparison to some reference system, such as LPC-resynthesized utterances) the achieved quality of the system, whereas a further diagonistic analysis of the results seldom leads to subsequent modifications of the system.

Especially in the recently started joint Dutch research program for developing a high-quality text-to-speech synthesis system (Pols, 1988), we hope to be able to follow a different line. In an initial evaluation the speech quality at the start of the project is specified (van Bezooijen and Pols, 1987; van Bezooijen, 1988). During the run time of the project, subjective tests will be performed regularly to evaluate the progress, but even more so to derive information about how to proceed. At the completion of the project a final test will be performed to measure the improvement and to compare, if possible, the results with similar systems in other languages.

A somewhat similar approach is followed in the ESPRIT project SPIN (Speech Interface at Office Workstation) in as far as rule synthesis for French and Italian is concerned (Pols et al., 1987; van Son et al., 1988). In ESPRIT project SAM (Mulitilingual speech input/output assessment, methodology, and standardization) (SAM-partners, 1988) the methodologies for evaluating speech recognizers and speech synthesizers will be further developed and, whenever possible, standardized.

Of course one must realize that, so far, most tests for evaluating the speech quality of text-to-speech synthesizers operate at the segmental level only. We will review those segmental tests, but we will also indicate how tests at the supra-segmental level are going to be developed for sentence intelligibility, global speech quality judgement, and prosodic evaluation.

Another level of evaluation and testing of course involves linguistic processing where results on paper generally suffice to indicate the performance, such as text preprocessing, syntactic analysis, or morphological decomposition. However, even here an acoustic realization and a listening test are sometimes required, for instance to find out whether an incorrect segmentation of a word in morphological components will nevertheless result in a correct pronunciation.

2. SHORT OVERVIEW OF TEST METHODS FOR SYNTHESIS EVALUATION

2.1. Purpose

Once the purpose of a test is identified it will also be easier to choose the appropriate speech material and the method of evaluation. I would like to distinguish four different purposes for developing speech quality tests:

- global testing
- diagnostic testing
- objective testing
- application-orientated testing

Global tests are mainly executed to describe and compare system characteristics in general terms, whether or not in comparison with a reference system or a competing system. The frequently used Mean Opinion Score (MOS) in telecommunication (Goodman

and Nash, 1984) is the ultimate example of this, but also a preference judgement by paired comparison, or a magnitude estimation on an 'acceptability' or a 'naturalness' scale are examples belonging to this category.

Diagnostic tests are performed with specific aims in mind and require a careful choice of the test material. An intelligibility test at the segmental level requires an approach totally different from an acceptability judgement about a number of different algorithms to generate prosodic contours in long sentences.

Objective tests, implying the use of physical means without using listener judgements, are presently virtually non-existent in speech synthesis evaluation. However, in evaluating the performance of analog and digital speech communication channels this approach is quite common. I just have to refer to the Articulation Index (AI) (Kryter, 1962), the Speech Transmission Index (STI) (Steeneken and Houtgast, 1980), or the signal-to-noise measure used in coding evaluation. The ESPRIT-SAM project intends to start research in this area of objective synthesis testing.

Application-specific tests represent again a different line of performance evaluation. Most laboratory conditions will then have to be abandoned; it frequently implies the use of task-specific test material, naive, untrained listeners, probably noisy, reverberant listening conditions, perhaps interaction in a dialog-type application with or without automatic speech recognition, etc. Examples so far are scarce (Hampshire et al., 1982)

2.2. Test Method

From telecommunication testing, psycholinguistics, psychoacoustics, speech audiometry, speech perception, language acquisition, and probably other areas, we have good knowledge about a great variety of subjective test methods. I give a brief overview here only.

Segmental intelligibility method. This method involves the phonemic level but is generally measured at the word level by using simple syllable or word forms of the type CV, CVC, or VCV (V=vowel , C=consonant). Well-known examples of this method are the Modified Rhyme Test (MRT) (House et al., 1965), the Diagnostic Rhyme Test (DRT) (Voiers, 1977), and the use of Phonetically- Balanced (PB) CVC words. There are many aspects of these tests that require careful consideration, e.g. :

- the use of words in isolation, or in a (fixed or variable) carrier phrase
- the use of closed (MRT 6 alternatives, DRT 2 alternatives) or open response sets
- the word type (e.g. CV, CVC, or VCCV)
- the use of meaningful or nonsense words
- equal phoneme probability or phonetically balanced
- language dependence, especially relevant for rhyme tests.

Supra-segmental intelligibility requires more complex test stimuli, such as word stress, word duration, syllable structure, sentence accent, and intonation. From speech audiometry, sets of carefully designed short sentences in various languages are available. However,

for synthesis evaluation these sentences are less appropriate because the set is fixed and sentences are easily remembered, whereas also the grammatical structure is too simple and with insufficient variation. In the next paragraph we will discuss some alternative structures for this sentence material.

Paired comparison allows for a direct judgement of pairs of stimuli that only differ in one specific attribute, such as duration rules or intonation contours.

Magnitude estimation involves the judgement of stimuli according to one or more attributes along, say a seven-point scale. Semantic scaling theory can be applied to process this type of data.

Psycholinguistic tests are also used sometimes to evaluate the quality of synthetic speech. Some examples are

- word recall in fixed or free order
- lexical decision (word vs. non-word; sentence vs. non-sentence)
- word monitoring
- phoneme monitoring
- word gating

Speech interference test are tests in which word or sentence intelligibility is measured against a level of masking noise (Nakatani and Dukes, 1973). Sorin (1982/83) calls this the Equivalent Signal-to-Noise Ratio method (ESNR) in her study about the contribution of pitch contours to the identification of resynthesized sentences. Other similar approaches are the speech reception threshold SRT (Plomp and Mimpen, 1979), and the monosyllabic adaptive interference test MASIT (Eggen, 1988).

Subjective ratings and questionnaires can be used to evaluate the 'linguistic' and 'psychological' aspects of speech understanding: can the sentences be reproduced, how large is the memory load, can one listen to synthetic speech for extended periods of time, can one reproduce the gist of a story, and what about the surface properties, can one comprehend the prose, do children have more difficulty with synthetic speech? Dave Pisoni and his co-workers at Indiana University certainly have most experience with this type of testing, although it is still in its infancy.

2.3. Test Material

From the short overview of the test methods given above it will be clear that these various tests use a great variety of speech material, ranging from syllables and words to sentences and paragraphs. Above (under 'segmental intelligibility') we have given already some characteristics of word material, here we will concentrate on sentence material.

Phonetically-balanced short, simple, and meaningful sentences have been developed for English (Egan, 1948), French (Combescure, 1981), and Dutch (Plomp and Mimpen, 1979). The English ones became known as the Harvard Psychoacoustic Sentences (Example: Cook the corn in a large pot of water). In order to lower the predictability and in order to make them more difficult to remember with repeated presentation, Nye and Gaitenby

(1974) developed syntactically correct, but semantically anomalous sentences of the type 'The late voice knew the table'. These sentences were called the Haskins sentences.

Pisoni and colleagues have used both types of sentences repeatedly to measure the word recognition in sentence context for various synthesis systems. For an overview, see Pols (1987).

For sentence verification tasks, 3- and 6-word sentences have been used, such as 'Mud is dirty' and 'Birds fly south for the winter', representing true sentences, and 'Rockets move slowly' and 'Beer is a popular contact sport' representing false sentences. Both a true-false reaction and a transcription were required from the subjects (Manous et al., 1985).

Various partners in the ESPRIT-SAM project have recently started a renewed discussion on the structure of sentence test material for synthesis evaluation. The idea of anomalous sentences is attractive since:

- it is a far more natural task than nonsense word identification although it is of course no real language communication either.
- it hopefully allows for controlled predictability.
- it allows for controlled complexity, for instance in terms of number of words per sentence, number of syllables per word, word frequency, grammatical structure, etc..
- it creates a very large and always different reservoir of sentences by starting from a (fixed) vocabulary from which words are randomly selected to create specific grammatical structures.
- it might be possible to develop really comparable sentences in different languages, at least in terms of word type and grammatical structure.

Presently, within SAM, we consider five different grammatical structures, instead of just one as in the Haskins sentences. Each grammatical structure will also require a different intonation contour, so, also in that respect we can run a more thorough test. Since presently none of the rule synthesizers is able to use semantic knowledge, it does not matter that the sentences are meaningless. Because of memory overload for the listener we probably will have to limit the number of words per sentence to seven.

One must keep in mind the purpose of the sentence material discussed here: evaluating word intelligibility in sentence context. So it would not be very appropriate to start studying phoneme confusions from the misidentified words. On the other hand the sentence intelligibility for real meaningful sentences will be higher than for these anomalous sentences because of semantic and pragmatic knowledge that normally can be effectively used by the listener.

It is interesting to realize that once this sentence material will be fully developed, it probably means that this test method is ahead of rule synthesizer development itself, since I do not know yet of any text-to-speech synthesis system able to extract from text, and able to generate, a number of different and appropriate prosodic realizations. This situation is contrary to that for speech recognition, where already connected word and continuous speech recognizers are available, at least as laboratory prototypes, whereas no evaluation methods at that level are available yet.

3. SOME EXAMPLES OF SYSTEMATIC EVALUATION

3.1. Segmental Intelligibility

None of the presently available rule synthesizers, whether they are diphone-based or allophone-based, have such a good segmental quality that one could further neglect this level and concentrate completely on higher level processing. All present systems will gain speech quality by improving segmental intelligibility. This was true for every system that we evaluated so far:

- the dyadic rule synthesizer (Olive, 1980). By systematic evaluation and subsequent improvement of a great number of CV and VC dyads, both the initial (58.2%) and final consonant (73.5%) intelligibility could be raised to 83% (Pols and Olive, 1983).

- the phoneme intelligibility scores for various diphone-based synthesis systems in several different languages (French, Dutch, Italian) all show room for further improvement (Pols et al., 1987; van Bezooijen and Pols, 1987; van Sol et al., 1988). The absolute scores (see Table 1) are not really important since these strongly depend upon the exact experimental conditions (such as word structure (CVC vs. VCCV and CVVC), and presentation rate), but also the listeners, specific characteristics of the synthesizer (such as prediction order, window size, and band-width) and the complexity of the language. But as long as the intelligibility scores for rule synthesis are quite a bit lower than those for the same words resynthesized, one knows that further progress can be made. More specifically, one of the synthesizers required improvement of r-diphones, whereas for certain consonant clusters it might be better to use tri-phones or quadro-phones.

- in an interactive process the segmental intelligibility of the Dutch allophone-based system will be improved step by step. The initial intelligibility was unacceptably low (van Bezooijen and Pols, 1987), but by modifying the rules and by running small specific tests, for instance for medial plosives only, the system will gradually improve.

Considering the overall consonant error rates reported for DECtalk (13.2 and 17.5% for Paul and Betty, respectively), while using the modified rhyme test with an open response set (Logan et al., 1985), I am almost certain that even this system would benefit substantially from further improvements at the segmental level.

Both for a French (van Son and Pols, 1988) and for two Dutch systems (van Bezooijen, 1988), the intelligibility of consonant clusters is almost unlimited, so the test was restricted here to initial and final clusters. However, for French medial (within-word) clusters were taken into account. See Table 2 for some overall results. These data still have to be studied in more detail in order to specify in which way the necessary improvements can be made more effectively.

3.2. Supra-Segmental Intelligibility

Relatively few results have so far been achieved with this level of speech quality evaluation. Greene et al. (1984) used the Harvard and Haskins sentences to evaluate

Table 1. Phoneme and word intelligibility scores (averaged over 8 subjects) for VCCV and CVVC words, for PCM speech, LPC-resynthesized speech and Italian rule-synthesized speech.

	V	C	C	V	VCCV
PCM	89.1%	86.9%	94.0%	79.7%	57.8%
LPC-15 resynth.	90.2%	79.4%	91.2%	79.9%	52.5%
Italian rule synth.	89.5%	68.3%	78.1%	87.8%	45.0%
	C	V	V	C	CVVC
PCM	94.3%	90.4%	84.1%	88.2%	63.2%
LPC-15 resynth.	88.4%	90.8%	85.3%	87.5%	60.2%
Italian rule synth.	74.4%	86.5%	84.9%	76.4%	44.4%

Table 2. Some overall intelligibility results for initial, medial, and final consonant clusters for French. Scores are averages for 8 subjects.

	PCM	LPC-resynth.	rule-synth.
Initial clusters(72)	92.0%	86.7%	62.5%
Medial clusters(70)	85.8%	84.5%	76.6%
Final clusters(48)	98.2%	96.3%	70.7%

Table 3. Word intelligibility scores in 'Haskins-type' sentences for natural and synthetic speech.

	natural	synthetic	
Nye and Gaitenby (1974)	95%	78%	Haskins lab. system
Pisoni and Hunnicutt (1980)	97.7%	78.7%	MITalk-79
Greene et al. (1984)	97.7%	86.8 / 75.1%	Paul / Betty DECtalk
Manous et al. (1984)	97.7%	64.0%	Speech Plus
Hazan and Grice (1988)	98.1%	76.6%	JSRU synth.-by-rule

DECtalk (two voices: Paul and Betty). The same did Manous et al. (1984) for Speech Plus Prose-2000 prototype. In 1980, Pisoni and Hunnicutt had already done this for MITalk-79. Very recently, Hazard and Grice (1988) ran a pilot test with newly developed English sentences with the same grammatical structures as the Haskins sentences: 'The ADJ $NOUN_1$ VERB the $NOUN_2$'. Table 3 summarizes the results of these various studies. It will be possible to do more interesting tests as soon as sentences with several different grammatical structures become available; these will require different prosodic characteristics and will introduce more variations for the listeners.

3.3. Quality Judgement of Intonational Aspects in Speech

In recent attempts to improve substantially the prosodic characteristics of rule-synthesized speech, Terken systematically studied natural speech and came up with better rules for intonation. These were evaluated by listening experiments with rule-synthesized diphone speech (Collier and Terken, 1987). For French, and meanwhile for several other languages as well, a set of 20 sentences were created. These sentences, in principle, should allow for testing several text-to-speech modules such as phonetic rules, diphone catenation, and prosodic processing (SAM Extension phase report, 1988). The corpus contains simple as well as complex sentences, with words of various complexity in terms of length, stress, affix structures, morphological struc- ture, phoneme realization, etc.

3.4. Quality Judgement of Prosodic Analysis from Text

Kager and Quene (1987) are developing an algorithm that, directly from Dutch text, derives pause locations and can indicate which words should get sentence accent. A first performance check was done by comparison with actual realizations of a specific speaker. However, a better check would be to run listening experiments on acceptability in order to study perceptual tolerance. These experiments are presently in preparation.

4. CONCLUSION

Although phoneme and word intelligibility of most rule synthesis systems is not yet good enough, there is a growing need for intelligibility and acceptability tests at the sentence level. The use of unpredictable, anomalous, short and rather simple, sentences seems to be a good choice at the intelligibility level. Grammatically more complex and longer sentences are generally required for naturalness and acceptability judgements. Only multilingual standardization will allow for comparison of performance figures.

References

1. J. Allen, M. S. Hunnicutt and D. H. Klatt: "From text to speech," The MITalk system, Cambridge Univ. Press, p.216, 1987.

2. R. van Bezooijen: "Evaluation of the quality of consonant clusters in two synthesis systems for Dutch," *Proc. SPEECH '88, 7th FASE Symp.*, 2, pp.445-452, 1988.

3. R. van Bezooijen and L. C. W. Pols: "Evaluation of two synth- esis-by-rule systems for Dutch," *Proc. Eurospeech.*, vol. 1, pp.183-186, 1987.

4. R. Collier and J. Terken: "Intonation by rule in text-to-speech applications," *Proc. Eurospeech.*, vol. 2, pp.165-168, 1987.

5. P. Combescure: "20 listes de dix phrases phonétiquement equilibrées," *Revue d'Acoustique*, 56, pp.34-38, 1981.

6. J. P. Egan: "Articulation testing methods," *Laryngoscope*, 58, pp.955-991, 1948.

7. B. Eggen: "Evaluation of speech communication quality with a Monosyllabic Adaptive Speech Interference Test," *Speech Comm.*, 1988.

8. B. G. Greene, L. M. Manous and D. B. Pisoni: "Perceptual evaluation of DECtalk: A final report on Version 1.8," *Research on Speech Perc.*, Progress Report 10, Indiana Univ., pp.77-127, 1984.

9. D. J. Goodman and R. D. Nash: "Subjective quality of the same speech transmission conditions in seven different countries," *IEEE Trans. Comm.*, 30, pp.642-654, 1984.

10. B. Hampshire, J. Ruden, R. Carlson and B. Granström: "Evaluation of centrally produced and distributed synthetic speech," *STL-QPSR*, 2-3, pp.18-23, 1982.

11. V. Hazan and M. Grice: "Intelligibility tests for the assessment of synthetic speech using semantically anomalous sentences," *SAM-report*, University College London, 1988.

12. A. S. House, C. E. Williams, M. H. L. Hecker and K. D. Kryter: "Articulation testing methods: Consonantal differentiation with a closed response set," *J. Acoust. Soc. Amer.*, 37, pp.153 -166, 1965.

13. D. H. Klatt: "Review of test-to-speech conversion for English," *J. Acoust. Soc. Amer.*, 82, pp.737-793, 1987.

14. K. D. Kryter: "Methods for calculation and use of the articulation index," *J. Acoust. Soc. Amer.*, 34, pp.1689-1697, 1962.

15. J. S. Logan, D. B. Pisoni and B. G. Greene: "Measuring the segmental intelligibility of synthetic speech: Results from eight text-to-speech systems," *Research on Speech Perc., Progress Report*, 11, Indiana Univ., pp.3-31, 1985.

16. L. M. Manous, B. G. Greene, and D. B. Pisoni: "Evaluation of Prose-The Speech Plus Text-to-Speech System. I. Phoneme intelligibility and word recognition in meaningful sentences," *Speech Research Lab. Technical Note*, 84-04, Indiana Univ., 1984.

17. L. M. Manous, D. B. Pisoni, M. J. Dedina and H. C. Nusbaum: "Comprehension of natural and synthetic speech using a sentence verificaion task," *Research on Speech Perc., Progress Report*, 11, Indiana Univ., pp.33-57, 1985.

18. L. H. Nakatani and K. D. Dukes: "A sensitive test of speech communication quality," *J. Acoust. Soc. Amer.*, 53, pp.1083-1092, 1973.

19. P. W. Nye and J. H. Gaitenby: "The intelligibility of synthetic mono-syllabic words in short syntactically normal sentences," Haskins Labs. SR-37/38, pp.169-190, 1974.

20. J. P. Olive: "A scheme for concatenating units for speech synthesis," *Proc. ICASSP-80*, pp.568-571, 1980.

21. D. B. Pisoni: "Perception of speech: The human listener as a cognitive interface," *Speech Techn.*, vol. 1, No. 2, pp.10-23, 1982.

22. D. B. Pisoni and S. Hunnicutt: "Perceptual evaluation of MI-Talk: The MIT unrestricted text-to-speech system," *Proc. ICASSP-80*, pp.572-575, 1980.

23. R. Plomp and A. M. Mimpen: "Improving the reliability of testing the speech reception threshold for sentences," *Audiology*, 8, pp.43-52, 1979.

24. L. C. W. Pols: "Quality evaluation of text-to-speech synthesis systems," *Deliverable of ESPRIT-project 1541 SAM, also IFA-report*, No. 94, p.31.

25. L. C. W. Pols: "Joint Dutch research program for developing a high quality text-to-speech synthesis system," *ASA/ASJ Meeting*, 1988.

26. L. C. W. Pols, J. P. Lefevre, G. W. Boxelaar and N. van Son: "Word intelligibility of a rule synthesis system for French," *Proc. Eurospeech*, vol. 1, pp.179-182, 1987.

27. L. C. W. Pols and J. P. Olive: "Intelligibility of consonants in CVC utterances produced by dyadic rule synthesis," *Speech Comm.*, 2, pp.3-13, 1983.

28. L. R. Pratt: "Quantifying the performance of text-to-speech synthesizers," *Speech Techn.*, vol. 3, No. 4, pp.54-64, 1987.

29. SAM-partners, "Multilingual speech assessment methods (SAM)," *Proc. SPEECH '88, 7th FASE Symp.*, 1, pp.137-143, 1988.

30. C. Sorin: "Evaluation de la contribution de F0 á l'intelligibilitè," *Recherches Acoustiques, CNET*, vol. 7, pp.141-155, 1982/83.

31. N. van Son, L. C. W. Pols, S. Sandri and P. L. Salza: "First quality evaluation of a diphone-based speech synthesis system for Italian," *Proc. SPEECH '88, 7th FASE Symp.*, 2, pp.429-436, 1988.

32. H. J. M. Steeneken: "Ontwikkeling en toetsing van een nederlands-talige diagnostische rijmtest voor het testen van spraakkommunikatiekanalen," *TNO Inst. for Perception, report IZF*, 1982-13, p.30, 1982.

33. H. J. M. Steeneken and T. Houtgast: "A physical method for measuring speech-transmission quality," *J. Acoust. Soc, Amer.*, 67, pp.318-326, 1980.

34. W. D. Voiers: "Diagnostic evaluation of speech intelligibility," In: M. Hawley (Ed.), Speech intelligibility and speaker recognition, Dowden, Hutchinson and Ross, Stroudsburg, pp.374-387, 1977.

This Page Intentionally Left Blank

Chapter 10

SPEECH DATABASE

This Page Intentionally Left Blank

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Considerations on a Common Speech Database*

Shuichi Itahashi

Institute of Information Sciences and Electronics, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305 Japan

Abstract

This paper discusses various aspects which should be considered when speech databases are created. It includes choosing the word sets and speakers, presentation of text and utterance timing, recording medium, microphone, editing, and labeling. It also mentions utilization of speech databases, i.e., recognition performance indeces, choosing the vocabulary subsets, and controlled distribution of the database.

1. INTRODUCTION

As information processing technology develops, the associated input/output modalities have changed from being totally dependent on the characteristics of the machine to accomodating the characteristics of human beings. Speech is the principal human input/output modality.

It was in about 1960 when we first began to use synthetic speech as an output modality. It was in 1972 when speech recognition devices were first commercialized. However, automatic recognition of continuous speech uttered by an unknown speaker and the synthesis of continuous speech with natural voice quality remain to be developed in the future.

To promote speech processing studies, a lot of speech data of various kinds spoken by many people are required; to develop speech processing systems, it is necessary to compare and estimate the performance of various analyses, syntheses, and recognition methods. The best way to do so, known today, is to analyze, synthesize, and recognize common speech data according to each method and compare the results. A collection of speech data used for this purpose is generally called a speech database or specch corpus.

2. PROGRESS OF SPEECH DATABASE WORK IN JAPAN

The necessity of common speech data has been pointed out, but it took a long time to realize such a data corpus in Japan.

*A revised version of a paper in Preprints of the First Symposium on Advanced Man-Machine Interface Through Spoken Language, Tokyo, (Jan. 1988).

ETL (Electrotechnical Laboratory, Agency of Industrial Science and Technology, Ministry of International Trade and Industry) initiated research on a speech database in Japan [1]. It developed a speech database with labeling on each subphonemic unit [2]. Tohoku University keeps discrete word utterances on optical discs [3].

A working group of 15 persons coming from various Japanese research institutes and private companies has been involved in the development of a speech corpus for common use. This work has been supported by JEIDA (Japan Electronic Industry Development Association). Their efforts resulted in the JEIDA Japanese Common Speech Data Corpus. The corpus is composed of 323 items uttered by 75 male and 75 female speakers. All items are uttered four times by each speaker, producing 193,800 samples in all, contained on 68 video cassettes [4-8]. The data corpus has been distributed to 50 organizations involved in speech research in Japan. The corpus has been transferred recently to 76 DAT cassettes.

ATR Interpreting Telephony Research Laboratories started developing speech databases in 1986. They plan to include 5000 important words as well as telephone conversational speech spoken by professional news announcers. The data are labeled by phonemes [9].

A research project on "Advanced Man-Machine Interface through Spoken Language" was started in 1987 as one of the priority areas supported by the newly created Grant-in-Aid for Scientific Research from the Ministry of Education, Science and Culture of Japan. The project considered a speech database as one of the impotant research areas. Continuous speech data were collected [10].

Another priority area research project on "Prosodic Features of Spoken Japanese" was started in 1989. The project aims to collect samples of various Japanese dialects and create speech databases which are expected to be useful for speech research and education. A compact disc which contains the famous Japanese folk tale "Peach boy" uttered in standard Japanese and in various dialects, and a weather forecast uttered in standard Japanese was produced. A TV announcer and speakers from 20 dialectal districts utter 61 items in all.

Trends in other countries can be found elsewhere [11-14].

3. OUTLINE OF SPEECH DATABASES

A speech database would be utilized for speech synthesis and speaker identification as well as speech recognition. Speech synthesis can be divided into two major sub-areas: synthesis-by-rule or text-to-speech synthesis and the analysis-synthesis method. The former accepts as its input written text and it is not directly concerned with the speech database. The latter needs to rely on a common speech database for its performance evaluation. In this case, the content of the database should be organized so that it is suitable for analysis-synthesis system evaluation. It would be possible to use a speech database which is prepared for speech recognition, at least in part for that purpose. Speaker recognition involves two major areas, i. e. that which is performed irrespective of the utterance content and that which uses predetermined key words. It seems that a speech database for speech recognition can be used for both areas, at least in part. In the following, a speech database for speech analysis and recognition will be described.

A speech database is necessary and important for the following two reasons. First, speech researchers need to evaluate various speech analysis methods and speech recog-

nition algorithms in order to develop better ones. Second, potential users of speech recognizers need to evaluate the performance of available speech recognizers in order to choose the most suitable one. Such evaluation entails the availability of a speech database for common use.

The following is required for a speech database to be easy-to-use and valuable. The database must be unbiased in the sense that the utterance text, the speakers, the recording conditions, etc. must be comprehensive; it must contain enough speech samples. Speech databases for recognizer performance assessment may be divided into two categories: those for speaker-dependent recognizers and those for speaker-independent recognizers. The database for the former use must contain two or more tokens per item, whereas one token per item would suffice for the latter.

4. CREATING SPEECH DATABASES

As is well known, there are several levels of speech utterance modes.

(1) Syllablewise utterance.

(2) Discrete words: digits, city names, family names, basic words, etc.

(3) Connected words: telephone numbers, etc.

(4) Sentence and a collection of sentences.

There could be other utterance modes such as phrasewise speech and continuous speech. Let us concentrate on item (2) in this section.

4.1. Choosing the Word Sets

There are alternative ways of selecting the sets of words to be used for database purposes. The following three ways are typical with regard to their practical use for speech recognition and their familiarity to many people.

(1) Last names: For example, we can select those last names which have a high frequency of occurrence in Japan, such as Tanaka and Sato.

(2) Place names: Mainly we can include the 47 prefecture names, 646 city names and the names of the 23 wards in Tokyo. By elimination of duplicate entries, the total is 676 place names.

(3) Frequently used applications vocabulary: There are two main examples of this category.

   (a) Bank services: Besides the digits, the following words and phrases are used frequently: Yes, No, Begin, End, Repeat, Correction, Please, etc.

   (b) Word processing: If we consider the case of spoken control operations, we need several dozen control words. These may include: Transfer, Insert, Delete, Correction, New Line, New Page, Left (open) parenthesis, Right (close) parenthesis, Space, Frame line, Word enrollment, Execute, Cancel, etc.

If we compare the frequency of occurrence of the items of (1), (2), and (3) in novels and newspapers, we find many syllables that do not occur in (1), composed of last names. In that sense, place names are better for a database. Frequently used applications vocabularies (3) are used in practical applications and are especially suitable to evaluate speaker independent speech recognition. However, the words used in each system are not necessarily common, and that makes it difficult to choose a fairly standardized word set, especially for word processors. Further discussion is needed about word processors concerning how to select the phrases or words for the function.

4.2. Choosing the Speakers

The database should be varied yet comprehensive; therefore the speakers have to be from several categories (not all young, not all old, etc.). The following categories should be represented:

(1) Sex: We need both male and female speakers.

(2) Age: The data should include speakers in their 20's, 30's, 40's, 50's, and 60's.

(3) Occupation: It should be varied.

(4) Region: Place of birth and the place where the speakers lived prior to the age of 12 should be varied.

(5) Standard and non-standard language: We cannnot avoid having dialect effects, but we wish to have the speech database consist of standard Japanese as a first step.

Among these items, the most important factors which make a difference in voice quality are sex and age. In order to take these factors into account, the following is suggested for the speaker dependent database: 50% male and 50% female voices.

It is desirable that the distribution of age is roughly proportional to the Japanese population statistics .

5. RECORDING AND EDITING OF SPEECH SAMPLE

5.1. Presentation of Text and Utterance Timing

There are several ways to display speech text and utterance timing.

(a) Word list and no timing display.

(b) Word list and timing through headphones.

(c) Displaying text and utterance timing on the display terminal.

Any one of these ways can be used, depending on the facilities of the recording place.

5.2. Recording Medium

Special care must be taken to determine the recording medium. There are several kinds of recording media:

(1) Analogue audio cassettes:
Analogue cassettes are most popular but they are not suitable for speech database storage, because they have many problems such as print through and sound quality deteorioration with multiple dubbing.

(2) DAT cassettes:
DAT cassettes have many advantages over analogue cassettes, including a digital dubbing facility; a weak point is the difficulty in recording label information.

(3) Compact discs (CD):
Compact discs have properties similar to DATs and the durability and ease of handling are much better. CDs are suitable for mass production, while its recording must be ordered to a professional company.

(4) CD-ROMs:
CD-ROMs have properties similar to CDs with the additional capability of storing text and label information, but they require AD/DA conversion.

(5) Optical Discs (MO):
Some types of optical discs can be produced in a laboratory, but they are not suitable for mass-production.

(6) Magnetic tape for computer use

Computer magnetic tape is suitable for speech database storage; however, the durability is not so good as imagined, if kept unused for a long time. Open reel magnetic tape has high compatibility among various computers, but cartridge tapes are mostly machine dependent, though easier to handle. Also it is desirable to use a high enough sampling rate, so that the user can choose any sampling rate s/he wants.

At present DAT, CD and CD-ROM seem to be suitable storage media for speech databases. Speech databses on DATs or CDs are especially suitable for performance assessment of speech recognizers. Utilizing a DAT/CD interface to a workstation or a personal computer makes it more useful to use DAT/CDs as speech database media. Such an interface often has a hardware down-sampling function, which is quite useful.

5.3. Microphone

Condenser microphones have good frequency characteristics, but dynamic microphones are preferred for speech recording at broadcast stations. So a dynamic microphone is recommended for the recording. Since two-channel recording is possible with DAT recorders, a condenser microphone can be used for the second channel.

5.4. Editing

The raw recorded data could be used directly as the data corpus, but it usually contains such undesirable signals as erroneous utterances, noise of shuffled paper and coughing, and, often, corrected utterances appear at the end of the recording. It must be edited and only the desired speech kept in the recording. The editing process is one of the most time consuming aspects of the speech database preparation. In order to simplify and shorten the editing process as well as to decrease the editing costs, simple editing can be performed removing noises and erroneous utterances, and inserting correct utterances on the right position. Extremely long silent intervals must also be shortened.

5.5. Labeling

Labeling speech segments is one of the significant aspects of a speech database. Although the speech data corpus itself is useful for speech research, labeing makes it far more useful. Generally the unit of labeling is chosen from such possible ones as paragraph, sentence, phrase, word, syllable, or phoneme. It is not so difficult to detect boundaries between paragraphs, sentences or phrases, since there would be breath intervals; and the labeling is easy. Word boundaries are not often clear, but it would be easier to detect those boundaries than those between syllables or phonemes. It is fairly difficult to detect syllable or phoneme boundaries and operators who perform the segmentation and labeling are not always in agreement. Moreover, it is impossible to do the process automatically using a machine. It seems plausible to divide the process into two parts. First let a machine segment and label speech automatically, as far as possible, and then an experienced operator can correct the errors. This semi- automatic process seems realistic.

Smaller segments than a phoneme could be used as a unit of segmentation and labeling. As a rule a phonemic interval can be said to be composed of three segments, these are the transition-steady-transition parts, and each segment is adopted as a unit of segmentation and labeling. This smaller unit is claimed to be acoustically more compact than a phonemic segment. However, it becomes necessary to transform ordinary phonemic transcription to this subphonemic transcription, though this process is relatively uncomplicated [2].

Another proposal is to transcribe speech in multiple ways using acoustic phonetic symbols for various data access requests and for the convenience of fine acoustic phonetic data analysis. Three types of categories are proposed for multiple transcription, i.e., linguistic and phonemic categories, acoustic event categories, and some allophonic variation categories [10].

6. OUTLINE OF THE SPEECH DATABASE OF THE "SPOKEN LANGUAGE" PROJECT

The "Spoken Language" project considered a speech database as one of the important research areas. The outline of the database is as follows.

(1) Vowels and numerals (37 items similar to the JEIDA auxiliary list) for utterance practice.

(2) 216 phonetically balanced words: Same as the ATR list.

(3) 110 monosyllables: Same as the JEIDA list.

(4) Continuous speech:

(a) University of Tsukuba list (70 short sentences): The sentence head word is composed of two to four syllables including all possible accent types (5 min.).

(b) Interrogative sentences (11 items): Same as the JEIDA auxiliary list (45 sec.).

(c) Sentences for speech quality test (7 sec.): Sentences composed of those syllables chosen from a set of 20 syllables with high occurrence in daily conversation (1 min.).

(d) Story: Aesop's fable "North wind" (50 sec.).

(e) Weather report: Containing basic words (50 sec.).

(f) Narrative sentences: For intonation studies (1: 30 sec.; 2: 40 sec.).

Quantity of the data:

(a) Four utterances per item.

(b) Two hours of speech per person.

(c) 10 male and 10 female speakers: 20 to 60 years of age.

The speech data has been recorded on video cassette tapes with PCM recording, which has been carried out using a common dynamic microphone in a sound proof room or an ordinary quiet room. The speech data has been converted to DAT recently.

7. UTILIZING THE SPEECH DATABASE

7.1. Performance Indexes

It is necessary to define suitable performance indexes in order to compare the results of performance evaluations based on the speech database. There are many ways to define performance comparison indexes. However, among these, the most important ones are the following: recognition accuracy, recognition time, and the variability of performance data for different speakers. Performances within and beyond the limits of the vocabulary should be defined separately.

7.1.1. Basic Recognition Performance Indexes:

(1) For discrete words:

Index A: Performance indexes within the vocabulary

Correct response rate:	(C1/Ta)*100	(1)
N-rank correct response rate:	(Cn/Ta)*100	(2)
Error rate:	((Ta-C1-Ra)/Ta)*100	(3)
N-rank error rate:	((Ta-Cn-Ra)/Ta)*100	(4)
Rejection rate:	(Ra/Ta)*100	(5)
Recognition rate:	(C1/(Ta-Ra))*100	(6)
Error recognition rate:	((Ta-C1-Ra)/(Ta-Ra))*100	(7)

Index B: Performance indexes beyond the vocabulary:

Correct rejection rate:	(Rb/Tb)*100	(8)
False acceptance rate:	((Tb-Rb)/Tb)*100	(9)

where,
Ta, Tb: the number of test samples in Index A and B, respectively,
Cn: the number of correct responses within rank n,
Ra, Rb: the number of rejected samples in Index A (undesirable) and in Index B (desirable), respectively.

(2) For connected word utterances:

Basically, it is necessary to compute the performance ratios as for isolated words, but using the connected series of words as a string (treated as one unit). This amounts to computing string error rates; we must take deletions and insertions as mistakes. The following figure is suggested as "word-wise recognition rate", including deletions and insertions:

Word recognition rate = ((C1-D)/(Ta+I-Ra))*100

where,

D: the number of deletions,

I: the number of insertions.

For both isolated and connected words, the setting of the rejection parameter is important. If it can be set such that there are no rejections (Ra=0), then we do so, and we use the figure n=1 as a comparative figure. Other procedures will not be useful for comparative purposes.

7.1.2. Recognition Time

The recognition time is to be measured after the speech utterance, and before we get output from the recognizer. We cite the average figure for comparative purposes, and also the largest and smallest figures as well as the variance.

7.1.3. Recognition Data for Speakers

Usually is cited the performance data averaged over several speakers for comparison purposes, as well as the largest and smallest figures to indicate the range and the variance. Besides the measures described above, we can consider other comparisons, such as those that deal with the effects of noise or the influence of level fluctuations. However, these are secondary items, discussed later.

7.2. Choosing the Vocabulary Subset

Assuming that the database user wants to use a partial database, there are several alternative procedures for choosing subsets:

(a) Choosing the vocabulary subset according to a particular rank order, regardless of the character of the speech. For example, by place location, north to south, or by size of the population, large to small.

(b) Choosing the subset by the length of the word. For example, by the number of phonemes or the number of syllables.

(c) Choosing the subset by considering the frequency of occurrence of the phonemes or syllables. For example, (c1) seek to distribute the frequency of phonemes or syllables uniformly or (c2) provide the same distribution as in a large vocabulary set.

(d) Choosing the subset by clustering based on the distance between words. For example, creating a cluster of similar vocabulary words based on the distance measured when using DTW matching of speech patterns, or on the Hamming distance of a series of phonemes.

The choice of the most appropriate procedure will depend on the purpose of the use of the database. For example, (a) and (c) are suitable for small vocabulary subsets, and (b) and (d) are suitable for subsets which have similar vocabulary, and (c) is considered applicable in the case of basing the recognition algorithm on units of phonemes.

In procedure (c1), the numbers are determined based on a criterion for the information entropy of the occurrence frequency of the syllables. To keep the entropy as large as possible, we eliminate words, one after another, from the 676 word vocabulary. The first eliminated word is set as that with number 676, and the last to be eliminated would be number one [15].

7.3. Choosing the Speaker Subset

Another consideration is the preferred procedure for choosing the speaker subset. Attributes of the speakers are listed on the speaker's card. We choose a subset based on considerations of the speaker's age and sex; the details are to be discussed later.

8. CONTROLLED DISTRIBUTION OF THE DATABASE

Currently the speech databases are maintained on DAT tape. In the future they will be transferred to compact discs and digitized for CD-ROMs.

(1) Care of the tapes: Several submaster tapes are prepared in addition to the set of master tapes. The primary storage location for the master tapes is considered to be in Tsukuba, but submaster tapes and tapes to be circulated are stored in several places (Tokyo, Osaka, etc.) for convenience. It is necessary to consider the matter of deterioration of the recorded material, but we still feel that it will be adequately durable.

(2) Lending: In the case of lending the tapes, it is necessary to have some rules to have smooth circulation of the database.

(3) Other: It is desirable to establish some committee responsible for overall management and control.

9. FUTURE PROSPECTS

Up to now, we have focused mainly on isolated-word speech and have also considered some connected digits. In the future we should look more at connected speech; also research is needed concerning speech over telephone lines and in noise.

Three years ago, when we started this research, DAT processors were not available, but they are now readily available, and they have become a very strong recording medium for speech databases. The compact audio disc, which has noncontacting playback, is even more promising. Currently there are some problems as the difficulty of recording and the associated costs. However, it seems to be a promising recording medium. CD-ROMs will be more widely use soon.

We could not fully examine speech synthesis. As speech response units become more widely used, promoting research regarding the standardization of performance comparisons for them becomes necessary.

10. CONCLUSION

Problems in creating common speech databases have been discussed. Specifications for speech databases were presented. A database has been partially realized. There remain many tasks to be done, as shown below.

(1) Expansion of the specifications:

  (a) Increase in number of speakers, tokens and words.

  (b) Choice of continuous speech texts and data collection.

(2) Consideration of environmental conditions.

  (a) Environmental noise.

(b) Telephone line speech.

(c) Stress of the recognizer users.

(3) Improvement of the process of data collection

(a) Presentation of the utterance text.

(b) Recording medium choice.

(c) Microphone placement to remove respiratory noise.

(4) Standardization concerning speech synthesis.

(5) Semi-automatization of the editing and labelling process.

(6) Accessibility of the speech database.

Speech database efforts in Japan have hitherto concentrated on the word level; continuous speech databases need to be prepared.
Recently, the Committee for the Investigation of Continuous Speech Databases of the Acoustical Society of Japan was established in 1990. The members are mostly from universities with some from national research institutes, and private enterprises. The objective of the committee is to investigate design methods for continuous speech databases and to propose a task setting for the databases, choosing suitable texts and programs for actual database creation.

References

1. T. Nakajima, T. Suzuki and H. Ohmura: "Data file control system for speech research," *Tech. Group Speech, Acous. Soc. Japan*, S73-07, 1973.

2. K. Tanaka, S. Hayamizu and K. Ohta: "A demiphoneme network representation of speech and automatic labeling techniques for speech database construction," *Proc.ICASSP8* 7.1, pp.309-312, 1986.

3. J. Miwa and K. Kido: "Spoken word data collecting system," *Proc. Spring Meeting Acous. Soc. Japan*, 1-4-21, 1982.

4. JEIDA(Japan Electronic Industry Development Association): *Research Report on Standardization of Japanese Language Information Processing*, 1982, 1983, 1984, 1985.

5. JEIDA: *Research Report on Standardization of Office Automation Equipments*, 1986, 1987, 1988, 1989, 1990.

6. S. Itahashi: "Speech database of discrete words," *J. Acous. Soc. Japan*, **41**, 10, pp.723-726, 1985.

7. S. Itahashi: "A Japanese Language Speech Database," *Proc. ICASSP86*, 7.4, pp.321-324, 1986.

8. S. Itahashi: "Speech database," *J. Inst. Electron. Inform. Commun. Japan*, **70**, 4, pp.433-438, 1987.

9. K. Takeda, Y. Sagisaka and S. Katagiri: "Acoustic-phonetic labels in a Japanese speech database," *Proc. Eurospeech*, pp.13-16, 1987.

10. H. Fujisaki: "Overview of the Japanese National project on advanced man-machine interface through spoken language," *Proc. Eurospeech*, pp.1-8, 1987.

11. Proc. Workshop on Standardization for Speech I/O Technology, National Bureau of Standards, 1982.

12. J. M. Baker, D. S. Pallet and J. S. Bridle: "Speech recognition performance assessment and available databases," *Proc. ICASSP83*, pp527-530, 1983.

13. R. Carre, R. Descout, M. Eskenazi, J. Mariani and M. Rossi: "The French language speech database," *Proc. ICASSP84*, 42.10, 1984.

14. Proc. ESCA Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, the Netherlands, 1989.

15. K. Shikano: "Phonetically balanced word list based on information entropy," *Proc. Autumn Meeting Acous. Soc. Japan*, 3-3-10, 1984.

Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, H. Fujisaki (Editor)

Transcription and Alignment of the TIMIT Database

Victor W. Zue and Stephanie Seneff

Spoken Language Systems Group, Laboratory for Computer Science
Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Abstract

The TIMIT acoustic-phonetic database was designed jointly by researchers at MIT, TI and SRI. It was intended to provide a rich collection of acoustic phonetic and phonological data, to be used for basic research as well as the development and evaluation of speech recognition systems. The database consists of a total of 6,300 sentences from 639 speakers, representing over 5 hours of speech material, and was recorded by researchers at TI. This paper describes the transcription and alignment of the TIMIT database, which was performed at MIT.

1. BACKGROUND

When the DARPA Strategic Computing speech program was first formulated in 1984, the consensus of the research community was that the amount of speech data available is woefully inadequate. As a result, a significant effort on database development was mounted in order to provide the research community with a large body of acoustic data for research, system development, and performance evaluation. One such database is the so-called TIMIT acoustic-phonetic database. The TIMIT database was designed jointly by researchers at MIT, TI, and SRI. It consists of a total of 6,300 sentences from 630 speakers, representing over 5 hours of speech material, and was recorded by researchers at TI. This paper describes the transcription and alignment of the TIMIT database, which was performed by researchers at MIT.

Each speaker in the TIMIT database recorded 10 sentences drawn from three different corpora as follows. Each speaker read two sentences, designated as S1 and S2, which were designed by Jared Bernstein of SRI in order to compare dialectal and phonological variations across speakers. Five sentences, designated as SX sentences, were drawn from a small set of sentences designed at MIT. The remaining three sentences for each speaker, designated as SI sentences, were selected from the Brown corpus by Bill Fisher of TI [1].

There are a total of 450 "MIT" sentences used in the TIMIT database. These were generated by hand in an iterative fashion, with the goal that they should be phonetically rich. Care was taken to have as complete a coverage of left- and right- context for each

frequently-occurring low-level phonological rules were adequately represented. To aid in the sentence generation process, we made use of an on-line, Webster's Pocket Dictionary containing nearly 20,000 words. Words or word-sequences containing particular phone pairs could be accessed from this dictionary automatically, which greatly facilitated the database design process. We performed a detailed analysis of the resulting sentence set, as well as the SI sentences that make up the remainder of the database . The interested reader should consult Lamel et al. [3] for further information about the corpora.

2. THE ACOUSTIC PHONETIC LABEL SET

All of the recorded sentences were provided with a time- aligned sequence of acoustic-phonetic labels. The label set is intended to represent a level somewhat intermediate between phonemic and acoustic. Our motivation was that clear acoustic boundaries in the waveform should all be marked, and that the criteria for positioning the boundaries between units should in part be based on our ability to mark them consistently. Table 1 lists all of the acoustic-phonetic labels that were used. Most of these labels are phonemic, although several symbols have been included for labelling acoustically distinct allophones as well as other special acoustic events.

2.1. Stops

Stops are characterized by a sequence of two events: a closure and a release. This departure from phonemic form is, we believe, important in order to preserve a boundary marking the onset of the release. There are six closure symbols for the stops. The closure region for affricates is identical with that of the corresponding alveolar stop. (e.g., the /č/ in "chat" is represented as [$t^{\square}$č]).

There are two major allophones for the stops. The glottal stop, [ʔ], is often inserted preceding a word-initial vowel. Sometimes a /t/ can also be realized as a glottal stop, as in "co*t*ton". The symbol [ɾ] is used to label a flap, which can either be an underlying /t/ or /d/. We make a separate flapping decision for every phonemic /t/ and /d/, based on listening and the spectrographic evidence. We allow flapping to occur in environments for which theory is violated, if in fact we believe that flap is what was heard/seen.

2.2. Nasal and Semivowels

We recognize four allophones for the nasals, three of them are the syllabics, [m̩, n̩, ŋ̩]. If there is any evidence of a preceding schwa, the non-syllabic form is preferred. The alveolar nasal, /n/ can be realized as a nasal flap, denoted by the symbol [ɾ̃]. Sometimes an underlying /nt/ sequence is realized as a nasal flap, as in "*en*tertain".

The liquid, /l/, has a syllabic allophone, denoted as [l]. Again, a non-syllabic form is preferred whenever a preceding schwa is observed.

Table 1. A list of the acoustic phonetic symbols used for the transcription of the TIMIT database.

Phonetic Symbol Mapping					
IPA	Char	Notes	IPA	Char	Notes
Stops					
p	p		b	b	
t	t		d	d	
k	k		g	g	
$p^{\square}$	⊕	Symbol-+	$b^{\square}$	δ	Symbol-shift-D
$t^{\square}$	∞	Symbol-i	$d^{\square}$	†	Symbol-g
$k^{\square}$	∂	Symbol-p	$g^{\square}$	±	Symbol-:
ɾ	F		ʔ	?	
Nasals					
m	m		m̥	M	
n	n		n̥	N	
ŋ	G		ŋ̊	π	Symbol-shift-P
ɾ̃	ϵ	Symbol-shift-E			
Fricatives					
s	s		š	S	
z	z		ž	Z	
č	C		ǰ	J	
θ	T		ð	D	
f	f		v	v	
Liquids, Glides, Silence, and h					
l	l		l̥	L	
ɾ	r		w	w	
y	y				
ɬ	λ	Symbol-shift-L	ɬ̊	⊂	Symbol-t
h	h		ĥ	H	
Vowels					
ɛ	E		ɪ	I	
ɔ	c		æ	@	
ɑ	a		ʌ	^	
u	u		ʊ	U	
ɝ	R		ʉ	:	
a^{y}	Y		$ɔ^{y}$	O	
$ɛ^{y}$	e		$ɪ^{y}$	i	
a^{w}	W		o^{w}	o	
ə	x		ɚ	X	
ɨ	|		ʔ̞	ɣ	Symbol-shift-G

2.3. Vowels

Two vowels, /i o/, are represented by symbols that included their corresponding off-glides. This is because they are usually realized as diphthongs in American-English. The four diphthongs, /α^{y}/, /α^{w}/, /$\mathfrak{o}^{y}$/, and /e^{y} /, are each represented as a single label, with no separate region defined for the off-glide portion. The retroflexed vowel /$\mathfrak{z}$ / is also represented as a single unit. This represents a departure from the International Phonetic Alphabet, which would represent this steady-state vowel as the sequence /Λr /.

Reduced vowels are represented by four separate allophones: back schwa ([$\mathfrak{o}$]), front schwa ([$\bar{\imath}$]), retroflexed schwa ([$\mathfrak{r}$]), and voiceless schwa ([$\mathfrak{h}$]). The decision for [$\mathfrak{o}$] vs [$\imath$] is based on whether the second formant is closer to the first or to the third. A low third formant leads to /$\mathfrak{r}$/. Schwas can often be devoiced in words such as "secure".

English does not distinguish phonemically between the fronted vowel /ü/ and the standard back /u/; however the difference in F_2 for the two forms can be as much as 800 Hz. We felt it was unsatisfactory to group two forms with such diverse formant frequencies into the same vowel category. The decision is made as for schwa: if F_2 is closer to F_1, it's considered a back /u/. Similar trends of fronting are also observed for /o/ and /υ/ in certain environments; however, the effect is most dramatic for /u/.

At present, we make no attempt to provide further sub-phonemic characterizations for vowels other than this front/back distinction for /u/ and the four schwas. For instance, many vowels are nasalized when they are followed by a nasal, or lateralized when followed by an /l/. Such information would surely be useful, but the decision-making process is prone to judgement error, and would require a significant increase in time and effort.

2.4. Others

We make a distinction between two types of /h/: voiced ([$\hat{\mathfrak{h}}$]) and unvoiced ([h]). The decision is based mainly on an examination of the waveform for clear low-frequency periodicity, and spectrogram for voicing striations. The voiced form is most common between two vowels.

Our label set includes a category "epenthetic silence,"[], which we use to mark acoustically distinct regions of weak energy separating sounds that involve a change in voicing. These short gaps are typically due to articulatory timing errors. The most common occurrences of such gaps are between an /s/ and a semivowel or nasal, as in "small", "swift", or "prince". Two other non-phonetic symbols are included: # is used to mark regions preceding and following a sentence, and [] is used to mark pauses within a sentence.

3. CRITERIA FOR BOUNDARY ASSIGNMENTS

The acoustic-phonetic transcription for the TIMIT sentences is time aligned with the speech waveform. The alignment is useful in that specific acoustic events can be accessed conveniently based on the transcription. We must stress, however, that the aligned transcription is intended to establish a *correspondence* between the transcription and important acoustic landmarks. One should not directly associate a region between

two time markers as a distinct phonetic unit, since the encoding of phonetic information in the speech signal is extremely complicated.

In most cases, the boundaries between two acoustic-phonetic events are clear and well-defined, such as that between a stop closure and its release. However, there are a number of cases where the exact placement of a boundary is problematic (as is the case between a semivowel and a vowel), or cases where it's not clear whether a region should be represented as one or two acoustic-phonetic units (as is the case for diphthongs). In these cases, we tried to define a set of criteria that would be systematic and least subject to human error, in order to produce boundary positionings that were as consistent as possible.

As mentioned previously, we decided that the boundary between the closure interval and the release of a stop is an important one that should be assigned. It is certainly a very distinct landmark in the waveform. Anyone interested in studying the burst characteristics of a stop would then be able to focus on just that region that includes only the released portion. In a strictly phonemic representation, the closure and release would be represented as a single unit, and therefore that critical boundary would remain unmarked.

A problematic boundary is one that separates a prevocalic stop from a following semivowel, as in "*t*ruck." Typically, part of the /r/ is devoiced, and therefore is absorbed into the aspiration portion of the stop. If listening were the only criterion, then the left boundary of the /r/ would occur somewhere in the aspiration, and the right boundary would occur somewhere after voicing the onset. A clear acoustic boundary at the point of voice onset would remain unmarked. It would also be difficult to decide where to mark the boundary between the stop burst and the aspirated /r/ portion. Since voice-onset time (VOT) is a parameter that has been a focus of many research efforts, it seems unsatisfactory not to include a reliable mechanism for measuring VOT based on the labelled boundaries. Therefore, we adopted the policy of always absorbing into the stop release all of the unvoiced portion of a following vowel or semivowel.

The boundary between many semivowels and their adjacent vowels is rather ill-defined in the waveform and spectrogram, because transitions are slow and continuous. It is not possible to define a single point in time that separates the vowel from the semivowel. In such cases, we decided to adopt a simple heuristic rule, in which one-third of the vocalic region is assigned to the semivowel, thus giving the vowel twice the duration of the adjacent semivowel. Previous investigators have also made use of such consistent rules for defining acoustically ambiguous boundaries [4].

One obscure condition is a /ts/ or /dz/ sequence, where typically there is little or no spectral change to characterize a boundary between the homorganic stop and fricative, yet the onset of acoustic energy of the unit is sufficiently abrupt such that a /t/ is heard. Our convention here is that, if a clear /t/ is heard, the early portion of the /s/ is marked as a /t/ release.

When gemination occurs, we do not attempt to mark a boundary between the two units. This situation occurs exclusively at word boundaries, as in "so*me* *m*oney." Furthermore, in the case of a stop-stop sequence where the first stop is unreleased, the closure interval is assigned to the first stop and the release to the second one.

4. PROCEDURES FOR TRANSCRIPTION AND ALIGNMENT

The transcription and alignment process involves three stages:

1. An acoustic-phonetic sequence is entered manually by a transcriber as a string.
2. The speech waveform is aligned automatically with the acoustic-phonetic sequence, using an alignment program developed at MIT.
3. The boundaries generated automatically are then hand- corrected by experienced acoustic phoneticians.

4.1. Transcription

In both stages 1 and 3, the labeller makes her/his acoustic-phonetic decision based on careful listening of portions of the speech waveform, as well as visual examination using displays such as the spectrogram and the original waveform. The process takes place within the SPIRE software facility for speech analysis, a powerful interactive tool that is well-matched to this task [2]. Stage 1 requires less intensive use of SPIRE than Stage 3, because it is only necessary to record what was heard, without identifying the time locations of the events. Furthermore, minor errors of judgement made at this stage can be readily corrected in stage 3. The labels can be entered either by typing or by mousing a displayed set. Figure 1 shows the SPIRE layout used for entering the transcription. The completed transcription is shown in the top window of this display.

In general, we try to label what we hear/see, rather than what we expect. Thus, if a person says "imput" for "input", the nasal will be marked as an /m/. However, in conditions of ambiguity, the underlying phonemic form is selected preferentially.

4.2. Automatic Alignment

The alignment of a phonetic transcription with the corresponding speech waveform is essential for making use of the database in speech research, since time-aligned phonetic transcriptions provide direct access to specific phonetic events in the waveform. Traditionally, this alignment is done manually by a trained acoustic-phonetician. This is an extremely time-consuming procedure, requiring the expertise of one or a very small number of people. Therefore, the amount of data that can be labelled is limited. In addition, manual labelling often involves decisions which are highly subjective, and thus the results can vary substantially from one person to the other.

Transcription alignment of the TIMIT database makes use of CASPAR, an automatic alignment system developed at MIT. Descriptions of preliminary implementation of the system can be found elsewhere [5, 6]. Basically, the alignment is accomplished by the system in three steps. First, each 5 ms frame of the speech data is assigned to one of five broad-class labels: *sonorant, obstruent, voiced-consonant, nasal/voicebar*, and silence, using a non-parametric pattern classifier. The assignment process makes use of a binary decision tree, based on a set of acoustically motivated features. Each sequence

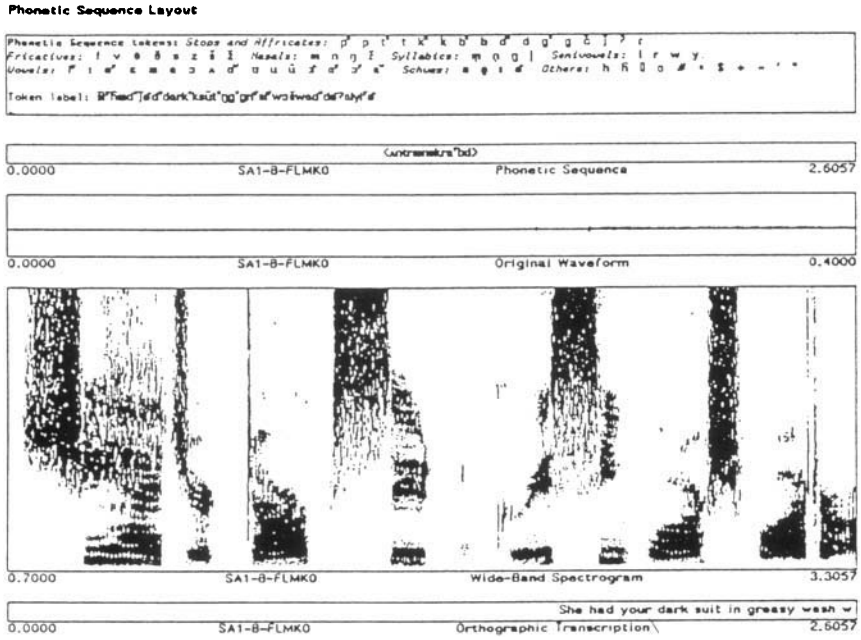


Figure 1. SPIRE layout for entering the acoustic-phonetic transcription.

of identically-labelled frames is then collapsed into a segment of the same label, thus establishing a broad-class segmentation of the speech. The output of the initial classifier is then aligned with the phonetic transcription using a search strategy with some look-ahead capability, guided by a few acoustic phonetic rules. For those segments which correspond to two or more phonetic events after preliminary alignment, further segmentation is done using specific algorithms based on knowledge of the phonetic context. In some cases heuristic rules are invoked (as between a vowel and a semivowel) to assign consistent, but somewhat arbitrary boundaries.

Over the past two years, two major modifications of CASPAR have taken place. First, the alignment of the broad-class acoustic labels with the phonetic symbols has been cast into a probabilistic framework. By using a large body of training data, a set of robust, context-dependent and durational statistics were obtained. Second, a fourth module has been added to the system to improve the resolution of the boundaries. This module com putes appropriate acoustic attributes at a high analysis rate using different window shapes that depend on the specific context. The boundaries are then adjusted on these attributes.

In a formal evaluation, it was found that CASPAR can correctly perform over 95% of the labelling task previously done by human transcribers. The boundary locations produced by the system agree well with those produced by human transcribers. For

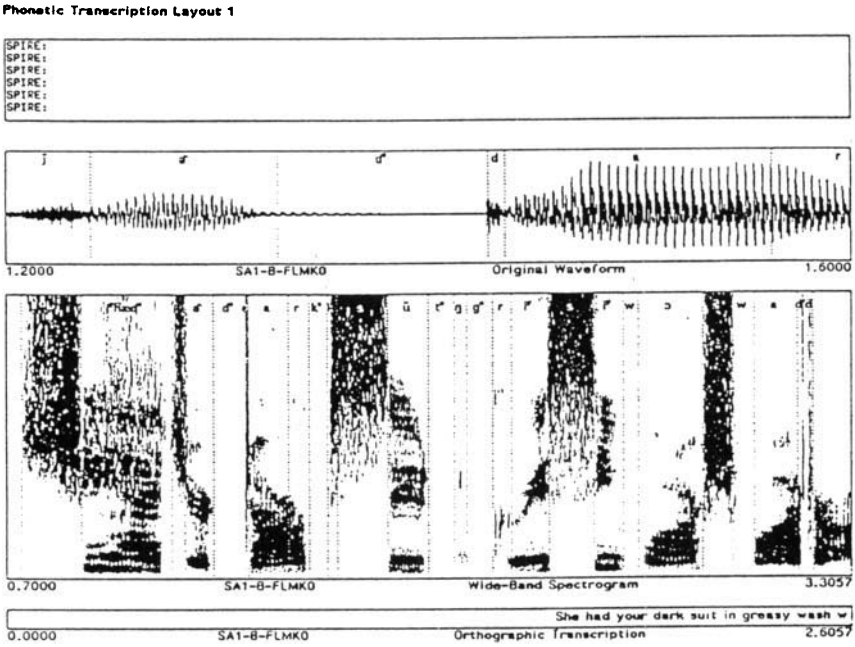


Figure 2. SPIRE layout showing the alignment produced by CASPAR.

example, over 75% of the automatically generated boundaries were within 10 msec of a boundary entered by a trained phonetician.

Figure 2 displays the output for the sentence, "She had your dark suit in greasy wash water all year." For this example, most of the boundaries have been found correctly by CASPAR. Note, however, that boundaries are missing in the [iɦæ] sequence of "She had." The waveform displays the word "dark" and the [s] of "suit." Note that the initial boundary of the first [d] is slightly too far forward in time.

4.3. Post-Processing

The final step is to correct by hand any errors in the automatically aligned acoustic-phonetic sequence. Some of the errors are due to the fact that CASPAR is not able to determine certain boundaries, such as some of those between two vowels. In other cases the boundaries may have been misplaced.

Hand correction of the aligned transcription is based on critical listening of portions of the utterance as well as visual examination of the spectrogram and the waveform. The spectrogram covers close to 3 seconds worth of speech at one time, whereas the waveform is displayed on a much more expanded time scale. For example, to accurately mark the onset of the release of a stop, the cursor is first positioned on the spectrogram at the

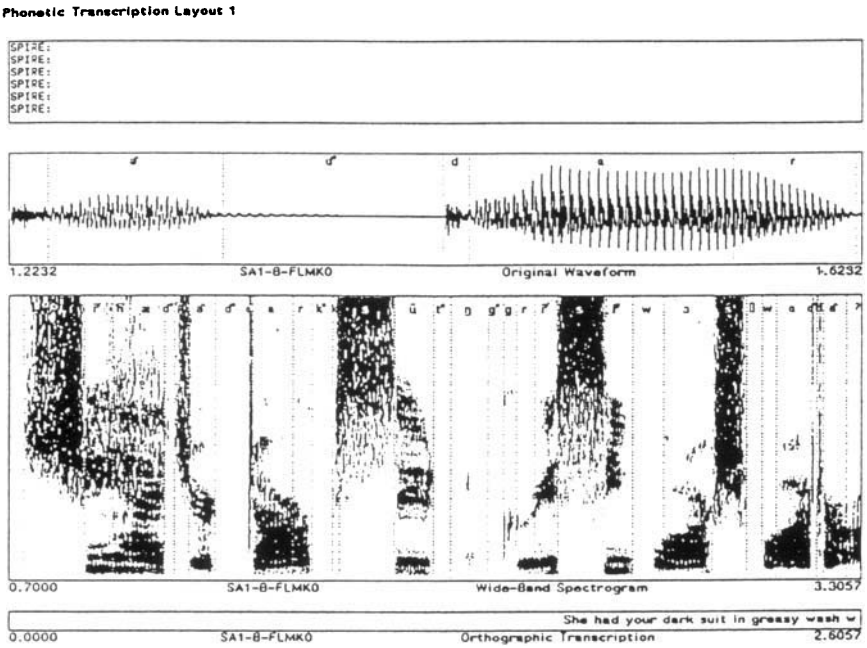


Figure 3. SPIRE layout showing the aligned transcription following post-processing.

approximate point in time. The waveform display automatically moves to synchronize in time with the cursor, and a fine-tuning of the boundary can be achieved by mousing the exact time point in the waveform.

The mouse can be used with ease to move an existing boundary to a new point in time, to erase a boundary, or to insert a boundary. Furthermore, a specified mouse click on any segment allows the labeller to change the acoustic-phonetic label associated with that segment. This step is sometimes necessary to correct an error of judgement in stage 1.

An example of the screen layout used for the correction process is shown in Figure 3. The boundary for the [d] burst onset has been corrected. Missing boundaries were inserted for the [iñæ] sequence. In addition, the boundaries associated with the first [w] were extended on both sides, and an epenthetic silence was inserted between the [š] and the following [w].

5. CONCLUDING REMARKS

Once the acoustic-phonetic transcription has been aligned, it is rather straightforward to propagate the alignment up to the orthographic transcription as well as the

intermediate phonemic transcription. A time-aligned orthographic transcription is useful when searching for a specific word, while a time-aligned phonemic transcription can be used to relate the lexical representation of words to their acoustic realizations. For example, the lexical representation of the word sequence "gas shortage" contains a word-final /s/ and a word-initial /š/, whereas its acoustic realization may simply be a long [š]. In this case, the time-aligned phonemic transcription will map the long to [š] both the underlying fricative. Researchers interested in studying the frequency of occurrence of certain low-level phonological rules will thus be able to derive the information from these transcriptions.

We have developed a system that maps a time-aligned acoustic-phonetic transcription to the phonemic and orthographic transcriptions [7]. However, the alignment effort for these transcriptions lags somewhat behind the phonetic alignment. In the interest of expeditiously making as much data available to the interested parties, we have decided to provide these other transcriptions in future releases.

The transcription and alignment of the TIMIT database is a sizable project. At this writing, all of the sentences have been processed and delivered to the National Bureau of Standards. A significant portion of the database is now available to the general public via magnetic tapes, and plans for distributing them by way of compact disc is well under way. Despite our best intention to provide as correct a set of transcriptions as possible, however, errors undoubtedly exists. We urge users of this database to communicate errors to us whenever possible, so that future users can benefit from this effort.

Finally, we would like to thank Dave Pallett, Jim Hieronymus, and their colleagues at NBS for the cooperation, patience, and good humour that they provided. Their help, particularly regarding data transfer, verification, distribution, and fending off eager inquiries, have been indispensable to this project.

The development of the TIMIT database at MIT was supported by the DARPA-ISTO under contract N00039-85-C-0341, as monitored by the Naval Space and Warfare Systems Command. Major participants of the project at MIT include Corine Bickley, Katy Isaacs, Rob Kassel, Lori Lamel, Hong Leung, Stephanie Seneff, Lydia Volaitis, and Victor Zue.

References

1. W. M. Fisher and G.R. Doddington: "The DARPA Speech Recognition Research Database: Specification and Status," *Proc. DARPA Speech Recognition Workshop*, pp.93-99, 1986.

2. V. W. Zue, D.S. Cyphers, R.H. Kassel, D.H. Kaufman, H.C. Leung, M.A. Randolph, S. Seneff, J.E. Unverferth, III and T. Wilson: "The Development of the MIT LISP-Machine Based Speech Research Workstation," *Proc. ICASSP-86*, 1986.

3. L. F. Lamel, R.H. Kassel and S. Seneff: "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop*, pp.100-109, 1986.

4. Peterson, G. and I. Lehiste: "Duration of Syllable Nuclei in English," *J. Acoust. Soc. Am.*, vol. 32, pp.693, 1960.

5. H. C. Leung and V.W. Zue: "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," *Proc. ICASSP 84*, pp. 2.7.1-2.7.4, 1984.

6. H. C. Leung: "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," S.M. Thesis, Department of Electrical Engineering and Computer Science, MIT, 1985.

7. R. H. Kassel: "Aids for the Design, Acquisition, and Use of Large Speech Databases," S.B. Thesis, Department of Electrical Engineering and Computer Science, MIT, 1986.

This Page Intentionally Left Blank