TEXT, SPEECH AND LANGUAGE TECHNOLOGY

THE INTEGRATION OF PHONETIC KNOWLEDGE IN SPEECH TECHNOLOGY

Edited by William J. Barry and Wim A. van Dommelen

Springer

The Integration of Phonetic Knowledge in Speech Technology

Text, Speech and Language Technology

VOLUME 25

The titles published in this series are listed on www.springeronline.com

The Integration of Phonetic Knowledge in Speech Technology

Edited by

William J. Barry
Universität des Saarlandes,
Saarbrücken, Germany

and

Wim A. van Dommelen
Norwegian University of Science and Technology,
Trondheim, Norway

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-2636-6 (PB) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-13 978-1-4020-2636-2 (PB) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-10 1-4020-2635-8 (HB) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-10 1-4020-2637-4 (e-book) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-13 978-1-4020-2635-5 (HB) Springer Dordrecht, Berlin, Heidelberg, New York
ISBN-13 978-1-4020-2637-9 (e-book) Springer Dordrecht, Berlin, Heidelberg, New York

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Printed on acid-free paper

Printed in the Netherlands.

TABLE OF CONTENTS

FOREWORD

The goal of this book is to present a discussion of the ideas arising from the European Special Event (ESE) on the Integration of Phonetic Knowledge in Speech Technology at Eurospeech 2001 in Aalborg. Where there is discussion, there must be unresolved questions, doubts must exist, integration is not a *fait accompli*. The different questions asked, methods applied and goals pursued are often invoked to explain why two sciences concerned with the same phenomenon – namely spoken language – do not generate more collaborative actions. Shared terminology can also divide rather than unite if meanings diverge. Yet for those who have ventured to look into both camps, either by becoming scientifically Janus-headed, or by practising cross-border communication, it is clear that each community can gain a lot from knowledge of the other.

The issue of whether Speech and Language Technology can profit from Linguistics is an old one, and there are a number of anecdotes from the 1980s, particularly in connection with automatic speech recognition which express decided opinions against involving Linguistics. The area of Linguistics we address is Phonetics/Phonology – the sound structures of spoken language. Whether or not formal models of sound structure can be incorporated in Speech Technology, it is clear from the contributions to this volume that phonetic knowledge already permeates many parts of it, though the extent to which this is the case may not be clear to many who work in the field.

Since there are indications that Speech Technology is about to change direction – some even speak of a change of paradigm – a renewal of interest in the possible contributions of Phonetics to Speech Technology is in the air.

One could, of course, attribute the apparent shift in interest simply to the Congress plan devised by the organisers – Paul Dalsgaard's interest in scientific integration is well documented. But one person's interest still needs communal acceptance. There was also Louis Pols' choice of topic for his keynote paper – not to mention the reception his address was given – a fitting sequel to his keynote paper at International Congress of Phonetic Sciences in San Francisco in 1999 and a clear symbolic bridge between the Speech Technology and Phonetics communities, a bridge which he also represents in persona, of course. Finally, there was the response to the ESE itself. Even more encouraging than the lively reaction to the call for contributions was the amount of interest that was apparent in the numbers that participated in the symposium.

This volume contains the reworked and extended versions of the six papers presented at the ESE (by Anton Batliner and Bernd Möbius – jointly, Frédéric Bimbot, Julie Carson-Berndsen, Heidi Christensen, Steven Greenberg (also discussant), and by Moisés Pastor) plus reflective discussions of the issue from three of the invited discussants on the symposium panel (Bill Ainsworth, Helmer Strik and Jan van Santen). A chapter by the editors expresses their own thoughts on the matter, brings together and comments on the other nine.

A sad loss to the Speech Community overshadows the otherwise happy event and its published sequel. Just four months after Eurospeech 2001, Bill Ainsworth died, a terrible blow to his family and a great loss to all his friends. It is to the memory of Bill that this book is dedicated.

Saarbrücken and Trondheim

WILLIAM J. BARRY*, WIM A. VAN DOMMELEN† and JACQUES KOREMAN‡,[1]

PHONETIC KNOWLEDGE IN SPEECH TECHNOLOGY – *AND PHONETIC KNOWLEDGE FROM SPEECH TECHNOLOGY?*

ABSTRACT. The contributions to this volume are considered within the framework of the question: "What sort of phonetic knowledge is relevant to speech technology?" This discussion throws light on the existing and the potential relationship between speech technology and the phonetic sciences, the possibilities for mutual gain and the need, ultimately, for researchers to emerge who combine the interest and expertise needed in both areas.

KEYWORDS. Phonetic knowledge, speech synthesis, speech recognition

1. INTRODUCTION

Although speech recognition and speech synthesis started out as much the territory of phoneticians and other linguists as of engineers, the linguistic approach soon lost terrain, in recognition applications at least, to (nonlinguistically orientated) engineers who were less concerned with formal linguistic insights, treating the signal as a pattern just like any other, and this with outstanding success. But the successes of engineering approaches are seen to have limits, most clearly in the challenges of spontaneous speech recognition and expressive speech synthesis, and once more the question arises whether the inclusion of additional linguistic, and more specifically phonetic knowledge is warranted. Of course, it is the degree of success so far which raises our sights to higher targets and exposes the limitations of techniques which were devised for the tasks

Address for Correspondence:
*Institut für Phonetik, Universität des Saarlandes, Germany
†Department of Language and Communication Studies, NTNU, Trondheim, Norway and
‡Institut für Phonetik, Universität des Saarlandes, Germany
[1]Although not an official organizer of the ESE nor an editor of this volume, Jacques Koreman has been WJB's discussion partner in matters that concern the symposium theme, and has, as such, contributed directly to the discussion presented here.

W. J. Barry and W. A. van Dommelen, The Integration of Phonetic Knowledge in Speech Technology, 1–12.

already (more or less) accomplished. With continuous (read or rehearsed) speech recognition systems commercially available, the drastic drop in performance found with *spontaneous speech* suggests that a ceiling may have been reached with the current processing methods. Similarly, with the intelligibility of speech synthesis systems no longer causing basic problems, the interest in and the calls for increased *naturalness* and *expressivity* in speech synthesis have become stronger. Here too there exists the realisation that the most successful approach to commercial synthesis, namely (fixed or variable) speech unit concatenation, while impressively natural within restricted domains, cannot provide the *flexibility of expression* together with the naturalness that is ultimately required. It is therefore practically limited. Finally, from the research point of view, i.e., in terms of learning how the production and perception of speech works, it is not theoretically satisfying.

The call for the inclusion of phonetic knowledge, however, presupposes both that the required knowledge is available, and that it exists in a form which is exploitable by the speech technology application. The question whether it is correct to make that assumption can be answered for both aspects with "partly", and the degree to which it is correct varies with the application being considered. In other words, there is certainly a lot of phonetic knowledge available which is not being used and which is relevant to speech recognition or synthesis (Pols 1999), but much of it does not exist in a form in which it can be used immediately. But of course there is also a great deal about the phonetic structuring of speech which is *not* understood, which could be of help to those speech technologies, and which has come to the notice of phoneticians as a result of contact to the field of speech technology. Thus, a simple answer to the question in our sub-title is "yes".

Before expanding on this issue we consider the individual contributions to this volume in terms of the way they include phonetic knowledge in the application they are presenting or considering. Alternatively, in the case of the discussion papers we consider their stance with regard to the potential integration of phonetic knowledge.

2. PHONETIC KNOWLEDGE IN THIS VOLUME

The six papers presented at the symposium comprise five which were selected from among the abstracts submitted because they appeared to promise studies that represent different approaches to the theme of integrating phonetic knowledge in speech recognition or speech synthesis. The sixth paper, by panel member Steve Greenberg, brings together

empirical results and general discussion. The final form of the papers does in fact reveal a diverse understanding of that theme. Four of them are reports of experimental applications. But Batliner & Möbius present a discussion of principle rather than details of a practical application (though they point to an example instantiation reported elsewhere), and Greenberg presents both a discussion of principle and concrete analysis examples from his own work. Among the discussants too, there was a healthy spread of opinion on the issue. The reworking and expansion of the papers and the post-hoc formulation by the panel members of the opinions developed during the ESE discussion session and presented in this volume underline the different perspectives. The following review of the chapters does not strictly separate the papers presented at the symposium from the panel members' discussions, but rather progresses from the purely empirical analyses to the theoretical discussions, removing the line between speakers and discussants.

Two of the empirically orientated papers were concerned with decoding the linguistic structure from the acoustic signal (Carson-Berndsen & Walsh; Christensen et al.) and two addressed the question of relating the decoded structure to representations in the lexicon, namely the problem of multiple pronunciations (Gravier et al.; Pastor & Casacuberta).

In their "Phonetic Time Maps", Julie Carson-Berndsen & Michael Walsh indicate how ASR can be made more robust by implementing phonetic and phonological constraints in a computational linguistic speech recognition model. The constraints can be used to guide the interpretation of multilinear event representations and to provide top-down predictions. The *Time Map Model* contains representations of the phonotactic constraints in a language. A special feature of the *Time Map Model* is that the phonotactic automata are defined with respect to the *syllable* domain. *Phonetic Time Maps* model phonetic details like the realisation of plosives (e.g., with/without release) or neutral vowels (e.g., elision before a nasal). The knowledge invoked in this approach is, in the first instance, of the type derived from traditionally established observations about allophonic variants and post-lexical modifications to the phonetic string that are captured in context-sensitive statements on assimilation and elision processes. What is particularly interesting about their processing framework is that it can operate both at the level of categorical, constraint-based representations of this knowledge and with a quantitative, probabilistic input to determine the ranking of such constraints.

Heidi Christensen, Børge Lindberg & Ove Andersen describe an ASR system for which the central issue is the exploration of multi-source recognition, which they term "heterogeneous processing". That is, the extraction of complementary phonetic information in different processing streams to provide more robust decoding. So-called "Expert MLPs" supplement the core (fullband; multiband) MLP systems; these are a "*voicing* expert" and a "*broad class* expert". The phonetic insight behind this approach is similar to that which motivates Carson-Berndsen & Walsh, namely the contextually determined change in the segmental identity of an underlying phonetic string. It also rests on the fact that a coarser definition of a segment can be more helpful for lexicon access than an incorrect decision at the phonemic level. In addition it appeals explicitly to parallels with human processing, which has recourse to different temporal and frequency granularities in order to cope with signal degradation.

There is no top-down component in the system other than the choice of "expert"; the whole process is data-driven. A number of other experts could have been chosen, but the two "experts" that were defined are plausible candidates in that the phonological voiced-voiceless opposition is extremely varied in its phonetic realisation, and changes to the phonetic properties of phonemic categories often result in shifts within the same broad class. It is presumably in this sense that the experts are seen as complementary to the *stem* system. In common with all stochastically orientated models, of course, the broad-class decisions and the voiced-voiceless decisions are as dependent on global probabilities as the phoneme decisions made by the stem system. It has no means of specifying the different contextual factors that are known to influence the changes, though it might be argued that this is catered for in the 7-frame ($\sim$ 100 ms) time base used for training.

Guillaume Gravier, François Yvon, Bruno Jacob & Frédéric Bimbot model contextual constraints on the phonetic forms of words at the search level to limit the search space to permissible pronunciation sequences. Using existing French lexicon resources containing pronunciation variants, they derive morpho-syntactically and phonologically context-sensitive rules to predict liaison, mute-e deletion and liquid consonant truncation.

A slight improvement in performance is found, a success in the light of the reduced search space that the approach offers. More interesting than this modest applicational success within the frame of this volume is, however, the concluding discussion of possible reasons why the results were

not more convincing. It highlights the interactions between phonetic factors (production task and speaking style), phonetic modelling complexity, the lexicon resource and the constraint definition.

Moisés Pastor & Francisco Casacuberta derive word-pronunciation variants using stochastic finite state automata to relate the phoneme output of a recognizer to the canonical pronunciation. Pronunciation alternatives are chosen on the basis of three different criteria: number of pronunciations, cumulative percentage, and threshold percentage. The results support the viability of the threshold-percentage criterion. Rather than theoretically discussing the possible use of phonetic knowledge in speech recognition, the authors experimentally show that pronunciation modeling should take into account articulatory reality. Whereas canonical models fail to do justice to the strong pronunciation variation due to deletions, assimilations and reductions, etc., modeling of frequently occurring pronunciation variants can (as also found by others) lead to improved recognition rates. In terms of the added value from this result, either for or from phonetic knowledge, the study confirms that multiple use of the same word will result in a variety of forms, and that the more a word is used, the more likely it will be to deviate from the canonical form.

The four contributions discussed so far all take "phonetic knowledge" at the general level of contextually based phonetic variation into consideration, but they vary considerably in the degree to which differentiated phonetic observations are or can be included. Also, they are all primarily and explicitly involved with the automatic *recognition* process, although contextually differentiated word forms may be one of the crucial aspects, so far neglected, for achieving more natural speech *synthesis*.

Coming now to the two more discussion-orientated and reflective of the six papers, Anton Batliner & Bernd Möbius address the question of knowledge integration in both recognition and synthesis. They are specifically concerned with the different demands placed on *prosodic* knowledge in automatic speech *understanding* (ASU) and text-to-speech *synthesis* (TTS). They introduce the distinction between phonetic-phonological knowledge and phonetic-phonological models and argue for the use of prosodic knowledge rather than prosodic models within ASU. Their standpoint is that models are an abstraction from phonetic reality and therefore introduce a quantisation error into the relationship between the phonetic form and the syntactic or semantic function. Rather than using subtle theoretical concepts, clear and stable prosodic markers need to be identified in order to define phrase boundaries and intonationally (and thus also informationally) important elements.

For synthesis, phonetically detailed prosodic events need to be generated (such as timing of tonal peaks in accented words dependent on consonant features, number of syllables, etc.), but though these events are clearly functional in demarcative, sentence-modal and information-structural terms, there seems to be no way of circumventing the intermediate phonological representation. Different ideologies behind these representations are also seen as a problem, as is the relationship between text and information- or discourse-structure which determines the prosodic form. With regard to a unified solution for intonation modeling in ASU and TTS, which is seen as ultimately desirable, the authors conclude that a common basis is not yet in sight. However, they do go on to discuss the sort of activities that are necessary in the phonetics and speech-technology community to move towards this goal.

In a more generally orientated discussion paper, Steve Greenberg discusses the fundamental importance of the two-way relationship between speech science and technology, i.e., of melding phonetic insight with speech technology to improve both the applications and the basic science. He sees the study of the large, naturally produced speech databases used in speech technology as a way to correct the largely unrealistic picture of speech and language projected by traditional phonetic and linguistic research, which is based largely on small-scale, carefully controlled and read material. In other words, speech and language science can and will improve. But he also sees that the successes in speech technology applications rest, in part, on imperfect scientific foundations, and that increasing demands, driven by the successes so far, are uncovering the limitations.

Greenberg illustrates this conviction with analyses of the Switchboard spontaneous speech database which uncover systematic relations between the prosodic-phonological category of stress accent and acoustic phonetic properties like duration and amplitude. The analysis results are presented both as a relationship of quantitative-phonetic properties to phonological categories, i.e., in terms of enhanced scientific insights, and as technologically exploitable facts. He presents the dramatic effect of stress-accent differences on the recognition performance (in terms of deleted words) of eight different recognition systems. Other relationships which are shown are those between word error rate and syllable structure on the one hand, and between stress accent and vowel identity on the other, which can also be of applicational importance. Importantly, within the framework of this volume, Greenberg illustrates both the gains in phonetic knowledge that come from asking phonetic questions of large

databases – the relationships he uncovers are by no means predictable from current phonetic or phonological theory – and the vital role that the careful phonetic labelling of such databases plays in that process.

Two of the panel members (Jan van Santen and Helmer Strik) take the fundamentally separate worlds of Speech Technology and Phonetics as their point of departure, van Santen concentrating on the implications for speech synthesis, while Strik's discussion is implicitly directed towards speech recognition.

Van Santen presents a relatively optimistic picture of the potential for integrating phonetic knowledge in speech synthesis, particularly with respect to making text-to-speech domain independent, even though there is little evidence of real cross-fertilization to date. On the contrary, developments in speech synthesis technology over the decades indicate a steady divergence from the level of phonetic theory: rule based methods gave way to fixed inventory concatenative techniques, and these appear to be in the process of being superceded by large-corpus based, variable-unit TTS; i.e., with apparently less emphasis on phonetic knowledge. However, linguistics may provide the type of knowledge that is needed to handle unseen unit types, which are still a problem in concatenative systems. This is illustrated by van Santen with reference to the different parts of the Bell Labs text-to-speech system that have been informed by phonetic knowledge: text analysis (computing phonemes; prosodic tags); duration modelling; intonation modelling; signal processing (special coarticulatory facts; segment lengthening details, etc.). He identifies the types of phonetic knowledge as: speech production/perception studies; architectural design; language dependent details (phonotactics, coarticulation, etc.); parameterized mathematical models.

One area in which van Santen particularly sees the need for phonetic support is in the perceptual evaluation of concatenation and signal manipulation techniques, e.g., thresholds for spectral and F0 discontinuities; subsegmental timing; vowel reduction; JND's for pitch contours. But there is also a clear knowledge deficit in the multidimensional modelling of prosodic features, particularly in relation to the definition of the properties covarying in emotional speech. Like others, he sees the potential for a phonetic contribution in a modified paradigm for phonetic research, in the development of a bridging field for research between phonetics and speech technology which he terms *computational phonetics*.

Helmer Strik has a rather less optimistic expectation for bringing the two different worlds of Speech Technology and Phonetics together. As negative examples of potentially useful, but in practice unusable phonetic

knowledge, he takes segment duration and lexical stress to show the difference between quantitatively supported insights and computationally usable analytic data. More positively, he shows the possibility of a phonetically oriented point of departure in pronunciation variation modelling: Rule knowledge is used, but the essential probabilities have to be derived from the data. This underlines his view that the existing phonetic knowledge is not complete and, above all, that it needs to be presented in probabilistic terms. As further illustration of the incompleteness of phonetic, and more generally linguistic knowledge he points out that prosodic models are rarely used in speech recognition, among other things because of the almost exclusive focus on F0, and that language models are based on written rather than spoken language. His conclusion is that using phonetic/linguistic knowledge in Speech Technology *can* be useful – improvement at the signal-processing level, for example, has been achieved due to knowledge about human auditory perception – but he clearly sees its use restricted to achieving a last few percent improvement.

Bill Ainsworth's discussion takes a long-term view of the Speech Technology scene, registering the divergence over the decades of speech technology methodology from the phonetic foundation, which focused on the facts of production and perception. To underline this he points out that hidden Markov models, the dominant approach in ASR, are very unrealistic models of speech production. Despite the positive point made by Strik (see above), a neglected aspect of speech science knowledge in terms of speech technology exploitation is the human auditory system, though it is partly modeled in modern ASR (multi-layer perceptrons; multi-band processing). Fundamental research into the physiology and neuro-anatomy of hearing has progressed greatly in recent years without its potential for speech signal processing having been exploited. To integrate phonetic knowledge in Speech Technology we need to base recognition and synthesis on realistic models of audition and speech production. Alone among the contributors he stresses the need to develop new mathematical models – though he does not claim to know what form they are likely to take – to cope, e.g. with the crucial fact that the *underlying control gestures* in speech production overlap.

3. QUO VADIS PHONETIC KNOWLEDGE

The contributions to this volume can be viewed both as a reflection of existing limits to the integration of phonetic knowledge in speech technology applications and as pointers towards ways in which more

knowledge can be of use in the future. We would like to bring those pointers together to a more general statement, and to give the reader a backdrop against which to consider the message of the individual contributions.

In his San Francisco ICPhS XIV keynote address, Louis Pols (1999) also addressed the question of Phonetics being of use to Speech Technology and vice versa. He took the difference between human and machine decoding as a point of departure, not because machine recognition should orientate itself in terms of processing principles on human recognition, but merely because it highlighted the potential for improvement. Imitation of human functionality, not duplication of human processing should be the aim. Understanding the limitations of the machine system and what makes it less "flexible, robust and efficient" (p. 9) than humans might contribute to improvement of the system. Within the present volume, Ainsworth is most explicit in taking this line of argument and pointing the finger at the Hidden Markov approach as an example of very powerful modelling which diverges fundamentally from the functionality of human speech production (and, one should add, of speech perception). None of the authors points the finger at concatenative synthesis as being perhaps *even further removed* from that functionality, lacking, as it does, the basic independence of the source from the filter characteristics of the system. Without that independence, naturally expressive synthesis is practically impossible. Thus, the message would appear to be that the courage to backtrack and reassess is necessary in both the main areas of speech technology. Pursuing established and hitherto very successful approaches might just be leading into a cul-de-sac.

Understanding what makes human speech decoding flexible, robust and efficient is, in broadest terms, *psycholinguistic* knowledge, part of which is more strictly *phonetic* knowledge. But that knowledge is certainly not normally couched in terms that can be directly integrated into automatic speech recognition, as many people in the past, including several in this volume, have pointed out. For speech technology and phonetics to have direct mutual benefit from each other, comparable data and comparable data representation are necessary. Viewing this from the speech technology vantage, Roger Moore (1995) used the term "computational phonetics" (cf. also van Santen in this volume), but a common data representation is perhaps an illusory aim. For one thing, even within speech technology, many different forms of data representation are required, depending on the task at hand. As Pols (1999, p.9) points

out, average formant values for vowels are of no use for vowel recognition in different contexts though they may be sufficient for formant synthesis. One might add that for building a diphone synthesis system not even vowel formant values are necessary, merely the knowledge that the quality of a vowel changes systematically along its time course as a product of local context, making the diphone a sensible building block.

This differentiation should highlight the difference between knowledge in the form of an *insight* into a phenomenon, the quantitative *specification* of that phenomenon (which may have led to the insight), and the *format* (e.g., average values, probability density functions or CART-trees) in which the quantitative data can be used for a particular application.

What is presumably meant by mutual benefit to the two different disciplines is (re)presentation in terms of the other discipline's problems, questions and aims. A phonetician always looks at data in terms of trying to develop an explanatory model for a human's ability to produce or perceive speech. However, she/he cannot be expected to deliver analysis results in the form required e.g. for a particular recognition algorithm or a particular synthesis system. This would be equivalent to expecting a speech technology engineer to ask phonetic questions of a database to gain his/her own insights. In fact, if the results are a new insight which could be important for speech technology, there is possibly no ready form of representation available; its exploitation might well require a new algorithmic approach. What *can* be expected, however, is that the analysis is carried out on data that is *relevant* for a particular application, and that the observation is at a *level of delicacy* that is relevant for the task. Finding an effect e.g. of a particular contextual factor, when the speech material has been carefully controlled and all possible confounding factors excluded, will not generalize to any realistic ASR task. Finding a robust effect in a large continuous-speech database is something else, however, and there should then also be an interest to communicate the implications of the observation in a manner which members of both communities can understand.

Communicating, on the one hand, what phenomena are clearly functionally important, and, on the other hand, saying how they should be dealt with in a speech technology application are two very different things. While the linguist is implicated in the former task, we suggest that it is the task of the speech technology scientist to undertake the latter. An example of the former is the well established simultaneous global and local importance of duration (cf. Pols 1999, p. 12). Within any given

tempo (varying locally within a global frame) there are globally calculable durational differences between phonemically long and short segments. In many languages, there are locally determined allophonically longer and shorter variants and local durational increases related to accentuation (which in turn is related to information structure). Finally, there is the local phrasal function of final lengthening. With regard to dealing with such functionally important variation within ASR, these insights present a strong challenge because they certainly cannot be exploited within present-day stochastical methods, dependent as they are on global probabilities. However, their functional and communicative importance has to be understood and accepted otherwise the challenge will not be recognized.

What emerges very clearly from this and other discussions is the need to understand both sides of the problem. Viewed within the present structures of science, the need for interdisciplinary interest and cross-disciplinary activities is undeniable. The greater access phonetically trained researchers have to the databases and tools used in mainline technology applications, the more likely it is that quantitative answers to phonetic questions can be presented in a way which can be useful for speech technology applications. Conversely, speech technology engineers will be increasingly prepared to look for innovative processing solutions, the more contact they have with quantitatively supported statements about the complex relationships between the relatively simple signal parameters (duration, intensity, frequency and spectral energy distribution, and their derivatives) and the communicative functions they are trying to decode (ASR) or encode (synthesis). What is certainly not to be expected as a rule at present is the phonetician who can develop new processing algorithms or the speech technology engineer who can ask new phonetic questions of a speech database.

However, a certain indication of the developing contact in the two areas of science can be gained from looking at the change in phonetically orientated contributions to Eurospeech conferences during the twelve years from Eurospeech I in Paris, 1989 to Eurospeech Scandinavia in Aalborg 2001. Although weaker than the growth in purely technology-orientated papers, the papers dealing with phonetic questions or integrating phonetics in technological applications grew by a very substantial 45% from 93 to 135. Ultimately, as a product of this increasing contact between the two disciplines more exemplars of the currently rare hybrid scientist should appear: the "linguist speech-technology engineer" and the "speech-technology linguist", or to borrow Roger Moore's and Jan van Santen's term, the *computational phonetician*.

Ultimately, progress towards and in interdisciplinary research, like other human interactions, is the product of the individuals involved. They must be interested and committed. But we echo van Santen's comment (this volume) that changes are sociologically determined, and a framework for contact and interaction is needed. A symposium and a published discussion are a first step in the right direction, inter-departmental courses and inter-disciplinary degrees are a further goal. However, any change of socio-scientific climate must also be triggered and established by individuals.

REFERENCES

Moore, R. (1995). *Computational phonetics*. In: *Proceedings ICPhS 1995*, Stockholm, Vol. 4, pp. 68–71.

Pols, L. (1999). *Flexible, robust, and efficient human speech processing versus present-day speech technology*. In: *Proceedings of ICPhS 1999*, San Francisco, Vol. 1, pp. 9–16.

WILLIAM A. AINSWORTH

CAN PHONETIC KNOWLEDGE BE USED TO IMPROVE THE PERFORMANCE OF SPEECH RECOGNISERS AND SYNTHESISERS?

ABSTRACT. A chronological survey of the development of machine recognition of speech is contrasted with the beginnings of speech synthesis, and the advantages and disadvantages of the different systems and approaches as well as their changing degrees of dependency on phonetic knowledge are sketched. The unsolved fundamental problem of concatenation quality in present-day synthesis is discussed and a knowledge based solution mooted which can be projected onto recognition: A mathematical model of the relationship between temporally overlapping underlying articulatory gestures and the resulting surface acoustic signal.

KEYWORDS. speech synthesis, speech recognition, concatenation, articulatory gestures

1. INTRODUCTION

Phonetic knowledge can be defined as knowledge derived from studying speech production and perception and the analysis of speech signals. Early speech recognition devices attempted to exploit such phonetic knowledge as was available at the time. Davis et al. (1952) built a device to discriminate between spoken digits which filtered the speech signal into the first and higher formant frequency bands then counted the zero-crossings in each band; Olson and Belar (1956) used a bank of filters to produce a crude spectrogram, similar to the 'visible speech' analyses of Potter et al. (1947); and Wiren and Stubbs (1956) attempted to detect acoustic correlates of the distinctive features of Jakobson et al. (1952). Unfortunately none of these devices produced acceptable results.

Speech synthesis based on phonetic knowledge was more successful. A series of perception experiments at the Haskins Laboratories (Liberman et al., 1952; 1954; 1956; 1958) established the basic relationships between the acoustic structure of speech sounds and their perception (Liberman et al., 1967). The results of these, and other, experiments led Holmes

W. J. Barry and W. A. van Dommelen, The Integration of Phonetic Knowledge in Speech Technology, 13–19.

et al. (1964) to develop a rule-based phoneme-to-speech synthesis system, which produced intelligible, if not completely natural sounding, speech. Context-dependent pronunciation rules (orthographic text to phonetic symbols) were later added to produce the first text-to-speech systems (e.g. Ainsworth, 1973). Refinements to this basic structure led to the MITalk (Allen et al., 1987) and commercial systems such as DECtalk.

2. AUTOMATIC SPEECH RECOGNITION

The subsequent history of automatic speech recognition up to the mid 1980s has been reviewed by Ainsworth (1988). First pattern recognition techniques were investigated (Forgie and Forgie, 1959; Denes and Mathews, 1960) employing normalisation to reduce the inherent variability of speech signals. Artificial neural networks which attempted to simulate the pattern recognition abilities of humans (Nelson et al., 1967) were also applied to speech signals during this period.

A concentrated effort to employ not only phonetic but also linguistic knowledge was made in the 1970s (see Klatt, 1977 for a review). One of the most successful projects which developed a technique for integrating different sources of knowledge was the Dragon system of Baker (1975).

In parallel with this investigation of large vocabulary systems, isolated word recognisers based on dynamic programming (Velichko and Zagoruyko, 1970; Sakoe and Chiba, 1978) were developed. This non-linear time normalisation plus pattern recognition, dynamic time warping (DTW), gave good results without the use of phonetic knowledge. DTW recognisers were shown to be useful for small vocabulary, speaker dependent, isolated word recognition but could not be generalised to large vocabulary, speaker independent, continuous speech recognition.

3. STATISTICAL SPEECH RECOGNITION

The use of algorithms based on stochastic models, using speech databases but not explicit phonetic knowledge, was investigated by Jelinek (1976) and these gave encouraging results. A system based on hidden Markov models (HMMs) was developed by Levinson et al. (1983) which confirmed this. In many ways HMM recognisers are complementary to DTW recognisers, giving them advantages which have led to their adoption in most current systems. Whereas DTW recognisers are computationally expensive during recognition, HMM recognisers are computationally expensive to train, which is acceptable, but efficient for recognition. They can also be used for speaker independent, large vocabulary recognition.

The theoretical disadvantage of HMM recognisers is that they are unrealistic models of speech production. They model speech as a sequence of static acoustic elements whereas speech is produced by a set of overlapping dynamic gestures. The next gesture begins before the previous one has ended. This is incorporated into HMM recognisers by employing context-dependent sub-word units (usually triphones). As most phones can occur in many contexts a great number of units are required. In order for these to be trained large speech databases are required. It is in fact likely that it is the collection of these databases which has led to the enhanced performance of HMM recognisers.

This enhanced performance, however, is only obtained in conditions which approximate the conditions in which the databases were recorded, usually read speech in a quiet environment. Recognition of spontaneous speech (e.g. the Switchboard corpus, Godfrey et al., 1992) and speech in noisy conditions can produce high error rates. One way of coping with variability of pronunciation is by employing dictionaries which include these alternatives (Pastor and Casacuberta, 2001) but a more realistic model of speech production may be able to deal better with these different conditions. What is required is a mathematical model, statistical or otherwise, which captures more nearly the underlying phonetics of speech production.

4. AUDITORY MODELLING

As mentioned earlier phonetics has traditionally been concerned with speech production, analysis and perception. The missing factor in the speech chain is speech processing in the human auditory system. During the last few decades physiological and neuroanatomical knowledge concerning the human auditory system has increased to a point where realistic computational models of the peripheral auditory system (the outer, middle and inner ear and the cochlear nerve) can be made and some of the processing from the cochlear nucleus to the auditory cortex is understood. Little of this knowledge has been incorporated into speech recognisers, and whether this will ever be advantageous is still an open question.

Almost the only auditory knowledge which is accepted as useful is the relationship between critical bands and frequency. This has been incorporated into many recognisers by employing MFCCs (mel frequency cepstral coefficients) as features.

The structure and function of the regions of the central nervous system concerned with audition, from the cochlear nucleus to the auditory cortex, are gradually being elucidated. General models of these regions

have been incorporated in speech recognisers in the form of neural networks such as multi-layer perceptrons (e.g. Christensen et al., 2001) but specialised models which reflect the specific structures of auditory regions have rarely been explored.

A new departure in speech recognition in recent years is multi-band processing (Hagen et al., 2000). This is based on Fletcher's ideas of processing in separate frequency channels (Fletcher, 1953; Allen, 1994). Such systems can tolerate noise interference in some frequency bands because the overall error rate is determined by the product of the individual error rates in each band. These systems are robust with regard to narrow band noise noises, but not for wide band interference. Perceptual studies with narrow bands (Greenberg et al., 1998; Crouzet and Ainsworth, 2001) should help to produce a solid scientific basis from which recognisers can be designed.

Other noise robust recognisers based on processing in the auditory system have been proposed. For example, amplitude modulation maps (Kollmeier et al., 1994; Berthommier and Meyer, 1995) which have been used for separating voiced speech from noise may perform similar processing to that in the inferior colliculus where units sensitive to amplitude modulations of certain frequencies have been found. Amplitude modulation maps are formed by passing the speech signal through a bank of band-pass filters then performing a Fourier analysis on each channel. The resulting two-dimensional map contains high energy ridges at frequencies corresponding to the harmonics of the pitch of the voice. Wide band noise is distributed in the valleys between the ridges. By sampling the ridges the speech can be separated from the noise (Meyer et al., 2000).

5. CONCATATIVE SPEECH SYNTHESIS

Just as bigger databases and more powerful computers have enabled more realistic units (triphones, etc.) to be employed in speech recognition, the same resources have enabled more and larger units to be employed in speech synthesis (see van Santen et al., 1997). This has allowed more natural sounding speech to be synthesised but the problem of joining the units together has remained. Sometimes two segments appear to be joined together smoothly, at least acoustically, but a perceptual discontinuity is heard between them (van Santen et al., 1993). This remains a current problem.

The basic model of concatative speech synthesis is a string of discrete units. A more realistic model of speech production is a sequence of overlapping speech gestures. It is, however, not the waveforms produced by

the gestures which must be overlapped, but the underlying control signals. If these are then applied to control a dynamic vocal tract model, a seamless flow of speech should be generated. Feasibility studies have been undertaken by Carré et al. (2001). The phasing of the control signals required to generate vowel sequences can be varied over quite wide limits and the same vowel sequences are perceived. Outside these limits additional, intrusive vowels are heard.

If this technique can be used to synthesise longer, seamless speech utterances the question will arise as to whether a similar approach should be incorporated into speech recognition systems. The principle would be to invert the speech signal to determine the vocal tract gestures, then estimate the control signals which would have generated these vocal tract gestures. However, this may be impossible in practice because several different sequences of control signals may give rise to the same acoustic signals.

6. CONCLUSIONS

We have argued that the way to integrate phonetic knowledge into speech technology is not by deriving the detailed acoustic structure of phones from sets of phonetic rules, but by basing both speech recognition and speech synthesis on more realistic models of speech production. The details are probably best derived from speech databases as at present. What is required is a mathematical framework which treats speech signals as the result of overlapping gestures, rather than as a sequence of overlapping discrete units.

In speech synthesis this can be derived by means of sequences of control signals applied to a dynamic vocal tract model with realistic constraints over the allowed shapes. Concatative speech synthesis generates high quality synthetic speech except at some joins. It is, therefore, unnecessary to employ the vocal tract model to generate all of the utterance. The speech signals from the database could be used to estimate the vocal parameters near to the 'join', then a transition representing an allowable gesture could be synthesised to cover the join.

In speech recognition a new model is required which reflects the underlying speech gestures. Some more realistic method of modelling the transitions between the phones is needed.

REFERENCES

Ainsworth, W.A. A system for converting English text speech. In: *IEEE Transactions*, AU-21, 1973: 288–290.

Ainsworth, W.A. *Speech Recognition by Machine*. Peter Peregrinus Ltd., London, 1988.

Allen, J.B. How do humans process and recognize speech? In: *IEEE Transactions on Speech and Audio Processing*, 2 (4), 1994: 567–577.

Allen, J., Hunnicutt, S., and Klatt, D. *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge, 1987.

Baker, J.K. The DRAGON system – an overview. In: *IEEE Transactions*, ASSP-23, 1975: 24–29.

Berthommier, F. and Meyer, G.F. Source separation by a functional model of amplitude demodulation. In: *Proceedings of Eurospeech'95*, 1995: 135–138.

Carré, R., Ainsworth, W.A., Jospa, P., Maeda, S., and Pasdeloup, V. Perception of vowel-to-vowel transitions with different formant trajectories. *Phonetica*, 58 (2001): 163–178.

Christensen, H., Lindberg, B., and Andersen, O. Introducing phonetically motivated information into ASR. In: *Proceedings of Eurospeech'01*, Volume 4, 2001: 2289–2292.

Crouzet, O. and Ainsworth, W.A. Envelope information in speech processing: acoustic-phonetic analysis vs. auditory figure-ground segregation. In: *Proceedings of Eurospeech'01*, 1, 2001: 477–480.

Davis, K.H., Biddulph, R., and Balashek, S. Automatic recognition of spoken digits. *Journal of the Acoustical Society of America*, 24 (1952): 637–642.

Denes, P. and Mathews, M.V. Spoken digit recognition using time-frequency pattern-matching. *Journal of the Acoustical Society of America*, 32 (1960): 1450–1455.

Fletcher, H. *Speech and Hearing in Communication*. Van Nostrand, New York, 1953.

Forgie, J.W. and Forgie, C.D. Results obtained from a vowel recognition computer program. *Journal of the Acoustical Society of America*, 31 (1959): 1480–1489.

Godfrey, J.J., Holliman, E.C., and McDaniel, J. Switchboard: Telephone speech corpus for research and development. In: *Proceedings of ICASSP-92*, 1, 1992: 517–520.

Greenberg, S., Arai, T., and Silipo, R. Speech intelligibility derived from exceedingly sparse spectral information. In: *Proceedings of ICSLP*, Sydney, 1998: 2803–2806.

Hagen, A., Morris, A., and Bourlard, H. From multi-band full combination to multi-stream full combination processing in robust ASR. In: *Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium, ISCA ITRW ASR2000*, 2000: 175–180.

Holmes, J.N., Mattingly, I.G., and Shearme, J.N. Speech synthesis by rule. *Language and Speech*, 7 (1964): 127–143.

Jacobson, R., Fant, G.C.M., and Halle, M. *Preliminaries to speech analysis*. MIT Tech. Report 13, 1952.

Jelinek, F. Continuous speech recognition by statistical methods. In: *Proceedings of IEEE*, 64 (1976): 532–556.

Klatt, D.H. Review of the ARPA speech recognition project. *Journal of the Acoustical Society of America*, 62 (1979): 1345–1366.

Kollmeier, B., Kock, R., and Kohlrauch, A. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *Journal of the Acoustical Society of America*, 95 (1994): 1593–1602.

Levinson, S.E., Rabiner, L.R., and Sondhi, M.M. Speaker-independent digit recognition using hidden Markov models. In: *Proceedings of ICASSP 1983*, 1983: 1049–1052.

Liberman, A.M., Delattre, P., and Cooper, F.S. The role of selected stimulus-variables in the perception of unvoiced stop consonants. *American Journal of Psychology*, 65 (1952): 497–516.

Liberman, A.M., Delattre, P., Cooper, F.S., and Gerstman, L.J. The role of consonant-vowel transitions in the perception of stop and nasal consonants. *Psychological Monographs*, 68 (1954): 1–13.

Liberman, A.M., Delattre, P., Gerstman, L.J., and Cooper, F.S. Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology*, 54 (1956): 358–368.

Liberman, A.M., Delattre, P., and Cooper, F.S. Cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, 1 (1958): 153–167.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 74 (1967): 431–461.

Meyer, G.M., Edmonds, B.A., Yang, D., and Ainsworth, W.A. Amplitude modulation maps for robust speech recognition. In: *Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, *ISCA ITRW ASR2000*, 2000: 168–174.

Nelson, A.L., Herscher, M.B., Martin, T.B., Zadell, H.J., and Falter, J.W. Acoustic recognition by by analog feature-abstraction techniques. In: *Models for the Perception of Speech and Visual Form*, MIT Press, 1967: 428.

Olson, H.F. and Belar, H. Phonetic typewriter. *Journal of the Acoustical Society of America*, 28 (1956): 1072–1081.

Pastor, M. and Casacuberta, F. Automatic learning of finite state automata for pronunciation modelling. In: *Proceedings of Eurospeech'01*, 4, 2001: 2293–2296.

Potter, R.K., Kopp, G.A., and Green, H.C. *Visible Speech*. Van Nostrand, New York, 1947.

Sakoe, H., and Chiba, S. Dynamic programming algorithms optimization for spoken word recognition. In: *IEEE Transactions*, ASSP-26, 1978: 43–49.

van Santen, J.P.H. Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language*, 7 (1993): 49–100.

van Santen, J.P.H., Sproat, R.W. Olive, J.P., and Hirschberg, J. (Eds.). *Progress in Speech Synthesis*. Springer-Verlag, New York, 1997.

Velichko, V.M., and Zagoruyko, N.G. Automatic recognition of 200 words. *International Journal of Man-Machine Studies*, 2 (1970): 223–234.

Wiren, J., and Stubbs, H.L. Electronic binary selection system for phoneme classification. *Journal of the Acoustical Society of America*, 28 (1956): 1082–1091.

ANTON BATLINER* and BERND MÖBIUS†

PROSODIC MODELS, AUTOMATIC SPEECH UNDERSTANDING, AND SPEECH SYNTHESIS: TOWARDS THE COMMON GROUND?

ABSTRACT. Automatic speech understanding and speech synthesis, two major speech processing applications, impose strikingly different constraints and requirements on prosodic models. The prevalent models of prosody and intonation fail to offer a unified solution to these conflicting constraints. As a consequence, prosodic models have been applied only occasionally in end-to-end automatic speech understanding systems; in contrast, they have been applied extensively in speech synthesis systems. In this chapter we aim to make explicit the reasons for this state of affairs by reviewing the role of prosodic modelling in these two fields of speech technology. Subsequently, possible strategies to overcome the shortcomings of the use of prosodic modelling in automatic speech processing are discussed. In particular, the question is raised whether or not there is a common framework for prosodic modelling in automatic speech understanding and speech synthesis systems, and if so, whether any particular model or theory of prosody can serve as a common ground. Finally, a catalogue of tasks in prosody research is proposed that ought to be relevant to both automatic speech understanding and speech synthesis and that might stimulate joint research activities.

KEYWORDS. prosody, intonation model, automatic speech understanding, speech synthesis

0. INTRODUCTION

The application of prosodic models in automatic speech understanding (ASU)[1] and speech synthesis (TTS)[2] is strikingly different. In the latter, such models have been extensively applied, but there is still no generally agreed upon approach to prosodic modelling. In the former, they have been applied only occasionally, rather in basic research, but almost never within an existing end-to-end system. In this chapter, we discuss the

Address for Correspondence:
* Chair for Pattern Recognition, University of Erlangen-Nuremberg, Germany
† Institute of Natural Language Processing, University of Stuttgart, Germany

W. J. Barry and W. A. van Dommelen, The Integration of Phonetic Knowledge in Speech Technology, 21–44.

reasons for this state of affairs and possible strategies to overcome the shortcomings of the use of prosodic modelling in automatic speech processing.

This chapter consists of three parts: the first part deals with the role of prosodic modelling in ASU, the second part concerns the role of prosodic modelling in TTS. These two parts are written from an inside perspective and focus on different aspects–simply because the use of models is considerably different in the two branches of speech technology discussed here. In the section on ASU, the argumentation is thus more general, dealing mainly with models as incarnations of theories, whereas in the section on speech synthesis, more details are given, dealing mainly with models as more or less concrete algorithmic formulations of theories. In the third part we present different possibilities for a closer co-operation of ASU and TTS; eventually this might lead to new types of prosodic models that are more adequate for automatic processing than the present ones.

In the title of this chapter, we speak of three different things: models, ASU, and TTS, and of one type of relationship: the common ground. Thus, first we have to know what prosodic /intonation models look like. For obvious reasons, we cannot give a detailed survey of the models that were developed during the last three decades. Instead we sketch common traits and principles that constitute models as such. One important characteristic is that a model is a considerable, sometimes even extreme, reduction of parametric values and, thereby, a mapping of these values onto a small number of units that can be compared with the classic distinctive features on the phone or phoneme level–all other properties may differ. The foundations of the model and, therefore, the philosophy behind it, can be physiological (Fujisaki:1988), or perceptual ('t Hart et al., 1990), or linguistic (Silverman et al., 1992, "ToBI"), just to mention a few well-known models. ToBI (*Tones and Break Indices*) is by far the most well-known model. There are at least two reasons for this fact: first, it was the first model by way of which researchers from different disciplines attempted to find a common standard and common evaluation procedures; and second, it was developed for (American) English, a fact which in itself enhances wide dissemination. The ToBI transcription system is a formalisation of the tone sequence theory of intonation (Pierrehumbert, 1980). It may be characterised as a broad phonemic system, consisting of *High* (H) and *Low* (L) tones and some few, additional diacritics. The phonetic details of fundamental frequency (F_0) contours in a given language have to be established in a second step. ToBI labels,

in conjunction with F_0 generation rules, are also frequently used in the intonation components of TTS systems.

Out of the theoretically possible relationships between models, ASU, and TTS, we can imagine four different types:

- Type 1: ASU ← model → TTS
- Type 2: ASU ← model 1 | model 2 → TTS
- Type 3: ASU ↔ TTS
- Type 4: ASU | TTS

Type 1 meets the ideas of generality put forward in most intonation theories that there is only one model that accounts for all possible applications. Type 2 is a weaker formulation that there have to be models especially tuned for different applications. With Type 3, a direct relationship between ASU and TTS, not mediated by any model, is imagined, and Type 4 (no relationship at all) might not be desirable but mirrors, in fact, the present situation quite closely.

There is a striking difference between ASU (many-to-one) and TTS (one-to-many): in ASU, many speakers/features/feature values have to be mapped onto few units (from parameters to categories), whereas in TTS, it is the other way round: one speaker/category has to be mapped onto many features/feature values (from categories to parameters). It has not been settled yet whether this is a one-way-trip or a round-trip–and by that, whether there is any common ground for these two fields at all, as far as prosody is concerned.

0.1. Caveat and Further Reading

This chapter is not intended to be an introduction into any of these three topics: models, ASU, and TTS. We hope, however, that it will be useful for experts in one of these fields who wonder why the state of affairs is the way it is. At the same time, we want to provide the readers with a sufficient degree of 'meta-knowledge' without presenting them all the basics. This chapter is thus not written as an in-depth treatise but rather as a *set of postulates* intended to provoke discussion.

To our knowledge, there is no up-to-date standard introductory textbook on intonation models. A comprehensive review of current intonation models is presented in Ladd (1996), albeit from the perspective of a proponent of the tone sequence approach. The language-specific use of some of these models is described in Hirst and Di Cristo's survey of intonation systems (Hirst and Di Cristo, 1998a). The computational analysis and modelling of prosody for the automatic processing of speech

is the topic of Sagisaka et al. (1997). A state-of-the-art account of prosody in ASU is given in Batliner et al. (2001c), whereas the intonational concepts used in several models, and synthesis approaches based on these models, are dealt with in Botinis et al. (2001).

1. AUTOMATIC SPEECH UNDERSTANDING

For contemporary prosodic theory, subtle changes in meaning that are potentially triggered by prosody are interesting. These are, however, not good candidates to start with in ASU: they will be classified rather poorly because of the many intervening factors, because of sparse data, and because they can only be observed in the laboratory. Therefore, we should start with a clear prosodic marking; the marking of boundaries is probably the most important function of prosody and thus most useful for ASU.

Information retrieval dialogues have been the standard application within ASU for many years. Recently, less restricted dialogues, for instance in the context of the Verbmobil system[3], had to be processed where turns are, on the average, three times longer than in an information retrieval application (Nöth et al., 2000). Segmentation is thus more important in the relatively new field of automatic processing of rather free dialogues–a chance to prove the impact of prosody! The contribution of prosody is not equally evident in other applications.

In the last two decades, a growing body of work on intonation and prosody research in general and on intonational modelling in particular has been conducted. (Note that we use *prosody* for all phenomena above the segmental level, whereas *intonation* only deals with pitch/F_0.) Researchers on these topics agree that ASU would benefit from the integration of this work. However, only in the last few years has prosody really begun to find its way into ASU, most of the time within *offline*, i.e., *in vitro*, research. The only existing end-to-end system that really uses prosody is, to our knowledge, the Verbmobil system (Batliner et al., 2000).

This state of affairs might be traced back to the general difficulty of carrying over theoretical work into practice as well as the well-known differences between the two cultures: on the one hand, humanities, on the other hand, engineering. In this section, we want to have a closer look at some of the most important factors that are responsible for this state of affairs, and with that, we want to make this general statement more concrete. First we discuss the shortcomings of current intonation models, as seen from an ASU perspective (Section 1.1). Then, we will show what

can be done to overcome these shortcomings by sketching our own *functional* prosodic model (Section 1.2), and we will outline the common ground of prosodic models on the one hand and ASU on the other hand (Section 1.3).

1.1. The Reasons Why (Occam's Razor Still Matters)

If one speaks of suprasegmental models that meet the standards of a theory, one very often speaks only of *intonation models*, which are almost always *production models*. (Transcription, labelling, and annotation are more down to earth and their topic is thus broader.) Production models might be good for synthesis but not for recognition. Too much emphasis is put on intonation in particular, i.e., too much emphasis on *pitch* in comparison to *other prosodic* features, and too much emphasis on *prosody* in comparison to *other linguistic* features. This is, of course, conditioned by the general approach to constructing intonation models as *stand-alone models*, and by the unfortunate notion of *pitch accent*, which prevents a more realistic view where all relevant features–be they intonational, other prosodic or other linguistic features–are considered in the analysis on the same level.

There is too much emphasis on *theoretical concepts* and on the discussion of which one is better suited for the description of a special language or of languages in general. Consider the old debate pertaining to whether levels or movements, local events or global trends, are *the* 'correct' units of description: a speech recognizer does not care whether it is trained with levels (F_0 maximum, F_0 minimum) or with movements (F_0 range, F_0 slope) as long as the training database is large enough and the labels are annotated correctly. After all, what goes up must come down: it does not matter whether there is an H tone at 200 Hz and a following L tone at 100 Hz or whether there is a movement between 200 Hz and 100 Hz (Batliner et al., 2001a).

Very often it is stressed that one cannot do prosody research or apply prosody within ASU without a 'real' phonological level of description and modelling, and that speech technologists should pay attention to the work of phonologists (Ladd, 1997). We fully agree with the view that phonological and prosodic *knowledge* should be used within ASU, but we fully disagree if it is about the direct use of intonation *models* in ASU. All these models introduce a phonological level of description that is intermediate between *(abstract) function* and *(concrete) phonetic form*: tone sequences, holistic contours, etc. It is our experience that one always gets better results if one can do without such an intermediate level, i.e., if one

can establish a direct link between (syntactic/semantic) function and phonetic form. (Here, we speak 'simply' of classification performance, not of theoretical interest or adequacy.)

After all, if such a mapping can be done automatically, we can map *level A (phonetic form)* onto *level C (linguistic function)* without an intermediate *(phonological) level B*; with such a level, we have to map *A* onto *B*, and *B* onto *C*. If this can be done automatically, we do not need *B* any longer. Sometimes it will do no harm to provide level B, but often results will get worse. Phonological systems like the ToBI approach (Silverman et al., 1992) only introduce a *quantisation error*: the whole variety of F_0 values available in acoustics is reduced to a mere binary opposition L vs. H, and to some few additional, diacritic distinctions. This fact alone prevents tone levels (or any other *phonological prosodic* concepts such as, e.g., the one developed within the IPO approach) from being a meaningful step that automatic processing should be based on; it seems better to leave it up to a large feature vector and to statistical classifiers to find the form to the function. To our knowledge, no approach exists that actually uses such phonological units for the recognition of prosodic events. Of course, there are many studies that describe *offline* classifications of such phonological prosodic concepts; this has to be distinguished from the successful *integration* in an existing end-to-end-system, as we have shown within the Verbmobil project (Batliner et al., 2000; Nöth et al., 2000).

Studies which compare the performance of intonational models for the automatic classification of prosodic events are rare; Siepmann (2001) assesses several models on the task of the classification of contrastive accents in German. He finds that classification performance is roughly a function of the number of predictor variables. It increases with the number of these predictor variables made available by a model. These findings fit nicely with our notion of quantisation described above. Evidently, a theoretically and phonologically 'adequate' description–in terms of a minimal inventory of units–on the one hand, and classification performance on the other hand, are simply two conflicting goals.

The difference between phonetic/prosodic knowledge and phonological concepts can be demonstrated with the following example: the prosodic 'default' feature that indicates questions in many languages is a final rise (or 'high boundary tone'), even though, at least in English and German, an accent pronounced in a non-final position can disambiguate sentence mood as well (Studdert-Kennedy and Hadding 1973; Batliner, 1989b). The same holds for Italian, where "[the] primary cue to interrogation in the Southern varieties is the pitch accent: $L + H^*$ in Bari Italian

and $L^* + H$ in Palermo and Neapolitan, after which there is usually a final fall'' (Grice et al., 2004). This is a very interesting fact in itself, but it is of course not a special tone that is the primary cue but something that can be *described as* a special tone within a special intonational model. This is actually not a nicety but crucial for our argumentation. Thus we want to distinguish between *basic* knowledge about the facts one observes, and knowledge that is *transformed into* and *mediated by* a specific model. The units of such a model might provide a convenient way to make oneself understood. The problem is that, by using such terminology, one tends to disregard those aspects that are not modelled by this concept; for instance, by using the terminology of a tone *level* model one disregards *movements*, and vice versa, and might end up with a mere *reification* of this concept.[4]

The classic phonological concept of the Prague school has been abandoned in contemporary intonation models, namely that phonemes–be they segmental or suprasegmental–should only be assumed if these units make a difference in meaning. This functional point of view has given way to more formal criteria such as economy of description. Thus, the decision on the descriptive units is not based on differences in meaning but on formal criteria, and only afterwards are functional differences sought that can be described with these formal units. In Hirschberg and Pierrehumbert (1986) for instance, the meaning of a tune, which is defined as a structure consisting of accents and tones, can be interpreted compositionally from the meanings of the individual accents and tones that the tune contains. It has been supposed that if phonological concepts could be motivated from theoretical reasons, then ASU should use them[5]–irrespective of whether they really make sense as units of ASU or not: this can only be determined empirically, not by theoretical considerations.

In conclusion, *Occam's razor* (i.e., the law of economy) should thus be followed here as well: *non sunt multiplicanda entia praeter necessitatem (entities are not to be multiplied beyond necessity)*; for 'entities' read: levels of description or processing.

1.2. A Functional Prosodic Model

In this section, we sketch an alternative model that puts emphasis on *function*, not on phonological *form*–actually, all other working approaches towards using prosodic information in ASU we know of are along these lines (cf. Shriberg et al., 1998; Nöth et al., 2000 and the references given in these papers). The prosodic functions that are

generally considered to be the most important ones on the linguistic level are the marking of boundaries, accents, and sentence mood; boundaries can delimit syntactic, semantic, or dialogue units. For these phenomena, the first step is the annotation of a large database. Annotation should be as detailed as possible, but more detailed classes can–if necessary–be mapped onto higher classes. We still do not know how many classes are most appropriate for the pertinent linguistic phenomena; it is, however, our experience that quite often, the higher linguistic modules can work fairly well with only two binary classes: present vs. not present.[6] The phonetic form is modelled directly with a large feature vector which uses all available information on (appropriately normalized) F_0, energy, and duration; other linguistic information pertaining to, for instance, part of speech classes is used as well. It is not a theoretical question but one of practical reasoning, availability, implementation, and recognition performance whether all this information is processed sequentially or in an integrated procedure. The model, classification results, and the use of prosodic knowledge in higher linguistic modules are described in Batliner et al. (2000), Nöth et al. (2000), and Batliner et al. (2001c).

1.3. Which Common Ground for ASU and Prosodic Models?

Mainstream ASU nowadays means statistical processing. For this approach, large databases and a standardization of different annotation concepts are needed. ToBI has been a step in the right direction but is still based too much on (one specific) phonology; it is not an *across models* but a *within model* approach; cf. the standardization efforts for dialogue act annotations described in Klein (1999). Only if they are based on a successful standardization, can the labels of different (intonation) models be used together in order to overcome the sparse data problem. The *primacy of phonology* has to give way to more practical considerations; models should take into account the requirements–and limitations–of speech processing modules. For instance, even if word recognition computes phone segment boundaries, these are often not available afterwards: the output is a word hypotheses graph with word boundaries only. An additional computation of phone segment boundaries would mean a considerable overhead.[7] Thus, intonation models that require an exact alignment with phones cannot be used. Therefore, we only used word boundaries in the final version of our prosody module in Verbmobil (Batliner et al., 2000)–without a decrease in performance!

The two cultures, viz. the humanities and engineering approaches, are still rather remote from each other. As in politics, one should begin with small steps, and with steps that pay off immediately. This means that subtle theoretical concepts are not well suited, but prosodic markers are, which are visible and stable enough to be classified reliably even in a realistic, *real life* setting. Thus it can be guaranteed that prosody really finds its way into ASU, because speech engineers can be convinced more easily that the integration of prosody indeed pays off. Later, it will be simply a matter of conquering or not: if more subtle differences can be modelled with prosodic means and classification performance is good enough, it will be no problem to incorporate them into ASU.

2. SPEECH SYNTHESIS

Prosodic models have been extensively applied in speech synthesis, simply because there is an obvious need for every TTS system to generate prosodic properties of speech if the synthesis output is to sound even remotely like human speech. However, the *necessity* of synthesizing prosody has as yet not resulted in a *generally agreed upon* approach to prosodic modelling. This statement holds for the assignment of segmental durations as well as for the generation of F_0 curves, the acoustic correlate of intonation contours.[8]

Intonation research is extremely diverse in terms of theories and models. On the *phonological* side, there is little consensus on what the basic elements are: tones, tunes, uni-directional motions, multi-directional gestures, etc. Modelling the *phonetics* of intonation is equally diverse, including interpolation between tonal targets (Pierrehumbert, 1981), superposition of underlying phrase and accent curves (Fujisaki, 1988), and concatenation of line segments ('t Hart et al., 1990).

Modelling *speech timing* for synthesis is less diverse. The important role of the syllable as a central processing unit in speech production and perception is widely accepted, but there is an ongoing controversy about how to best implement the pertinent effects in a model of speech timing; cf. the *syllabic timing model* proposed by Campbell (Campbell and Isard, 1991; Campbell, 1992), on the one hand, and the sums-of-products model of *segmental duration* proposed by van Santen (1993; 1994), on the other hand.

In natural speech, tonal and temporal prosodic properties are coproduced, and there is an increasing body of evidence that tonal and temporal as well as spectral properties of speech are jointly planned by the speaker in a way that prosodic events can be optimally perceived by

the listener (House, 1990, House, 1996, Dogil and Möbius, 2001). The conventional solution in speech synthesis systems, in contrast, embodies a unidirectional flow of information instead of synergy: first, the duration of speech sounds and syllables is assigned and then the F_0 contour of the utterance is computed.

One pivot in our discussion of prosodic models in automatic speech processing is the relevance of a phonological level of description.[9] This aspect is rather indistinct with respect to models of speech timing. The remainder of this section therefore concentrates on the use and usability of intonation models in speech synthesis.

2.1. Intonation Synthesis: A Two-Stage Process

Intonation synthesis can be viewed as a two-stage process, the first aimed at representing grammatical structures and referential relations on a *symbolic* level and the second at rendering *acoustic* signals that convey the structural and intentional properties of the message. Intonation models differ in terms of the interface that they provide between the higher linguistic components and the acoustic prosodic modules.

In many TTS systems sophisticated methods, such as syntactic parsing and part-of-speech tagging, are applied in the service of providing sufficient information to drive the acoustic prosodic components of the system, in particular, the intonation model. The intonationally relevant information comprises sentence mood as well as the location and strength of phrase boundaries and the location and type of accents.

Establishing the relation between syntactic structure and intonational features is among the most challenging subtasks of TTS conversion, and its imperfection contributes to the perceived lack of naturalness of synthesized speech. This shortcoming is unavoidable because TTS systems have to rely on the computation of linguistic structures from orthographic text, a level of representation that is notoriously poor at coding prosodic information in many languages.

The task of the acoustic-phonetic component of an intonation model in TTS is to compute continuous acoustic parameters (F_0/time pairs) from the symbolic representation of intonation. A large variety of models have been applied in TTS systems to perform this task, including implementations of the major frameworks of intonation theory: phonological models that represent the prosody of an utterance as a sequence of abstract units (e.g., tones), viz. *tone sequence* models; and acoustic-phonetic models that interpret F_0 contours as complex patterns resulting from the superposition of several components, viz. *superposition* models.

Besides these prevalent models, several other approaches have been taken, in particular *perception-based*, *functional*, and *acoustic stylization* models. For instance, the INTSINT system (Hirst and Di Cristo, 1998b) performs an automatic analysis and generation of F_0 curves by deriving a sequence of target points, specified in time and frequency, that represents a stylization of the F_0 curve.

All of these approaches rely on a combination of *data-driven* and *rule-based* methods: they all systematically explore natural speech databases, but vary in terms of what is derived from the analysis to drive intonation synthesis. For instance, there are two different approaches to acoustic stylization modeling. In one approach, continuous acoustic parameters are interpreted as directly representing intonation events (Taylor, 2000); in the other approach, intonation events are related to phonological entities such as tones or register via prototype building (Möhler, 1998). The abstract tonal representation provided by phonological intonation models is converted into F_0 contours by means of phonetic realization rules. The phonetic rules determine the F_0 values of the (H and L) targets, based on the metric prominence of syllables they are associated with, and on the F_0 values of the preceding tones. The phonetic rules also compute the temporal alignment of tones with accented syllables. Fujisaki's classic superpositional model computes the F_0 contour by additively superimposing phrase and accent curves and a speaker-specific F_0 reference value. Phrase and accent curves are generated from discrete commands, the parameter values of which are usually derived by generalization of values statistically estimated from a speech database. While this model can be characterized as primarily acoustically oriented (and physiologically motivated), it is possible to find phonological interpretations of its commands and parameters (Möbius, 1995).

2.2. Intonation Synthesis and Phonetic Detail

F_0 contours as *acoustic realizations of accents* vary significantly depending on the structure (i.e., the segments and their durations) of the syllables they are associated with. For example, F_0 peak location is systematically later in syllables with sonorant codas than in those with obstruent codas (*pin* vs. *pit*), and also later in syllables with voiced obstruent onsets than with sonorant onsets (*bet* vs. *yet*). Moreover, the F_0 peak occurs significantly later in polysyllabic accent groups than in monosyllabic ones (van Santen and Möbius, 2000).

Intonation models need to generate as much of this phonetic detail as possible. The quantitative model of F_0 alignment proposed by van Santen

and Möbius (2000), for instance, explains the diversity of surface shapes of F_0 contours by positing that accents belonging to the same phonological (and perceptual) class can be generated from a common *template* by applying a common set of *alignment parameters*. The templates are representatives of phonological intonation events of the type predicted by intonation theories, i.e. accents and boundaries.

Acoustic stylization models (Möhler, 1998; Taylor, 2000) also synthesize F_0 contours from a small number of *prototypical patterns*. They learn and predict phonetic details of F_0 movements from a set of features comprising segmental, prosodic and positional information. While the F_0 prototypes are defined as being phonetically distinct, they are also intended to be related to phonologically motivated intonation events.

2.3. What is the Common Ground for TTS and Prosodic Models?

In Section 1 we have argued that the most appropriate type of intonation model for ASU would be one that provides a *functional* representation of the positions of accents and phrase boundaries; any intermediate phonological level only introduces a quantisation error. In the ToBI notation (Silverman et al., 1992) such a functional representation would consist only of the location of accents (the stars) and phrase boundaries (the percents). In the following we discuss to what extent, or whether at all, the conclusions drawn for the ASU domain are valid for the TTS domain too; in doing so we consider both the state of the art in intonation synthesis and the feasibility of alternative designs.

In state-of-the-art TTS systems, such as Festival (Black et al., 1999), Bell Labs (Sproat, 1998), AT&T (Syrdal et al., 2000), and others, the only symbolic prosodic information used–apart from sentence mood–is the *location of accents* and *boundaries*. This design can be characterized as the bare-bones minimum of prosodic modelling, because phrase structure and accentual structure are surface reflections of the underlying semantic and syntactic structure of the sentence, and at least a coarse representation of phrasing and accenting needs to be achieved by any self-respecting TTS system.

However, it has been demonstrated that models which use more detailed and more precise input information, for instance ToBI *accent type* labels in addition to accent location alone, can generate F_0 contours that are perceptually more acceptable than models which use accent location alone (Syrdal et al., 1998). The problem is that computing from text such detailed intonational features as accent type is difficult and unreliable. It should therefore come as no surprise that even the very

same research group that so convincingly demonstrated the importance of detailed input information, came up with the solution (the 'ToBI Lite' approach) of collapsing ToBI accent labels onto merely two categories and of mapping only edge tones marking major phrases onto just one category (Syrdal et al., 2000). Note, however, that strictly speaking, these results are an indication that a greater variation of accent types will result in a higher degree of acceptability; they are no proof that a ToBI-like accent representation is the best or the only possibility of modelling variation.

The degree of potential improvement to synthesized prosody can also be illustrated by manually marking up the text or by providing access to semantic and discourse representations (Prevost and Steedman, 1994). In practice and in existing end-to-end systems, however, the situation in intonation synthesis appears to be similar to the one described for the ASU domain. But it is still worth noting that relying for the most part only on accent and boundary location is not a *judicious design decision* made by speech synthesis researchers but one made by system developers *bowing to necessity*. It is evident that much more information than just the stars and the percents is needed to achieve the kind and degree of improvement to intonation synthesis that has been demonstrated in fragmentary research systems.

Can we do without a phonological representation of intonation in speech synthesis? Certain synthesis strategies beyond the classic TTS scenario offer more immediate interfaces between symbolic and acoustic representations of intonation. *Concept-to-speech* systems, in particular, provide a direct link between language generation and acoustic-prosodic components. A concept-to-speech system has access to the complete linguistic structure of the sentence that is being generated; the system knows *what* to say, and *how* to render it. Such a system may potentially incorporate semantic and discourse representations like those used in the experiment by Prevost and Steedman (1994).

Yet, even in concept-to-speech systems, it is still necessary to specify the mapping from semantic to symbolic features and from symbolic to acoustic features. The issue of how much, and what kind of, information the language generation component should deliver to optimize the two mapping steps (i.e., the definition of a semantics/syntax-prosody interface) is a hot research topic. Once the two mapping steps are optimized, we may be able to advance one step further and get rid of the intermediate level (i.e., a phonological prosodic representation) just as hypothesized for ASU (see Section 1.1).

The most drastic redesign of intonation synthesis would be to avoid synthesizing intonation in the first place. Consider the early *unit selection*

synthesis approach implemented in the CHATR TTS system (Black and Taylor, 1994). Unit selection generates speech by concatenating speech segments of varying length (as short as half-phones and as long as entire utterances) that are extracted at runtime from a large speech database. CHATR follows the strategy of simply *resequencing* speech segments without performing any modifications by signal processing. The underlying assumption is that the listener will tolerate occasional spectral or prosodic mismatches in an utterance if the quality of the output speech in general approaches that of natural speech.

The unit selection algorithm attempts to minimize two types of cost, one for *unit distortion* and one for *continuity distortion*. The former is a measure of the distance of the candidate unit from the desired target, whereas the latter is a measure of the distance between two adjacent units at the concatenation point. Each target is specified by a feature vector that comprises positional, spectral, and prosodic features, and the values of these features for a given target are specified on the basis of some kind of model. In the case of prosodic features, the desired F_0 contour is usually predicted by an intonation model. Thus, even in the most extreme version of corpus-based synthesis, the mapping from a target specification to acoustic-phonetic details of candidate units is mediated by a model that relies on a symbolic representation of intonation, which customarily amounts to a phonologically based or motivated intonation model.

A phonological approach is even advocated explicitly in an interesting recent approach to unit selection termed *phonological structure matching* (Taylor and Black, 1999), where phonological information, such as canonical pronunciation, positional factors and accentuation, is used for unit selection, instead of narrow phonetic transcriptions and absolute duration and F_0 values. The key idea in this approach is that most of the variability in the speech signal is predictable and that units selected from the appropriate context are likely to have the right specifications, including prosodic ones. This means that intonation contours generated by models may not be necessary anymore. But what will still be relevant is the knowledge about the factors and their respective quantitative effects on observed contours; this knowledge can be used to develop powerful unit selection criteria.

3. WHICH COMMON GROUND FOR ASU AND TTS–WITH OR WITHOUT PROSODIC MODELS?

We have illustrated that the basic problems connected with the use of prosodic models in speech processing are similar for ASU and TTS. One of these problems is the lack of an appropriate annotation concept.

We have argued that ToBI–while representing a step in the right direction–is based too much on one specific intonational phonology and does not generalize across models. We have further argued that in the ASU context, ToBI provides a special layer of representation that is both too abstract (i.e., too far from the signal to be useful as input to classifiers) and at the same time not abstract enough, with some of its notational units lacking a linguistic counterpart. A mirror image of this situation is evident in the context of TTS, where ToBI lacks the required granularity.

3.1. Shared Models for ASU and TTS?

In our view, the most appropriate type of intonation model for ASU would be one that provides a functional representation of the positions of accents and phrase boundaries without any intermediate prosodic-phonological level. At present, such a type of model is widely used in intonation synthesis, albeit with some intermediate prosodic-phonological representation. This apparent similarity between ASU and TTS requirements is brought about by very different motivations. In ASU a finer-grained level of description has not yet been shown to model reliably the linguistic function that it presumably corresponds to. In TTS, in contrast, more detailed input information is required to generate F_0 contours that are perceptually more acceptable than those based on accent and phrase boundary locations alone. While computing such features is extremely hard in a TTS framework, it may be accessible in different speech synthesis strategies such as concept-to-speech.

Recent advances in TTS can be partly attributed to the use of statistical methods for detecting relevant features in large databases, learning them, and modelling them. A standardized annotation concept would be an additional advantage. However, the prevalent annotation convention, viz. ToBI, misses the required granularity: it is confined too much within one type of intonation model; it is too elaborate and specific in terms of its descriptive inventory to lend itself as a generic interface to higher-level linguistic-prosodic analysis; at the same time it is far too abstract to facilitate a computation of the rich phonetic detail and precise alignment that F_0 contours require in order to sound natural. Data-driven intonation models, on the other hand, can learn to synthesize these details. For the integration in a TTS system, a complete intonation model needs to provide a mapping from categorical phonological elements to continuous acoustic parameters. Quantitative models such as those presented recently (Möhler, 1998; Taylor, 2000; van Santen and Möbius,

2000) offer feasible solutions to the F_0 generation task. However, it is not clear yet whether these two approaches can be integrated into existing TTS systems without any additional phonological representation.

We believe that no intonation model equally appropriate for both tasks, ASU and TTS, is currently available. The requirements are, for the time being and for some time to come, too different. They might converge in the future, giving rise to a unified solution to prosodic modelling, but we simply do not know when and whether this will be the case.

3.2. Multilinguality

One aspect that we have not discussed in this chapter yet is *multilinguality*. Both ASU and TTS have gone multilingual:

In the Verbmobil system (Batliner et al., 2000), prosodic information is computed for ASU for three languages, viz. German, English, and Japanese. The *multilingual prosody module* facilitates the sharing of prosodic feature extraction and classification procedures, which are considered to be language independent. Note, however, that it is not clear yet whether or not the same set of features is appropriate for typologically different languages, for instance tone and non-tone languages. Language specific data, such as duration normalization tables, are kept in separate structures and are loaded as needed. Similarly, separate classification parameters, such as different n-gram sizes, can be specified by means of configuration files (Batliner et al., 2000).

In remarkably the same spirit, the multilingual intonation component in the Bell Labs TTS system (Sproat, 1998) is used for a number of intonationally quite diverse languages, including American English, French, German, Italian, Japanese, Mexican Spanish, and Russian; this component implements the quantitative model by van Santen and Möbius (2000). One of the key assumptions of this model is that phonological accent classes can be mapped onto a corresponding number of distinct F_0 templates by means of alignment parameter matrices (see Section 2.2). Language specific adjustment pertains to transformations of these parameter matrices, which can be handled offline and stored in configuration files. Again similar to the ASU prosody design presented above, one of the most intriguing research questions is to what extent the inventory of templates can be shared across languages: notice that Mandarin Chinese is not currently handled by this multilingual intonation approach (van Santen et al., 1998).

3.3. No Panacea: the Database Argument

Sufficiently large and discourse-rich, (prosodically) *annotated databases* are of course a desideratum (Botinis et al., 2001). They are necessary for a good and robust classification of prosodic events in ASU, and they are necessary for modelling variability in TTS. They are, however, definitely no panacea; in particular, they cannot make up for the lack of relevance of intonation models like ToBI for ASU: First, an elaborate prosodic annotation is very *time-consuming* (Batliner et al., 1998) and therefore simply too expensive. This might not be a 'scientific' argument but is nonetheless a decisive obstacle. Second, it is too complicated and thus prone to *low inter-labeller correspondence*, cf. ToBI vs. 'ToBI Lite' (Syrdal et al., 2000). Third, it is doubtful whether 'real-life' spontaneous speech is always *prosodically rich* to the extent that special and/or rare functions are indicated by prosodic means; an extrapolation of constructed examples to spontaneous speech might turn out to be mere wishful thinking. Fourth, it is said that *more data* is always better than less data; but on the other hand, with 'only' 90 minutes of annotated speech material for German, and less than half the amount of data for English, we obtained the following overall classification rates (two-class problems): German boundaries, 87%; German accents, 81%; English boundaries, 92%, English accents, 79% (Batliner et al., 2000). Given the fact that inter-labeller reliability has not been proven to be very high for such tasks, it might not be possible to improve on these results to any considerable extent, even with a much larger training database. The benefit of larger training databases might thus not be the possibility to obtain much better classification rates but the possibility to model variability much better. That means, in turn, that performance will not fall drastically if one has to deal with new tasks, new scenarios, or new applications. (Regarding the portability of speech recognizers to new tasks, cf. Lamel et al., 2001.) Finally, and most importantly, more labels cannot be a remedy for the missing link to clear functions, cf. Section 1.1.

3.4. A Catalogue of Shared Tasks

A straightforward way for ASU and TTS to co-operate would be to exchange knowledge, concepts, rules, algorithms and special databases between colleagues and research sites. Such a sharing of methods and resources is already a reality in several subdomains of speech processing, cf. efficient search algorithms (Viterbi search), signal representations (e.g., HMM), or the use of linguistic or phonetic information (language

models, duration models). This kind of exchange and sharing, as well as joint future work, would be a Type 3 approach (cf. the Introduction), a direct link between ASU and TTS, not mediated by traditional phonological models.

It might be argued that the tasks for ASU and TTS differ because TTS normally focuses on a formal speaking style, whereas ASU has to deal with a more casual, informal style. In our opinion, this is not a categorical but only a gradual difference which might diminish in the future: with more elaborate synthesis, some computer speech will surely be more casual in order to approximate human-human communication. On the other hand, if large-scale content extraction has to be performed automatically from, e.g., radio news, ASU will have to deal with formal speech as well.

In the context of prosody, we would propose the following catalogue of shared tasks:

- Inventory of relevant linguistic prosodic functions: marking of accents, phrases, discourse structure, etc. This can be illustrated by the rules for accent assignment that have been developed independently within ASU (Batliner et al., 1999) and TTS (Hirschberg, 1993; Widera et al., 1997), to mention just a few.
- Inventory of relevant paralinguistic prosodic functions: emotions/user states, individual speaker traits, etc. (Batliner et al., 2001c).
- Inventory of structured prosodic features: these features pertain to linguistically relevant units of speech, for instance phonemes, syllables, words, phrases, etc. Structured prosodic features are derived from basic acoustic-prosodic features, such as F_0, energy values, etc. This typology of prosodic features is described in Kießling (1997) and Nöth et al. (2000).
- Inventory of lexical prosodic features: word accent position, part-of-speech information, etc.
- Inventory of syntactic/semantic prosodic features: sentence mood, syntactic structure and boundaries, positional and counting factors, centres of information.
- Annotation system, oriented towards function (not form), motivated by practical (not phonological) considerations.
- Procedures for detecting, learning, and modelling of prosodic features from speech databases. In state-of-the-art TTS, prosodic features are learned from single-speaker databases. It might be

feasible to train models on multi-speaker corpora to obtain prototypes via clustering or averaging (Batliner and Nöth, 1989); the prototypes might each represent one possible (virtual) and plausible speaker, but they do not have to represent any particular speaker in the corpus.

- Integration of *all* prosodic parameters and features, not just F_0, in TTS, following the ASU approach and acknowledging the fact of co-production of prosodic features in natural speech by co-modelling them.

A common task for ASU and TTS is to learn the mapping from acoustics to categories. In ASU the direct mapping of acoustic features onto functions without any intermediate phonological level is standard. In TTS, such a direct mapping might be feasible as well, for both offline training and runtime synthesis; hopefully, this will be a research avenue for the near future.

4. CONCLUDING REMARKS

Coming back to the title of this book, 'Integration of Phonetic Knowledge in Speech Technology', we would like to refer to the distinction made in Section 1.1 between basic knowledge on the one hand and transformed/mediated knowledge on the other hand. This is, of course, a gross distinction; there is, in reality, a continuum from pure basic, acoustic, knowledge (e.g., about concrete F_0 values) to transformed, very abstract knowledge. Phonetic knowledge is thus never purely basic but always transformed to a certain degree. The decisive step is, however, when it comes to a considerable reduction of information in order to achieve a level of 'phonological adequacy', cf. the quantisation error mentioned in Section 1.1. In our opinion, phonetic/prosodic knowledge that has not yet crossed this rubicon is of course necessary for speech technology. Actually, it has always been used even if speech engineers might not have been aware of this fact. As for transformed/mediated phonological knowledge, we are not that sure and opt rather for those kinds of co-operation between ASU and TTS that are described in the catalogue of shared tasks in Section 3.4.

By successively working through this catalogue, we might eventually end up with something that might be called a new type of prosodic model, capable of explaining and predicting variability, and which can connect phenomena and their processing by automatic means more directly than current intonation models do.

ACKNOWLEDGMENTS

The authors wish to thank Gregor Möhler, Elmar Nöth, and Antje Schweitzer, as well as the two editors, for valuable comments and discussion on earlier versions of this chapter. We also acknowledge the helpful suggestions by two anonymous reviewers. This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the SmartKom project under Grants 01IL905D and 01IL905K7. The responsibility for the content lies with the authors.

NOTES

[1] In our understanding, automatic speech recognition (ASR) comprises 'only' word recognition, which is a necessary prerequisite for automatic speech understanding (ASU). Thus, if we have to choose one of these two terms, we prefer ASU because it covers the whole story and not only some part of it. Moreover, 'understanding' is more directly connected with higher linguistic levels such as syntax, semantics, and pragmatics. If we consider the work on automatic processing of prosody conducted so far, it might be the case that the impact of prosody is much stronger for these higher levels, compared to the impact on word recognition.

[2] Strictly speaking, TTS is the customary acronym for *text-to-speech*, but in the context of this chapter we have opted to use it for any kind of speech synthesis, disregarding the exact type of input representation (e.g., text, concept, or structured document), unless explicitly indicated otherwise.

[3] The Verbmobil system was developed in a large-scale German research project focusing on automatic speech-to-speech translation in appointment scheduling dialogues (Wahlster, 2000).

[4] In Batliner (1989a) we have discussed the problem of reification from a slightly different point of view. An evident analogy on the segmental level is the famous *rabid/rapid* distinction (Lisker, 1978): it might be possible for a strictly phonological approach to work with only one distinctive feature, whereas for automatic speech processing, this would be a rather suboptimal approach.

[5] "Probably, it will be very difficult to detect [automatically] a boundary marker that takes the form of a declination reset. ... [If its identification] in the acoustic signal cannot take place until a close-copy stylization has first been made, and that is the present situation, one can imagine that its automatic detection will only become a possibility once the technique of automatic stylization has been sufficiently mastered" ('t Hart et al., 1990, page 182). That simply means to beg the question–there is ample evidence nowadays, that boundaries can be detected without the help of such phonological concepts as declination (Batliner et al., 1998; Nöth et al., 2000).

[6] Of course, linguists would like to get information from prosody for more subtle distinctions; maybe such distinctions can be provided and used successfully in the future, but not with the present state of the art and, especially, of the databases available (sparse data problem).

[7] It would of course be no problem in principle for a word recognition module to store computed segment boundaries. In distributed systems, however, if prosody has to use

the output of some existing word recognition module, this would mean rewriting the module accordingly–which could not be done in the Verbmobil system due to project-internal constraints. Instead, in the first phase of the project, phone segments were re-computed in the prosody module, which caused a significant overhead. Thus, in the second phase, we computed only word based prosodic features–without any reduction of recognition performance (Batliner et al., 2000).

[8] Notice that TTS systems do not usually provide a prosodic model for the amplitude profile of the synthetic utterance.

[9] We do not argue against any phonological level as such. If we consider the well-established phonological level for word recognition, then there is a clear relationship between distinctive form and function; such a clear relationship, however, has not yet been proven for the prosodic level. Note that a *prosodic* phonological level might still be relevant for language typology, second language learning, etc., even if it might be irrelevant for the automatic processing of speech.

REFERENCES

Batliner, A. Eine Frage ist eine Frage ist keine Frage. Perzeptionsexperimente zum Fragemodus im Deutschen. In: Altmann, H., Batliner, A., and Oppenrieder, W., (eds.), *Zur Intonation von Modus und Fokus im Deutschen*, Niemeyer, Tübingen, 1989a: 87–109.

Batliner, A. Wieviel Halbtöne braucht die Frage? Merkmale, Dimensionen, Kategorien. In: Altmann, H., Batliner, A., and Oppenrieder, W., (eds.), *Zur Intonation von Modus und Fokus im Deutschen*, Niemeyer, Tübingen, 1989b: 111–162.

Batliner, A., Buckow, A., Niemann, H., Nöth, E., and Warnke, V. The prosody module. In: (Wahlster, 2000), 2000: 106–121.

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., and Niemann, H. Boiling down prosody for the classification of boundaries and accents in German and English. In: *Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark)*, 4, 2001a: 2781–2784.

Batliner, A., Kompe, R., Kießling, A., Mast, M.,Niemann, H., and Nöth, E. M = Syntax + Prosody: A syntactic-prosodic labelling scheme for large spontaneous speech databases. *Speech Communication*, 25(4) (1998): 193–222.

Batliner, A., Möbius, B., Möhler, G., Schweitzer, A., and Nöth, E. Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground. In: *Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark)*, 4 (2001b): 2285–2288.

Batliner, A. and Nöth, E. The prediction of focus. In: *Proceedings of the European Conference on Speech Communication and Technology (Paris)*, 1989: 210–213.

Batliner, A., Nöth, E., Buckow, J., Huber, R., Warnke,V., and Niemann, H. Whence and whither prosody in automatic speech understanding: A case study. In: Bacchiani, M., Hirschberg, J., Litman, D., and Ostendorf, M., (eds.), *Proceedings of the Workshop on Prosody and Speech Recognition 2001 (Red Bank, NJ)*, 2001c: 3–12.

Batliner, A., Nutt, M., Warnke, V., Nöth, E., Buckow,J., Huber, R., and Niemann, H. Automatic annotation and classification of phrase accents in spontaneous speech. In: *Proceedings of the European Conference on Speech Communication and Technology (Budapest)*, 1, 1999: 519–522.

Black, A.W. and Taylor, P. CHATR: a generic speech synthesis system. In: *Proceedings of the International Conference on Computational Linguistics (Kyoto, Japan)*, 2, 1994: 983–986.

Black, A.W., Taylor, P., and Caley, R. *The Festival speech synthesis system–System documentation.* CSTR Edinburgh. Edition 1.4, for Festival version 1.4.0.[http://www.cstr.ed.ac.uk/projects/festival/manualfestival/_toc.html], 1999.

Botinis, A., Granström, B., and Möbius, B. Developments and paradigms in intonation research. *Speech Communication*, 33(4) (2001): 263–296.

Campbell, W.N. Syllable-based segmental duration. In: Bailly, G., Benoît, C., and Sawallis, T.R., (eds.), *Talking Machines: Theories, Models, and Designs*, Elsevier, Amsterdam, 1992: 211–224.

Campbell, W.N. and Isard, S.D. Segment durations in a syllable frame. *Journal of Phonetics*, 19 (1991): 37–47.

Dogil, G. and Möbius, B. Towards a model of target oriented production of prosody. In: *Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark)*, 1 (2001): 665–668.

Fujisaki, H. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In: Fujimura, O., editor, *Vocal Physiology: Voice Production, Mechanisms and Functions*, Raven, New York, 1988: 347–355.

Grice, M., D'Imperio, M., Savino, M., and Avesani, C. Strategies for intonation labelling across varieties of Italian. In: Jun, S. A., editor, *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, Oxford, UK, 2004.

Hirschberg, J. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63(1–2) (1993): 305–340.

Hirschberg, J. and Pierrehumbert, J. The intonational structuring of discourse. In: *Proceedings of the 24th Annual Meeting of the ACL (New York)*, 1986: 136–144.

Hirst, D. and Di Cristo, A., (eds.), *Intonation Systems–A Survey of Twenty Languages*. Cambridge University Press, Cambridge, UK, 1998a.

Hirst, D. and Di Cristo, A. A survey of intonation systems. In: (Hirst and Di Cristo:1998a), 1998b: 1–44.

House, D. *Tonal Perception in Speech*. Lund University Press, Lund, 1990.

House, D. Differential perception of tonal contours through the syllable. In: *Proceedings of the International Conference on Spoken Language Processing (Philadelphia, PA)*, 1 (1996): 2048–2051.

Kießling, A. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Berichte aus der Informatik. Shaker, Aachen, 1997.

Klein, M. Standardization efforts on the level of dialogue act in the MATE project. In: *Proceedings of the ACL Workshop "Towards Standards and Tools for Discourse Tagging" (University of Maryland)*, 1999: 35–41.

Ladd, D.R. *Intonational Phonology*. Cambridge University Press, Cambridge, UK, 1996.

Ladd, D.R. Introduction to part I. Naturalness and spontaneous speech. In: (Sagisaka et al., 1997), 1997: 3–6.

Lamel, L., Lefevre, F., Gauvain, J.-L., and Adda, G. Portability issues for speech recognition technologies. In: *Proceedings of the Human Language Technology Conference HLT-2001 (San Diego, CA)*, 2001: 9–16.

Lisker, L. Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction. Haskins Laboratories: Status Report on Speech Research SR-55/56 1978: 127–132.

Möbius, B. Components of a quantitative model of German intonation. In: *Proceedings of the 13th International Congress of Phonetic Sciences (Stockholm)*, 2 (1995): 108–115.

Möhler, G. Describing intonation with a parametric model. In: *Proceedings of the International Conference on Spoken Language Processing (Sydney)*, 7 (1998): 2851–2854.

Nöth, E., Batliner, A., Kießling, A., Kompe, R.,and Niemann, H. Verbmobil: The use of prosody in the linguistic components of a speech understanding system. *IEEE Transactions on Speech and Audio Processing*, 8 (2000): 519–532.

Pierrehumbert, J. *The phonology and phonetics of English intonation*. PhD thesis, MIT, Cambridge, MA, 1980.

Pierrehumbert, J. Synthesizing intonation. *Journal of the Acoustical Society of America*, 70 (1981): 985–995.

Prevost, S. and Steedman, M. Specifying intonation from context for speech synthesis. *Speech Communication*, 15(1–2) (1994): 139–153.

Sagisaka, Y., Campbell, N., and Higuchi, N., (eds.), *Computing prosody–Computational models for processing spontaneous speech.* Springer, New York, 1997.

Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky,D., Ries, K., Cocarro, N., Martin, R., Meteer, M., and Ess-Dykema, C. V. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41 (1998): 439–487.

Siepmann, R. Phonetische Intonationsmodelle und die Parametrisierung von kontrastiven Satzakzenten im Deutschen. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation (München)*, FIPKM, 38 (2001): 3–111.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M.,Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. ToBI: A standard for labeling English prosody. In: *Proceedings of the International Conference on Spoken Language Processing (Banff, Alberta)*, 2 (1992): 867–870.

Sproat, R., (ed.), *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.* Kluwer, Dordrecht, 1998.

Studdert-Kennedy, M. and Hadding, K. Auditory and linguistic processes in the perception of intonation contours. *Language and Speech*, 16 (1973): 293–313.

Syrdal, A., Möhler, G., Dusterhoff, K., Conkie, A., and Black, A. Three methods of intonation modeling. In: *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, 1998: 305–310.

Syrdal, A.K., Wightman, C.W., Conkie, A., Stylianou, Y., Beutnagel, M., Schroeter, J., Strom, V., Lee, K.-S., and Makashay, M.J. Corpus-based techniques in the AT&T NextGen synthesis system. In: *Proceedings of the International Conference on Spoken Language Processing (Beijing)*, 3 (2000): 410–413.

't Hart, J., Collier, R., and Cohen, A. *A Perceptual Study of Intonation–An Experimental-Phonetic Approach to Speech Melody*. Cambridge University Press, Cambridge, UK, 1990.

Taylor, P. and Black, As. W. Speech synthesis by phonological structure matching. In: *Proceedings of the European Conference on Speech Communication and Technology (Budapest)*, 2 (1999): 623–626.

Taylor, P.A. Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America*, 107(3) (2000): 1697–1714.

van Santen, J.P.H. Exploring *N*-way tables with sums-of-products models. *Journal of Mathematical Psychology*, 37(3) (1993): 327–371.

van Santen, J.P.H. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8 (1994): 95–128.

van Santen, J.P.H. and Möbius, B. A quantitative model of F0 generation and alignment. In: Botinis, A., (ed.), *Intonation–Analysis, Modelling and Technology*, Kluwer, Dordrecht, 2000: 269–288.

van Santen, J.P.H., Möbius, B., Venditti, J., and Shih, C. Description of the Bell Labs intonation system. In: *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, 1998: 293–298.

Wahlster, W., (ed.), *Verbmobil: Foundations of Speech-to-Speech Translations*. Springer, Berlin, 2000.

Widera, C., Portele, T., and Wolters, M. Prediction of word prominence. In: *Proceedings of the European Conference on Speech Communication and Technology (Rhodes, Greece)*, 2 (1997): 999–1002.

JULIE CARSON-BERNDSEN and MICHAEL WALSH

PHONETIC TIME MAPS

Defining Constraints for Multilinear Speech Processing

ABSTRACT. This paper presents a constraint-based model for the interpretation of multilinear representations of speech utterances which can provide important fine-grained information for speech recognition applications. The model uses explicit structural constraints specifying *time maps*–overlap and precedence relations between features in both the phonological and the phonetic domains–in order to recognise well-formed syllable structures. In the phonological domain, these constraints together form a complete phonotactic description of the language, while in the phonetic domain, the constraints define the internal structure of phonological features based on phonetic realisations. The constraints are enhanced by a constraint relaxation procedure to cater for underspecified input and allow output representations to be extrapolated based on the phonetic and phonological information contained in the constraints and the rankings which have been assigned to them. This approach thus describes the integration of explicit phonetic knowledge into a computational linguistic model to improve robustness in speech recognition.

KEYWORDS. phonotactic models, phonetic time maps, finite state transducers, multilinear representations

1. INTRODUCTION

This paper presents a computational linguistic model which has been developed for the explicit purpose of providing fine-grained structural information for speech technology applications. The model has been described in detail elsewhere (Carson-Berndsen, 1998, 2000) but we review the model below with explicit reference to the types of constraints it assumes and discuss how these have been enhanced to address the notion of robustness in speech recognition. Our primary concern in this paper is to highlight areas in which we believe explicit phonetic and phonological knowledge constraints can contribute to speaker- and

Address for Correspondence:
University College Dublin

W. J. Barry and W. A. van Dommelen, The Integration of Phonetic Knowledge in Speech Technology, 45–66.

corpus-independent speech recognition and reduce the need for training data.

Pols (1999) highlights a number of areas in which insights of computational phonetics should be able to contribute to a more robust speech technology improving system performance. For example, humans have the ability both to compensate for noisy or underspecified information in speech utterances and to predict what will come next. Speech recognisers must be flexible and adaptable and should model more explicitly the functionality of human behaviour, although not necessarily attempt to replicate it exactly. There are varying systematic types of phonetic knowledge which could be incorporated more explicitly into speech recognition systems. Such consistent (predictable) characteristics of speech include durational variability, coarticulation and communicative expectation, to name but a few (cf. Pols, 1999 for a more comprehensive list and explanation). Experimental phonetics has concerned itself for many years now with such characteristics but, as yet, only very few results of these studies have been integrated into speech recognisers explicitly. The motivation for the research presented in this paper has grown out of a desire to more explicitly model phonetic and phonological knowledge for use in the speech recognition process. Our approach proposes modelling such knowledge in terms of constraints on the well-formedness of phonological and phonetic representations, modelling the ability of the native speaker to decipher legal from illegal structures in a language, to predict what fully specified set of structures corresponds to some underspecified (or noisy) representation, and to interpret the overlapping gestures found in coarticulation.

In what follows, we discuss how phonological and phonetic constraints can be modelled and used by a computational linguistic model for speech recognition. This is very much in line with parallel research by Deng (1997, 1998) who proposed an autosegmental feature-based approach to generating word pronunciation models represented as finite state automata which were then interfaced with a trended HMM. While our motivation is very similar, the constraint-based model is based on a complete phonotactic description of a language which is used to provide top-down constraints on such multilinear (autosegmental) feature representations. The fine grained information developed in connection with this model may also be used for fine tuning of alternative stochastic approaches (cf., for example, Jusek et al., 1994).

The next section discusses the motivations for the constraint-based model, in particular the multilinear representation of speech utterances

which serves as input to the model. Section 3 briefly reviews the model in the context of speech recognition and section 4 discusses the application of phonotactic constraints in the phonological domain. Section 5 demonstrates how this technique is being extended to the phonetic domain to allow for a representation which is more closely related to the speech signal. Section 6 describes how the constraints in each domain can be relaxed in order to extrapolate the output representations given noisy input data. Section 7 concludes with some discussion of future work.

2. TIERS, FEATURES AND CONSTRAINTS

The original motivation for the design of the constraint-based computational linguistic model was to address specific problems in the area of speech recognition below the level of the word. In particular, the problem of out-of-vocabulary items, also termed the "new word" problem, is addressed explicitly in the model. This is done by including *complete* phonotactic descriptions of a language which describe not only those forms which are described in some corpus lexicon, but also all potential forms which adhere to the constraints imposed by the language. We discuss this issue in more detail in section 4. Another specific problem addressed by this approach is the modelling of coarticulation phenomena. This is done by assuming a non-segmental approach to the description and interpretation of speech utterances which avoids having to segment an utterance into non-overlapping units at any level of representation.

Speech utterances are defined in the model in terms of a multilinear representation of tiers of features which are associated with signal time. The notion of tiers of features is not new in the area of phonology (cf., for example, Goldsmith, 1976, 1990; Browman and Goldstein, 1989) and indeed the non-segmental approach underlies the YorkTalk speech synthesis system (cf., for example, Coleman and Local, 1992); however, it is only more recently that researchers in speech recognition are beginning to consider this type of representation more explicitly. Our research has focussed on optimising the constraint-based model for interpreting multilinear representations of speech utterances and we have, therefore, placed little emphasis on developing the front-end feature extractors assumed by our model, working instead with bootstrapped labelled data for test purposes. However, recently there has been a significant upsurge in phonetic feature extraction and classification, and automatic transcription using the type of features proposed in our model (e.g. Salomon and Espy-Wilson, 1999; Koreman, Andreeva, and Strik, 1999). Ali et al. (1999) discuss a method for automatic segmentation and categorisation

into phoneme groupings, in particular the obstruent grouping. For example, both stops and fricatives are detected as part of a two-stage process, voicing detection followed by place of articulation detection. Each of these stages contributes specific information to the overall detection of each of these manner classes. Voicing detection in turn relies on several specific features which are characteristic to the particular manner class as does the place of articulation detection.

Chang, Greenberg, and Wester (2001) propose an elitist approach to articulatory-acoustic feature extraction, based on multilayer perceptron (MLP) classifiers, known as ARTIFEX. This approach restricts frame selection to those for which the MLP classifiers are highly confident. They demonstrate that by training articulatory-place classification in a manner-specific way, place-feature extraction can be improved significantly. Their primary motivation for ARTIFEX has been to devise an efficient method for automatic phonetic annotation of large data sets which does not rely on a classification into phonetic segments thus making such annotated material more easily accessible to researchers in phonetics and speech technology. It is exactly this type of feature extraction which is assumed by the model presented in the next section.

Results of an articulatory-acoustic feature classification can be represented in terms of a multilinear event representation, a tiered structure of such phonological features analogous to an autosegmental score (Goldsmith, 1990) and not unlike that used in the synthesis model of articulatory phonology (Browman and Goldstein, 1989). Shawn Chang has kindly provided us with articulatory-acoustic feature data from the TIMIT corpus for evaluation purposes which is extracted using the techniques described in Chang, Greenberg, and Wester (2001) and the feature set defined in Chang, Shastri, and Greenberg (2000); this data was the input to the example in Figure 1. We have selected a subsection of a longer utterance here for purposes of illustration. The details of the complete utterance would not be legible in a figure of this size. The utterance subsection depicted in Figure 1 is the word *pace*; the spectrogram representation is shown for comparison purposes. The feature set is based loosely on the International Phonetic Alphabet classification of features with the addition of a tier associated with a static versus dynamic spectrum.

As can be seen from the figure, each feature in a multilinear event representation is associated with a specific tier (on the vertical axis) and with a specific time interval in terms of milliseconds (on the horizontal axis). The features do not all start and end simultaneously. An overlap of

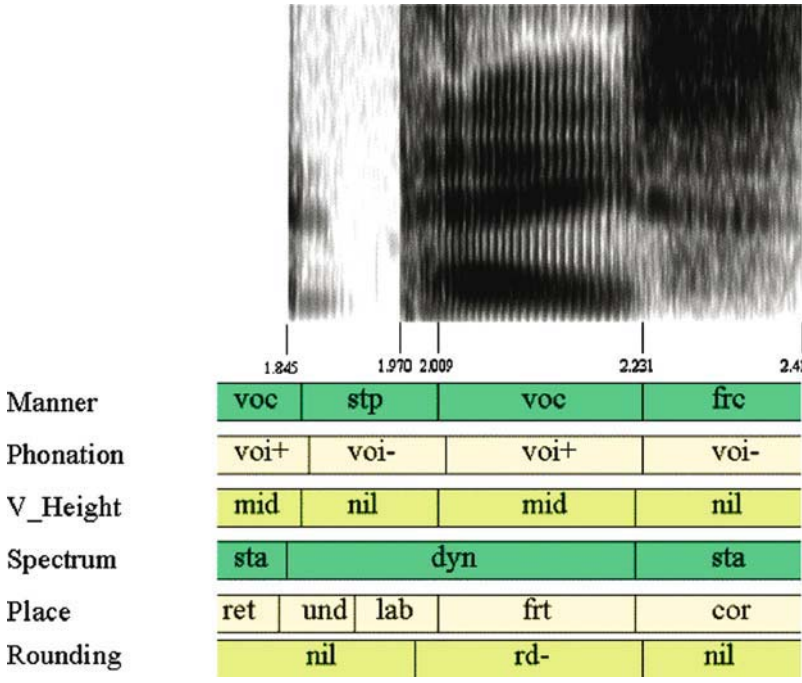

Figure 1. Multilinear event representation of *pace.*

properties exists in any time interval; for example, in Figure 1 below the feature *rd-* begins before the *voc* feature indicating that the lips have been spread during the plosive (*stp*) anticipating the following nonround vowel. The multilinear event representation captures coarticulation phenomena and these are then interpreted by the computational linguistic model as described in Sections 3 and 4.

As will be explained in further detail in the next section, the computational linguistic model assumes that these features are autonomous (i.e. independent of each other), although the feature extraction process clearly does assume dependencies (see Chang, Greenberg, and Wester, 2001). Feature detection is not an autonomous process; the realisation of particular features may be dependent on the presence or absence of other features in the signal. From the point of view of the model described in the next section, however, we are only concerned with these dependencies to a certain extent. The model treats the features thus extracted as autonomous events (i.e. feature-interval pairs). Furthermore, the computational linguistic model also ignores both undefined (*und*) and

nil information (i.e. a feature which is not relevant for a particular sound such as vowel height for consonants) and treats both these features as gaps on that tier in the representation. The resulting multilinear event representation (without the spectrogram) which serves as input to the computational linguistic model is that given in Figure 2.

A multilinear event representation of a speech utterance is in fact highly constrained. It is not the case, that any combination of features can occur in any order. The allowable combinations of features are dictated partly by the phonological structure of the language, as defined by the phonotactics, and partly by predictable phonetic variation, which often results from limitations associated with human speech production (e.g. maximal communication with minimal effort; cf. Boersma, 1998) leading to apparent deletions, insertions and substitutions in the speech stream. Indeed allophonic information has been shown to be very valuable for speech segmentation (Church, 1987) but systematic modelling of this type of constraint has been consistently avoided in current speech technology. Furthermore, both the phonological and many of the phonetic constraints referred to here are restricted to the domain of a syllable (cf. Carson-Berndsen, 1990; Greenberg, 1999 and a wealth of literature on what have been traditionally termed phonological processes). Although speech technology has proposed very efficient models within the triphone domain, this domain does not provide enough context to fully avail of these constraints. That is not to say that the only relevant unit for speech recognition is the syllable, but rather that units of varying granularity can be constrained within the syllable domain and can therefore provide important information for segmenting speech signals. The constraints which are applied in the syllable domain are discussed in Sections 4 and 5 below.

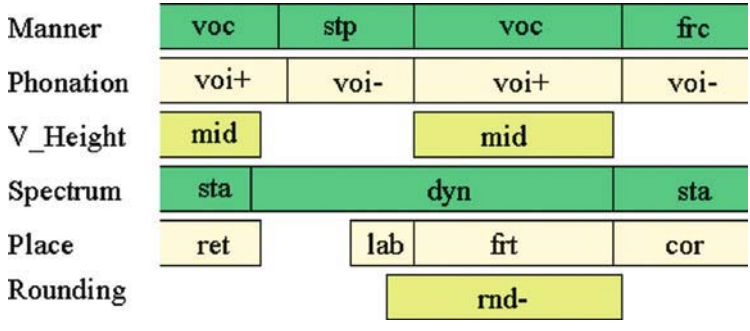

Figure 2. Multilinear event representation of *pace* with gaps for *nil* and *und.*

Once phonetic and phonological constraints are modelled explicitly in a speech recognition system, they can be used to not only guide the interpretation of multilinear event representations but also to provide top-down predictions both for missing information and for expectations of what has yet to come. In Section 6 of this paper, we will illustrate how an underspecified multilinear representation (i.e. a representation with missing information), can be interpreted by a computational linguistic model known as the *Time Map* model and how the output representation can be extrapolated using phonetic and phonological constraints to attain a representation which is more fully specified. We first sketch the model briefly in Section 3. A more comprehensive description can be found in Carson-Berndsen (1998, 2000).

3. THE TIME MAP MODEL

The *Time Map* model was proposed as a computational linguistic model for speech recognition by Carson-Berndsen (1998) and has been tested within a speech recognition architecture for German. The model has recently been extended to English and has been provided with an interface which allows users to define and evaluate phonotactic descriptions for other languages and sublanguages. This generic development environment is known as the *L*anguage *I*ndependent *P*honotactic *S*ystem (Carson-Berndsen and Walsh, 2000a, b). *LIPS* aims to provide a diagnostic evaluation of the phonotactic descriptions in the context of speech recognition. That is to say, rather than just providing recognition results, partial analyses can be output indicating which constraints have or have not been satisfied and where the parsing breaks down.

The *Time Map* model uses a finite-state network representation of the phonotactic constraints in a language, known as a phonotactic automaton (cf. Section 4 below), together with axioms of event logic to interpret multilinear representations of speech utterances. These axioms are defined in detail in Carson-Berndsen (1998) building on Bird and Klein (1990); they are used to infer temporal relations (overlap, precedence, immediate precedence and temporal inclusion) between features in a multilinear representation. For example, a precedence relation ($>$) between two features is defined with respect to the temporal endpoint of the first feature and the temporal start point of the second feature; since the relation is transitive if feature a precedes feature b and feature b precedes feature c then feature a also precedes feature c ($\mathrm{e}_a > \mathrm{e}_b \wedge \mathrm{e}_b > \mathrm{e}_c \Rightarrow \mathrm{e}_a > \mathrm{e}_c$). In order to provide speech utterances with a phonological interpretation, this approach encompasses both an

absolute time level, in which speech signals are related to speech events by temporal annotations (in terms of millisecond values), and a relative time level, in which speech events are related to each other in terms of overlap and precedence relations. The architecture of the model in the context of speech recognition is depicted in Figure 3.

Input to the model is a multilinear representation of a speech utterance in terms of absolute time events, i.e. features with start and end points which are extracted from the speech signal. Phonological parsing in the *Time Map* model is guided by the phonotactic automaton which provides top-down constraints on the interpretation of the multilinear representation, specifying which overlap and precedence relations are expected by the phonotactics. If the constraints are satisfied, the parser moves on to the next state in the automaton. Each time a final state of the automaton is reached, a well-formed syllable has been found which is passed then to a corpus lexicon which distinguishes between actual and potential syllables. The corpus lexicon is compiled from the generic lexicon described in Carson-Berndsen (1999).

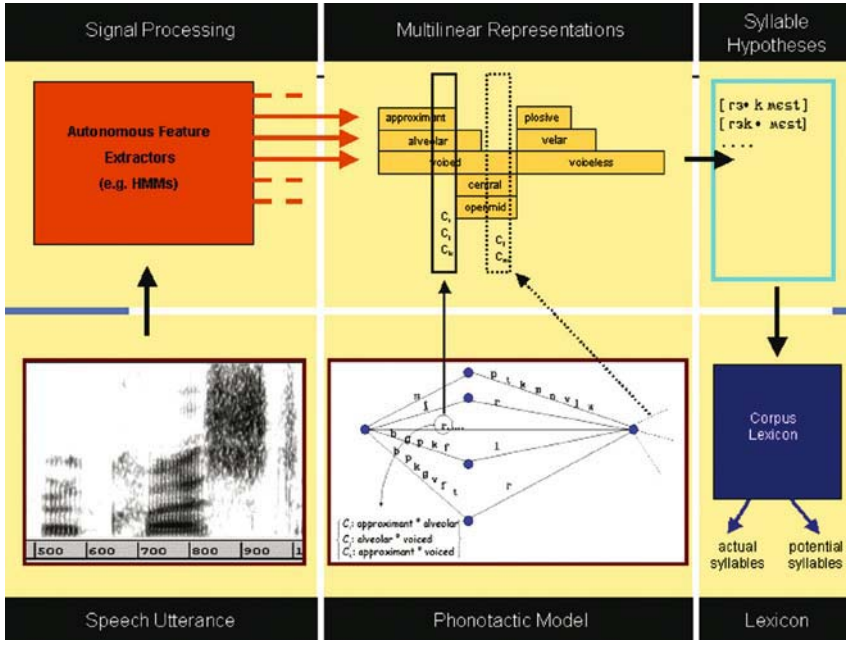

Figure 3. Time map architecture.

4. PHONOTACTIC CONSTRAINTS

The primary knowledge component of the model is a complete set of phonotactic constraints for a language which is represented in terms of a finite state automaton. A subsection of a phonotactic automaton for CC-combinations in English syllable onsets can be seen in Figure 4. The arcs in the phonotactic automaton define a set of constraints on overlap relations which hold between features in a particular phonotactic context (i.e. the structural position within the syllable domain).[1] The phonological features used in this figure are the features of Chang, Greenberg, and Wester (2001) which are defined with respect to a tier model where the tiers define phonation, manner of articulation, place of articulation, vowel height, rounding, tenseness and static/dynamic spectrum. In the phonotactic automaton of Figure 4, the constraint C_1: *stp* ° *voi-*, for example, states that the feature *stp* (a plosive) on the manner tier should overlap the feature *voi-* (voiceless) on the phonation tier. The millisecond values refer to the average durations for the sounds in this particular phonotactic context.

The phonotactic automata of the *Time Map* model are defined with respect to the syllable domain. While traditional speech recognition systems do model a restricted phoneme or phone context, it is not with respect to the syllable domain, but in terms of a statistical dependence on immediate neighbouring units. However, this misses a significant amount of contextual information which is used by native speakers to distinguish illegal from well-formed structures of their language. The phonotactic automaton

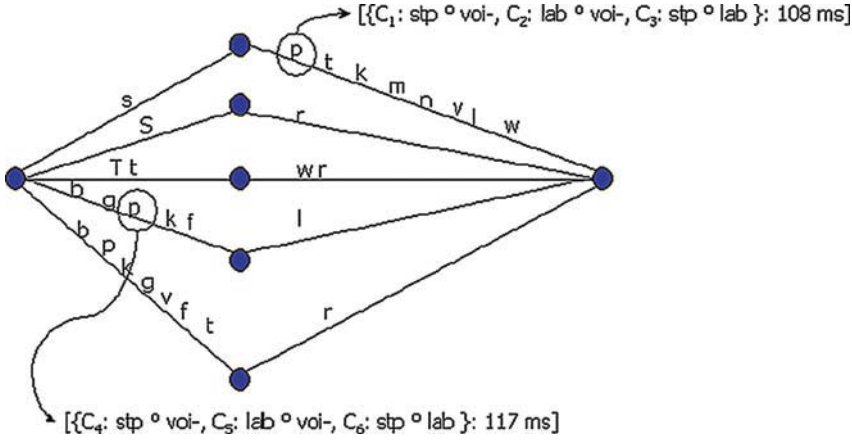

Figure 4. English CC-onsets.

is based on the fact that the number of syllables in a language is finite and therefore a complete description of relevant constraints can be provided. However, a phonotactic automaton can also be extended to include other levels of granularity such as phonological words specifying "syllable-tactics" (see below) and other information types such as graphemes, allophones etc. This new construct is termed a *multilingual time map*, a multilevel finite state transducer which facilitates portability of the Time Map model to other languages (see Carson-Berndsen, 2002).

Since it is a multilinear input representation which is constrained by this phonotactic automaton, there is no strict segmentation of the input at the level of the phone or phoneme and the constraints apply not to the actual temporal annotations of the input but to the temporal relations which exist between them. Therefore, when applying this automaton to the example multilinear event representation of Figure 2, it is not necessary that the start and end points of the features *stp* and *voi-* be the same. Also, the fact that the *rd-* feature also overlaps with the *stp* and *voi-* features (modelling coarticulation) does not prevent this arc in the automaton being chosen. This is illustrated in Figure 5 using a subsection

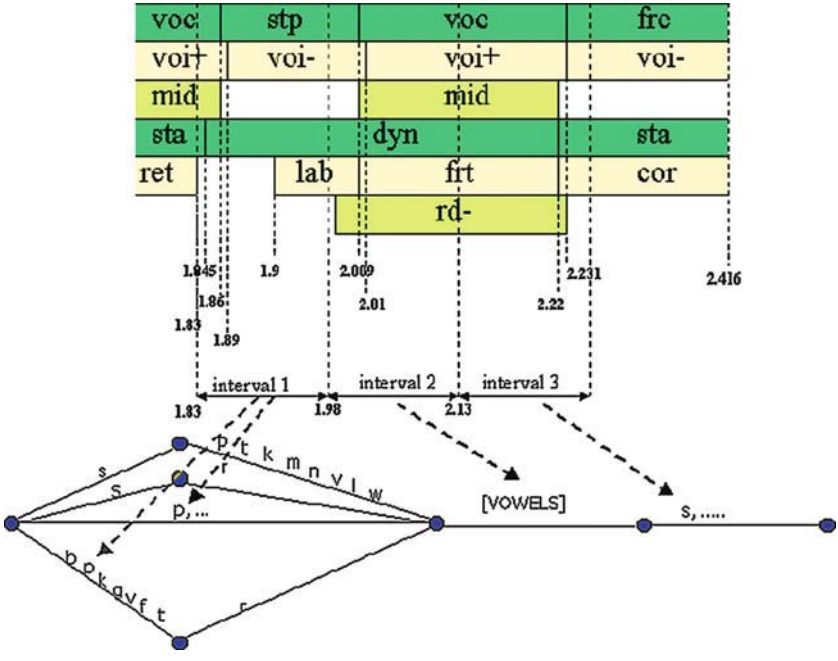

Figure 5. Application of constraints to the multilinear event representation.

of the complete phonotactic automaton showing only the relevant arcs.

As defined in Ashby, Carson-Berndsen, and Joue (2001), a phonotactic automaton can be constructed for whatever language is to be recognised by the system on the basis of whatever feature set has been chosen to classify the speech signal. Thus the model has been equipped with multilingual functionality. Furthermore, since a phonotactic automaton models all possible combinations of sounds in the syllable domain of a language, at each position in the automaton, it is possible for the model to predict what will come next. The constraints, therefore, provide top-down information for the interpretation of multilinear event representations of speech utterances. The advantage of the phonotactic constraints is that they restrict all outputs of the model to structures which are well-formed in the language and are, therefore, a means of treating out-of-vocabulary items and modelling the predictive power of a native speaker of a language. Partial analyses may also be output for diagnostic purposes and extrapolation. While this addresses the notion of robustness in speech recognition to a certain extent, the main criticism we would have of our original model, is that it did not take any statistical knowledge into account and thus provided no means of ranking the output hypotheses. For this reason, we have extended the phonotactic automaton and the parsing procedure to incorporate a number of additional constraints.

Firstly, each feature participating in a constraint is now augmented by an average duration parameter with respect to the particular phonotactic context in which it appears. These average durations are calculated on the basis of a large body of data, but are not intended to be corpus- or speaker-specific and may need to be tuned to reflect speech rate. The duration parameter is used to define the prediction space for the next arc in the phonotactic automaton during processing and thus merely serves as a rough temporal guideline for parsing. Clearly, higher-level constraints, such as syllable position in the phonological word or position in the phrase, will affect how this duration parameter will be interpreted.

Secondly, phonotactic automata are defined with respect to different syllable types (e.g. stressed vs. unstressed) and a "syllable-tactics"[2] constrains how these types are realised in phonological words stipulating how stressed and unstressed syllables can be combined in larger domains. Lexical stress, acoustic realisations of accent and other types of prosodic information are assumed to form additional tiers in the multilinear representation and thus are treated as events which temporally overlap with

the phonological features defined above. While the model does provide explicitly for the integration of other higher-level information, this has not been the main emphasis of the work thus far; it represents the next logical step in the development of the model, however.

Thirdly, we have integrated a constraint ranking methodology into the model by allowing constraints on the arcs to be ranked and an overall threshold to be defined which provides the basis for constraint relaxation. The aim of constraint relaxation is to cater for imperfect and noisy input by allowing constraints to be relaxed and output representations to be extrapolated based on structural information defined in the phonotactics. This will be discussed in more detail in Section 6.

Despite the completeness of the phonotactic constraints and the fact that they are not dependent on segmentation of the input into non-overlapping phonemic units, the model has in the past been viewed as a primarily phonological approach based on features which are not apparent in the signal. We now address this issue in Section 5.

5. TOWARDS PHONETIC CONSTRAINTS

The *LIPS* generic development environment for the *Time Map* model is independent of any particular feature set and allows users to define the phonotactic automaton with respect to any feature set. However, as discussed above, the features in the input representation are treated autonomously. In this context, we distinguish between two different types of event within a knowledge domain: simplex events and complex events. A simplex event is atomic i.e. has no internal structure with respect to a knowledge domain. For example, the phonotactic automaton in Section 4 assumed that the plosive feature *stp* is a simplex feature in the phonological domain. It becomes a simplex event when coupled with a temporal annotation. A syllable is a complex event in the phonological domain, however, as it has an internal structure based on the composition of the simplex events defined in the phonotactic automaton.

This notion can be extended to the phonetic domain (cf. Carson-Berndsen, 1998). We assume that, at the phonetic level, a feature such as *plosive* is indeed complex consisting of combinations of articulatory movements or acoustic manifestations. A complex *plosive* feature may consist of simplex events such as *closure*, *release*, *frication*, each of which will be realised differently depending on the context in which it occurs. In English, a voiceless plosive in syllable initial position before a vowel will realise all three of these features. However, in syllable final position, the *release* may not be apparent at all. This information may provide a useful

cue in parsing (Church, 1987) but is usually neglected in speech recognition. Furthermore, information on the transition phase from *plosive* to *vowel* should also be included as this provides useful indications: an increase in F_1 frequency during the transition to the vowel is characteristic of all plosive-vowel transitions and a cue to manner of production (Kent and Read, 1992; Stevens, 1998).

In order to cope with this type of predictable variation in the model, we extend the phonotactic automaton by a transduction relation which maps between phonological feature automata, known as *phonetic time maps* and a set of constraints on their overlap relations.[3] A *phonetic time map* defines the internal structure of a complex event in the phonetic domain. The transduction relation defines the many-to-one mapping between the phonetic time maps and the phonological feature itself. An example of a possible phonetic time map for a plosive in initial syllable position before a vowel is depicted in Figure 6.

The phonetic time map for plosive in this position consists of two alternative paths. One path models the internal structure of the complex event *stp* as a closure (*clo*) event, followed by a release (*rel*) event, followed by a frication (*frc*) event followed by a increase in F_1 ($F_1 inc$) event. The second path allows for the fact that if the events *rel* and *frc* are detected independently, then they may not be temporally distinct. The *Time Map* model also defines a transduction relation between the

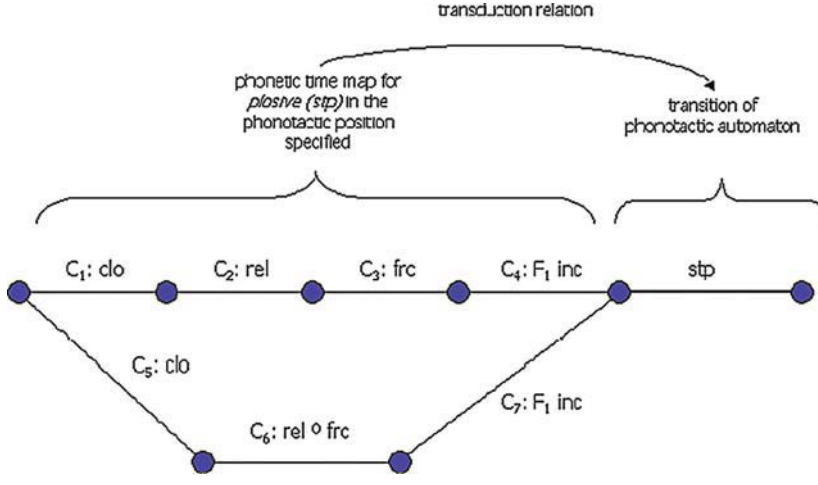

Figure 6. An example phonetic time map for *plosive* (*stp*).

phonetic time map and the complex event *stp* as it is defined in this position in the phonotactic position.

Similarly, the place of articulation feature, for the plosive in this phonotactic context before a vowel, will be heavily influenced by the formant transitions of the following vowel. Thus, a phonetic time map for the place feature of the plosive would expect to use a constraint on the F_2 and possibly F_3 transitions which seem to be sensitive to the place of articulation. In particular, the starting frequency (or locus) of F_2 may be used.

Other types of predictable phonetic variation which can also be captured within phonetic time maps are what are traditionally termed phonological processes. For example, the neutral vowel in an unstressed syllable is elided in spontaneous speech before a nasal causing the latter to become syllabic. This is traditionally defined in terms of two phonological processes: elision of the neutral vowel, and nasal becoming syllabic at the end of a word when preceded by an obstruent. This is modelled by a phonetic time map for the neutral vowel in the phonotactic context of a nasal is depicted in Figure 7. The phonetic time map caters for both the neutral vowel followed by the nasal and the syllabic nasal in the phonetic domain, allowing presence or absence of the neutral vowel in this context. The context for application of this transduction is similarly defined by the phonotactic automaton which models the syllable domain. The feature *neut_V* occurs in the phonological domain primarily in an unstressed syllable.

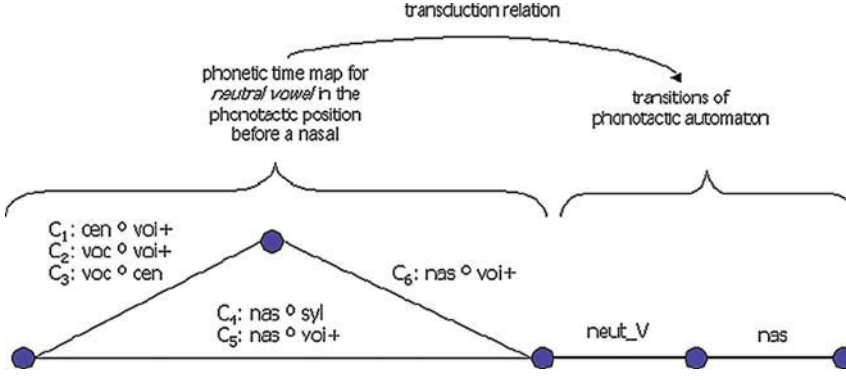

Figure 7. An example phonetic time map for *neutral vowel.*

The phonetic time maps aim to integrate finer phonetic detail into the model by defining constraints which go beyond the discrete phonological features to allow a degree of gradualization to be included. Each phonological feature in a particular phonotactic context is represented in terms of a phonetic time map which maps the different degrees of a feature onto a generalisation of that feature. For example, a range of degrees of voicing in particular syllable contexts can be modelled in a phonetic time map by allowing voicing to be scaled rather than assuming the binary distinction of *voi-* and *voi+* seen at the phonological level. Since, during parsing, the phonetic time maps are used to guide the interpretation of the multilinear representation, all that is necessary in this case is to model a precedence relation between one degree of voicing and another (a decrease or an increase); the actual point of change (boundary between one degree and another) becomes irrelevant. Of course, this is dependent on being able to distinguish between various degrees of voicing during feature extraction.

Experiments with various phonetic time maps modelling predictable variation show that the set of phonetic time maps must be constructed with respect to phonetic or acoustic features which can be detected in the signal. In some cases, the feature sets already being used are sufficient for this task. However, as was seen above, in other cases, it is necessary to include a finer level of granularity such as the closure and release phases of stops. Since it has now been demonstrated that an extensive feature set can be classified very accurately from the signal using the elitist approach (Chang, Greenberg, and Wester, 2001) referred to in Section 2, a more fine-grained representation should now also prove possible. Since the *Time Map* model is not restricted to any one technique, it is open to combining feature input from multiple sources. *LIPS* serves as an experimentation environment for testing and diagnostically evaluating constraints which can be modelled in the syllable domain.

Clearly the phonetic time maps and the phonotactic automata define sets of top-down constraints which may not always be fulfilled for every multilinear event representation input to the system. The input may be underspecified or noisy, which may lead to some features not being recognised or to the wrong features being recognised during feature extraction. Rather than propagate the same degree of underspecification further up the recognition process, the *Time Map* model attempts to resolve as much of the underspecification as possible by applying constraint relaxation and output extrapolation procedures in both the phonetic and the phonological domains. This is described in the next section.

6. CONSTRAINT RELAXATION

Constraint relaxation should be performed in the model if only *some* of the constraints specified by either the phonotactic automaton or phonetic time maps can be satisfied. As it stands, this is a very arbitrary statement. However, when coupled with a constraint ranking, it becomes a method for dealing with variability and underspecification in the input representation. Constraint ranking is a data-oriented ordering of constraints in particular phonotactic contexts. For example, constraints may be ranked with respect to frequency, average duration and percentage overlap of features in specific structural contexts. This information can either be specific to a single corpus or may be based on data from several different corpora. Based on this ranking, constraint relaxation can be applied, for example, when an infrequent feature is encountered or a duration is outside a given standard deviation. Furthermore, it is possible to combine this type of ranking with cognitive factors in order to go beyond a corpus-dependent ordering (Carson-Berndsen and Joue, 2000). Constraint relaxation can then be regarded as a means by which particular constraints on an input representation can be ignored. We illustrate constraint relaxation simply using the following three constraints on overlap relations on a particular arc of a phonotactic automaton:

C_1 : *stp*$^\circ$ *voi-* with ranking 0.6
C_2 : *voi-*$^\circ$ *lab* with ranking 0.4
C_3 : *stp*$^\circ$ *lab* with ranking 0.3

Assuming the arc has a threshold in this phonotactic context of 0.7, then at least two of these constraints must be satisfied (i.e. the sum of the rankings must reach the threshold) in order for this arc to be taken. A multilinear event representation such as that given in Figure 8 would satisfy the constraints C_1 and C_3 in the interval being examined and exceed the threshold and therefore, this arc in the phonotactic automaton can be taken.

Output extrapolation, on the other hand, is performed to further specify the output representation if the constraints specify expectations that do not conflict with information found in the input. Here again, a ranking of the constraints, which can participate in output extrapolation, is required. We use same constraints C_1, C_2, C_3, for illustration but this time but choose a threshold value of 0.6. Given an input representation

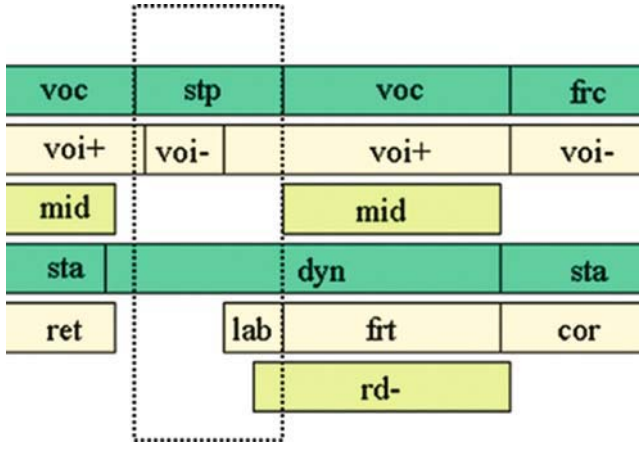

Figure 8. Illustration of constraint relaxation in a particular interval.

which is underspecified with respect to place of articulation as depicted on the left of Figure 9, then constraint relaxation will allow the arc to be taken with only C_1 satisfied. Output extrapolation can then be performed using C_2 and C_3 if there is no information in the input representation which explicitly conflicts with these constraints such as another place of articulation, for example. So although the input representation was underspecified with respect to place of articulation, the extrapolated output representation on the right side of Figure 9 will contain this information as it was augmented using the structural knowledge and temporal information contained in the phonotactic automaton.

The application of output extrapolation does not guarantee that the output syllable structures are fully specified, however, only that they are well-formed. Should the output representation still be underspecifed, the corpus lexicon will be able to further resolve some of the underspecification by ranking those multilinear representations which can be subsumed by a fully specified entry in the lexicon higher than those which do not. Figure 10 shows a later interval in the processing of the multilinear event representation where output extrapolation has been unable to augment the underspecified representation any further. The voicing specification for the fricative is missing from the input but both voiced and voiceless coronal fricatives are possible in this phonotactic position. The extrapolated output subsumes two forms, [peIs] and [peIz]. However, the corpus lexicon only contains one of these forms and therefore, the representation can be augmented with the *voi-* feature and this corpus

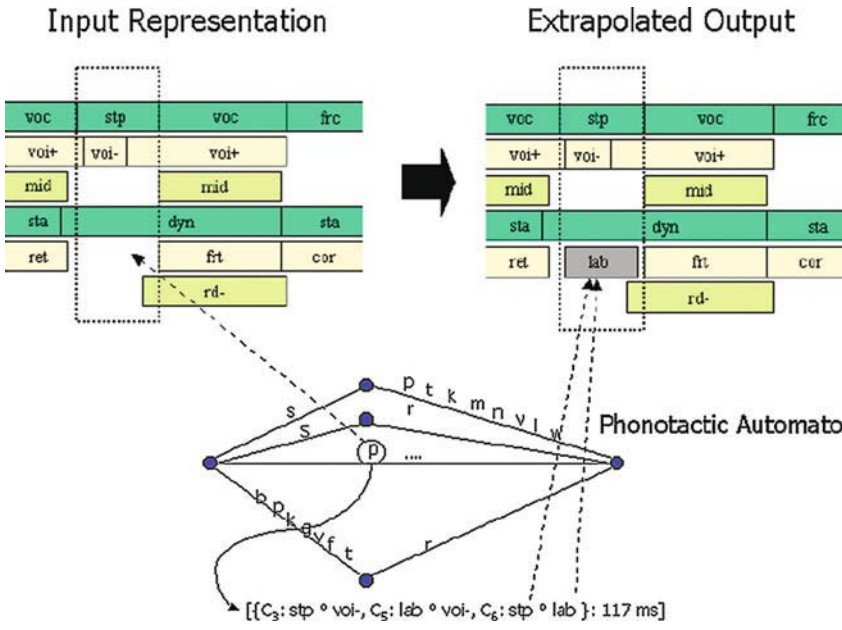

Figure 9. Illustration of output extrapolation in a particular interval.

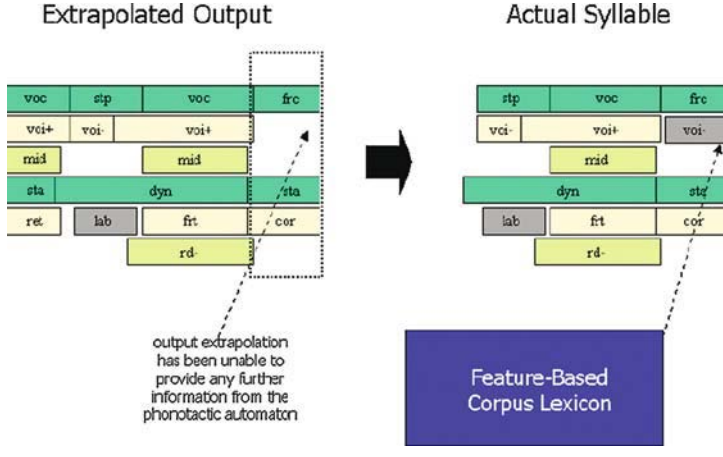

Figure 10. Illustration of output after lexicon consultation.

form ranked higher than its well-formed counterpart which is not included in the corpus.

Although output extrapolation can also be applied at the finest level of granularity of the *Time Map* model, using the constraints defined in phonetic time maps, it is currently unclear whether extrapolating predictable phonetic information will be of real value in speech recognition. The reason for this is that if predictable phonetic information is to be used to segment the multilinear event representation, and this information is missing in the input representation, then it may be incorrect to augment the representation to include it. The phonotactic automaton already provides top-down constraints which will provide a coarser segmentation if the predicted phonetic information is not found in the signal. More experimentation remains to be done on this topic whenever a more comprehensive set of phonetic time maps and their relevant constraint rankings have been developed.

In LIPS, a distinction is made between *online* processing where speech utterances are interpreted using the constraints and constraint rankings, and *offline* processing, which is concerned with finding the optimal parameters and constraint rankings for the system (see Figure 11). While the constraint rankings refer to individually ranked constraints on temporal overlap relations between phonological or phonetic features, taken collectively these rankings also provide the basis for weighting in the phonotactic automaton and the phonetic time maps respectively, through the

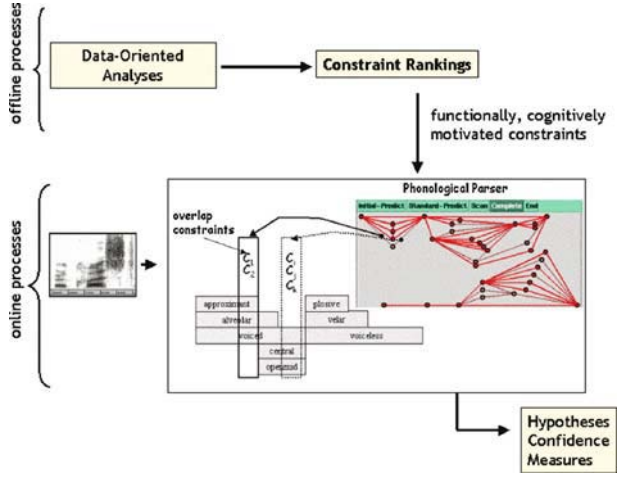

Figure 11. Offline vs. Online Processing in LIPS.

use of arc thresholds. A more detailed discussion of the role of constraint rankings, arc weightings and constraint relaxation can be found in Carson-Berndsen, Joue, and Walsh (2001).

7. CONCLUSION

This paper has presented a constraint model for interpreting multilinear representations of speech utterances which can provide important fine-grained information for speech recognition applications. It was demonstrated that the model integrates phonotactic and phonetic information in a non-segmental fashion. Currently, we are working on developing a wider range of phonetic time maps and investigating how they contribute to robustness in the model through the use of the constraint relaxation and output extrapolation techniques which were described above. While our ongoing research is directed towards optimising this model through the use of statistical information and cognitive constraint rankings, the model also provides useful constraints for fine-tuning more stochastic approaches for robust speech applications. Future directions for this work are to investigate the application of the computational model in the context of speech synthesis whereby multilinear phonetic representations are generated rather than interpreted by the model and serve as parameters to a synthesis module. We believe that the use of the same model for both recognition and synthesis will provide insights into the different levels of granularity of information required for truly robust speech applications and lead to approaches which combine linguistic and statistical knowledge more explicitly. Furthermore, the integration of phonotactic descriptions of other languages adds a dimension of multilinguality to our model which will support the incorporation of phonetic features which are common among languages.

NOTES

[1] The monadic symbols written on the arcs in Figure 5 are purely mnemonic for the feature overlap constraints they represent; the $^{\circ}$ symbol represents the overlap relation.
[2] The syllable-tactics is also defined in terms of a finite state automaton.
[3] This is not unrelated to the allophonic parser described in Carson (1988) except that the parser defined there assumed a segmentation of the input into a string of allophones.

REFERENCES

Ali, A.M..A., Van der Spiegel, J., Mueller, P., Haentjaens, G., and Berman, J. An Acoustic-Phonetic Feature-Based System for Automatic Phoneme Recognition in Continuous Speech. In: *IEEE International Symposium on Circuits and Systems (ISCAS-99)*, 1999: III-118–III-121.

Ashby, S., Carson-Berndsen, J. and Joue, G. A testbed for the development of multilingual phonotactic descriptions. In: *Proceedings of Eurospeech 2001*, Aalborg, 2001: 321–324.

Bird, S. and Klein, E. Phonological Events. *Journal of Linguistics* 26 (1990): 33–56.

Boersma, P. *Functional Phonology*. LOT, Netherlands Graduate School of Linguistics, The Hague. 1998.

Browman, C.P. and Goldstein, L. Articulatory gestures as phonological units. In: *Phonology 6*, Cambridge University Press, Cambridge, 1989: 201–251.

Carson, J. Unification and Transduction in Computational Phonology. In: *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 1, 1988: 106–111.

Carson-Berndsen, J. Phonological Processing of Speech Variants. In: *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)* Helsinki, 3, 1990: 21–24.

Carson-Berndsen, J. *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition*. Kluwer Academic Publishers, Dordrecht, 1998.

Carson-Berndsen, J. A Generic Lexicon Tool for Word Model Definition in Multimodal Applications. In: *Proceedings of EUROSPEECH 99, 6th European Conference on Speech Communication and Technology*, Budapest, September 1999: 2235–2238.

Carson-Berndsen, J. Finite State Models, Event Logics and Statistics in Speech Recognition. In: G. Gazdar, K. Sparck Jones, and R. Needham (eds.): *Computers, Language and Speech: Integrating formal theories and statistical data*. Philosophical Transactions of the Royal Society, Series A, 358(1770), 2000: 1255–1266.

Carson-Berndsen, J. Multilingual Time Maps: Portable Phonotactic Models for Speech Technology. In: *Proceedings of the LREC Workshop on Portability Issues in Human Language Technology*. Las Palmas, May 2002.

Carson-Berndsen, J. and Joue, G. Cognitive constraints in a computational linguistic model for speech recognition. In: *Proceedings of the 11th Irish Conference on Artificial Intelligence and Cognitive Science*, Galway, Ireland, 2000.

Carson-Berndsen, J. and Walsh, M. Generic techniques for multilingual speech technology applications. In: *Proceedings of the 7th Conference on Automatic Natural Language Processing*, Lausanne, Switzerland, 2000a: 61–70.

Carson-Berndsen, J. and Walsh, M. Interpreting Multilinear Representations in Speech. In: *Proceedings of the 8th Australian Conference on Speech Science and Technology*, Canberra, Australia, 2000b: 472–477.

Carson-Berndsen, J., Joue, G. and Walsh, M. Phonotactic Constraint Ranking for Speech Recognition. In: W. Daelemans, K. Sima'an, J. Veenstra, and J. Zavrel (eds.): *Computational Linguistics in the Netherlands 2000*, Editions Rodopi b.v. Amsterdam, New York: 2001: 16–29.

Chang, S., Shastri, L. and Greenberg, S. Automatic Phonetic Transcription of Spontaneous Speech (American English). In: *ICSLP-2000*, Beijing, October 2000.

Chang, S., Greenberg, S. and Wester, M. An Elitist Approach to Articulatory-Acoustic Feature Classification. In: *Proceedings of Eurospeech 2001*, Aalborg, 2001: 1725–1728.

Church, K.W. *Phonological Parsing in Speech Recognition*. Kluwer Academic Publishers, Boston, 1987.

Coleman, J.S. and Local, J.K. Monostratal Phonology and Speech Synthesis. In: P. Tench (ed.): *Studies in Systemic Phonology*. London, Pinter Publishers. 1992: 183–193.

Deng, L. Speech Recognition Using Autosegmental Representation of Phonological Units with Interface to the Trended HMM. *Free Speech Journal*, 1997.

Deng, L. A dynamic feature-based approach to the interface between phonology and phonetics for speech modelling and recognition. *Speech Communication* 24 (1998): 299–323.

Goldsmith, J. *Autosegmental Phonology*. Indiana University Linguistics Club, Bloomington Indiana. 1976.

Goldsmith, J. *Autosegmental and Metrical Phonology*. Basil Blackwell, Cambridge, MA, 1990.

Greenberg, S. Speaking in shorthand-a syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29(2–4) (1999): 159–176.

Jusek, A., Rautenstrauch, H., Fink, G.A., Kummert, F., Sagerer, G., Carson-Berndsen, J. and Gibbon, D. Detektion unbekannter Wörter mit Hilfe phonotaktischer Modelle. In: *Mustererkennung 94*, 16. DAGM-Symposium Wien Berlin, Springer Verlag, 1994: 238–245.

Kent, R.D. and Read, C. *The Acoustic Analysis of Speech*. Whurr Publishers, 1992.

Koreman, J., Andreeva, B. and Strik, H. Acoustic Parameters Versus Phonetic Features in ASR. In: *Proceedings of ICPhS 99*, 1999: 719–722.

Pols, L. Flexible, Robust, and Efficient Human Speech Processing Versus Present-day Speech Technology. In: *Proceedings of ICPhS 99*, 1999: 9–16.

Salomon, A. and Espy-Wilson, C. Automatic Detection of Manner Events based on Temporal Parameters. In: *Proceedings of EUROSPEECH 99, 6th European Conference on Speech Communication and Technology*, 1999: 2797–2800.

Stevens, K.N. *Acoustic Phonetics*. The MIT Press, Cambridge MA, London, 1998.

HEIDI CHRISTENSEN*,†, BØRGE LINDBERG†,
and OVE ANDERSEN†

INTRODUCING PHONETICALLY MOTIVATED, HETEROGENEOUS INFORMATION INTO AUTOMATIC SPEECH RECOGNITION

ABSTRACT. This chapter investigates a way to introduce more heterogeneous information into an existing ASR system. A phonetic *expert* is implemented which is specifically targeted at correcting the errors made by an existing ASR system. This gives a heterogeneous system, where the individual items are designed to be complementary. To avoid the curse of dimensionality problem, the expert information is introduced at the level of the acoustic model. Two types of *expert* configurations are used, each providing discriminative information regarding groups of phonetically related phonemes. The phonetic *expert* is implemented using an MLP. Experiments show that, when using the *expert* in conjunction with both a fullband and a multiband system speech recognition performance is increased and noise robustness improved for a range of noise levels.

KEYWORDS. data-driven information extraction, heterogeneous processing, multiple classifiers, noise robustness, phonetics, speech technology

1. INTRODUCTION

Within the area of speech recognition the paramount cause of the discrepancies between the performance of humans and machines is the lack of immunity of Automatic Speech Recognition (ASR) systems to variation in the acoustic signal not affecting the linguistic message; for example, variation stemming from a change of speaker or speaking style, environmental noise, or channel distortions (Lippmann, 1997; Pols, 1997).

Conventional ASR systems rely on a single source of information, i.e. extracts a single type of spectral feature[1] from the speech signal. This is in sharp contrast to what we know about the way humans process speech.

Address for Correspondence:
* Speech and Hearing Group, University of Sheffield
† Department of Communication Technology, Institute of Electronic Systems, Aalborg University

W. J. Barry and W. A. van Dommelen, The Integration of Phonetic Knowledge in Speech Technology, 67–86.

Both physiological and psycho-acoustic studies have shown that human speech recognition is based on the extraction of parallel information from the speech signal. In particular Greenberg (1997) described how different types of speech segments (e.g. by their intensity) determine the type of spectral processing they are given in the human ear. Work done by Fletcher (reviewed in Allen, 1994) led to a suggestion of a model of speech intelligibility in band limited speech. The model says that in human recognition, phoneme identification errors in a given frequency band are independent of the errors in another band. Further improvements to this model were suggested by Steeneken (1992), who found that in restricted transmission conditions, human subjects require different optimal frequency ranges for the correct recognition of independent phonemes. Investigating human perception, Ghitza's (1997) experiments led him to conclude that different phonetic features are transmitted in different frequency regions. These experiments all indicate that incorporating more heterogeneous processing into ASR systems might be a way to escape from the limitation of the local performance maxima of current ASR systems.

The term 'heterogeneous' means *consisting of dissimilar or diverse ingredients or constituents*. In speech recognition systems there are numerous ways to obtain heterogeneity, ranging from combining multiple competing, self-contained, high-performance systems, based on different architectures and methodologies, to combining complementary types of features. The incentives for employing heterogeneous processing in ASR are various:

Theoretical motivation can be derived from research in classic pattern recognition. By fusing different information streams it is possible to exploit the strengths and weaknesses of the different features. Obviously, the nature of the added information is crucial; adding more information to a system is only beneficial to the degree that it is not redundant with respect to the information already contained in the system. Two systems exhibiting exactly the same response behaviour to a given input signal pattern will not benefit from being combined. Theoretically, if systems are mutually independent, and each of the systems is more right than wrong, then combining the systems can decrease the overall error rate (Turner, 1996; Bishop, 1995). However, in practice even less distinct architectures, methodologies and features can be successfully combined.

Heterogeneous speech processing draws further empirical motivation from the results of numerous physiological and psycho-acoustic studies. It has been repeatedly demonstrated that the human brain is highly

dependent on heterogeneous processing. Specifically, human speech recognition is based on heterogeneous processing of the acoustic signal received by the ear. Gold and Morgan (2000) argue:

> Although we don't know exactly what signal processing occurs in the auditory system, we do know that processing occurs with a range of time constants and bandwidths. Given the robustness of human listening to many signal degradations, each of which would severely degrade an individual representation, it is likely that many maps of the input signal are available to the brain.

In evolutionary terms, the speech production system is newer than the auditory perception system, and it is therefore generally believed that the different speech segments have developed their distinguishable spectral characteristics to increase the chances of discrimination within a wide range of natural auditory environments. For example sonorants and non-sonorants have very different levels of intensity and this acoustic distinction triggers different processing in the auditory system (Greenberg, 1997). Studies that suggest the use of specific processing in different frequency bands are also widely reported (Fletcher, 1953; Allen, 1994). Experiments on the intelligibility of word pairs have shown that different phonetic features are transmitted in different temporal-frequency slots (Ghitza, 1994). A related conclusion is drawn by Steeneken (1992) who showed that the optimal frequency range for recognising a phoneme in restricted transmission conditions is very dependent on the type of phoneme. So although it is highly debatable whether speech recognisers should mimic the human brain (Hermansky, 1997), some principles might still be worth adopting for the speech recognition researcher.

Relying on multiple sources of information for pattern recognition tasks can increase accuracy and efficiency of the application by taking advantage of inherent weaknesses and strengths of the individual classifiers (Ho, 1992; Kittler et al., 1998). The concept of combining classifiers has been analysed recently (Turner, 1996; Hansen, 1990), and within speech applications there have been several studies on the use of sets of classifiers to increase acoustic modelling in speech recognition tasks ranging from large vocabulary speech recognition to classification of syllables, phonemes or groups of phonemes. The *multi-stream* framework is one particular type of multiple classifier system that, in recent years, has proved useful for experimenting with the use of heterogeneous features and information sources (Bourlard, 1996). In a multi-stream based system, parallel streams are processed independently before being combined at a later stage.

In the area of multiple classifier systems, the crucial question to address is of course: What type of extra information to add? Christensen et al. (2000a) showed that combining several standard features in both multi-stream and multi-band[2] type systems could significantly improve performance. We further showed (Christensen et al., 2000b; Christensen et al., 2000c) that designing the feature types so as to be particularly tuned towards a particular phonetic group (such as voiced phones or consonants) also helped improve performance. These experiments have confirmed our hypothesis that introducing in particular phonetically motivated information into ASR can help increase performance, and have encouraged this further work in that direction.

One of the main questions arising in the multi-stream approach concerns the nature of the feature streams to combine. Nearly all previous multi-stream research has employed features designed primarily for conventional, single-stream systems. Typically, the features that have been chosen are those that have highest performance in isolation, under the assumption that this will lead to the highest performance when the features are combined. However, this assumption is not necessarily valid and many multi-stream approaches, although often demonstrating good performances, may appear rather *ad hoc*.

This chapter explores the clean speech performance and noise robustness of an approach aimed at adding more complementary information, specifically targeted at discriminating between larger groups of phonemes. Central to the approach has been the extraction of the phonetically motivated information in a *data-driven* fashion.

As a starting point the investigation will focus on a particular configuration of a heterogeneous system: A single feature type ASR system (a *stem* system) which is combined with an *expert* system. Such a heterogeneous system is sketched in Figure 1. As the name indicates, the knowledge extracted by the *expert* is rather specific, and a key issue is that the *expert* is guided towards being complementary to what is modelled by the *stem* system. In other words, the overall design idea is to target the *expert* towards a particular type of errors occurring in the *stem* system.

Two fundamentally different *stem* systems are investigated: a fullband and a multi-band system. Initially the chosen *stem* systems is informally analysed with respect to any patterns arising from the type of confusion errors that are prominent. The overall error rates for the *stem* systems conceal a large variation in the individual errors from one phoneme to another (Christensen, 2001). Examining the confusion matrices[3] shows that a significant number of confusions occur between larger groups of

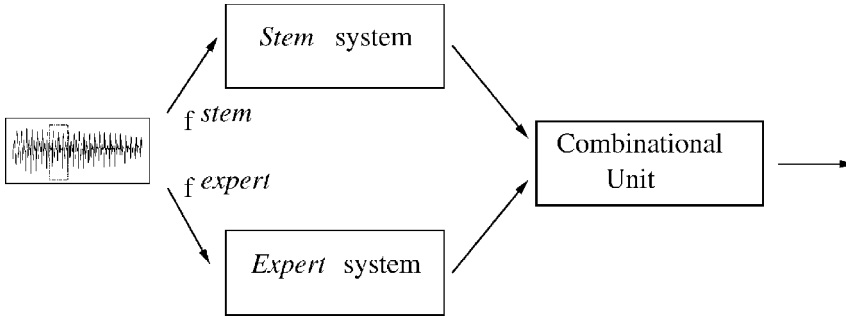

Figure 1. Schematic representation of a heterogeneous system comprising a *stem* system (e.g. a single feature conventional fullband or a homogeneous multi-band system) and an *expert* system.

phonemes possessing largely *phonetic* similarities. This is in accordance with Halberstadt (1998).

Two different ways of examining the confusion matrices are employed, each collapsing the phonemes into different groups: either by dividing the phonemes into broad classes {*vowel*, *consonant*, *nasal*, *liquid*, *silence*}, or by using a voicing criteria and dividing the phonemes into {*voiced*, *unvoiced*, *silence*}. It is of interest to see that confusions occur between the different broad groups of phonemes.

The analysis of the confusion matrices points towards the inclusion of more phonetic information into ASR systems. The approach adopted here is aimed at adding more phonetically motivated information, specifically targeted at resolving the observed types of confusions, and derived from the speech signal in a data-driven fashion. What is required is access to a phonetic *expert* that can provide additional, heterogeneous information to an already available, trained *stem* system like the current fullband and multi-band systems.

An important question is how to introduce the expert information into the speech recogniser. One approach, when adding information to a statistical pattern recognition system is to simply augment the feature vector with any additional features available. However, when training statistical models, increasing the dimensionality of the feature space increases the amount of data needed for securing a sufficient estimation of the parameters. The phenomenon is often referred to as the *curse of dimensionality* (Bishop, 1995). With a limited amount of training data available, other ways of introducing the extra information are of interest. Further, changing the composition of the feature vector requires a

retraining of the whole system, which is a cumbersome task and not always feasible. The experiments reported here avoid the *curse of dimensionality* by introducing the additional, heterogeneous information at the level of the acoustic models.

The following section briefly presents the theory and Sections 3 and 4 present details about the implementation of the *experts* employed. Sections 5 and 6 present the results from the experiments combining the *expert* system with a fullband and a multi-band *stem* system, respectively. Finally, Section 7 summarises and concludes.

2. THEORY

Assume one has access to information from an *expert* providing information on the presence of certain phonetic features in the observed data. Conventional ASR systems are defined within a statistical framework, and so, if the presence of phonetic evidence can be expressed in a statistical manner, i.e. with posteriors, the statistical formulation of the speech recognition problem can easily be modified to accommodate such knowledge.

In ASR the decoding task is aimed at finding the hypothesis or model sequence that is the most likely given the data, X:

$$M^* = \arg\max_{M \in \mathbf{M}} P(M|X) \tag{1}$$

where M is the model sequence (typically word models) and $\mathbf{M}$ is the set of possible model sequences given the vocabulary used.

The term to maximise when incorporating the *expert* information into this expression, is then the joint posterior probability of the model sequence M, the system parameters λ and the *expert* sequence $\mathbf{E}$:

$$P(\lambda, M, \mathbf{E} \mid X) = P(\lambda, \mathbf{M} \mid X) \cdot P(\mathbf{E}, M \mid X) \tag{2}$$

as λ and $\mathbf{E}$ are independent. The first term, the posterior probability of the model sequence, is obtained from the conventional acoustical model (an MLP classifier in the current work) in the system. The second term is modelled by a separate MLP in the following work.

Figures 2 and 3 show an overview of how the posterior probabilities are combined in the implemented systems using the above equations. The *expert* MLP classifiers contribute the posterior probabilities and are then multiplied with the per-phoneme posteriors obtained from the fullband or multi-band system according to Equation (2).

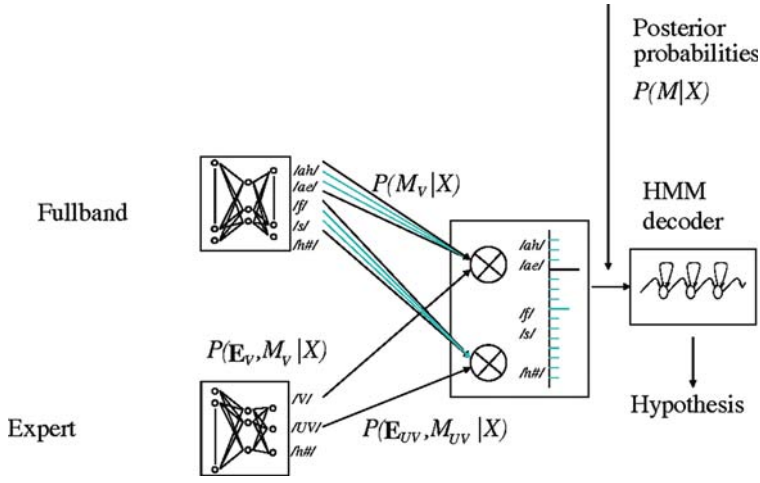

Figure 2. Overview of the combination of a fullband baseline system with a voicing *expert* system.

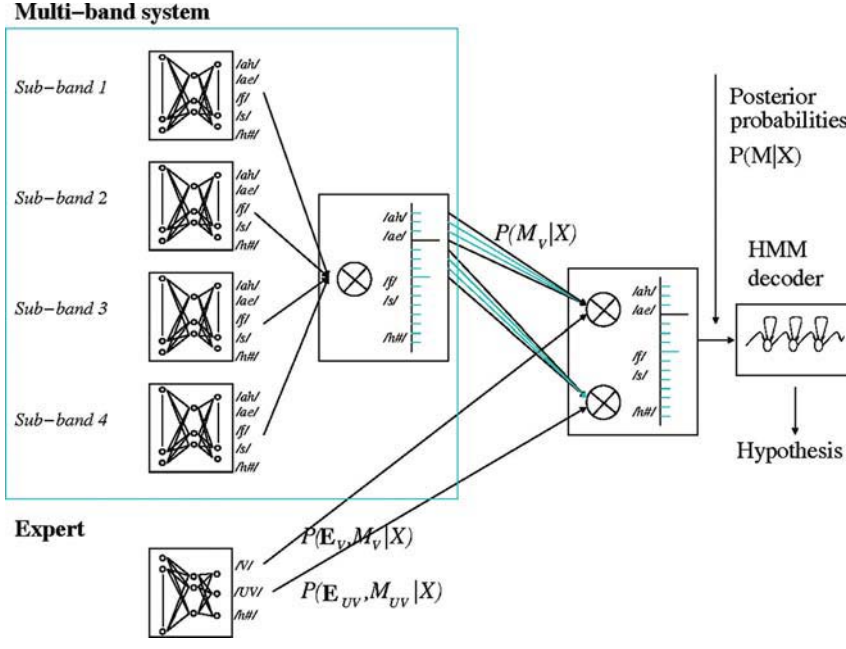

Figure 3. Overview of the combination of a multi-band baseline system with a voicing *expert* system.

3. EXPERIMENTAL SETUP AND DATA

Two series of experiments are carried out. The first series investigates the effect of adding expert information to a conventional connectionist HMM/MLP system; in this case, a conventional fullband system. The second series of experiments focusses on a multi-band system. Each series of experiments is conducted with two types of expert: based on voicedness criteria or classifying broad phonetic classes.

The data for training and testing the systems is taken from the Oregon Graduate Institute Numbers95 database of recordings of American English speakers uttering continuous digit and number sequences over the fixed telephone network (Numbers95, 1995). 112 minutes of speech (13873 words in 3590 utterances) are used for training (of which a 10% are used for cross-validation), and 38 minutes of speech (4670 words in 1206 utterances) from non-overlapping sets of speakers are used for development testing purposes. The vocabulary size is 32 words. For testing the noise robustness of the systems, noise samples from the NOISEX database (Varga et al., 1992) are added per utterance at SNR levels of 0, 6, 12 or 18 dB. The *car noise* and *factory noise* are chosen for their different spectral characteristics.

Three different feature processing methods are used for extracting basic features plus the energy: Mel frequency cepstral coefficients (mfcc) (Davis and Mermelstein, 1980), Perceptual linear prediction coefficients (plpc) (Hermansky, 1990) and J-rasta filtered plpc's (j-rasta-plpc) (Hermansky, 1994). A feature vector is extracted on 25ms Hamming windowed frames, each overlapping 50%. Delta and delta-delta coefficients (regressing over windows of 5 and 7 frames respectively) are added.

A fullband and a multi-band system, each based on connectionist MLP/HMM entities, are used both individually to provide baseline results and in conjunction with the phonetic *expert* systems described above. All MLPs are trained on feature vectors derived from 9 frames centered around the current frame and each MLP has 33 outputs representing 32 phonemes and a silence label.

The **FullbandBaseline** system uses 12 basic features yielding a 39 dimensional feature vector. The MLP has 351 (9×39) input units and 1500 hidden units. The **MultibandBaseline** system comprises four bands with frequency ranges [216–778 Hz], [707–1632 Hz], [1506–2709 Hz] and [2122–3769 Hz][4]. 5, 5, 3 and 3 basic features are derived respectively yielding corresponding vector dimensions of 18, 18, 12 and 12. The MLPs have 162 (9×18), 162, 108 (9×12) and 108 input units and 1000,

1000, 660 and 660 hidden units per band respectively. The two baseline systems have a comparable number of parameters.

All experiments are evaluated in terms of Word Error Rates (WERs).

4. MODELLING OF PHONETIC INFORMATION

An *expert* MLP is employed to model the phonetic information to be supplemented to the *stem* system, and thus provide an estimate for the term $P(\mathbf{E},\mathrm{M}|\mathbf{x})$ in Equation (2). The *expert* MLP is trained to distinguish between a set of expert labels, i.e. phonetic classes. In the experiments reported here, two different sets of expert labels are tried out: one classifying speech segments into *voiced*, *unvoiced* and *silence* and another dividing the phoneme set into five broad phoneme classes: *vowel*, *consonant*, *liquid* (comprising traditional liquids and the approximant [w]), *nasal* and *silence*. The mappings used to generate the appropriate target values are given in Tables 1 and 2. Please note that only phonemes represented in the data labelling of the number strings are included.

It was decided to train the *expert* MLPs with the same training data, employing the same feature processing methods as the other MLPs used in the experiments. The *expert* MLPs also have the same number of hidden units as the MLP used for the fullband classifier, i.e. 1500. The labels are obtained by mapping the training labels to the target value of the appropriate phonetic group. Figure 4 illustrates this training process.

Table 1. Definition of mapping from phoneme classes to voiced and unvoiced phoneme classes.

Voiced/unvoiced	Numbers 95 phone labels
Voiced	iy, ih, eh, ey, er, ay, ah, ao, ow, uw, d, dcl, z, v, n, l, r, w
Unvoiced	t, k, tcl, kcl, s, f, th, hh
Silence	h#

Table 2. Definition of mapping from phoneme classes to broad phoneme classes.

Broad class	Numbers 95 phone labels
Vowel	iy, ih, eh, ey, ay, ah, ao, ow, uw
Consonant	d, t, k, dcl, tcl, kcl, s, z, f, th, v, hh
Nasal	n
Liquid	w, r, l, er
Silence	h#

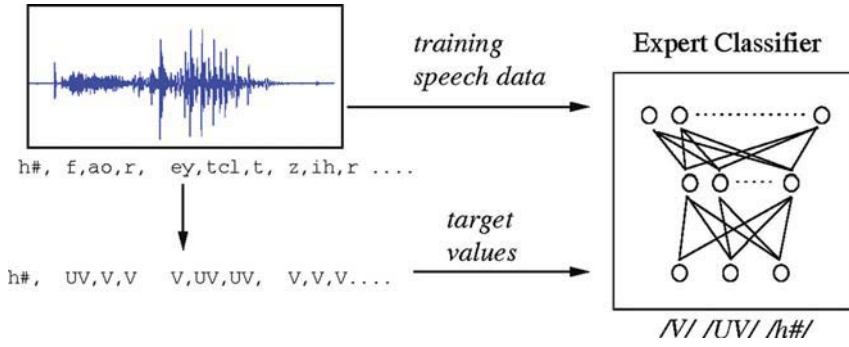

Figure 4. Overview of the training of a phonetic *expert* MLP. The target values are obtained from the annotations of the database by a direct mapping of the labels, keeping the alignments.

So, do the *expert* MLPs succeed in classifying their designated phoneme sub-sets? Figure 5 illustrates the accuracy of the trained broad phoneme class MLP *expert* classifier. The output posterior probabilities of two utterances from the Numbers95 database are shown with the correct labels superimposed. The darker the colour, the higher the posterior probability is for that particular broad phoneme class and frame. The top utterance is taken from the training and cross-validation part of the database and the lower utterance is taken from the development test set of the database. In both cases the classifier does, in general, succeed in producing high posteriors for correct classes. As expected, the utterance from the development test set has a few more discrepancies between the labelling (marked with the superimposed line) and the distribution of the output posteriors. The Frame Error Rates (FERs) for the training+cross-validation and development test set utterances are 6% and 16% respectively.

In Table 3 the frame error rates are shown for the training, cross-validation and development test set of the Numbers95 database. For all three feature types the frame error rates for the *experts* are lower than for the corresponding fullband systems. That is, the *experts* do contribute added discriminant power to the fullband system for both types of phoneme division and for all three types of feature.

5. ADDING PHONETIC INFORMATION TO A FULLBAND ASR SYSTEM

This section describes experiments conducted with a fullband *stem* system in combination with an *expert* system, which has been trained to provide

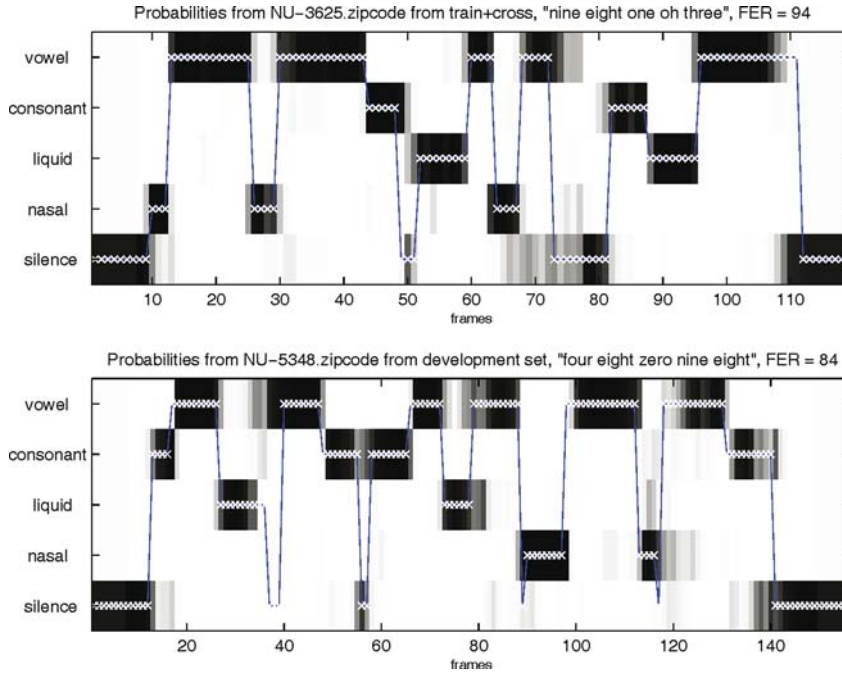

Figure 5. Posterior probabilities for two utterances from the training and development test part of the Numbers95 development test set respectively. The correct labels are superimposed. FER = Frame Error Rate.

Table 3. Frame error rates for the train, cross and development test part of the Numbers95 database for the *expert* MLPs.

Feature type	Type of system	Training (%)	Cross-validation (%)	Development test (%)
j-rasta-plpc	Broad *expert*	7.53	10.02	14.13
	Voicing *expert*	5.84	7.79	9.93
	Fullband	8.32	12.74	23.32
plpc	Broad *expert*	8.68	11.92	14.77
	Voicing *expert*	7.90	10.39	15.07
	Fullband	9.77	14.16	16.85
mfcc	Broad *expert*	8.14	12.54	15.95
	Voicing *expert*	9.23	10.37	12.44
	Fullband	10.32	15.05	19.12

complementary discriminant power to help resolve particular classification confusions between phoneme sub-sets.

The different configurations of fullband *stem* systems combined with an *expert* system is described below. The performance of the systems will be compared to a baseline system, which is a conventional single-stream fullband system.

- The **FullbandBaseline** system is a full frequency range ASR system.
- The **Fullband + VoicingExpert** system is a *stem* fullband system combined with an *expert* system providing posteriors from its *voiced/unvoiced* classifier. The mapping from phoneme labels to voiced/unvoiced/silence expert labels was given in Table 1.
- The **Fullband + BroadExpert** system combines an *expert* system trained at classifying broad phoneme classes (mapping given in Table 2) with a fullband *stem*
  system.

5.1. Clean Speech Results

Table 4 lists the results obtained from testing all fullband based systems on clean speech. In general, it is clear that the systems where phonetic information is added perform better than the corresponding **FullbandBaseline** systems for the various feature types. This is true except for the j-rasta-plpc **Fullband + BroadExpert** system, which performs a little worse than the corresponding **FullbandBaseline** system.

Comparing the three features types, the j-rasta-plpc is the only feature type where the **FullbandBaseline** system outperforms one of the systems with phonetic experts. At the same time, it is also the feature type that exhibits the best relative improvement: with the **Fullband + VoicingExpert** system.

5.2. Noisy Speech Results

Figures 6 and 7 show plots depicting the word error rates obtained from testing the two systems and the corresponding baseline system in car and

Table 4. Results from combining baseline fullbands with fullband *expert* systems.

System	j-rasta-plpc (%)	plpc (%)	mfcc (%)
FullbandBaseline	7.26	7.39	8.22
Fullband + VoicingExpert	6.77	7.22	7.75
Fullband + BroadExpert	7.54	6.81	7.73

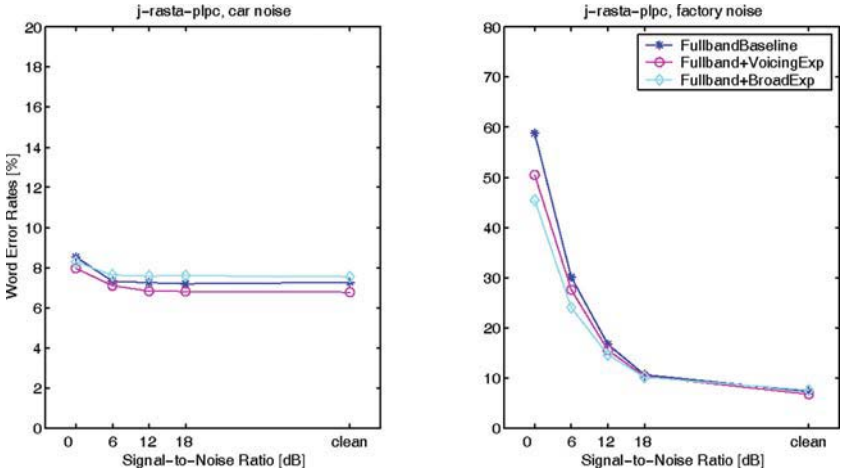

Figure 6. Word error rates plotted against SNR for the **FullbandBaseline**, the **Fullband + VoicingExp** and the **Fullband + BroadExp** systems for **j-rasta-plpc** features. The left and right panels are the results from testing in car and factory noise, respectively.

factory noise. The figures show the results from each of the three feature processing methods.

Comparing the different systems when tested in the two types of noisy environment, it is interesting to see that the relative performances of the

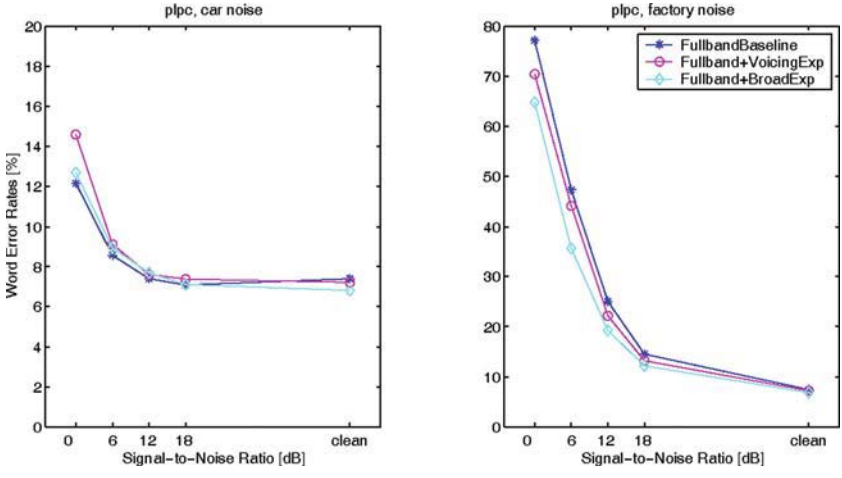

Figure 7. Word error rates plotted against SNR for the **FullbandBaseline**, the **Fullband + VoicingExp** and the **Fullband + BroadExp** systems for **plpc** features. The left and right panels are the results from testing in car and factory noise, respectively.

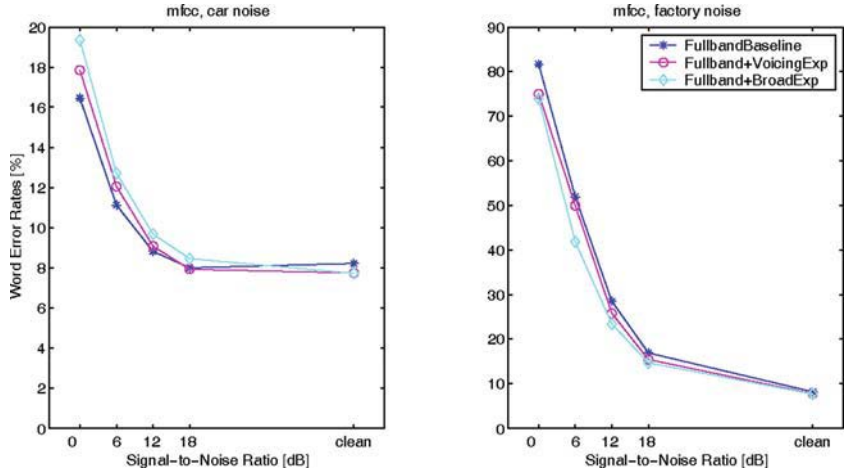

Figure 8. Word error rates plotted against SNR for the **FullbandBaseline**, the **Fullband** + **VoicingExp** and **the Fullband** + **BroadExp** systems for **mfcc** features. The left and right panels are the results from testing in car and factory noise, respectively.

systems seem to be dependent on the noise type. When testing with speech to which car noise samples are added, the **FullbandBaseline** performs better for both plpc and mfcc features as the noise level increases. For the noise robust j-rasta-plpc features, the **Fullband** + **VoicingExp** systems seems to be slightly better in both clean and car noise affected conditions. In the factory noise condition, all the feature types exhibit the same relative performances, and the **Fullband** + **BroadExp** can be seen to be consistently better than the other systems.

In the above described experiments the behaviour of the phonetic *expert* is crucial for the performance, since it is the complementariness of the errors of the *expert* which cause any observed increase in clean speech performance and noise robustness. The experiments here have been designed to demonstrate that adding phonetic information to a high-performance ASR system is beneficial. It was chosen to model the phonetic expert knowledge using a powerful MLP configuration. However, only limited effort was spent trying to further optimise the experts. Nevertheless, an improvement in performance was observed when phonetic information was added.

However, one must take into consideration that the phonetically augmented systems described above all use approximately twice as many parameters as their corresponding baseline systems. In (Christensen et al., 2000a) we showed that even when making fairer comparisons

(parameter-wise) between systems, it is possible to significantly increase performance when using heterogeneous signal processing in an ASR system.

6. ADDING PHONETIC INFORMATION TO MULTI-BAND SYSTEMS

This section concerns the inclusion of knowledge from *expert* systems to multi-band systems.

The different configurations of multi-band *stem* systems combined with an *expert* system is described below. The performance of the systems will be compared to a baseline system, which is a conventional multi-band system.

- The **MultibandBaseline** system is a multi-band system, where *the same* feature is used in all sub-bands.
- The **Multiband + VoicingExpert** system combines a multi-band *stem* system with an *expert* providing *voiced/unvoiced* classification information.
- The **Multiband + BroadExpert** system is a multi-band *stem* system is combined with an *expert* system providing estimates for broad phoneme posterior probabilities.
- The **Multiband + PhonemeExpert** system combines a multi-band *stem* system with an *expert* system which is in principle an individual fullband system on its own. That is, it provides individual posterior probabilities for each of the phoneme classes.

6.1. Clean Speech Results

Table 5 lists the clean speech results from testing the multi-band based systems augmented with phonetic *experts* for all three feature types. It is clear that considerable improvements are obtained over the baseline system results presented in the first row.

Table 5. Results (word error rates) from adding information from *expert* system to the multi-band *stem* system.

System	j-rasta-plpc (%)	plpc (%)	mfcc (%)
MultibandBaseline	13.19	13.32	15.07
Multiband + VoicingExpert	11.07	11.22	12.70
Multiband + BroadExpert	10.06	10.32	11.26
Multiband + PhonemeExpert	7.49	7.62	8.33

Not unexpectedly, the finer and more discriminative the phonetic information is modelled (moving downwards in the table), the greater the advantage of adding the *expert* system. In the most extreme case, where for the **Multiband + PhonemeExpert** basically a fullband is added to a conventional multi-band system, the relative improvement is over 40% for each of the feature types. Such a large advantage of combining a multi-band system with a fullband system is in accordance with results reported by Mirghafori and Morgan(1998) for plpc features.

Comparing the different feature types, the j-rasta-plpc based systems outperform the plpc based systems, and the mfcc based systems exhibit the highest word error rates. The j-rasta-plpc based system are significantly[5] better than the mfcc based systems (for all conditions).

It is informative to compare the results for the **Multiband + PhonemeExp** system to the **FullbandBaseline** system, see Table 4. For all feature types, adding the multi-band system does not help improve clean speech performance, despite the fact that the **Multiband + PhonemeExp** systems have twice as many parameters as the **FullbandBaseline** systems.

6.2. Noisy Speech Results

When looking at the results from testing on noisy speech, plotted for the different feature types in Figures 9–11, we find roughly the same pattern as was observed when testing the systems on clean speech. The effects of

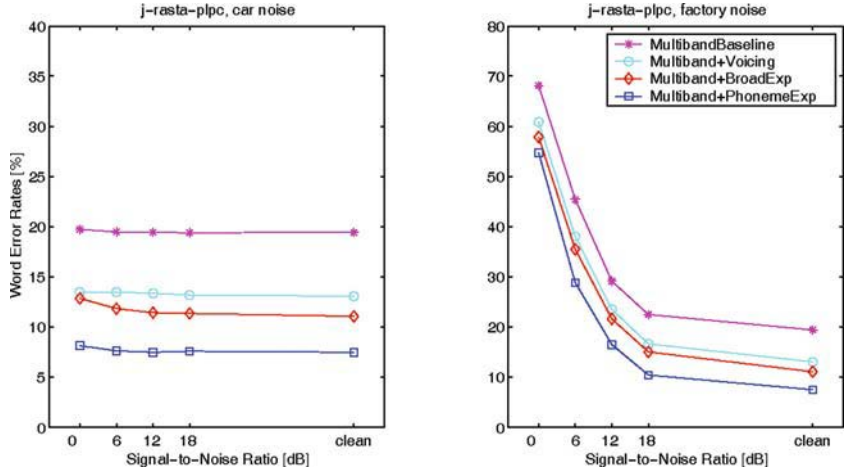

Figure 9. Plot of results (WER's) when testing the phonetic *experts* together with the multi-band systems, all based on **j-rasta-plpc** features. The systems are tested in car noise (left-hand side) and in factory noise (right-hand side).

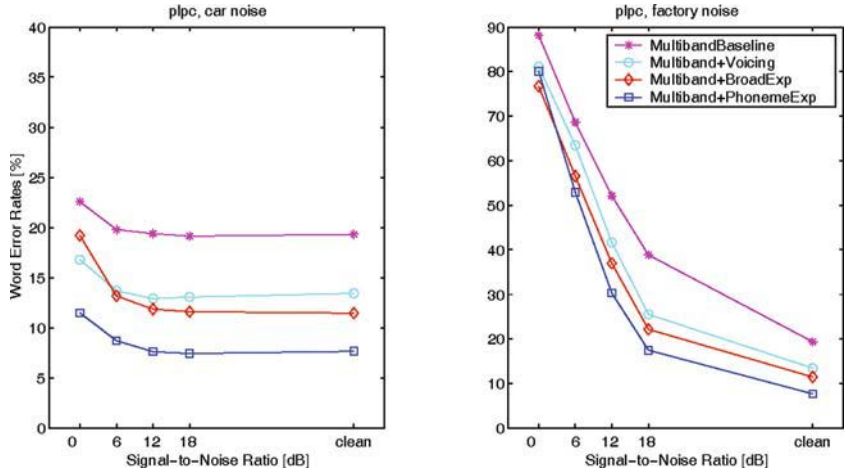

Figure 10. Plot of results (WER's) when testing the phonetic *experts* together with the multi-band systems, all based on **plpc** features. The systems are tested in car noise (left-hand side) and in factory noise (right-hand side).

the different *expert* systems can be ranked as: **PhonemeExpert** > **BroadExpert** > **VoicingExpert** > **Multiband** with no *expert*. For very low SNR's there is a small decrease in how much better the **Multiband** + **PhonemeExpert** performs.

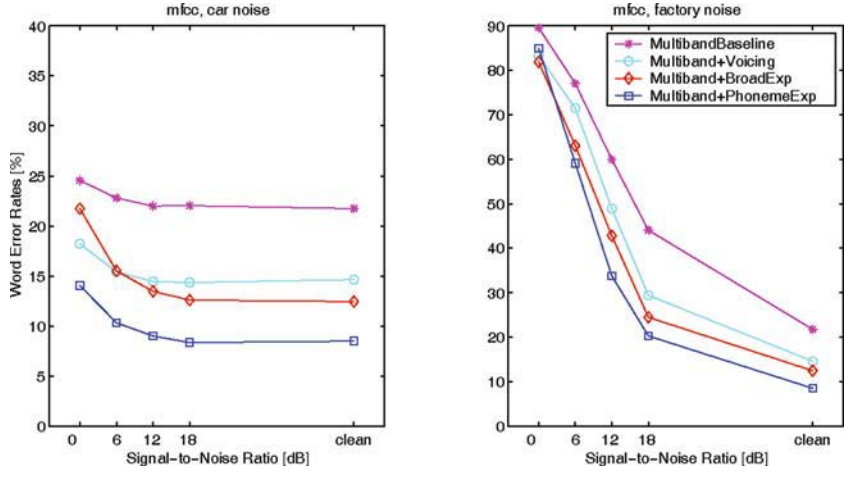

Figure 11. Plot of results (WER's) when testing the phonetic *experts* together with the multi-band systems, all based on **mfcc** features. The systems are tested in car noise (left-hand side) and in factory noise (right-hand side).

Comparing the performance of the multi-band based systems on the two different types of noise, it is evident that the system performance degradation is far less drastic when operating in car noise. The car noise sample from the NOISEX-92 database is low-frequency, and to a large degree a band limited noise type.

7. DISCUSSION AND CONCLUSIONS

It is interesting to compare the fullband and multi-band based systems with each other. Table 6 shows the clean speech results from the best multi-band based system (the **Multiband + PhonemeExp**), the best fullband based system (the **Fullband + BroadExpert**), as well as the two baseline systems. The WER's for the **FullbandBaseline** and the **Multiband + PhonemeExpert** are not significantly different.

Comparing the results from testing on noisy data, a similar pattern is found, although for the plpc and mfcc features in particular the **Fullband + BroadExpert** clearly shows the best noise robustness, in particular at higher noise levels for the factory noise.

This chapter investigated a way to introduce more heterogeneous information into an existing ASR system. An inspective analysis of the confusion matrices for the ASR *stem* systems led to the design of a phonetic *expert* system. The *expert* was specifically targeted at supplementing the errors committed by the *stem* system, giving a heterogeneous system, where the individual items were designed to be complementary.

To avoid the curse of dimensionality problem, the expert information is introduced at the level of the acoustic model. Two types of *expert* configurations are used, each providing discriminative information regarding groups of phonetically related phonemes. The phonetic *expert* is implemented using an MLP. Experiments show that when using the *expert* in conjunction with both a fullband and a multi-band system speech recognition performance is increased, and, in many cases, noise robustness improves for a range of noise levels.

Table 6. WER's for fullband- and multi-band based systems.

System	j-rasta-plpc (%)	plpc (%)	mfcc (%)	Parameters
FullbandBaseline	7.26	7.39	8.22	576,000
Fullband + BroadExpert	7.54	6.81	7.73	1,108,500
MultibandBaseline	13.19	13.32	15.07	576,120
Multiband + PhonemeExpert	7.49	7.62	8.33	1,152,120

NOTES

[1] In this chapter, the term *feature* will be used to mean a given representation of the acoustic information in data.
[2] A multi-band is a particular configuration of a multi-stream system, where adjacent frequency bands comprise the individual streams.
[3] A confusion matrix presents details of recognition results. It contains a set of counters, c_{ij} where $i,j \in \{1,\ldots,K\}$ each representing the number of times phoneme j was received/hypothesised when it was phoneme i that was transmitted/presented to the recogniser. Thus counts in the diagonal represent correctly recognised units, and off-diagonal counts are errors.
[4] The frequency bands are chosen so as to roughly capture the formant regions.
[5] In this chapter, when comparing experimental results, the term 'significant' is used to indicate that two WERs are different when considering a 0.05% significance level interval.

REFERENCES

Allen, J.B. How do humans process and recognize speech. *IEEE Trans. Speech and Audio Processing* 2(4) (1994): 567–577.

Bishop, C.M. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, England, 1995.

Bourlard, H. and Dupont, S. A new ASR approach based on independent processing and recognition of partial frequency bands. In: *Proceedings of ICSLP '96*, Philadelphia, PA, 1 (1996): 426–429.

Christensen, H., Lindberg, B., and Andersen, O. Employing heterogeneous information in a
multi-stream framework. In: *Proceedings of ICASSP '00*, Istanbul, Turkey, 2000a.

Christensen, H., Lindberg, B., and Andersen, O. Multi-stream speech recognition using heterogeneous minimum classification error feature space transformations. In: *Proceedings of NORSIG '00*, Norrköping, Sweden, 2000b.

Christensen, H., Lindberg, B., and Andersen, O. Noise robustness of heterogeneous features employing minimum classification error feature space transformations. In: *Proceedings of ICSLP '00*, Beijing, China, 2000c.

Christensen, H. *Speech Recognition using Heterogeneous Information Extraction in Multi-Stream Based Systems*. Ph. D. thesis, Aalborg University, Denmark, 2001.

Davis, S.B. and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP, 28(4) (1980): 357.

Fletcher, H. *Speech and Hearing in Communication*. Krieger, New York, 1953.

Ghitza, O. Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Trans. Speech and Audio Processing* 2(1) (1994): 115–132.

Gold, B. and Morgan. N. *Speech and Audio Signal Processing. Processing and Perception of Speech and Music*. John Wiley & Sons, Inc., 2000.

Greenberg, S. Auditory function. In: M. J. Crocker (ed.), *Encyclopedia of Acoustics*. John Wiley & Sons, Inc., 1997: 1301–1323.

Halberstadt, A.K. *Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition*. Ph.D. thesis, Massachusetts Institute of Technology, Massachusetts, USA, 1998.

Hansen, L.K. Neural network ensembles. *IEEE Trans. Pattern Analysis and Machine Intelligence* 12(10) (1990): 993–1001.

Hermansky, H. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America* 87(4) (1990): 1738–1752.

Hermansky, H. RASTA processing of speech. *IEEE Trans. Speech and Audio Processing* 2(4) (1994): 578–589.

Hermansky, H. Should recognizers have ears? In: *Proceedings of Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, 1997: 1–10.

Ho, T.K. *A Theory of Multiple Classifier Systems And Its Application to Visual Word Recognition*. Ph.D. thesis, Department of Computer Science, State University of New York at Buffalo, New York, USA, 1992.

Lippmann, R.P. Speech recognition by machines and humans. *Speech Communication* 22 (1997): 1–15.

Kittler, J. Hatef, M., Duin, R.P.W., and Matas, J. On combining classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence* 20(3) (1998): 226–239.

Numbers95. *Numbers corpus, release 1.0*. Department of Computer Science and Engineering. Oregon Graduate Institute, 1995.

Mirghafori, N. and Morgan, N. Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers. In: *Proceedings of ICSLP '98*, Sydney, Australia, 1998.

Pols, L.C.W. Flexible human speech recognition. In: S. Furui, B.-H. Juang, and W. Chou (eds.), *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*. 1997: 273–283.

Steeneken, H.J.M. *On Measuring and Predicting Speech Intelligibility*. Ph.D. thesis, University of Amsterdam, Amsterdam, 1992.

Turner, K. *Linear and Order Statistics Combiners for Reliable Pattern Classification*. Ph.D. thesis, Graduate School of The University of Texas at Austin, Austin, USA, 1996.

Varga, A., Steeneken, H.J.M., Tomlinson, M., and Jones, D. The NOISEX-92 CD-ROMs. The NOISEX-92 study on the effect of additive noise on automatic speech recognition, 1992.

GUILLAUME GRAVIER[*], FRANCOIS YVON[†], BRUNO JACOB[‡] and FRÉDÉRIC BIMBOT[§]

INTRODUCING CONTEXTUAL TRANSCRIPTION RULES IN LARGE VOCABULARY SPEECH RECOGNITION

ABSTRACT. This paper presents an approach to the integration of contextual phonological rules in the beam-search algorithm of a large vocabulary speech recognition system. The main interest of contextual transcription rules is that they implement well-formedness constraints on pronunciation of word sequences. These constraints complement the language model probabilities on word sequences. As such, they should decrease the average acoustic confusability between words and thereby reduce the recognition search space. This approach is evaluated on a dictation task in French for several different sets of contextual phonological rules. Our results show that, in the current setting, the introduction of contextual rules does not harm the overall performance, while effectively reducing the search space. We detail the algorithmic aspects of the introduction of contextual rules in the decoder, present our empirical results and discuss possible extensions of this work.

KEYWORDS. large vocabulary speech recognition, pronunciation variants, contextual transcription rules

1. INTRODUCTION

Modern large vocabulary continuous speech recognition (LVCSR) systems view the process of decoding the input speech signal as a *search* for the most likely word sequence, *i.e.* for the sequence which best matches the acoustic input, while being at the same time syntactically plausible (Jelinek, 1998). Given a closed vocabulary V, the search algorithm considers all the possible sequences of words from V, measuring the syntactic plausibility of each sequence based on stochastic language

Address for Correspondence:
[*] CNRS and IRISA/INRIA Rennes
[†] ENST Paris, Dpt. Informatique et Réseaux
[‡] Laboratoire d'Informatique de l'Université du Maine
[§] CNRS and IRISA/INRIA Rennes, Equipe METISS

W. J. Barry and W. A. van Dommelen, The Integration of Phonetic Knowledge in Speech Technology, 87–106.

models (LM). The acoustic match, on the other hand, results from computing the likelihood of the input signal in a probabilistic model of word acoustics. As the typical size of V makes it impossible to model each word individually, acoustic models are built by plugging "generic" models of acoustic units into a broad phonetic representation of the word pronunciation(s). The acoustic units are usually modeled by hidden Markov models (HMMs). Three levels of description are thus involved in the search: acoustic units, pronunciations and words.

As the focus shifts from dictation systems to recognition of more spontaneous speech, accurately modeling word pronunciations becomes increasingly important. Dealing with different speech styles and accents however requires an appropriate description of pronunciation variation (Strik and Cucchiarini, 1999). Although some variation can be captured using refined statistical models, the modeling of pronunciation variants is usually performed by enriching the pronunciation lexicon with new variants. Many approaches have been proposed to derive these pronunciation variants (see *e.g.* Strik et al., 1998), from the use of hand-crafted generative phonological rules (Wester et al., 1998; Pérennou and Calmès, 2000) to the automatic inference of variants from speech corpora (Ramabhadran et al., 1998; Amdal et al., 2000; Yang and Matens, 2000), sometimes abstracted as variation rules (Ravishankar and Eskenazi, 1997; Riley et al., 1999). Both the knowledge-based and the data-driven approaches result in an extended list of pronunciations. Whenever possible, the probability of each variant is estimated from the data and included in the lexicon.

However, adding more variants expands the search space and increases the average acoustic confusability among words. As a consequence, the results from recognition experiments using these larger lexicons have been inconclusive. As suggested for instance in Jurafsky et al. (2001), some of these problems might be alleviated by introducing more contextual dependencies in the way variants are considered. Though short range phonetic dependencies are already modeled in many systems under the form of so-called "cross-word triphones", this is usually not the case for the higher order constraints also involved in speech production. At the sentence level, for instance, the selection of a specific elocution strategy which determines the speech style, is bound to respect strong internal consistency constraints: speakers do not freely switch between different rates or accents, but rather tend to maintain some sort of intra-utterance consistency. However, there is no model of

such high-level linguistic constraints. Therefore, this paper rather focuses on the introduction of phonological and lexical constraints.

Short-distance constraints are common and are traditionally modeled by post-lexical phonological rules. For instance, in English, the degemination rule provokes the simplification of consonantic clusters at word juncture, as in *last time* where the two /t/s are often merged into one, in informal speech. In French, the phenomenon of *liaison*, *i.e.* the optional realization of a word-final latent consonant in specific contexts (see Section 2.2) offers another example of contextual variation. Introducing contextual constraints into an ASR system has an impact on the search algorithm, as it is no longer true that any sequence of pronunciations is valid. For instance, if liaison is handled properly, a variant where a latent consonant is pronounced cannot be followed by a consonant initial word. An illustration of this situation is given in the following figures. In Figure 1, all the pronunciation sequences are permitted, thus giving a total of 18 different possible paths; when contextual constraints apply, as illustrated in Figure 2, the number of possible sequences is reduced to two.

How can a search algorithm be modified in order to efficiently exploit such constraints? A simple solution could be to consider all the variants as non-contextual, leaving it to the acoustic match to select the appropriate variant. However, this results in the exploration of unlikely state sequences during the search, thus wasting the decoder's resources on invalid paths.

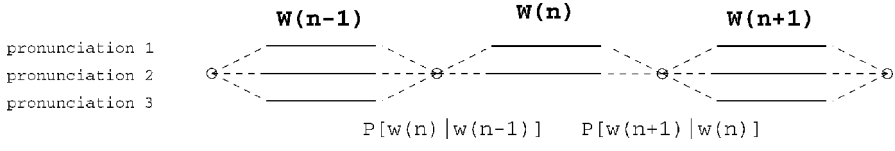

Figure 1. Unconstrained sequence of pronunciations.

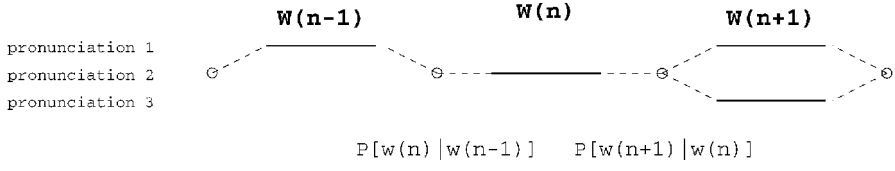

Figure 2. Constrained sequence of pronunciations.

Another solution could be to introduce contextual constraints through multi-word units (MWU). This approach consists of modeling MWU as words, listing all the possible pronunciations of a MWU directly in the lexicon (Riley et al., 1999; Kessens et al., 1999). As a consequence, the context of each sub-unit is easily taken into account. This approach is effective for frequent sequences involving short functional words whose pronunciations can be drastically affected through coarticulation. However, it cannot be applied to systematic patterns such as liaison, which apply to every instance of a plural determiner followed by a vowel initial noun. A more generic approach consists of building the stochastic language model over the pronunciation variants (rather than words) and in incorporating the contextual dependencies directly into the bigram probabilities (Cremelie and Martens, 1995; Schiel et al., 1998). This approach, however, requires large corpora annotated with pronunciation variants from which to estimate the LM probabilities.

In this paper, we present an initial attempt to incorporate contextual transcription rules directly at the search level, thus limiting the search space to permissible pronunciation sequences. The basic principle of our approach is to define lexical classes, *i.e.* subsets of (orthographic) words sharing common characteristics at various levels (syntactic, morphological, phonological, etc). Lexical classes are subsequently used by the speech-decoding algorithm, which checks the compatibility between adjacent pronunciations, based on constraints expressed in terms of these word classes. The main motivation for modeling contextual interactions directly at the search level is the intuition that the subsequent reduction of the search space should help reduce (acoustic) confusions and improve the overall recognition rate.

The paper is organized as follows: we first describe the basics of the tree-based search algorithm with a bigram LM (Ortmanns and Ney, 1997; Deshmukh et al., 1999). We then introduce contextual transcription rules and show how they can be used during the search. This methodology is used and evaluated on a dictation task for French. In the light of these results, we finally discuss our algorithm, pointing out its current limitations and possible extensions.

2. DECODING WITH CONTEXTUAL CONSTRAINTS

2.1. The Standard Search Algorithm

The *Sirocco* speech decoder[1] is based on a time-synchronous beam-search decoding strategy. We recall here some basic aspects of this algorithm, in particular those related to the processing of word ends. The reader is

refered to Ortmanns and Ney (1997) for a detailed description. In this approach, the search space is organised so as to factor out, in a tree, identical word-initial sequences of HMM states. This organisation is meant to reduce the number of hypotheses corresponding to word-initial states. However, it delays the identification of the word w currently being decoded until a leaf of the tree is reached. Consequently, it also delays the application of the bigram LM probability $P(w|v)$ and forces the algorithm to keep track of the history v until w is identified. A solution to this is to keep a separate tree copy for each distinct history (the preceding word in the case of a bigram LM), as illustrated in Figure 3 for a 3 word vocabulary.

Within a tree copy, acoustic hypotheses are propagated according to the classic dynamic-programming (DP) equation. At the end of a word, *i.e.* when a leaf of the tree is reached, the bigram score is added to the score of the current path and the DP maximization is performed over all the previous words. This is illustrated in Figure 3: upon starting the

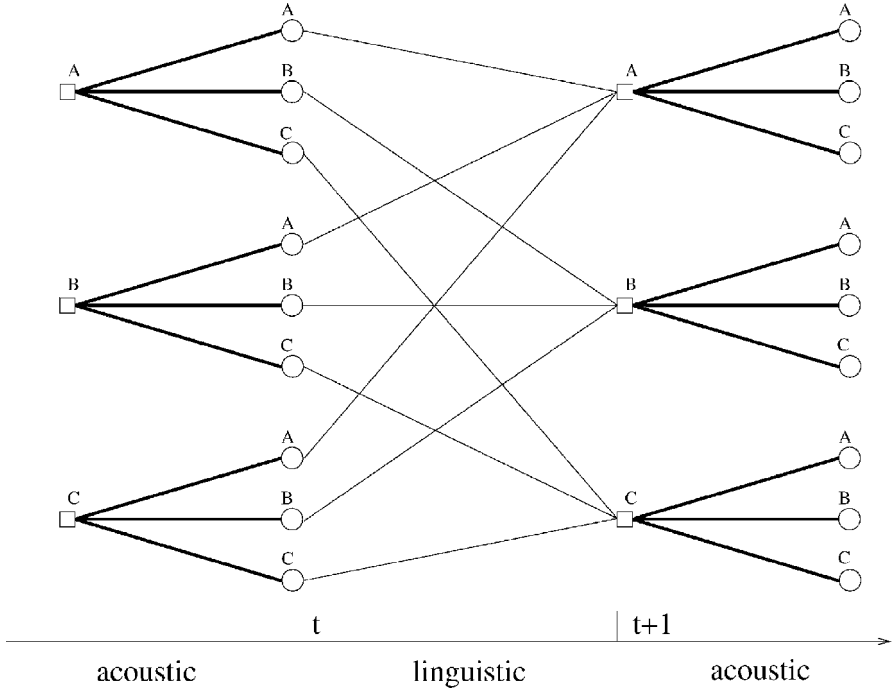

Figure 3. Principle of the search strategy for tree-based lexicons. Thin lines correspond to the LM score; thick lines correspond to the acoustic score.

right tree-copy for A, the maximization is performed over the 3 possible states ending A at time t as indicated by the thin lines. Formally, let us assume that an acoustic hypothesis reaches the leaf S_w of a tree copy at time t, thus hypothesizing the end of word w. Let us denote $h_w(t)$ the log-likelihood of the best word sequence ending with word w at t. The word level DP maximization is given by

$$h_w(t) = \max_{v \in V} q_v(t, S_w) + \beta \ln P[w|v], \tag{1}$$

where $q_v(t, S_w)$ denotes the log-likelihood of the best path ending at t in the node S_w of the tree copy corresponding to the preceding word v; $P[w|v]$ is the bigram score and β the LM scale factor. The maximization (1) is performed over the set V of all possible words. Finally, the word end hypothesis (w, t) is recorded with a back-reference to the word hypothesis (v, t') which maximizes (1): (v, t') is the *back-pointer* information propagated with the acoustic hypothesis. A new tree copy for w is then started, giving rise to new hypotheses having w as their preceding word and a back-pointer equal to (w, t).

2.2. Classes and Contextual Transcription Rules

As with any LVCSR system, *Sirocco* requires a set of lexical resources, consisting of a list of orthographic word[2] forms and a pronunciation dictionary. Two additional refinements to the lexical description are possible in *Sirocco*:

- word classes can be defined, for instance based on morpho-syntactical and phonological features.
- pronunciations can be restricted to specific contexts, where a context is defined as logical combinations of word classes, hence the term "contextual transcription rules" refers to the entries in the pronunciation dictionary.

Let us illustrate these concepts using the phenomenon of liaison in French (Lerond, 1980). Liaison is primarily dependent on the phonological environment; a latent consonant can only be pronounced in the context of a following vowel or a glide. We will refer to this context as *vocalic*. However, even when the phonological context allows it, the latent consonant is not necessarily pronouced. First, some vocalic contexts nonetheless prohibit liaison: these idiosyncratic contexts correspond to a closed list of words, notably including words starting with a mute 'h' *(h aspiré)*. In this case, the constraint is purely lexical. Second, a liaison

can either be mandatory (as between a plural determiner and the following noun or adjective), or optional. In the latter, the liaison is realized depending on a complex constellation of contextual factors, including the syntactic environment as well as additional lexical, prosodic and stylistic factors. In some contexts, liaison does not occur, such as between a plural nominal subject and the following verb.

To give an example: the plural determiner *les* ('the'), is either pronounced /lez/[3] if a liaison occurs, or /le/ otherwise. The rules governing the choice of one or the other pronunciation are, in this case, quite simple: if the following word starts with a vowel, and is a plural adjective or noun determined by *les*, the liaison variant is preferred. Otherwise, the transcription /le/ must be used. The corresponding contextual transcription rules are given in Table 1, where *Vinit* is the class of words whose pronunciations start with a vocalic sound, *Plural* is the class of plural nouns and adjectives, and * denotes the universal class (which contains the entire vocabulary). In fact, the actual rules for *les* are somewhat more complex, due to the fact that this word form corresponds both to a determiner and a clitic pronoun. If the pronunciation of the former essentially follows the rule discussed above, the latter will select the liaison variant when it occurs as the preposed direct object of a vowel-initial verb.

In the experiments reported in Section 4, we have mostly used classes and transcription rules to express linguistically motivated constraints. However, word classes can be used for different purposes. For instance, we also use them to automatically align the speech signal given an orthographic transcription. This is achieved by assigning each word i in the transcription to a class, C_i, whose only element is i. Transcription rules are then defined in such a way that the pronunciation of i can only occur if preceded by a word in class C_{i-1} and followed by a word in class C_{i+1}, thus making the reference word sequence the unique valid path with respect to the contextual transcription rules.

2.3. Using Rules During the Search

Using rules to constrain the search mainly affects the processing of end-of-word hypotheses. Indeed, our approach requires one to verify

Table 1. Example of contextual phonological rules for the word *les*. The operator '!' denotes the boolean negation.

(*)	*les*	(Vinit & Plural)	→	/lez/
(*)	*les*	(!(Vinit & Plural))	→	/le/

the left and right contexts of each pronunciation to check its validity. As our algorithm proceeds from left to right, the left context of any word w is known and can be easily checked. This is not true for the right context, which will not be known until the next word is decoded. Until then, the pronunciation of w will have to be stored and propagated.

This modification is performed as follows: the list of valid contexts for a pronunciation p is attached to the corresponding tree leaf, along with the corresponding word(s) identifier. In the previous example (see Table 1), the leaf corresponding to the pronunciation $p = /\text{lez}/$ would point to the word *les* and to the corresponding left and right contexts, *i.e.* (*) and (Vinit) respectively. Hypotheses reaching a tree leaf node have slightly different semantics than before: they now indicate that we have decoded a specific variant of a given word w, and not just any variant of w. End-of-word hypotheses are therefore characterized by the word, the word end time and the pronunciation rule.

Let us denote the end-of-word hypothesis associated with state S_w by (w, a, t), where a denotes a specific pronunciation rule. We also assume that the back-pointer associated with the acoustic hypothesis which generated the word end hypothesis (w, a, t) carries information concerning the previous word transcription rule a'. When a word end is reached in the tree copy corresponding to the preceding word v, one has to decide whether the transition $(v, a', t') \rightarrow (w, a, t)$ is valid with respect to the contexts of a and a'. This will be the case if and only if:

i. the classes of w match the right context for the transcription rule a' of v, and
ii. the classes of v match the left context for the transcription rule a of w.

In a more formal manner an end-of-word hypothesis (w, a, t) is valid if the classes of w match the right context of the previous word transcription rule taken along the best path (condition (i)) and if the classes of v match the left context of the current word end hypothesis (condition (ii)). These new constraints modify the word level maximization (1), which should now be performed only over valid end-of-word hypotheses. Given this modification, (1) can now be viewed as a two-stage process where the maximization is first carried out for each end-of-word hypothesis (w, a, t), over the set of possible predecessors (v, a'), according to:

$$h_w(t, a) = \max_{(v,a') \in V(w,a)} q_v(t, S_w) + \beta \ln P(v, w) + \ln P(w, a) \qquad (2)$$

where $V(w,a)$ is the set of all the valid word-rule pairs that can occur before (w,a). If a end-of-word hypothesis is not valid with respect to its left context, it is simply discarded along with the corresponding path. The second stage of the maximization aims at finding the best word (w,t) and is carried out over the set of rules that hypothesized the end of word w at t, according to:

$$h_w(t) = \max_a h_w(t,a) \tag{3}$$

New acoustic hypotheses are then generated with a back reference to $(w,\hat{a},t)$, where $\hat{a}$ is the rule which maximizes (3).

During the search, the validation of the right context of a rule is performed with a delay of one word. This is a consequence of the tree-based organization of the pronunciations, which postpones the identification of decoded words until a leaf is reached, thus delaying the application of contextual pronunciation constraints and of the bigram LM.

3. TYPOLOGY OF CONTEXTUAL RULES

In this section, we introduce the MHATLex lexicon (Pérennou and Calmès, 2000), which has been our primary lexical resource for the experiment reported in Section 4. We first describe the resource in general terms (Section 3.1), mainly focusing on the phonological representations contained in the lexicon. We then focus on three cross-word phenomena, namely liaison, mute-e elision and liquid truncation. These are discussed in Section 3.2.

3.1. MHATLex

In MHATLex, two levels of description are considered for word pronunciations: the *phonological* representation expresses the association between lexical items and pronunciation variants *in a specific linguistic context*. Context-free rewrite rules provide the capacity to generate the corresponding set of phonetic representations. Each phonetic variant inherits the contextual constraints associated with the corresponding phonological representation. For instance, the verb *prendre* ('to take') is represented at the phonological level by two entries /pRa ~ dR/, valid if followed by a vowel-initial word, and /pRa ~ (dR@)/, which requires a consonant-initial successor. The latter encodes several variants, which are derived by rewriting the variable group /(dR@)/ into any of the following: /d /, /n / or /dR@ /. The corresponding pronunciations /pRa ~ d /, /pRa ~ n /, and /pRa ~ dR@ / all require the same context as /pRa ~ (dR@)/, *i.e.* a consonant initial successor.

MHATLex also includes morpho-syntactic information, such as the main syntactic category and citation form of each entry. From these, a total of 20 lexical classes were built and used to define contexts: half of the classes encode morpho-syntactic information, while the remaining ones encode phonological information. A list of all those classes is given in Table 2. Word-to-class mappings were defined accordingly. Note however that, while MHATLex distinguishes among homographs on the basis of their morpho-syntactic properties, our system only handles orthographic strings. This leads to ambiguous class assignments for forms like *tombe*, which can be either a noun ('(a) grave') or a verb ('(I) fall').

3.2. Three Contextual Phonological Rules

In order to introduce contextual rules in a progressive and controlled manner, the variant generation rules used in Mhatlex were manually tagged according to the phonological phenomenon they account for. Three main contextual phenomena were considered: liaison, mute-e deletion and liquid consonant truncation (Dell, 1985). In this section, we present these phenomena and explain how the rules were tagged.

3.2.1. *Liaisons*

A description of liaison has already been given (see Section 2.2) and we focus here on the rule-tagging process. For each lexical item having a liaison variant, two entries occur in the phonological lexicon: the first one corresponds to the non-liaison case and requires a consonant-initial

Table 2. List of the lexical classes derived from MHATLex.

Morpho-syntactic classes	Phonological classes
conjunction	initial vowel
determiner	initial glide
adjective	initial aspirated h + vowel
noun	initial aspirated h + -vowel
interjection	initial non-nasal consonant
preposition	initial nasal consonant
pronoun	word-final consonant
verb	word-final vowel
acronym	phonological group boundary
proper noun	
adverb	

successor; the second one represents the liaison case and requires a vowel-initial successor. To account for optional liaisons the second representation, in fact, encodes two variants, corresponding to the pronunciation or non-realization of the latent phoneme. In this second representation, the rewrite rule generating the first variant is tagged 'L' (a latent phoneme is pronounced), while the rewrite rule generating the second variant is tagged 'l'.

Since there is no easy direct way to link the rewrite rule for the first phonological description to the fact that it corresponds to a non-liaison, this rule is not tagged as corresponding to a liaison phenomenon.

3.2.2. *Mute-e rules*

As far as mute-e is concerned, three main cases were covered: final mute-e deletion, mute-e deletion in a word-initial syllable and epenthetic mute-e.

In standard French, the deletion of a word-final mute-e is considered mandatory when the following word starts with a vowel. When the word starts with a consonant, the situation is more complex, and a mute-e may or may not be realized, depending on a complex set of variables in which the structure of consonant cluster resulting from the deletion plays a major role.

For a restricted set of words, a mute-e in the first syllable may optionally be dropped if the preceding word ends with a vowel. In any other context, the /@/ must be pronounced. For instance, the initial /@/ can be dropped in: *les chemins* ('the paths') then pronounced /leSme~/ / instead of /leS@me~/, but not in *quatre chemins* ('four paths'). Note that the precise rule is in fact a bit more complex and the mute-e deletion might be possible after a consonant-final word.

It is possible to insert an epenthetic mute-e at the juncture between two words when the concatenation of the word-final and -initial consonant clusters is too complex, as for instance in the sequence *un ours triste* ('a sad bear') which may be pronounced with an epenthetic /@/: /9~nuRs@tRist/.

The rewrite rules corresponding to these three phenomena were tagged 'E' or 'e', depending on whether the /@/ is pronounced or not.

3.2.3. *Liquid truncation*

Liquid truncation refers to the optional deletion of a word-final liquid when preceded and followed by a consonant. This often occurs in conjunction with a final mute-e deletion in informal speech. For example, liquid truncation is involved in the pronunciations /pRa~d/ and

/pRa~n/ derived for the word *prendre* ('to take'). The corresponding rewrite rules were tagged 'R' and 'r', respectively.

3.2.4. *Illustration*

An example of our tagging scheme is given by the verb *montrent* ('(they) show'). For this word, two contexts are considered:

- when the following word starts with a vowel, a liaison can optionally take place: if it does, though, the final mute-e will be realized, yielding a first variant: /mo~tR@t/, tagged 'LE' (liaison, mute-e maintained); if it does not, the mute-e has to be dropped, yielding a new variant /mo~tR/, tagged 'le'.
- when the following word starts with a consonant, no liaison can occur, but the final mute-e and the preceding /R/ can optionally be deleted, yielding the pronunciations /mo~tR@/ (tagged 'RE') and /mo~t/ (tagged 're'). Note that we could not mark these variants with 'l', as it should be, for there is no explicit way in MHATLex to relate this entry to the previous one.

Finally, note that due to the mismatch between the MHATLex notion of a word and ours, we had to assign all the (contextual) pronunciations of a set of homographs to the same word, sometimes resulting in a proliferation of (redundant or conflicting) pronunciation rules. *Rule compaction* has been applied whenever possible to factor out rules having similar left or right context, and to limit the number of rules for a given word. Each pronunciation may optionally end with an inter-word short-pause model.

4. EXPERIMENTAL RESULTS

Experiments using various sets of the contextual phonological rules described above were carried out on a dictation task in French using the BREF corpus (Lamel et al., 1991; Dolmazon et al., 1997).

4.1. The BREF Corpus

The BREF corpus is composed of read sentences extracted from the newspaper *Le Monde*. A set of 41,000 sentences uttered by 80 speakers was used to estimate the parameters of 40 monophone HMMs with 3 states and 32 Gaussians per state. A separate set of 300 sentences uttered by 20 speakers was used for testing, using the protocol defined in Dolmazon et al. (1997). The test set has a total of approximately 9,000 words. The vocabulary contains the 20,000 most frequent words occurring

in two consecutive years of *Le Monde* and this corpus was also used to estimate bigram and trigram LMs.

4.2. Lexical Resources

Several lexical resources were derived from MHATLex by varying the number of contextual rules. First, an unconstrained lexicon was generated, using all the pronunciation variants of MHATLex without any contextual restriction. Second, a lexicon for each of the three contextual phonological phenomena described in Section 3.2 was generated. In each of these, contextual constraints only apply to a specific subset of rules. These three lexical resources will be referred to as *E*, *L* and *R*, for mute-e, liaison and liquid-truncation rules respectively. To study the combination of those three phenomena, we generated a lexicon including all possible constraints for these rules; this lexicon will be referred to as ELR. Finally, a lexicon, *M*, corresponding to all the MHATLex contextual rules was generated.

Let us illustrate some of the differences between these resources by re-examining the word *prendre*, already discussed in Section 3.1. As mentioned earlier, *prendre* has two phonological representations. The first yields the pronunciation /pRa~dR/ in the context of a following word beginning with a vowel. The second representation yields the pronunciations /pRa~d/, /pRa~n/ and /pRa~dR@/ in the context of a following consonant-initial word. According to our tagging scheme, the first two pronunciations of the second phonological description correspond to a mute-e deletion (tag 'e'), while the last one corresponds to the realization of the /@/ (tag 'E'). However, it is impossible to tag 'e' the rule used to generate the pronunciation /pRa~dR/. Therefore, the right context for that rule will be omitted in the lexicon *E*, whereas it is kept in the lexicon *M*.

The lexicons described above were generated using two slightly different strategies. In the first, (A), all the MHATLex rules were used to generate the pronunciation variants. In the second, (B), some very infrequent rules were omitted, thus generating lexicons with slightly fewer pronunciation variants.

For both resources, Table 3 gives the number of entries for each lexicon. In the absence of contextual rules, the number of rules is equal to the number of pronunciation variants. This table also gives the branching factor of each lexicon, computed as the average ratio between the number of pronunciation variants that may follow a given variant, divided by the total number of variants. A value of 1 therefore corresponds to an uncon-

Table 3. Number of contextual transcription rules for each set of rules and branching factor as a measure of the decoding complexity.

	*	E	L	R	ELR	M
lex A, # rules	83,983	85,363	84,819	83,985	84,867	84,037
lex A, factor	1.0	0.86	0.84	0.81	0.75	0.74
lex B, # rules	82,223	83,593	83,059	82,225	83,097	82,275
lex B, factor	1.0	0.84	0.81	0.98	0.75	0.74

strained lexicon. This factor gives an indication of the reduction of the search space obtained by using more constrained pronunciation rules.

As an alternative to the MHATLex rules, contextual constraints were introduced using the following heuristic:

> Words whose orthographic form ends with the letters 'r', 's', 't', 'x', 'd' or 'n' (plus exceptions such as *beaucoup* or *franc*) are pronounced with a liaison if and only if followed by a word starting with a vocalic sound; otherwise, only the non-liaison variant can be used.

This heuristic is quite restrictive and does not pretend to be exact or entirely accurate. It is mentioned here mainly to allow the results presented in this study to be related to those reported in Gravier et al. (2001).

Let us illustrate the difference between this heuristic and the *L* resource using the word *très* ('very'). This word has two phonological descriptions: the first one yields the variant /tRe/ in the context of a following consonant; the second one yields two variants, /tRe/ and /tRez/, in the context of a following vocalic sound. The rewrite rules which generated the last two pronunciations are tagged as (l) and (L) respectively. In the *L* resource, /tRe/ is possible if the following word begins with a consonant *or* with a vowel, which amounts to having an unconstrained transcription. However, with the heuristic, the pronunciation /tRe/ is only possible in the context of a following consonant, which is a stronger and probably too restrictive constraint.

4.3. Results

Results are reported for the lexical resources A and B in Table 4. Results were obtained with a single-pass bigram based decoder and with a two-pass trigram based decoder in the case of lexicon B. For the two pass decoder, word graphs are generated using contextual rules with a bigram LM. Those word graphs are then rescored with a trigram LM using the acoustic scores generated during the first pass. In both cases, the first pass

Table 4. Recognition accuracy for various sets of contextual phonological rules for a bigram decoding on the lexical resources A and B, and for a linguistic trigram rescoring on the lexical resources B.

	*	E	L	R	ELR	M
lex A, bigram	57.9	58.1	58.2	57.8	58.4	46.2
lex B, bigram	58.1	58.2	58.3	58.0	58.6	46.1
lex B, trigram	63.3	62.8	63.4	63.2	62.9	48.8

decoding is performed with a narrow beam width allowing a maximum of 10,000 acoustic hypotheses. At each time frame, a maximum of 100 end-of-word hypotheses are considered for insertion into the word graph. Results are given in terms of accuracy, accuracy being defined as the percentage of correct words minus the insertion rate. Given that the insertion rate was almost constant across all the experiments, comparing accuracies is equivalent here to comparing recognition rates.

As can be seen, non-significant marginal improvements were obtained by adding contextual phonological rules in the bigram decoder. In particular, the liaison contextual rule gave the best results over the mute-e and liquid-truncation rules. However, the best result was observed when the three types of contextual rules ('ELR') are simultaneously used in the lexicon, yielding a 58.6% accuracy over 58.1% for context free rules. The slight advantage of contextual phonological rules in the single pass bigram decoder does not remain after trigram rescoring of the word graph. It therefore seems that the context-dependent rules help the beam search focus on the best path. However, alternate paths are also considered and stored in the word graph, which explains why no improvement is observed after trigram rescoring.

Using the whole set of MHATLex phonological rules as contextual rules degrades the results significantly. This may be a consequence of our too coarse definition of a word, which merges homographic MHATLex entries, thus degrading the accuracy of the original rules.

Our previous work uses the heuristic described in 4.2 to approximate contextual constraints for liaison. Table 5 compares results obtained using the heuristic (H) with the MHATLex-derived rules for the liaisons (L). The word-error rate for the heuristic is significantly higher than for both the context-independent rule and the MHATLex-derived liaison rules. This is because our heuristic is too severe and prohibits certain sequences of pronunciations that are nevertheless possible. This heuristic, for instance, enforces the realisation of a liaison /p/ in the word sequence

Table 5. Recognition accuracy with no context and with the MHATLex (L) and heuristic (H) contextual liaison rules.

	*	H	L
lex B, bigram	58.1	53.6	58.3
lex B, trigram	63.3	59.9	63.4

beaucoup à, the alternative pronunciation being entirely ruled out. Allthough a liaison is certainly possible in this context, it is probably not the most common realization. Such irrealistic constraints do not appear in the L lexicon.

5. DISCUSSION

Our extension of the classical beam-search decoding algorithm to accommodate contextual transcription rules resulted in a potential reduction of the search space without degradation of the (first pass) recognition accuracy. Yet, in many aspects, these preliminary experiments have proven unsatisfactory for reasons which relate to the task we have considered, the resources we have used, and the internal properties of our pronunciation model and search algorithm.

First, dictation is clearly not the ideal task for evaluating the benefits of improved pronunciation models, and this might well be one reason why we could not really improve on the baseline performance. We hope to be able to demonstrate more clearly the benefits of contextual rules on more spontaneous speech material in the future.

A second weakness of our experiments concern the resources used: the word-based language model released for the evaluation campaign reported in Dolmazon et al. (1997), and the relatively coarse monophone acoustic models. Though improving the language model is quite a separate issue, it is likely that the use of more elaborate, contextual acoustic models would have significantly improved performance.

In fact, the contextual constraints considered in this study have mostly been defined based on the phonological environement of a word-final phone, and could have been modeled directly using cross-word contextual acoustic models. On the other hand, we have not yet been able to use word classes as a way to introduce lexical and syntactic constraints, such as, for instance, enforcing a noun after the 'liaison' variant of adjectives such as *moyen* ('middle') or *léger* ('light'). A proper description of such constraints requires (i) the differentiation of homographs based

on their main syntactic category and (ii) a LM based on syntactically tagged words. Subject to the availability of tagged corpora for French, we expect to be able to implement these constraints in our future work.

Another area where the lack of corpora for French has proven detrimental is the pronunciation model itself. The pronunciation variants of a given word are not equally likely, and introducing non-uniform variant probabilities has been reported as an effective way to improve recognition accuracy (Wester et al., 1998). However, due to lack of resources from which to estimate those probabilities, we applied a uniform probabilistic model during the search. Again, the picture is slowly changing and we wish to consider more realistic probability distributions in the future.

Finally, these experiments have revealed some inadequacies of our pronunciation model and search algorithm. Regarding pronunciations, a major weakness concerns the definition of contexts based on the previous and following words, when they should be defined in terms of the previous and following pronunciations. To make this obvious, consider the case of a word such as *retour* ('come back'), whose variant /RtuR/ requires on its left a word ending with a vowel. However, for many words, we are unable to tell if they match this context, as both liaisons and mute-e elisions are likely to change the nature of word-final phoneme. Defining contexts based on classes of pronunciations is, in such cases, clearly required.

At the search level, the introduction of contextual constraints was done by introducing a heuristic in the search criterion which no longer guarantees that the best word or pronunciation sequence is found. Therefore, the algorithm needs to be revised in order to improve the decisions it makes at word boundaries. To see why this is necessary, let us recall that at each time t, for any given word w, at most one single end-of-word hypothesis for w is retained through the DP maximization process expressed in (3). As a result, only hypotheses corresponding to the most promising transcription rule a for w will be further developed. Such a decision is however premature, as the right-hand constraints on the pronunciations of w still remain to be applied, and will, in some cases, invalidate this decision. Consider, for instance, the following configuration, where the pronunciation, p, of a word, w, is licensed by two rules a_1 and a_2. During the search, upon reaching the leaf node for p, we have to decide "blindly" which of $a1$ or $a2$ (assuming that both are possible) will be developed, as both yield the same acoustic score. Suppose, for instance, that a_1 is selected: at the end of the next word,

upon matching the right-hand context of w, we might realize that a_1 (but not a_2) violates the constraint, causing a (valid) path to be discarded.

These errors thus directly result from an inconsistency in our search algorithm, which tries to search for the best word sequence instead of searching for the best joint (word, pronunciation) sequence, which is what the proper application of contextual constraints requires. As far as we can see, optimizing this new criterion would have two main consequences. First, it would significantly increase the search space, as a separate tree copy for each transcription rule, rather than for each word, would be needed. In our case, this would be equivalent to multiplying the size of the decoding lexicon by four. Second, it would require us to either reevaluate our pruning strategy or modify our scoring function: using the current resources and scoring functions, it is expected that hypotheses indexed by variants of the same word would have very similar scores, making pruning inefficient.

This is simply because our algorithm acts on pairs $(v, a) \rightarrow (w, b)$ and defines the probability of such a sequence as:

$$P[w, a|v, a'] = P[w|v] \;\; \delta((v, a') \rightarrow (w, a)) \tag{4}$$

where $\delta((v, a') \rightarrow (w, a))$ is a binary function taking the value 1 if the sequence (a, a') is valid, and 0 otherwise: all the valid possible pairs are given exactly the same score. While we consider this model an improvement over unconstrained models, which simply express the probability of a transition between two pairs (word, pronunciation variant) as:

$$P[w, a|v, a'] = P[w|v] \;\; P[a|w], \tag{5}$$

we think that our approximation, given in (4), is also too simplistic, and cannot be used to discriminate efficiently between hypotheses in a first pass decoding. Future work will consequently go toward the definition of more complex modeling of this probability, as for instance in

$$P[w, a|v, a'] = P[w|v] \;\; P[a|w, v, a'] \tag{6}$$

where $P[a|w, v, a']$ is a pronunciation pair model still to be specified.

Another line of work will be to implement the full search for the best joint (word, transcription) sequence in a restricted search space, using for instance contextual constraints to reevaluate a word lattice.

6. CONCLUSIONS

In this chapter, we have presented a method for, and discussed the implications of incorporating contextual-transcription rules into a large-vocabulary speech recognition system. This approach has been tested on a dictation task in French, yielding a potential significant reduction in the search space with no increase of the word error rate. Overall, our initial results are somewhat inconclusive and suggest that this approach could probably be more beneficial if applied to different data, using improved linguistic and acoustic models. These experiments have nonetheless allowed us to identify various improvements to our model, which should eventually result in the integration of more complex and flexible constraints on pronunciation sequences in the search algorithm and thus in a more effective use of pronunciation variants.

ACKNOWLEDGMENTS

We would like to thank the two reviewers of this paper for their careful reading and helpful comments. We would also like to thank G. Pérennou and M. de Calmès for providing access to MHATLex, A. Dauchy for his work on acoustic modeling and F. Antoine for his contribution to the word graph rescoring module.

Part of this work has been done within the framework of the *Sirocco* project, partially funded by the Institut National de Recherche en Informatique et Automatique (INRIA) as a Cooperative Research Action.

NOTES

[1] See `http : //www.enst.fr/~sirocco`.

[2] In this context, and unless otherwise specified, the notion of a word refers to an orthographic string and homographs are therefore treated as a unique word.

[3] Transcriptions are presented using the SAMPA phonetic alphabet (see http://www.phon.ucl.ac.uk/ home/sampa/home.htm).

REFERENCES

Amdal, I., Korkmazskiy, F., and Surendran, A.C. Data-driven pronunciation modeling for non-native speakers using as-so-cia-tion strength between phones. In: *Proceedings of the ISCA Workshop on Automatic Speech Recognition – Challenges for the New Millenium*, 2000: 85–90.

Cremelie, N. and Martens, J.-P. On the use of pronunciation rules for improved word recognition. In: *Proc. Eurospeech*, Madrid, Spain, 3, 1995: 1747–1750.

Dell, F. *Les règles et les sons*. Hermann, Paris, France, 1985.

Deshmukh, N., Ganapathiraju, A., and Picone, J. Hierarchical search for large-vocabulary conversational speech recognition. *IEEE Signal Processing Magazine*, 1999: 84–107.

Dolmazon, J.-M., Bimbot, F., Adda, G., Bèze, M.E., Caërou, J.C., Zeilinger, J., and Adda-Decker, M. Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale. In: *Actes des Journées Scientifiques et Techniques du Réseau Francophone d'Ingéniérie de la Langue de l'AUPELF-UREF*, 1997: 13–18.

Gravier, G., Yvon, F., Jacob, B., and Bimbot, F. Integrating contextual phonological rules in a large vocabulary decoder. In: *Proc. Eurospeech*, Aalborg, Danemark, 4, 2001: 2293–2296.

Jelinek, F. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, 1998.

Jurafsky, D., Ward, W., Zhang, J., Herold, K., Yu.X., and Zhang, S. What kind of pronunciation variation is hard for triphones to model? In: *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, UT, 1, 2001: 577–580.

Kessens, J.M., Wester, M., and Strik, H. Modeling within-word and cross-word pronunciation variation to improve the performance of a Dutch CSR. In: *Proc. Intl. Conf. on Phonetic Science*, San Francisco, CA, 1999: 1665–1668.

Lamel, L.F., Gauvain, J.-L., and Eskenazi, M. BREF, a large vocabulary spoken corpus for French. In: *Proc. Eurospeech*, Genoa, Italy, 1991: 505–508.

Lerond, A. *Dictionnaire de la prononciation*. Larousse, Paris, France, 1980.

Ortmanns, S. and Ney, H. A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech and Language* 11 (1997): 43–72.

Pérennou, G. and Calmès, M.D. MHATLex: Lexical resources for modeling the French pronunciation. In: *2nd International Conference on Language Resources and Evaluation*, 1(2000): 257–264.

Ramabhadran, B., Bahl, L., Souza, P.D., and Padmanabhan, M. Acoustic-only based automatic phonetic baseform generation. In: *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, WA, 1(1998): 309–312.

Ravishankar, M. and Eskenazi, M. Automatic generation of context-dependent pronunciations. In: *Proc. Eurospeech*, Rhodes, Greece, 5(1997): 2467–2470.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., and Zavaliagkos, G. Stochastic pronunciation modeling from hand-labelled phonetic corpora. *Speech Communication* 29(2–4) (1999): 209–224.

Schiel, F., Kipp, A., and Tillman, H.G. Statistical modeling of pronunciation: it's not the model, it's the data. In: *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Rolduc, The Netherlands, 1998: 131–136.

Strik, H. and Cucchiarini, C. Modeling pronunciation variation for ASR: a survey of the literature. *Speech Communication*, 29(2–4) (1999): 225–246.

Strik, H., Kessens, J.M., and Wester, M., editors. *ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Rolduc, The Netherlands, 1998.

Wester, M., Kessens, J.M., and Strik, H. Modeling pronunciation variation for a Dutch CSR: Testing three methods. In: *Proc. Intl. Conf. on Speech and Language Processing (ICSLP)*, Sydney, Australia, 6(1998): 2535–2538.

Yang, Q. and Martens, J.-P. Data-driven lexical modeling of pronunciation variation for ASR. In: *Proc. Intl. Conf. on Speech and Language Processing (ICSLP)*, Beijing, China, 1(2000): 417–420.

STEVEN GREENBERG

FROM HERE TO UTILITY

Melding Phonetic Insight with Speech Technology

ABSTRACT. Technology and science are often perceived as polar extremes with respect to spoken language. Speech applications rarely incorporate scientific insight and conversely, basic research is often viewed as oblivious to practical concerns of the real world. Melding phonetic insight with speech technology can, however, yield extremely productive results for both applications and basic science if performed within the appropriate theoretical framework. Such an approach is illustrated with respect to the relation between prosodic (stress accent) and phonetic properties of conversational telephone dialogues (American English) using the Switchboard corpus. Phonetic properties, such as vocalic identity and duration, are shown to reflect prosodic phenomena, and thus could be used to enhance the quality of automatic speech recognition performance, as well as provide detailed insight into the nature of spoken language.

KEYWORDS. Speech technology, automatic speech recognition, prosody, phonetics, spontaneous speech, syllable structure

1. INTRODUCTION

It is twelfth-century Japan, and a nobleman has been killed. A magistrate is charged with establishing the identity of the killer and delineating the sequence of events leading up to the murder. Several witnesses are called to testify – the victim's wife, the accused (a notorious bandit), a woodsman as well as the victim himself (through a spirit medium). Each witness provides a singular account of the man's death. They agree on but a single fact – that the nobleman is dead. How he died, and by whose hand, are very much in dispute.

The story of *Rashomon* (Ritchie, 1987) is cited often in philosophical discussions of "truth." As nothing is known (or knowable) with certainty, all knowledge is relative (and hence ephemeral). The concept of truth is a chimera and therefore unworthy of pursuit.

Address for Correspondence:
International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704

W. J. Barry and W. A. van Dommelen, The Integration of Phonetic Knowledge in Speech Technology, 107–131.

Yet, there is an alternative interpretation, one that questions not the concept of truth itself, but rather the capacity of its assimilation through a single vantage point. Perhaps the "true" message of *Rashomon* is that deep and ever-lasting knowledge can only be gained through exposure to a variety of perspectives, no single source providing sufficient depth and clarity to comprehend a situation as complex (and as tragic) as the murder of a man.

As in fiction, potentially in science ...

In *Rashomon* the testimony of each witness acquires new significance in light of alternative accounts (Figure 1). Can an intellectual domain as complex as *spoken language* be fully understood through a single perspective? Or must orthogonal forms of evidence be sought with which to reconstruct the "truth"?

Knowledge gained in the pursuit of "pure" research is often viewed as the pinnacle of scientific endeavor, unsullied by practical concerns of technological application and customer satisfaction. Science unfettered by pragmatic constraints is (from this perspective) the most noble of

Figure 1. A woodcutter, a priest and a peasant ponder the unfathomable nature of "truth" in their attempt to reconstruct the events leading up to a nobleman's death in twelfth-century Japan. From the film *Rashomon*, directed by Akira Kurosawa (reprinted from Ritchie, 1987).

objectives and should therefore serve as the principal deity in the temple of knowledge.

As in myth, potentially in science...

How does true insight proceed from "objective" study of spoken language? Is it possible to fully comprehend the multivocal nature of a scientific domain from the exclusive vantage point of a laboratory? Or does the spirit of *Rashomon* compel us to seek testimony from a wider variety of sources in the pursuit of objective knowledge?

2. THE STRUCTURE OF SCIENTIFIC EVOLUTION

The course of a discipline's intellectual evolution is often tortuous and of a curvilinear nature. Where does the domain of speech research lie with respect to its "great chain of being"? Is this community still engaged in determining the number of phonemes *on* a word? Or has the collective unconscious progressed to a higher plane of existence? What will the speech scientists of the *twenty-second* century write concerning the science of the *twenty-first*?

Scientific maturity is often marked by its close relation to technology. The great monuments of any age (whether they be pyramids, cathedrals or casinos) are often based on the most advanced science and technology of the age. And in turn, such monuments usually spur further progress in the domains upon whose foundations they are formed. The synergy between science and technology is simple to discern, for successful products are difficult to build on anything other than a strong and secure scientific foundation. And technology, in turn, provides a rigorous proving ground for the empirical and theoretical precepts of any discipline. Technology may thus serve as a "forcing function," driving a field beyond the bounds of traditional scientific inquiry, posing challenges to surmount by dint of technical (and often commercial) imperative. In tandem with technology comes a focus on empiricism. It is difficult to divine how well a product is likely to work purely on the basis of theory. For theory needs to be tempered with data representative of the environment in which the technology is deployed. In such fashion a field can mature quite quickly; and thus it may ultimately come to pass with respect to speech technology.

3. THE GALAPAGOS OF SPOKEN LANGUAGE

The voyage of the *Beagle* (Darwin, 1839) provided an effective forcing function for Darwin's thoughts on the origin of species (Darwin, 1859), particularly his trip to the Galapagos Islands, west of Ecuador. Among

the fauna of those islands are many varieties of finch, who by virtue of variation in color, size and shape (particularly of the beak) came to provide crucial clues as to the mechanism of natural selection (Weiner, 1994).

Speech, as a field, is still in search of its Galapagos. Somewhere, off the coast of the intellectual mainstream, lie the finches of language – if only we knew their form and function. Should we wait patiently for their emergence? Or should we embark on our own voyage of discovery, aggressively seeking the critical evidence required to solve the mystery of spoken language?

4. UNOBTRUSIVE MEASURES

Every academic discipline has a favored means of collecting data. Astronomers gaze into the heavens, high-energy physicists smash atoms, ethologists play peeping toms, and linguists either introspect or elicit citation forms from "informants."

Long ago, marketing researchers discovered some of the pitfalls associated with elicited data. A shopper, upon entering the supermarket, is asked to enumerate the items intended to be purchased in the store. At checkout a video camera enables a comparison of the shopper's original list with what has actually been bought – intention and deed turn out to bear scant relation to each other; for there is scarcely a product in the shopper's cart mentioned in the interview only a few minutes before (Ries and Ries, 1998).

Because most spoken-language data are derived from either introspection or elicitation the empirical foundations of linguistics are potentially forged on the scientific equivalent of quicksand. From a distance the foundation appears secure, only to collapse in a nebulous undertow upon closer inspection.

5. THE LINCHPIN OF FUTURE TECHNOLOGY

What is an ambitious field to do? Can a discipline reinvent itself with sufficient zeal and celerity as to accommodate the technological and societal transformations of the twenty-first century?

In this circumstance our *Beagle* (and hence salvation), is likely to emerge in the guise of scientific imperatives driven by the frenetic pace of technology. For speech is destined to serve as a technological linchpin of the twenty-first-century economy by virtue of its ability to facilitate and automate communication between humans and machines (cf. Greenberg, 2001). A unique opportunity potentially arises for a synergistic relationship between the science and technology of spoken language.

A solid empirical and theoretical foundation is generally required to develop reliable technology; speech communication is unlikely to be granted an exemption in this regard. Thus, the science of spoken language is likely to evolve quite rapidly over the coming decades as the demand for speech technology accelerates with the emergence of the "communication age."

Sophisticated technology depends on "getting the details right" to a degree that far exceeds what passes for knowledge and insight within the domain of "pure" science (which is why applied technology research is so much more costly than basic research). With respect to speech the contrast between "pure" and "applied" research is stark indeed. Linguists and phoneticians often view spoken language through a "glass menagerie" of abstract forms, which often bear but the faintest resemblance to language spoken in the "real" world. Current speech technology (whether it be in the form of automatic speech recognition or text-to-speech synthesis) relies heavily on training materials representative of the task domain for this very reason (cf. Figure 2). Such a training-intensive approach offers many advantages over a more abstract, rule-governed framework, particularly with respect to performance. But an emphasis on machine-learning algorithms and training regimes often comes at the expense of genuine insight into the nature of spoken language and not infrequently violates the precepts of the hypothetico-deductive method (cf. Greenberg, 1998; Popper, 1959).

Speech technology can proudly point to its *apparent* success with speech recognition and concatenative synthesis in defense of its machine-learning-centric approach. And indeed, imperfect science is capable of providing an effective foundation for technology – as long as the demands of the market place are not exceedingly stringent or profound. However, as commercial expectations rise, immature science is unlikely to suffice as the empirical and theoretical foundation of future-generation technology (Greenberg, 2001).

6. THE SCIENCES OF THE SUPERFICIAL

The academic perspective on language differs markedly from that of the technologist. The linguist is primarily concerned with abstraction and structure of what is normally hidden from view, while the technologist focuses on the more superficial aspects of language (such as the acoustic signal) most amenable to computation (Figure 2); each perspective has its pros and cons.

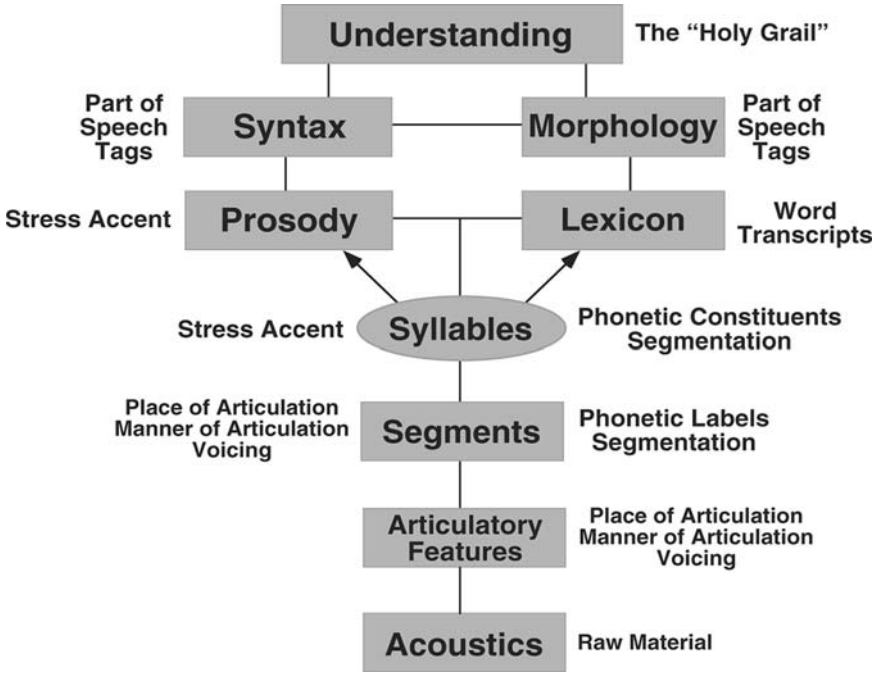

Figure 2. Corpus-centric perspective on spoken language. Manually annotated material forms the basis for statistical characterization of speech, as well as for training systems to perform automatic labeling for speech recognition. Currently, most manual annotation focuses on the lexical level and seeks to derive labels and segmentation for the lower tiers (particularly segments) via automatic methods using some form of Viterbi decoding. The quality of such automatically generated labels and segmentation boundaries is poor when applied to spontaneous corpora such as Switchboard (cf. Greenberg and Chang, 2000). There is precious little manually annotated material associated with non-lexical tiers for any language.

The linguist can use extensive knowledge to make great leaps of intuition that can, on occasion, derive significant insight into spoken language (e.g., Jakobson et al., 1961). But typically such insight is of limited utility to the technologist, saddled with the gory details of daily chatter. Under such circumstances it is unsurprising that speech technology relies mainly on methods designed to automatically divine structure through statistical analysis of surface forms. Does there somewhere lie a path, between the surface and the deep, that provides a plane of mediation between linguistics and technology?

7. INTO THE WILDS (OF SPONTANEOUS SPEECH)

Scholars of medieval Europe sought, in vain, to determine the number of angels residing on the head of a pin (Lovejoy, 1939), their efforts stymied through want of empirical data.

In the realm of spoken language we are more fortunate, for the world literally reeks of material with which to quantify virtually any (superficial) aspect of human discourse; it is merely a matter of recording an appropriate mix of speakers talking in ways representative of the "real" world and then taking the time to annotate the material for statistical characterization (cf. Figure 3).

Two corpora of spoken language are particularly germane to the present discussion. "Switchboard" (Godfrey et al., 1992) has served as a development corpus for evaluation of automatic speech recognition systems for more than a decade. The corpus contains hundreds of brief

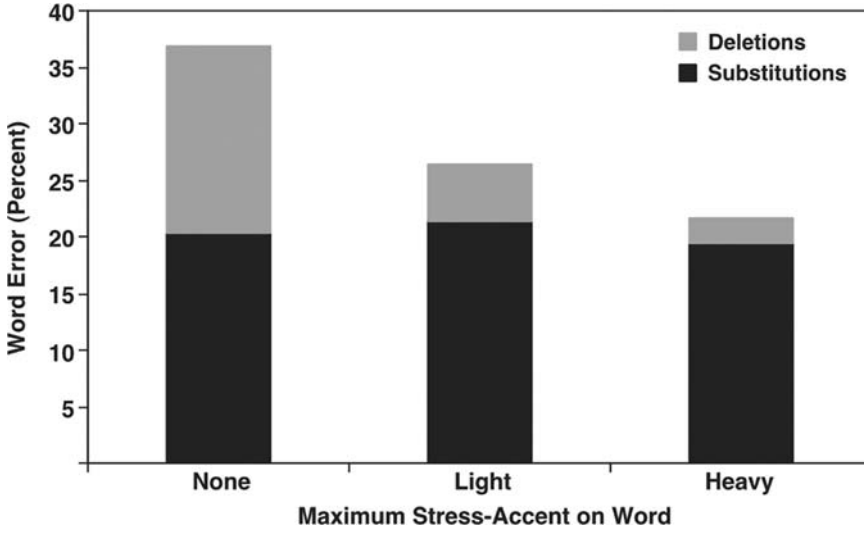

Figure 3. The relation between stress-accent level and word error in the Switchboard corpus for eight separate speech recognition systems (the data have been pooled, given the common pattern exhibited across sites). Word-deletion errors are highly correlated with stress accent level. In contrast, word-substitution errors appear unaffected by stress-accent level. Over 80% of the words are monosyllabic. Three quarters of the remainder consist of just two syllables. In polysyllabic words the maximum stress-accent level pertains to the syllable with the highest degree of accent, irrespective of the stress pattern associated with the other syllables in the word. From Greenberg and Chang (2000).

(5–10 minute) telephone *dialogues* representative of casual conversation, and is thus of great use in characterizing properties of spontaneous (American English) speech. A subset (ca. five hours) of this material has been phonetically annotated by linguistically knowledgeable transcribers at the International Computer Science Institute (Greenberg, 1999) and is electronically accessible over the web (http://www.icsi.berkeley.edu/real/stp).

A one-hour subset of Switchboard has also been manually labeled with respect to stress-accent by two individuals not involved in the phonetic annotation. The remaining four hours has been automatically labeled using an algorithm trained on hand-labeled material (cf. Greenberg et al., 2001).

These same two individuals also labeled two and a half hours of stress-accent material from a separate (phonetically annotated) corpus, "OGI Stories" (Cole et al., 1994), containing hundreds of telephone *monologues* (of ca. 60-seconds each). These two annotated corpora provide (but) one means with which to characterize spoken language (and thereby serve to bridge the gap between linguistics and technology).

8. THE ACOUSTIC BASIS OF STRESS ACCENT

Prosodic accent is an integral component of speech, particularly for languages, such as English, that so heavily depend on it for lexical, syntactic and semantic disambiguation (thereby providing important information concerning the focus of a speaker's attention). Languages mark accent in a variety of ways, utilizing such acoustic properties as duration, amplitude and fundamental frequency (f_o). Some languages, such as Japanese, tend to mark accent primarily in terms of fundamental frequency variation ("*pitch* accent" systems), while others, such as English and German, accentuate using a *constellation* of features (i.e., *stress*) including vocalic duration and identity, as well as fundamental frequency and other acoustic properties associated with the patterning of syllables within an utterance (Beckman, 1986; Clark and Yallop, 1990).

Traditionally, f_o (and its perceptual correlate, pitch) has been thought to serve as a primary cue for stress accent in English (Fry, 1955; Fudge, 1984; Gimson, 1980; Lehiste, 1970):

> "Pitch is widely regarded, at least in English, as the most salient determinant of prominence... when a syllable or word is perceived as 'stressed,' ... it is pitch height or a change in pitch, more than length or loudness that is likely to be mainly responsible..."
>
> (Clark and Yallop, 1990; p. 280)

However, it is unclear whether such statements truly apply to language spoken in the "real" world, free from constraints imposed by scripted or non-meaningful material recorded in the laboratory.

In an effort to resolve this thorny issue the acoustic basis of stress accent was examined as part of a project to incorporate such information into automatic speech recognition systems focused on spontaneous material from the OGI Stories corpus (Silipo and Greenberg, 1999; Silipo and Greenberg, 2000). These studies suggest that duration and amplitude appear to play a far more important role than f_o in accounting for the stress-accent patterns observed in this corpus. Several different automatic methods (based on neural networks, fuzzy logic, and signal-detection theory melded with a threshold model) were developed for simulating the stress-accent patterns labeled in the manual transcription of the prosodic patterns. Each computational method weighted duration and amplitude far more heavily than f_o in order to provide a faithful simulation of the stress-accent annotation (Silipo and Greenberg, 2000), consistent with recent studies examining this issue from the perspective of (American English) telephone voicemail (Koumpis and Renals, 2001) and Dutch spontaneous phone material (van Kuijk and Boves, 1999). Together, such studies suggest that pitch variation plays a much smaller role in the stress-accent pattern of spontaneous speech than has been generally believed (cf. Figure 11 and Table I, as well as Section 12, for additional material germane to this issue); thus caution is warranted in extending the conclusions of laboratory studies on stress-accent to the real world, particularly if technology is viewed as the ultimate arbiter of "truth."

9. STRESS ACCENT AND AUTOMATIC SPEECH RECOGNITION PERFORMANCE

Stress accent is likely to prove of critical importance for future-generation speech recognition systems. Not only does it provide a potential means of identifying key words in an utterance, but such material may also help to enhance recognition performance overall. In a diagnostic study of the linguistic and acoustic factors associated with recognition performance in ASR systems using the Switchboard corpus (telephone dialogues – cf. Godfrey et al., 1992) it was determined that the stress-accent pattern is highly correlated with a specific form of recognition error, namely word deletion (Greenberg and Chang, 2000). If a word contains a primary accent it is far less likely to sustain a deletion error in recognition than if it contains only unaccented syllables (Figure 3). This pattern, observed across all eight recognition systems examined,

Table I. Features used in developing the automatic stress-accent labeling (AutoSAL) system. Delta features refer to the *first* temporal derivative of the spectrum, while double-delta features are associated with the *second* temporal derivative of the same representation. Vocalic energy is normalized in terms of standard-deviation (Z) units relative to the mean. Features listed pertain to those associated with labeling performance shown in Figure 11.

Feature legend
1. Vocalic place (front-central-back) [Voc-Place]
2. Nucleus/Syllable Duration Ratio [N_S-Dur-Ratio]
3. Speaker gender [Gender]
4. Minimum-maximum (dynamic range) of vocalic f_0 [f_0-Range]
5. Mean vocalic f_0 [f_0-Mean]
6. Static/Dynamic Property of Nucleus (Diphthong/Monophthong) [Voc-Dyn]
7. Vocalic height (high-mid-low) [Voc-Height]
8. Average vocalic-segment spectrum [Voc-Spec]
9. Vocalic identity [Voc-ID]
10. Vocalic-segment duration [Voc-Dur]
11. Voc-Spec + delta features [Voc-Spec_D]
12. Normalized energy (of the nucleus relative to the entire utterance) [Z-Energy]
13. Voc-Spec + delta and double-delta features [Voc-Spec_D_DD]
14. f_0-Mean + f_0-Range
15. Voc-Height + Voc-Place
16. Voc-ID + f_0-Range
17. Voc-Dur + f_0-Range
18. Z-Energy + f_0-Range
19. Voc-Dur + Voc-ID
20. Voc-Dur + N_S-Dur-Ratio
21. Voc-Spec_D_DD + f_0-Range
22. Voc-ID + Z-Energy
23. Voc-ID + Voc-Spec_D_DD
24. Voc-Spec_D_DD + Z-Energy
25. Voc-Dur + Z-Energy
26. Voc-Dur + Voc-Spec_D_DD
27. Voc-Height + Voc-Place + Voc-Dyn
28. Voc-Height + Voc-Place + Voc-ID
29. Voc-Height + Voc-Place + Voc-Dur
30. Voc-Height + Voc-Place + Z-Energy
31. Voc-Height + Voc-Place + Voc-Spec_D_DD
32. Voc-Dur + N_S-Dur-Ratio + f_0-Range
33. Voc-Dur + Z-Energy + f_0-Range
34. Voc-Dur + Voc-ID + Z-Energy
35. Voc-Dur + Z-Energy + Voc-Spec_D_DD
36. Voc-Dur + Z-Energy + Voc-Height + Voc-Place
37. Voc-Dur + Z-Energy + Voc-Spec_D_DD + f_0-Range
38. Voc-Dur + Z-Energy + Voc-Spec_D_DD + Gender
39. Voc-Dur + Z-Energy + Voc-Spec_D_DD + Voc-ID
40. Voc-Dur + Z-Energy + Voc-Spec_D_DD + N_S-Dur-Ratio

(Continued)

Table I. (Continued)

Feature legend
41. Voc-Dur + Z-Energy + Voc-Spec_D_DD + Voc-ID + Gender
42. Voc-Dur + Z-Energy + Voc-ID + N_S-Dur-Ratio + f_0-Range
43. Voc-Dur + Z-Energy + Voc-ID + N_S-Dur-Ratio + Gender
44. Voc-Dur + Z-Energy + Voc-Sp_D_DD + Voc-ID + N/S-Dur + Gen + f_0-Mean + f_0-Range
45. Voc-Dur + Z-Energy + Voc-Spec_D_DD + Voc-ID + N_S-Dur-Ratio + Gender

suggests that stress-accent information could be used to improve recognition performance (particularly for large-vocabulary task domains, which generally contain a significant proportion of unaccented words) by utilizing such knowledge to interpret the acoustic signal with respect to phonetic classification and lexical segmentation.

Currently, stress accent is not commonly incorporated into ASR system design. Moreover, there is no general consensus as to the specific form and nature of the prosodic parameter, especially its acoustic correlates. Perhaps there is another property of the speech signal that garners a higher degree of agreement as to its linguistic manifestation and which bears a close affinity to stress accent?

10. SYLLABLE STRUCTURE AND AUTOMATIC SPEECH RECOGNITION PERFORMANCE

Words may be classified in terms of their constituent syllable structure. Most words in English are monosyllabic and their structure is likely to be one of several forms – consonant + vowel + consonant (CVC), consonant + vowel (CV), vowel + consonant (VC) and vowel (V). Together, these syllable types account for ca. 85% of the structural forms found in spontaneous (American) English (cf. Figure 4 and Greenberg, 1999). Consonant clusters occasionally occur at either the syllable onset (e.g., CCVC) or coda (e.g., CVCC), but such forms account for only ca. 15% of the syllable types in spontaneous English (Greenberg, 1999). And a relatively small proportion of words (ca. 19% in the Switchboard corpus) contain more than a single syllable (of this number, approximately three quarters are disyllabic in form).

Of interest, in the current context, is the relation between syllable structure and word-deletion errors for the Switchboard speech recognition systems. Monosyllabic words beginning with a vowel (i.e., V, VC and VCC forms) are far more likely to be mis-recognized in terms

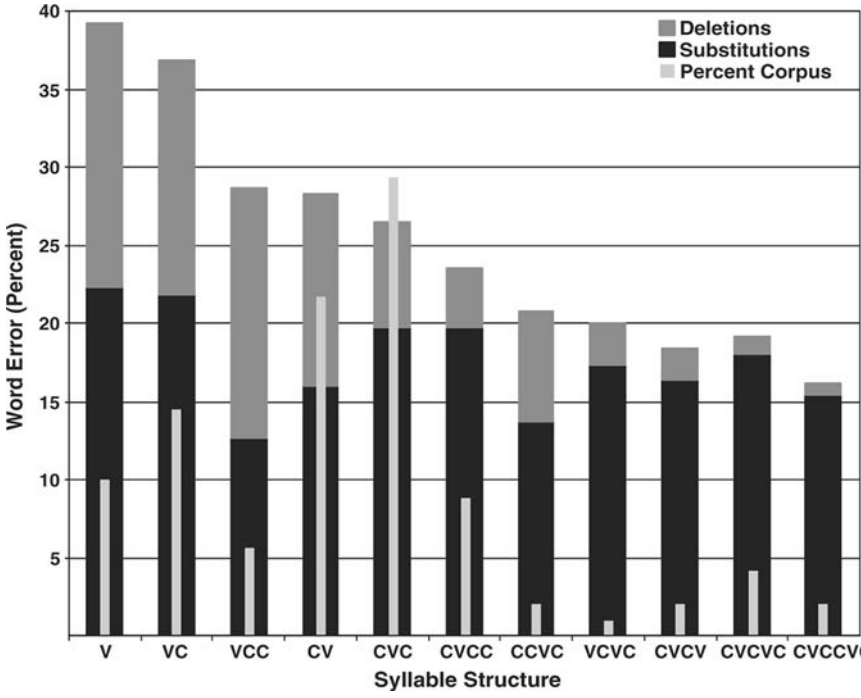

Figure 4. Relationship between word-error rate and syllable structure for Switchboard speech recognition systems. Monosyllabic words beginning with a vowel are far more likely to be mis- recognized in terms of word deletions than words beginning with a consonant or containing two or more syllables. From Greenberg and Chang (2000).

of word deletions than other syllable forms. The governing parameter does not appear to be vocalic-initial lexical forms *per se*, as VCVC words (such as "about") are rarely associated with word-deletion errors (Figure 4). Rather, the word-deletion rate appears linked to the stress-accent pattern associated with each syllabic form. Di-syllabic words usually carry a heavily accented syllable, typically the second when the initial syllable begins with a vowel. Words with consonantal onsets also tend to carry some measure of accent. Thus, syllable structure and accent pattern are in some sense inextricably linked – two sides of the same linguistic coin. Perhaps the philosophy of Rashomon is also relevant to understanding spoken language; the phenomena under study are multifaceted and far too complex to yield their secrets viewed from just a single perspective. And there may be other perspectives (such as vocalic identity) that are equally germane.

11. STRESS-ACCENT AND VOCALIC IDENTITY

In principle, stress accent is independent of vowel quality (with each vocalic segment capable of assuming any degree of stress), and therefore the distribution of accent should be relatively uniform across the vocalic inventory. From this perspective, stress accent is largely a lexical phenomenon, where each word has its distinctive accent pattern (as defined in a pronouncing dictionary) that is only marginally influenced by embedding within the context of spoken discourse. And as there is an arbitrary relation between sound (in this instance, vowels) and symbol (i.e., words) there should be little evidence of a systematic relationship between stress-accent and lexical form.

However, a rather different pattern emerges from analysis of the Switchboard corpus (cf. Figure 5). High vowels (e.g., [ih], [uh]) are far

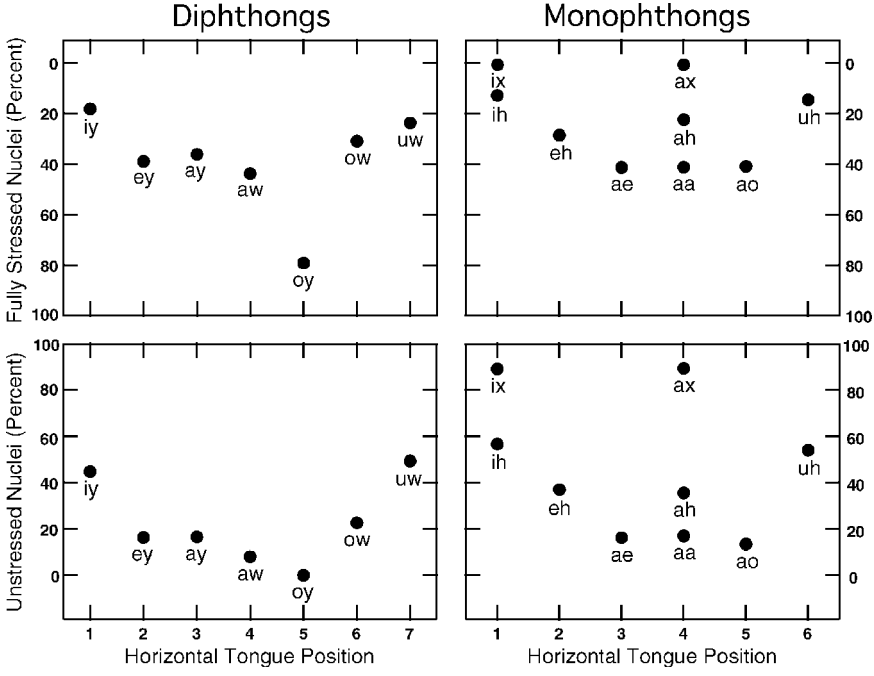

Figure 5. The proportion (in percent) of tokens for each vocalic class labeled as either completely accented (level-1 accent, top panels) or entirely unaccented (level-0 accent, bottom panels), partitioned into two broad classes, diphthongs and monophthongs (for clarity of illustration). Note reversal of scale for the ordinates associated with the top and bottom panels. This scale reversal is required to maintain the spatial relationship between vowel height and proportion of heavily accented (or unaccented) syllables. Adapted from Hitchcock and Greenberg (2001).

more likely to be unstressed than low vowels (e.g., [ae], [aa], [ao]); this relation between vowel height and stress accent extends to diphthongs as well. Thus, [iy] and [uw] are much less frequently accented than [aw] and [ay]. Moreover, the relation between vowel height and stress accent is graded. Mid-height vowels, such as [eh], [ey], [ah] and [ow] exhibit a stress-accent pattern intermediate between their low and high vocalic counterparts (Hitchcock and Greenberg, 2001; Greenberg et al., 2001).

The relation between vocalic identity and stress accent appears to go far deeper than a mere statistical association between parameters. Vocalic duration, for example, is highly correlated with stress accent. Stressed nuclei are often 50% to 100% longer in duration than their unstressed counterparts (cf. Figure 6). In consequence, duration and stress accent are highly correlated in spontaneous discourse (cf. Figure 6). Moreover, there is a close association between duration and vowel height (Figure 7; Hitchcock and Greenberg, 2001; Peterson and Lehiste, 1960) that is likely to be linked to stress accent as well. Duration may hence serve as a

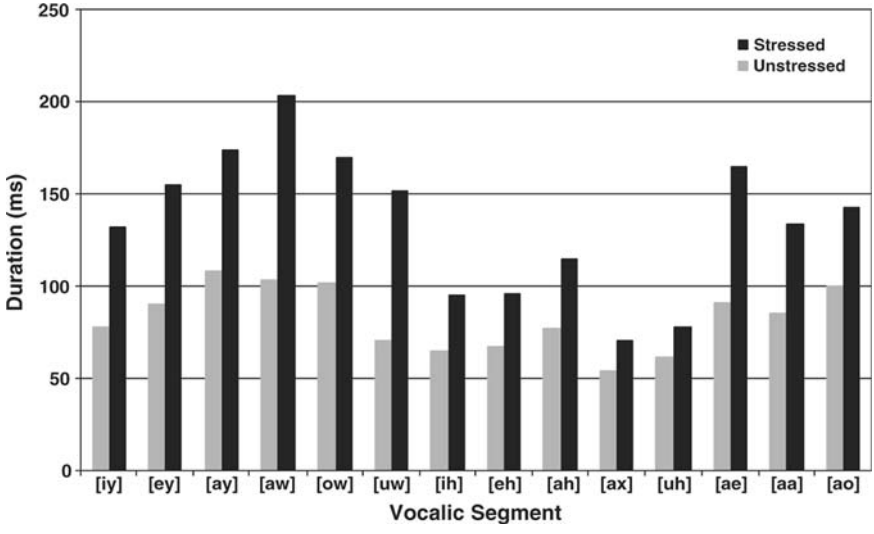

Figure 6. The relationship between segment duration and vocalic identity. Stressed nuclei are consistently longer in duration than their unstressed counterparts. The difference in duration is particularly marked for diphthongs and low monophthongs, and is smallest for the high monophthongs (which are rarely heavily accented). Only segments consistently labeled as fully stressed or entirely unstressed are included in the analysis. Fully stressed [ix] segments were too few to include in the analysis. From Greenberg et al. (2001).

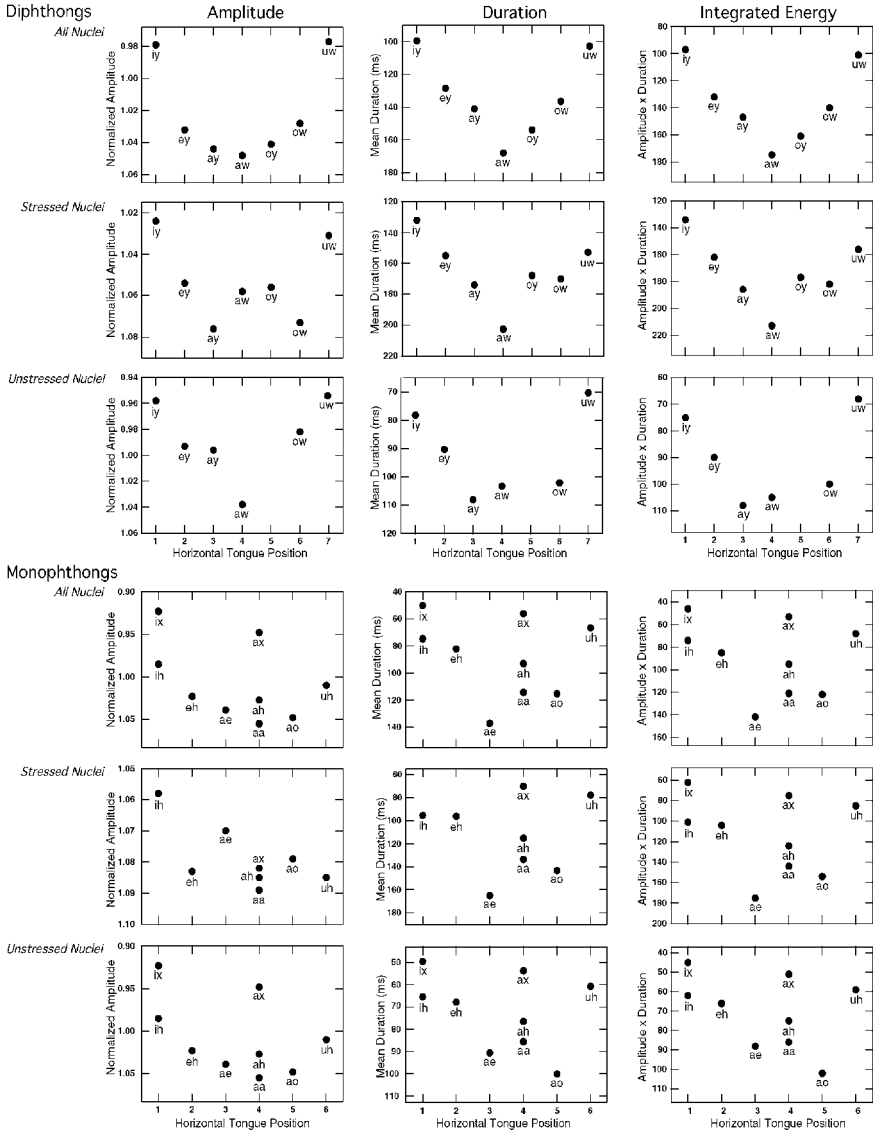

Figure 7. Spatial patterning of the duration, amplitude and integrated energy of vocalic nuclei as a function of stress level (heavy or none), as well as for occurrences averaged across all levels of accent. The data are partitioned into two classes, diphthongs and monophthongs, in order to highlight the patterns. The data points represent averages for each vocalic class. The standard deviations were relatively uniform and are therefore omitted (but are provided in a more extended account – Hitchcock, 2001). The vocalic labels are derived from the Arpabet orthography (cf. Greenberg, 1997 for a description of the phonetic inventory). Horizontal tongue position is schematic in nature and is not intended to denote articulatory measurement (but is *roughly* correlated with the frequency of the second formant). From Hitchcock and Greenberg (2001).

secondary (and under certain circumstances, even as a primary) cue to vowel height.

Vocalic amplitude is also correlated with vowel height (Figure 7), though not *at first glance* to the degree exhibited by duration. Vowel height is directly correlated with the frequency of the first formant; "high" vowels are associated with a low-frequency F_1 (225–350 Hz) while "low" vowels have a F_1 (700–800 Hz). The audibility function for human hearing changes markedly over this range, so that a component at 800 Hz is likely to be as much as 20 dB louder than one at 250 Hz. Thus, the seemingly small disparity in amplitude between high and low vowels may actually be considerably larger when perceptually relevant factors are taken into account.

In some very real sense stress-accent and vowel height may not be entirely distinguishable. Vocalic distinctiveness is, in principle, based on a pattern associated with formants one, two and three (Ladefoged, 1993); yet duration (bound with stress-accent) appears to play an important role as well (cf. Figures 6 and 7), reflected, perhaps, in the pattern of vocalic reduction observed in spontaneous speech (cf. Lindblom, 1990).

The consequence of such patterns is a systematic relation between vowel height and stress-accent pattern. Tongue height associated with vocalic forms in unaccented syllables is *far* more likely to be high than mid or low, for both canonical and non-canonical realizations of syllables and words (Figure 8). The distribution of vowels with respect to tongue height is of a far more even nature for syllables with some degree of stress accent (either light or heavy) relative to those without.

As a consequence of this relation between stress accent and vowel height the overall distribution of unaccented vocalic forms differs dramatically from those associated with heavily accented syllables (Figure 9). The overwhelming majority of vocalic forms in unaccented syllables are in the high-front and high-central regions of the vowel space. The number of low and mid vowels associated with such syllables is rather small. Many (but not all) of the words incorporating such unaccented syllables are "function" words (such as conjunctions, articles, pronouns and demonstratives) which occur with great frequency in conversational speech. Thus, a listener may be "primed" to interpret unaccented syllables as function words under many circumstances (barring evidence to the contrary).

There is a relatively even distribution of vocalic forms associated with fully accented syllables (particularly among the front and low/mid-central vowels). Certain vowels, such as [ao], [oy], [aa] [ae] and [aw], rarely

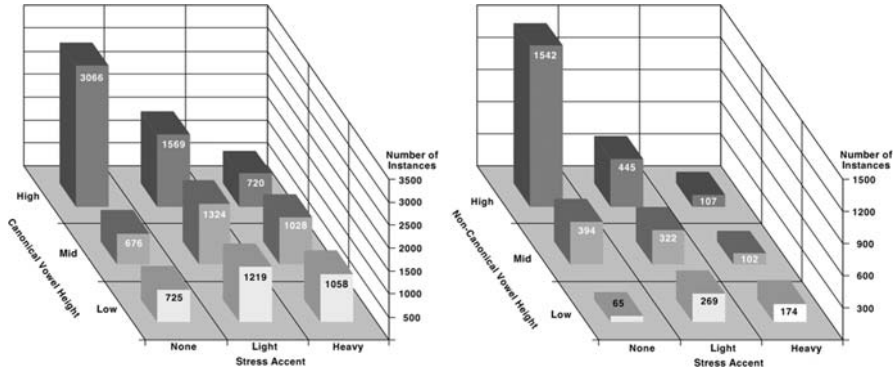

Figure 8. The impact of stress accent on the number of vocalic segments associated with high, mid and low articulatory height (cf. Figure 10 for the relation between segmental *identity* and vowel height), partitioned into canonical (left panel) and non-canonical forms (right panel). Note the difference in scale between the two panels. There is a pronounced skew towards the high vowels for both the canonical and non-canonical forms associated with unaccented syllables. From Greenberg et al. (2002).

occur in unaccented syllables and are typically associated with "content" words (such as nouns and their adjectival complements), particularly those that are relatively uncommon (and hence highly "informative" from a mathematical perspective).

The phonetic realization of vocalic forms is shaped to a certain degree by the (negative) entropy (or "information") associated with the syllables and words in which they are contained. The stress accent pattern can thus be thought of as the surface manifestation of local variation in information associated with the sequence of words and syllables within an utterance.

The intimate relationship between stress accent and vocalic identity in spontaneous discourse suggests that the two may also not be readily distinguishable at some (relatively high) level of linguistic abstraction. Accent may be as integral a component of vocalic identity as tongue height and horizontal tongue position (if not more so). Diphthongs are rarely found in unaccented syllables, regardless of the underlying canonical form, nor are low or back vowels frequently encountered in such contexts. In this sense the absence of accent is accompanied by a constriction of the articulatory space to mostly high-front and high-central vowels. Such a constriction is probably associated with the reduction in duration associated with unaccented syllables and is likely to reflect the "undershoot" phenomenon described by Lindblom (1963) and others (e.g., Öhman, 1966).

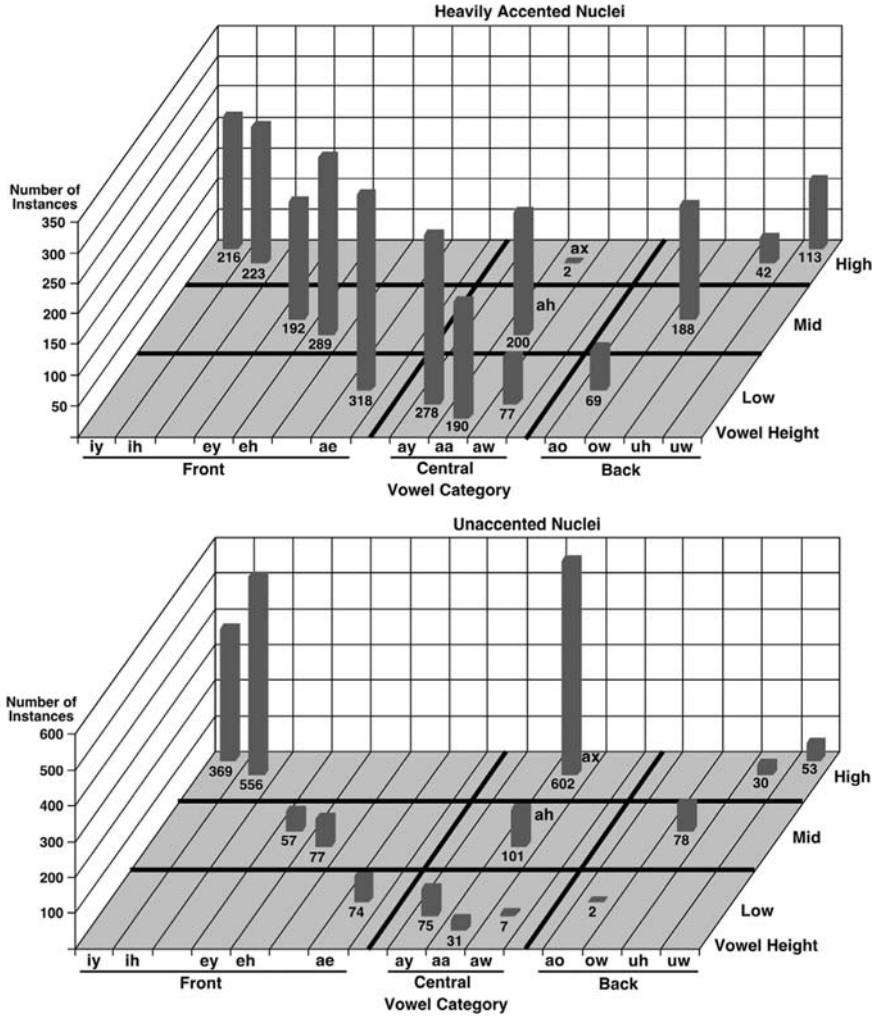

Figure 9. The impact of stress accent (''Heavily Accented'' versus ''Unaccented'') on the number of instances of each vocalic segment type in the corpus. The vowels are partitioned into their articulatory configuration in terms of horizontal tongue position (''Front,'' ''Central'' and ''Back'') as well as tongue height (''High,'' ''Mid'' and ''Low''). Note the concentration of vocalic instances among the ''Front'' vowels associated with ''Heavy'' accent and the association of high-front and high-central vowels with unaccented syllables. The data shown pertain solely to canonical forms realized as such in the corpus. The skew in the distributions would be even greater if non-canonical forms were included (cf. Figure 9 for additional data pertaining to this issue). From Greenberg et al. (2002).

The phonetic forms associated with consonantal segments in both onset and coda constituents of the syllable exhibit a comparable (though quite different) dependence on stress accent (Greenberg et al., 2002). The durational properties of onset (but not coda) consonants are highly sensitive to stress accent – the onsets of heavily accented syllables tend to be 50–60% longer than their unaccented counterparts. And coda constituents are far more likely to be "deleted" (or at least phonetically unrealized) in unaccented syllables than in syllables with some degree of stress accent (relative to their "canonical" pronunciation), particularly for alveolar and liquid segments. Such patterns of pronunciation variation provide yet additional evidence that prosodic factors are extremely important in understanding the phonetic properties of spoken language.

12. THE UTILITY OF PHONETIC INSIGHT

Knowledge of the relation between pronunciation and stress accent may be of utility for automatic speech recognition, particularly under conditions of acoustic interference where the low-frequency portion of the spectrum is degraded. For such knowledge to be of utility in technology applications automatic methods are required to computationally embed the kernel of insight within the confines of a functioning system.

Such an automatic stress accent labeling (AutoSAL) system has been developed for the Switchboard corpus. Multilayer perceptron (MLP) neural networks were trained on 45 minutes of manually labeled material and then applied to an additional four hours of data from the same corpus. The training material contains five distinct levels of stress accent (from entirely unaccented at one end of the spectrum to heavily accented at the other). The degree of machine-human concordance depends on the granularity of the accent labeling. For a very strict metric of concordance (an exact match between human and machine labels) there is precise agreement for 67.5% of the syllables. When the concordance metric is relaxed to a single level of accent disparity the concordance rises to 78%. And when the concordance criterion is further relaxed to 2 accent levels of disparity the agreement between human and machine is nearly 98%. Because the human transcribers were using a *three*-level system to mark accent (i.e., fully accented – 1, completely unaccented – 0, an accent in between the extremes – 0.5), the most realistic concordance metric to assess the reliability of AutoSAL provides for two levels of accent disparity. In this sense, the machine labels are as reliable (and as consistent) as those generated by highly trained human transcribers (Figure 10).

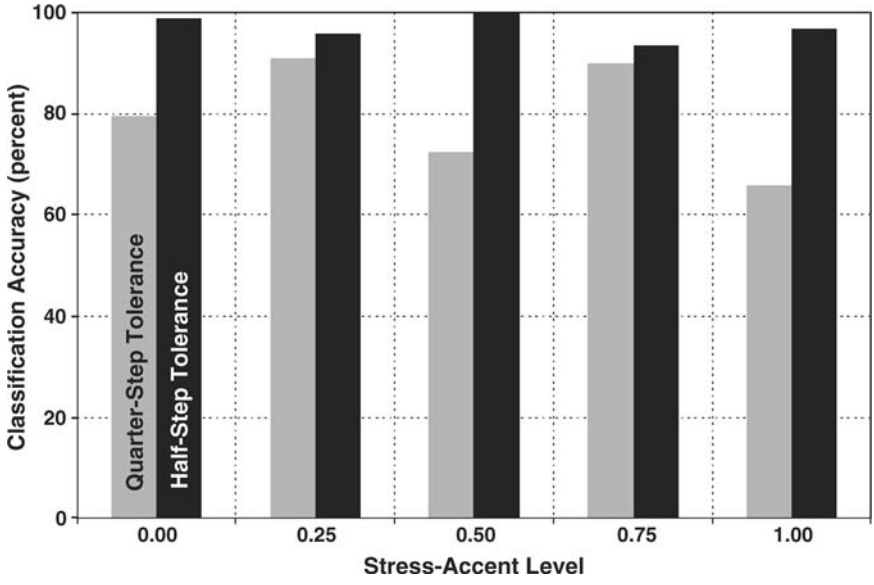

Figure 10. Classification accuracy of the automatic (MLP-based) stress-accent labeling (AutoSAL) system for the Switchboard corpus using two degrees of accent-level tolerance – quarter-step and half-step. The reference accent level is derived from the (average of the) material manually labeled by two transcribers. A syllable is scored as correctly labeled if the ASAL system output is within the designated tolerance limit. Such a metric is required to compensate for the inherent "fuzziness" of stress accent in spontaneous material, particularly for syllables with some measure of accent. For accented syllables there appears to be a gradation in stress; in contrast, unaccented syllables behave as a relatively homogeneous class. From Greenberg et al. (2001).

It is of interest to ascertain the specific acoustic, phonetic and linguistic features required to simulate stress-accent assignment performed by the human transcribers in order to understand the nature of the cues potentially used by human listeners when decoding spoken language. Forty-five distinct feature combinations were used as input to the AutoSAL system in an effort to determine the features mostly closely associated with human-like, stress-accent labeling performance (Figure 11 and Table I). These feature sets were derived from a variety of acoustic, phonetic and linguistic parameters thought to be of relevance to the perception of stress accent (e.g., Fry, 1955; Lehiste, 1970; Lehiste,1996; Silipo and Greenberg, 1999) – duration and amplitude of the syllabic nucleus, the fundamental frequency contour across syllables, as well as parameters believed to be germane to the task through statistical analysis

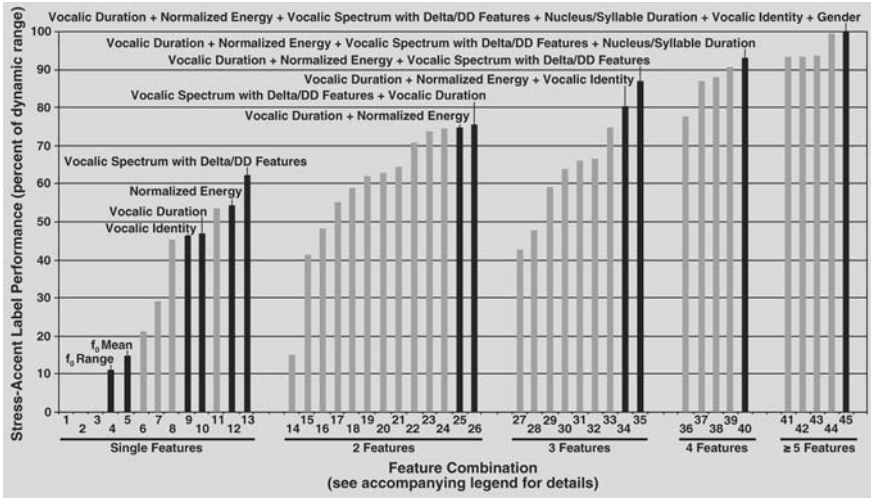

Figure 11. Features used in developing the automatic stress-accent labeling (AutoSAL) system. The final version is based on the features associated with set #45, hereafter defined as the baseline (100 percent performance), achieving performance equivalent to that of a human transcriber. The most poorly performing feature sets are those whose labeling accuracy is close to chance (40%; hereafter 0% of the dynamic range), equivalent to the prior probability of the most common stress-accent label (level-0). The magnitude associated with each feature set is the label accuracy transformed into dynamic-range-normalized units. The best performing feature combination (#45) achieves an accuracy of 67.5% with respect to *five* distinct levels of stress accent, comparable to the *overall* concordance between the two human transcribers. These results are based on a analysis using a tolerance step of 0 (i.e., an *exact* match between human and machine accent labels was required for a "hit" to be scored) and a five-accent-level system. The concordance between machine and human labels is 78% for the five-level system, and is 97.5% for a three-level version of the same system. The feature set is detailed in Table I. Revised version of a figure from Greenberg et al. (2001), in which additional details about the AutoSAL system are described.

of the Switchboard corpus (Hitchcock and Greenberg, 2001), such as the height and forward position of the tongue during vocalic articulation, vocalic identity and the dynamic properties of the nucleus (i.e., whether the segment is a diphthong or monophthong).

The traditional perspective on stress accent ascribes a prominent role to pitch (i.e., fundamental frequency) variation across syllables in a phrase (e.g., Fry, 1955; Fudge, 1984; Gimson, 1980); yet the AutoSAL system does not require such f_0-based features to achieve performance on par with an experienced human transcriber. Of the 45 feature-combination sets tested (Table I), parameters associated with vocalic identity

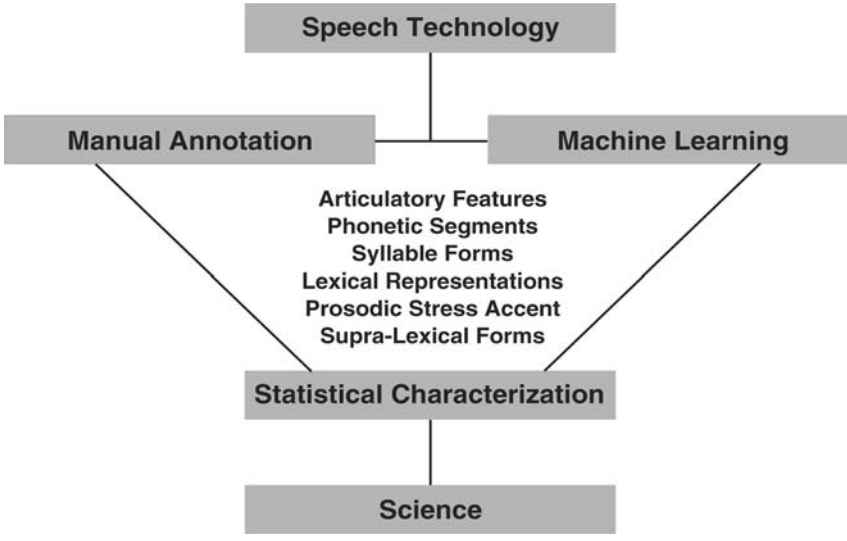

Figure 12. The "eternal pentangle" illustrates the essential tension between science and technology. Although the two poles are often considered exclusive domains, they are in fact complements of each other, providing synergistic relations that further the goals of both. Great technology generally depends on first-rate science and conversely, cutting-edge science often requires superb engineering. Moreover, insights garnered from activity in one pole often help to elucidate problems in the other.

and the attendant spectrum (in terms of the spectral contour over the duration of the segment) are consistently among the most effective cues, along with the duration and normalized energy associated with the syllabic nucleus. Thus, statistical analysis of a spoken-language corpus has proven to be a far better guide for developing classification algorithms of stress accent than perceptual studies using (rather) artificial stimuli. In this fashion speech technology can provide the sort of insight into the nature of spoken language that complements and extends knowledge gained from more traditional sources of scientific experimentation (cf. Figure 12).

13. THE ONCE AND FUTURE KINGDOM OF SPOKEN LANGUAGE RESEARCH

Many aspects of spoken language can be likened to the unicorns of yore – mythical in nature, with their sanctity especially esteemed. These mythical (and languid) creatures are often "sighted," yet ever fail to materialize, the ephemeral pot of gold at the edge of the linguistic

rainbow. Thus, spoken language, as seen through the "eyes" of phonetics and technology, may appear as a chimera, its form and substance in perpetual mutation, and its reification dependent on circumstance rather than on principle.

Scientific insight often stems from necessity, and in such circumstance technological imperatives are likely to serve as an effective catalyst in transforming phonetics (and the rest of linguistics) into a mature field of scientific endeavor. An essential tension exists between science and technology with respect to spoken language. Over the coming decades this tension is likely to dissolve into a collaborative relationship melding linguistic knowledge with machine-learning and statistical methods as a means of developing mature science and technology pertaining to human-machine communication. In the process many mysteries surrounding the form and substance of spoken language are likely to be resolved through the concerted efforts of scientists and engineers focused on the creation of "flawless" speech technology.

ACKNOWLEDGEMENTS

The author wishes to thank Hannah Carvey, Shuangyu Chang, Jeff Good, Leah Hitchcock and Rosaria Silipo for important contributions to the research described. The research was funded by the U.S. Department of Defense and the National Science Foundation.

REFERENCES

Beckman, M. *Stress and Non-Stress Accent*. Dordrecht, Fortis, 1986.

Clark, J. and Yallop, C. *Introduction to Phonology and Phonetics*. Oxford, Blackwell, 1990.

Cole, R., Fanty, M., Noel, M., and Lander, T. Telephone speech corpus development at CSLU, In: *Proceeding of the Third International Conference on Spoken Language Processing* 1994.

Darwin, C. *Voyage of the Beagle*. New York, Collier [reprinted, 1909] 1839.

Darwin, C. *On the Origin of Species*. Cambridge, MA, Harvard University Press (facsimile of the 1st edition, 1964) 1859.

Fry, D. Experiments in the perception of stress. *Language and Speech* 1 (1955): 126–152.

Fudge, E. *English Word-Stress*. London, Allen and Unwin, 1984.

Gimson, A. *An Introduction to the Pronunciation of English* (3rd ed.). London, Edward Arnold, 1980.

Godfrey, J.J., Holliman, E.C., and McDaniel, J. SWITCHBOARD: Telephone speech corpus for research and development. In: *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, 1992: 517–520.

Greenberg, S. The Switchboard Transcription Project. In *Research Report #24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop*

Technical Report Series. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 1997.

Greenberg, S. Recognition in a new key–Towards a science of spoken language. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1998: 1041–1045.

Greenberg, S. Speaking in shorthand–A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29 (1999): 159–176.

Greenberg, S. Whither speech technology?–A twenty-first century perspective. In: *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, 2001: 3–6.

Greenberg, S., Carvey, H., and Hitchcock, L. The relation between stress accent and pronunciation variation in spontaneous American English discourse. In: *Proceedings of the International Conference on Speech Prosody-2002*, 2002.

Greenberg, S. and Chang, S. Linguistic dissection of switchboard-corpus automatic speech recognition systems. In: *Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, 2000: 195–202.

Hitchcock, L. *Acoustic Properties of Vocalic Nuclei Associated with Prosodic Stress Accent in Spontaneous American English Discourse*, Undergraduate Honors Thesis, Department of Linguistics, University of California, Berkeley, 2001. (available from http://www.icsi.berkeley.edu/steveng/prosody).

Hitchcock, L. and Greenberg, S. Vowel height is intimately associated with stress-accent in spontaneous American English discourse. In: *7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, 2001: 79–82.

Jakobson, R., Fant, G., and Halle, M. *Preliminaries to Speech Anlysis: The Distinctive Features and Their Correlates*. Cambridge, MA, MIT Press, 1961.

Koumpis, K. and Renals, S. The role of prosody in a voicemail summarization system. In: *Proceedings of the ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001: 93–98.

Kuijk, D. and van and Boves, L. Acoustic characteristics of lexical prominence in continuous telephone speech. *Speech Communication* 27 (1999): 95–111.

Ladefoged, P. *A Course in Phonetics* (3rd ed.). New York, Harcourt, 1993.

Lehiste, I. *Suprasegmentals*. Cambridge, MA, MIT Press, 1970.

Lehiste, I. Suprasegmental features of speech. In: N. Lass (ed.). *Principles of Experimental Phonetics*, St. Louis, Mosby, 1996: 226–244.

Lindblom, B. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* 35 1963: 1773–1781.

Lindblom, B. Explaining phonetic variation: A sketch of the H and H theory. In: W.J. Hardcastle and A. Marchal (eds.), *Speech Production and Speech Modelling*, Dordrecht, Kluwer, 1990: 403–439.

Lovejoy, A.O. *The Great Chain of Being*. Cambridge, MA, Harvard University Press, 1939.

Öhman, S.E.G. Coarticulation in VCV-utterances: Spectrographic measurements. *Journal of the Acoustical Society of America* 39 1965: 151–168.

Popper, K. *The Logic of Scientific Discovery*. London, Hutchinson. [originally published in German, 1934] 1959.

Ries, A. and Ries, L. *The 22 Immutable Laws of Branding* New York, Harper, 1998.

Ritchie, D. (ed.) *Rashomon*. New Brunswick, NJ, Rutgers University Press, 1987.

Silipo, R. and Greenberg, S. Automatic transcription of prosodic prominence for spontaneous English discourse. In: *Proceedings of the XIVth International Congress of Phonetic Sciences*, 1999: 2351–2354.

Silipo, R. and Greenberg, S. Prosodic stress revisited: Reassessing the role of fundamental frequency. In: *Proceedings of the NIST Speech Transcription Workshop* 2000.

Weiner, J. *The Beak of the Finch*. New York, Knopf, 1994.

MOISÉS PASTOR and FRANCISCO CASACUBERTA

PRONUNCIATION MODELING

Automatic Learning of Finite-state Automata

ABSTRACT. The great variability of word pronunciation in spontaneous speech is one of the reasons for the low performance of the present speech recognition systems. The generation of dictionaries which take this variability into account may increase the robustness of such systems. A word pronunciation is a possible phoneme-like sequence that can appear in a real utterance, and represents a possible acoustic production of the word.

In this paper, word pronunciations are modeled using stochastic finite-state automata. The use of such models allows the application of grammatical inference methods and an easy integration with the other knowledge sources. The training samples are obtained from the alignment between the phoneme-like decoding of each training utterance and the corresponding canonical transcription.

The models proposed in this work were applied in a translation-oriented speech task. The improvements achieved by these new models ranged from 2.7 to 0.6 points depending on the language model used.

KEYWORDS. Automatic learning, pronunciation, spontaneous speech, canonical model, labeling, finite-state automata.

INTRODUCTION

In a speech recognition system, the mapping between the vocabulary words and phoneme-like models are known as pronunciation models. Usual pronunciation models are sequences of phoneme-like units that correspond to the standard pronunciation that can be found in a common dictionary (canonical pronunciation). The speech recognition systems based on such pronunciation models achieve a good performance in a laboratory environment. However, the performance of such systems decreases dramatically in spontaneous environments. This fact seems contradictory to the assumption that a spoken word should not present so large a difference in pronunciation from its canonical representation as to be misunderstood by human listeners. However, the human brain

Address for Correspondence:
Institut Tecnològic d'Informàtica, Universitat Politècnica de València

W. J. Barry and W. A. van Dommelen, The Integration of Phonetic Knowledge in Speech Technology, 133–148.

uses syntax, semantic and pragmatic knowledge to recover from partial information which is present in an utterance. For that reason, words with no semantic information and a high n-gram probability (usually short function words) do not need to have an accurate pronunciation to be understood by listeners with a good syntactic knowledge (Fosler-Lussier, 1999).

Canonical models would be a good model for some words with enough acoustic information (long words). A small variation of a phoneme pronunciation does not represent an important part of the acoustic score of the word. On the other hand, short words (pronouns, articles, etc.) are the most problematic ones. A small deviation from the canonical pronunciation can represent an important variation with respect to the canonical representation. These words are common in human language and usually do not carry important semantic information. They also have a high n-gram probability of occurrence.

There are several approaches to automatic pronunciation modeling. In our opinion, one of the most interesting approaches is the phoneme-based rule-learning technique (Fosler-Lussier et al., 1996; Fosler-Lussier, 1999; JHU Workshop 96 Pronunciation Group). On the one hand, the main problem that arises with these techniques is the over-generalization. On the other hand, one of their greatest advantages is their easy extension to infrequent or unobserved words. However, if a word is infrequent, the effect on the global performance of the system is small. For these words, we use only the canonical pronunciation as a model, as used for long words. Another alternative is using finite-state automata as pronunciation models (De Mori et al., 1995). The transitions of such automata are labeled by phoneme-like units. A path from the initial state to the final state represents a possible pronunciation of the word being modeled. One of the advantages of this type of model is the existence of a number of grammatical inference techniques to learn such models automatically from training pronunciation with different degrees of generalization (Garcia and Vidal, 1990; Oncina and Carrasco, 1994). Another advantage of the models is their easy integration with other knowledge levels and other levels of processing.

1. AUTOMATIC LEARNING OF WORD PRONUNCIATIONS

1.1. Pronunciation Model

In speech recognition we want to obtain the sequence of words, $\widehat{w}$, that maximize the probability of

$$\widehat{w} = \underset{w}{\operatorname{argmax}} \Pr(w|x) \tag{1}$$

x being the acoustic representation of an input sentence. Using Bayes rule, we can decompose this probability $\Pr(w|x)$ as,

$$\Pr(w|x) = \frac{\Pr(w)\Pr(x|w)}{\Pr(x)} \tag{2}$$

where $\Pr(w)$ is the probability of the word sequence, $\Pr(x|w)$ is the probability for this sequence to produce the acoustic observation and $\Pr(x)$ is the acoustic probability. As the factor $\Pr(x)$ does not affect the final result, we can modify the Equation (2) as:

$$\Pr(w|x) = \Pr(w)\Pr(x|w) \tag{3}$$

Usually, the second factor is decomposed into two new factors. In this way, we can decompose Equation (3) as:

$$\Pr(w|x) = \Pr(w)\Pr(s|w)\Pr(x|s) \tag{4}$$

Here, $\Pr(w)$ is the probability of the sequence of words, $\Pr(s|w)$ the probability for this sequence of words to generate the pronunciation, and $\Pr(x|s)$ the probability that this pronunciation produces the acoustic observation.

In practice, $\Pr(x|s)$ is modeled by acoustic models (HMMs) and $\Pr(w)$ by an n-gram model. We can see that the pronunciation model, $\Pr(s|w)$, acts as a mapping between words and subwords units.

The pronunciation model is efficiently specified using stochastic finite-state models.

A stochastic finite-state model (see Figure 1) is a finite-state network whose transitions are labeled by two items: an input symbol (a word of the vocabulary. In the figure: *a, b* or *c*) and a transition probability (between brackets in the figure). In the case of pronunciation models, the arcs are labeled with the transition probability and a sub-word unit. Each path through this automaton represents each possible modeled pronunciation. The probability of each variant is the result of the multiplication of probabilities of the path that generates the pronunciation.

Stochastic finite-state automata have been intensively studied over the years and we have a number of techniques at our disposal for automatically learning them (Jelinek, 1998; Rossmanith and Zeugmann, 2001; Casacuberta, 1990; Oncina and Carrasco, 1994).

1.2. Generation of the Training Pronunciations

The manual transcription of words is a difficult and expensive task. For this reason, it seems interesting to obtain the transcriptions in an

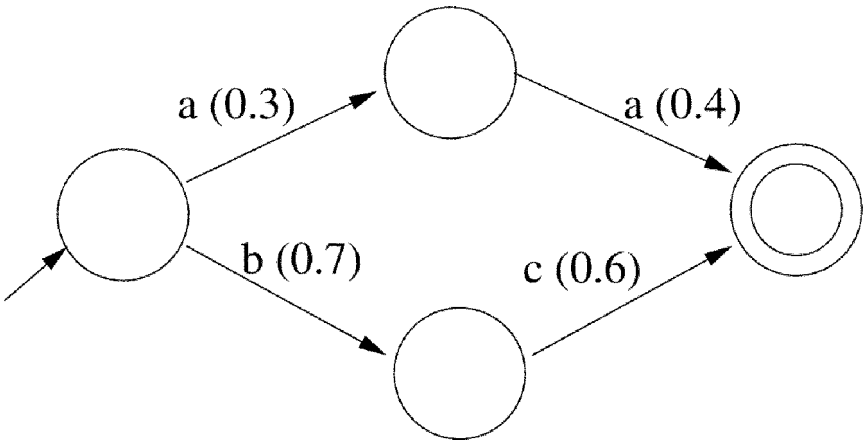

Figure 1. Example of stochastic finite-state automata.

automatic way. Training samples are obtained from the alignment between the phoneme decoding of each training utterance and the corresponding canonical transcription. The alignment between two sequences of phoneme-like units is a subproduct of the computation of the editing distance between both sequences (Hanna et al., 1999). There can be equivalent (same distance) alignments. Our editing distance algorithm gives a better score to those paths which include substitutions and deletions because, from observation, there are more substitutions and deletions than insertions.

For each word of the vocabulary, $w \in \Sigma$, let $\mathrm{Pron}(w)$ be a set of pairs $(P_w^i, n_w^i); 1 \leqslant i \leqslant m(w)$ where P_w^i is the i-th sequence of phoneme-like representation of an acoustic production of word $w, m(w)$ the number of different pronunciations for the word w, and n_w^i the number of times that P_w^i is obtained from the alignments.

Firstly, it is necessary to select the words that will be modeled by their canonical pronunciations and those that will be modeled by grammatical inference. Currently, the criterion for choosing these words is their frequency in the corpus. In this work, we used a more restrictive criterion: we chose those words, w, that have one or more pronunciations, P_w^i which appear(s) at least a given number of times σ:

$$\Theta_\sigma(\Sigma) = \{w | \exists (P_w^i, n_w^i) \in \mathrm{Pron}(w); n_w^i > \sigma\} \tag{5}$$

Table 1. Examples of alternative pronunciations for several Spanish words.

el = {(el,44), (e,18), (1,17), (o1,11), (al,6), (en,5), (r,3), (on,3), (er,2), (ei,2)}
de = {(de,399), (d,41), (do,25), (da,24), (be,21), (e,15), (di,15), (o,5), (le,5), (@,5)}
favor = {(fabor,217), (fabo@,33), (fobor,12), (fbor,12), (fabur,12), (fabo,8), (fabr,6), (faboa,5)}
por = {(por,220), (po,110), (pr,19), (pol,10), (or,7), (pu,3), (pur,2)}
una = {(na,116), (ona,83), (una,56), (ma,15), (rna,13), (ana,13), (@na,11), (gna,8), (lna,6), (bna,6), (ono,4)}
las = {(las,68), (los,20), (nas,11), (das,10), (uas,7), (nos,6), (dos,6), (bos,5)}

Note that the set defined by $\Theta_\sigma(\Sigma)$ is contained in the corresponding set defined by the word frequency. In practice, long words do not appear in $\Theta_\sigma(\Sigma)$.

Some examples of alternative pronunciations for some words are presented in Table 1.

The main problem that arises when obtaining pronunciations automatically from acoustic utterances is its low reliability due to the poor accuracy of phoneme decoders. Some pronunciations are not suitable to be used for training. Pronunciations that are far from the most representative and systematic pronunciations must be discarded. The next step, then, is to choose the pronunciations that are representative for a word. To do that we test with three different criteria:

Number of pronunciations

This is the simplest criterion. We choose a fixed number of the most representative pronunciations for each word:

$$\begin{aligned}\Gamma(w) = &\{(P_w^i, n_w^i) \in \mathrm{Pron}(w);\\ &n_w^1 \geqslant n_w^j \geqslant n_w^{j+1} | w \in \Theta(\Sigma) \wedge 1 \leqslant i \leqslant \delta\}\end{aligned} \tag{6}$$

Accumulative percentage

We have a set of pronunciations of a given word. This set is ordered from the highest to the lowest probability. We collect the productions until we reach a definite threshold:

$$\begin{aligned}\Gamma(w) = \Bigg\{ &(P_w^i, n_w^i) \in \mathrm{Pron}(w);\\ &n_w^1 \geqslant n_w^i \geqslant n_w^{i+1}; 1 \leqslant i \leqslant m | w \in \Theta(\Sigma) \wedge \sum_{i=1}^{m} \frac{n_w^i}{\sum_j n_w^j} = \delta \Bigg\}\end{aligned} \tag{7}$$

Threshold percentage

For a given word, $w \in \Theta_\sigma(\Sigma)$, we take into account those pronunciations which appear at least a percentage δ of the total occurrences for word w:

$$\Gamma(w) = \left\{ (P_w^i, n_w^i) \in \text{Pron}(w) | w \in \Theta(\Sigma) \wedge \frac{n_w^i}{\sum_j n_w^j} \geqslant \delta \right\} \tag{8}$$

1.3. Building Lexical Models

As mentioned above, each $w \notin \Theta_\sigma(\Sigma)$ is represented by its canonical pronunciation. These words either present a low frequency of occurrence or do not have even one systematic pronunciation.

For each $w \in \Theta_\sigma(\Sigma)$, an automaton is built (stochastic k-testable finite-state automaton; Garcia and Vidal, 1990). These models are a particular case of finite-state automata, and their integration into a finite-state language model (as n-gram) is straightforward. The production of lexical acoustic word models is also straightforward, requiring just a substitution of each phoneme transition by the corresponding acoustic model.

Three examples of the models inferred by the method proposed here are presented in Figure 2. In Figure 3, a canonical model for the Spanish word *vale* is also presented. The first three models presented here show a richer structure than the fourth one.

2. EXPERIMENTAL EVALUATION

2.1. The TRAVELER Task Corpus

The general aim was to cover common sentences usually needed by a traveler visiting a foreign country, whose language he/she cannot speak. This framework includes a great variety of different translation scenarios and, thus, it becomes appropriate for progressive experimentation with increasing complexity. In a first phase, the scenario was limited to some human-to-human communication situations at a reception desk of a hotel: asking for rooms, wake-up calls, keys, the bill, moving the luggage, asking for information about rooms (availability, features, price), confirming a previous reservation, signing the registration form, asking and complaining about the bill; notifying about the departure details and other common expressions. For this purpose, a corpus was acquired during the first phase of the EuTrans project.

A small seed corpus was created from several guide books with sentences considered useful for tourists. This corpus was used to help the

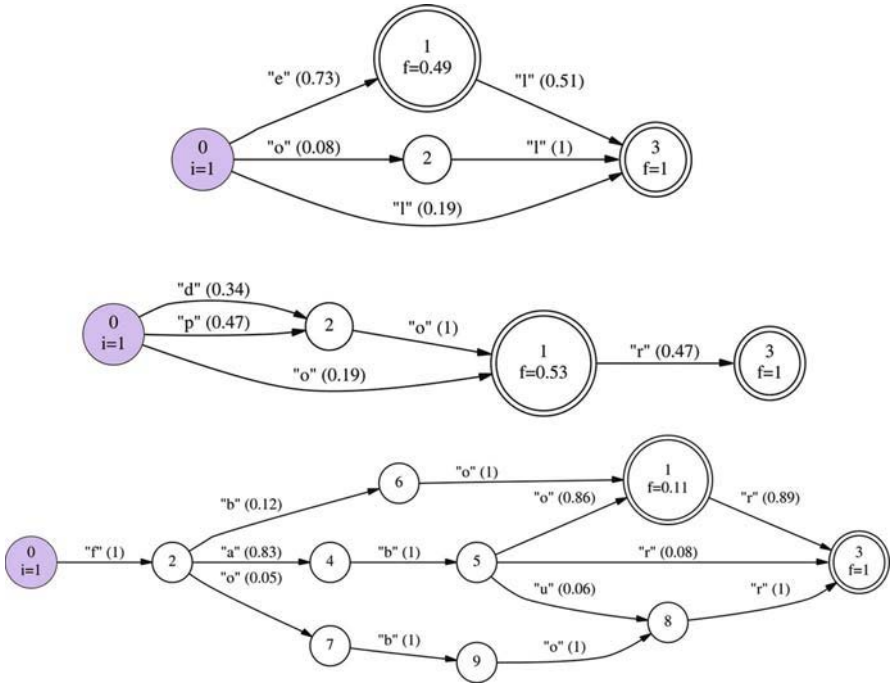

Figure 2. Three stochastic finite-state networks inferred for modeling the allowed pronunciations for the Spanish words ''el'', ''por'' and ''favor'', respectively. The initial state is represented by the darkened circle and the final states are marked by a double circle.

design of the *Traveler Task* corpus, which was automatically built by using a set of Stochastic Syntax-Directed Schemata (Gonzalez and Thomason, 1978) with the help of a data generation tool specially developed for the EuTrans-I project. This software allows the use of several syntactic extensions to these schemata in order to express optional rules, permutations of phrases, concordance (of gender, number and case), etc. The use of automatic corpus generation was convenient due to cost-effectiveness and time constraints in the first phase of the EuTrans-I project. Moreover, this procedure allows control of the level of task complexity.

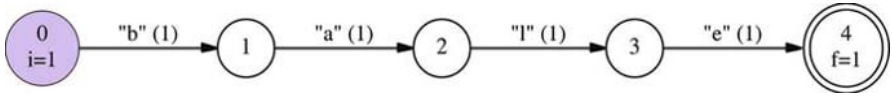

Figure 3. Stochastic finite-state network allowing only the canonical pronunciation for the Spanish word ''vale''.

Table 2. Examples of usual sentences from Traveler Corpus.

Reservé una habitación individual hasta el día seis a nombre del señor y la señora Arnau. *(I booked a single room until the sixth for Mr and Mrs Arnau.)*
Cunto cuesta por día una habitación doble incluyendo desayuno? *(How much does a double room including breakfast cost per day?)*
Hay caja fuerte en las habitaciones, por favor? *(Is there a safe in the rooms, please?)*
Tengo que firmar alguna hoja de registro? *(Should I sign a registration form?)*
Haría el favor de cambiarme a otra habitaci'on con menos ruido? *(Would you mind moving me to a quieter room?)*
Le importaría despertarme a las nueve en punto, por favor? *(Would you mind waking me up at nine o'clock, please?)*
Me puede dar las llaves de la habitación dos veinticinco, por favor? *(Can you give me the keys to room number two two five, please?)*
Lleve las bolsas al taxi. *(Send the bags to the taxi.)*
Nos tenemos que ir el día ocho a las seis y media de la tarde. *(We should leave on the eighth at half past six in the afternoon.)*
Por favor, prepárenos nuestra cuenta de la habitación cero quince. *(Could you prepare our bill for room number oh one five for us, please?)*
Me podría repasar la factura de la habitación nueve nueve cinco? *(Could you check the bill for room number nine nine five for me, please?)*
Me podría pedir un taxi para la habitación cuatro seis siete, por favor? *(Could you ask for a taxi for room number four six seven for me, please?)*

Some example pairs of the *Traveler Task* corpus are shown in Table 2. The acoustic training subset corpus consisted of 1264 sentences uttered by 16 speakers and the test data comprised 336 sentences produced by 4 speakers. The different language models were trained with the transcriptions of the acoustic training subset. The utterances and the speakers used to train the system were different from the ones used for testing. The vocabulary size of the Traveler task was 680 words. Test and training were gender-balanced.

2.2. System Overview

The experiments were performed using the EuTrans telephone speech input translation prototype. This is an engine capable of translating telephone calls from one language to another (Amengual et al., 2000; Pastor, Sanchis, Casacuberta, and Vidal, 2001). This recognition system is based

on ATROS (Automatically Trainable Recognizer Of Speech) engine (Llorens et al., 1999a; Llorens et al., 1999b; Sánchez et al., 1999). ATROS is a continuous speech recognition system which uses stochastic finite-state models at all its levels: acoustic-phonetic, lexical and syntactic. All these models can be obtained in an automatic way (Llorens et al., 1999a). This makes the system easily adaptable to different recognition tasks.

The acoustic front-end operates on 25 ms frames with an interframe distance of 10 ms. A filter bank of 21 trapezoidal filters with increasing widths according to the mel-frequency scale is applied to the 512-point FFT, producing 21 spectrally weighted mean values. A discrete cosine transform is applied to these coefficients producing 10 mel-frequency cepstral coefficients. Energy is also added. First and second derivatives of cepstrum coefficients and energy complete the 33-component frame.

The acoustic models of phoneme-like units were 24 left-to-right continuous-density context-independent HMMs. They were trained with the HTK Toolkit (HTK Book; Young et al.). The probability density functions of HMM states were modeled by Gaussian mixture densities with diagonal covariance matrices, and were estimated with the standard Baum-Welch algorithm. Bigrams and trigrams used in these experiments were trained using the k-testable training algorithm (Garcia and Vidal, 1990; Bordel et al., 1997).

For decoding, the acoustic and pronunciation models are dynamically integrated in the syntactic model: the transitions in the syntactic model automaton are substituted by the corresponding pronunciation model, and each transition on the pronunciation model is substituted by the corresponding acoustic model (see Figure 4). The decoding process is performed using the beam-search Viterbi algorithm (Ney, 1984) through the integrated network.

2.3. Results

All experiments were done using a Pentium II 233 MHz with 64 Mb of memory, running Linux operating system. First of all, we determined the optimal values for σ and δ (see Equations 6, 7 and 8). For this, bigrams, trigrams and the training corpus were used (see Figures from 5 to 10). The best result was obtained using the *Threshold percentage method* for $\sigma = 20$ and $\delta = 0.8$. This method deals with those pronunciations which appear at least a given percentage of the total occurrences for a word.

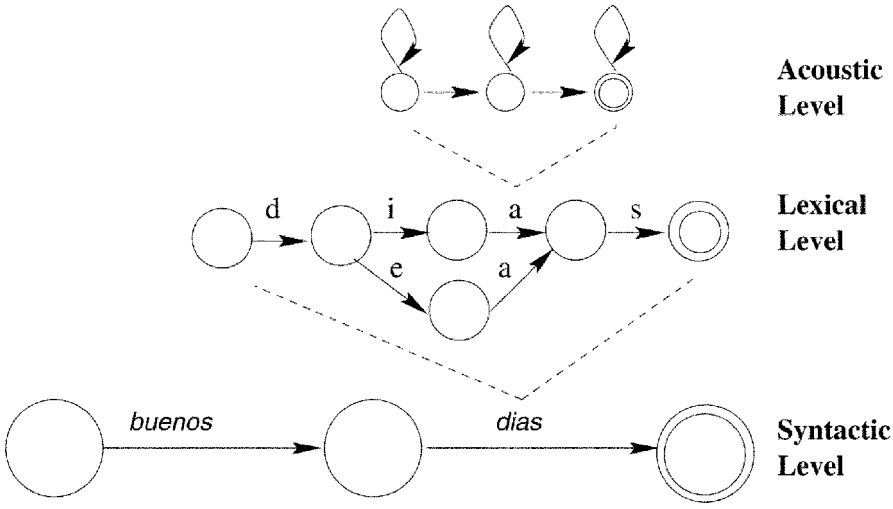

Figure 4. Integrated model.

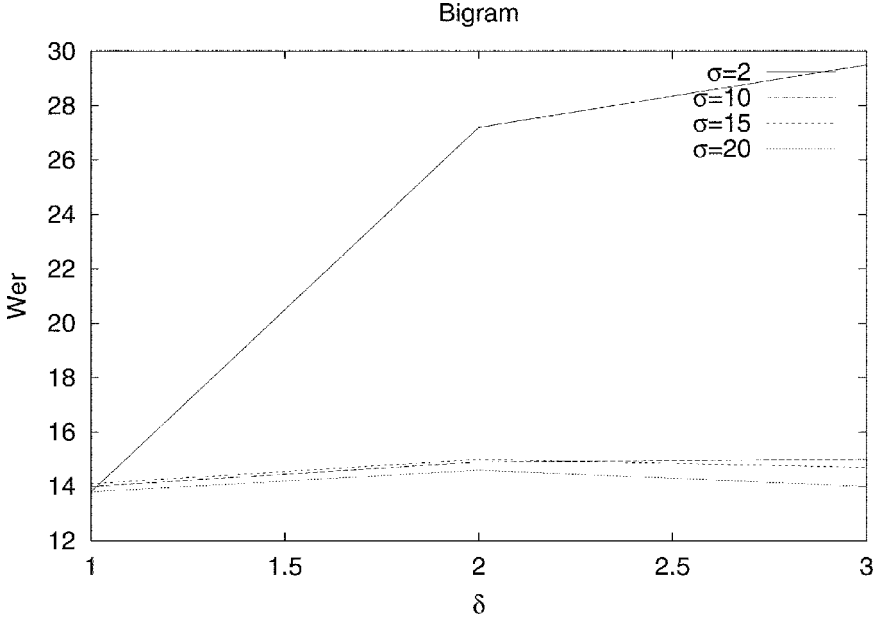

Figure 5. Relationship between the word selection criteria and the selection of representative pronunciations (rejection of noisy pronunciations). Criterion used: Number of Pronunciations. Language model: bigram.

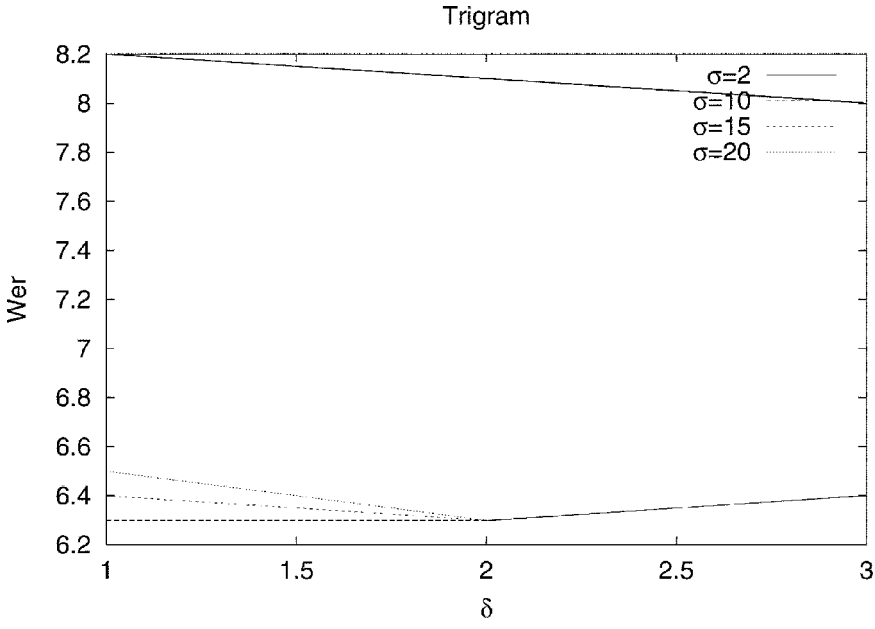

Figure 6. Relationship between the word selection criteria and the selection of representative pronunciations (rejection of noisy pronunciations). Criterion used: Number of Pronunciations. Language model: trigram.

The *Number of pronunciations* criterion (Figures 5 and 6) uses a fixed number, namely the most representative number of pronunciations (see Section 1.2) from the set of pronunciations proposed by the phoneme-like decoder. This obtains interesting results given that this is a very simple and not expensive method. The main *drawback* for this criterion is that it takes productions blindly, and does not evaluate the systematic nature of the pronunciations, and, possibly, leads nowhere.

The *Accumulative percentage* criterion (see Section 1.2) does not reject noisy pronunciations satisfactorily because it does not take into account whether each contribution to the total percentage is systematic enough. This method is intended as a refinement for the *Number of pronunciations* criterion. However, it obtains the worst results (see Figures 7 and 8). It has the same problems as the *Number of pronunciations* criterion and, in the case of unsystematic, noisy pronunciations it accepts more pronunciations.

The *Threshold percentage* criterion could be similar to the principle used to select words for modeling. As we can see in Figures 9 and 10, the second criterion is more restrictive. A marker is the fact that every

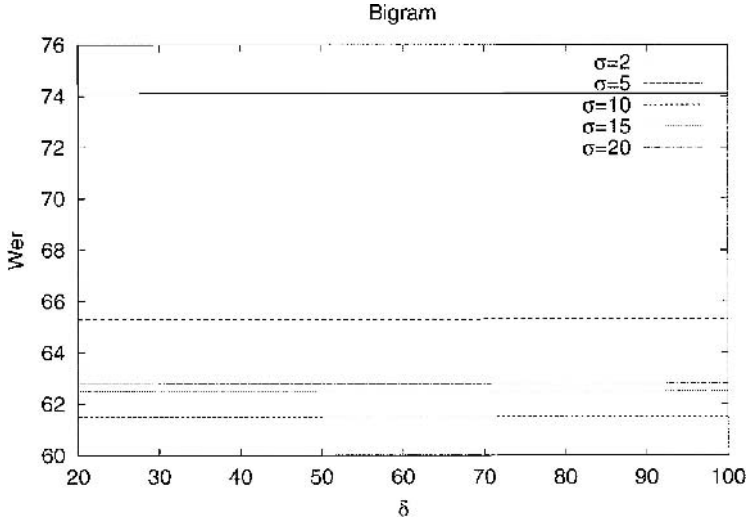

Figure 7. Relationship between the word selection criteria and the selection of representative pronunciations (rejection of noisy pronunciations). Criterion used: Accumulative Percentage. Language model: bigram.

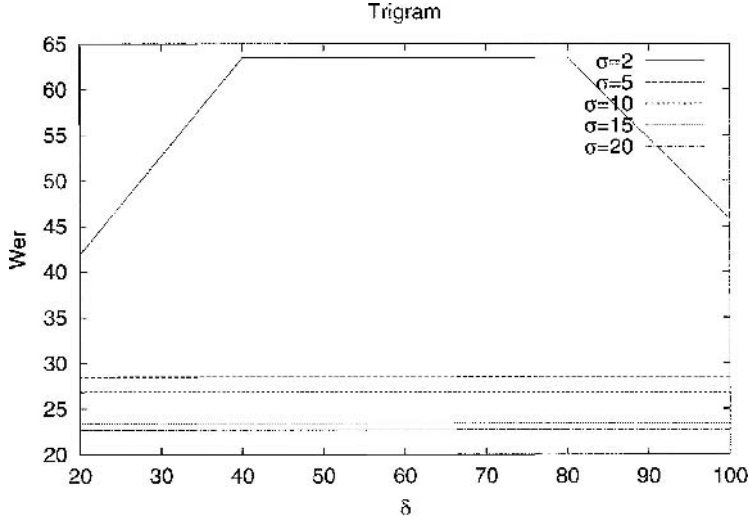

Figure 8. Relationship between the word selection criteria and the selection of representative pronunciations (rejection of noisy pronunciations). Criterion used: Accumulative Percentage. Language model: trigram.

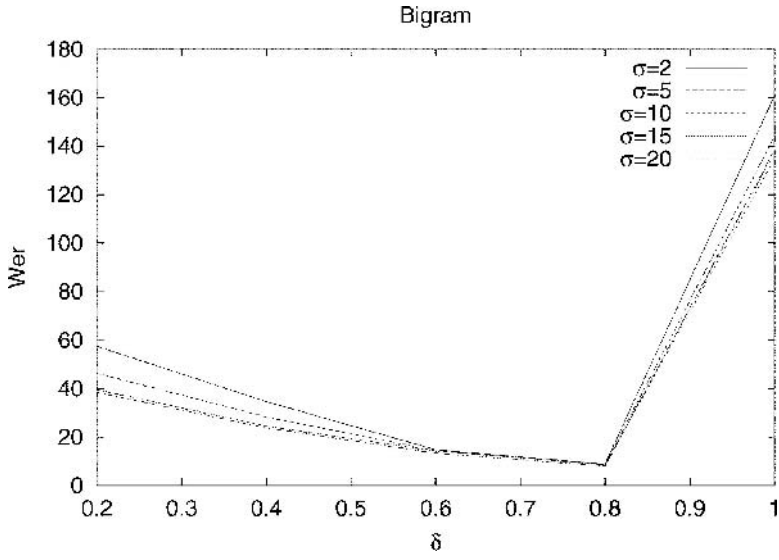

Figure 9. Relationship between the word selection criteria and the selection of representative pronunciations (rejection of noisy pronunciations). Criterion used: Threshold Percentage. Language model: bigram.

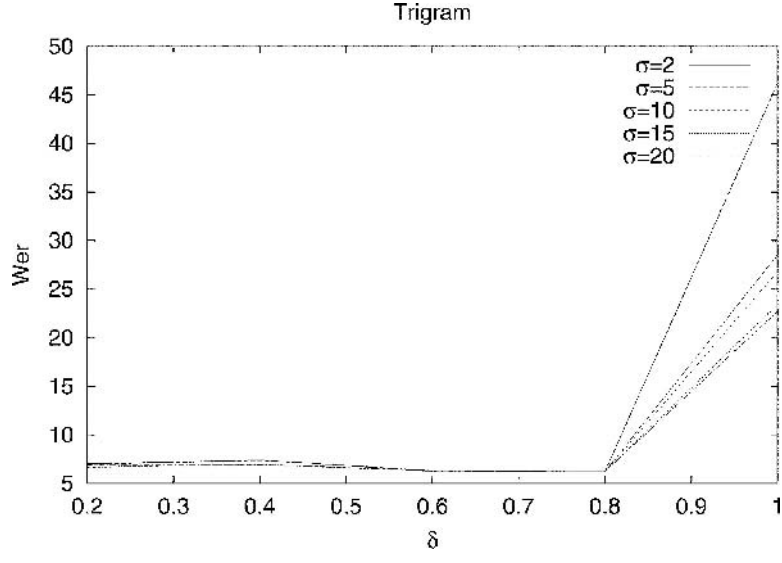

Figure 10. Relationship between the word selection criteria and the selection of representative pronunciations (rejection of noisy pronunciations). Criterion used: Threshold Percentage. Language model: trigram.

Table 3. Word error rate for different language models and different pronunciation models.

LangMod	LexLin	LexAlt	Impr
zerogr	43.25	40.8	5.7
bigram	9.16	6.43	29.7
trigram	3.11	2.55	18.1

Table 4. Real Factor Time for different language models and different pronunciation models.

LangMod	LexLin	LexAlt
zerogr	5.3	5.5
bigram	7.8	9.5
trigram	1.7	1.9

curve converges to the best δ value. If there is not a pronunciation remaining after the application of the *Threshold percentage* criterion, the canonical pronunciation would be taken. This criterion can be seen as a refinement of the word selection process.

We tested the system for different language models and the test corpus. The pronunciation model obtained from the training corpus using the *Threshold percentage* criterion with the best values for σ and δ calculated during the tuning phase. Better results were obtained when the new lexical models were used. The best performance improvement was obtained using bigrams with a 29.7% word error rate reduction (see Table 3). The increment of the real time factor was not relevant due to the small size of increment (see Table 4).

3. CONCLUSIONS

A method for automatically learning pronunciation models from speech data has been tested. This method allows an acoustic-phonetic decoder to propose pronunciations. This defines the training corpus. A more restrictive criterion than frequency of appearance in the corpus was applied. Then, three criteria for rejecting noisy pronunciations were tested. The only one which has proved to be satisfactory is the *Threshold percentage*. At the end, for each word, a stochastic finite-state automaton is automatically trained in order to model every allowed pronunciation.

The method proposed here enabled us to achieve better performances than with conventional canonical models. The greatest improvement in performance was obtained using bigrams.

ACKNOWLEDGEMENTS

This work has been partially funded by the European Union and the Spanish CICYT, under grants IT-LTR-OS-20268 and TIC97-0745-C02, respectively. See the project home page at http://hermes.zeres.de/Eutrans/

REFERENCES

Amengual, J.C., Benedí, J.M., Casacuberta, F., Castaño, A., Castellanos, A., Jiménez, V.M., Llorens, D., Marzal, A., Pastor, M., Prat, F., Vidal, E., and Vilar, J.M. The EuTrans-I Speech Tranlation System. In *Machine Translation*, 15(1–2) (2000): 75–103.

Bordel, G., Varona, A., and Torres, I. K-TLSS(S) Language Models for Speech Recognition. In *Proceedings of ICASSP'97*, 1997: 819–822.

Casacuberta, F. Some Relations Among Stochastic Finite–State Networks Used in Automatic Speech Recognition. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7) (1990): 691–695.

Casacuberta, F., Llorens, D., Martínez, C., Molau, S., Nevado, F., Ney, H., Pastor, M., Picó, D., Sanchis, A., Vidal, E., and Vilar, J.M. Speech-to-Speech Translation Based on Finite-State Transducers. In: *Proceedings of ICASSP'01*, 2001.

De Mori, R., Snow, Ch., and Galler, M. On the Use of Stochastic Inference Networks for Representing Multiple Word Pronunciations. In: *Proceedings of ICASSP'95*, 1995.

Fosler-Lussier, E., Weintraub, M., Wegmann, S., Kao, Y., Khudanpur, S., Galles, C., and Saraclar, M. Automatic Learning of Word Pronunciation from Data. In: *Proceedings of ICSLP'96*, 1996.

Fosler-Lussier, E. *Dynamic Pronunciation Models for Automatic Speech Recognition*. PhD thesis, U.C. Berkeley, 1999.

García, P. and Vidal, E. Inference of k-testables languages in the strict sense and applications to syntactic pattern recognition. In: *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12(9) (1990): 920–925.

Gonzalez, R. and Thomason, M.G. *Syntactic Pattern Recognition: An Introduction*. Addison-Wesley, Reading, Massachusetts, 1978.

Hanna, P., Stewart, D., and Ming, J. The application of an Improved DP Match for Automatic Lexicon Generation. In: *Proc. of EUROSPEECH'99*, 1999: 475–478.

Jelinek, F. Speech Recognition by Statistical Methods. MIT Press, Cambridge, MA, 1998.

Llorens, D., Casacuberta, F., Segarra, E., Sánchez, J.A., Aibar, P., and Castro, M.J. Acoustic and Syntactical Modeling in the Atros System. In: *Proceedings of ICASSP'99* 3 (1999a): 641–644.

Llorens, D., Casacuberta, F., Segarra, E., Sánchez, J.A., and Aibar, P. A Fast Version of the Atros System. In: *Proceedings EUROSPEECH'99*, 1999b: 1299–1302.

Ney, H. The Use of a One–Stage Dynamic Programming Algorithm for Connected Word Recognition. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984: 263–271.

Oncina, J. and Carrasco, R. Inference of Probabilistic Automata. In: *Lecture Notes in Computer Science*. ICGI'94, Springer-Verlag, 1994.

Pastor, M. and Casacuberta, F. Automatic Learning of Finite-State for pronunciation modeling. In: *Proceedings of EUROSPEECH'01*, 2001: 2297–2300.

Pastor, M., Sanchis, A., Casacuberta, F., and Vidal, E. EuTrans: a Speech-to-Speech Translator Prototype. In: *Proceedings of EUROSPEECH'01*, 2001: 2385–2388.

Rossmanith, P. and Zeugmann, T. Stochastic Finite Learning of the Pattern Languages, *Machine Learning* 44(1–2) (2001): 67–91.

Sánchez, J.A., Casacuberta, F., Aibar, P., Llorens, D., and Castro, M.J. Fast phoneme look–ahead in the Atros system. In: *Proceedings of VIII Spanish Symposium of Pattern Recognition and Image Analysis*, 1 (1999): 77–84.

JHU Workshop 96 Pronunciation Group. Automatic Learning of Word Pronunciation from Data. Project Report, April 1997.

Young, S., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. The HTK Book Cambridge University Department and Entropic Research Laboratories Inc.

JAN P. H. van SANTEN

PHONETIC KNOWLEDGE IN TEXT-TO-SPEECH SYNTHESIS

ABSTRACT. This chapter focuses on the value of phonetics for speech technology, specifically text-to-speech synthesis (TTS). After some general remarks about the two fields of research, we describe the linguistic, including phonetic, knowledge incorporated in certain TTS systems, with the goal of showing the diversity of such knowledge. Next, we argue that linguistic knowledge may play an important role in making TTS systems domain-independent. For this, close collaboration between linguists and speech technologists in several types of research is needed. Finally, we give some examples of phonetics research problems whose resolution could be of direct benefit for speech technology.

KEYWORDS. phonetics, speech technology

INTRODUCTION

Phonetics and speech technology focus on the same topic, spoken language. Yet, these fields remain quite separate and little cross-fertilization takes place. Why is that? Does this hurt either field and, if so, what can be done about it? The answers to these questions are fairly obvious. These two scientific communities have developed largely on separate paths as a result of the usual sociological and educational factors that create scientific communities in the first place. As a result of this separation, the phonetics community has not focused on questions most relevant for speech technology while the speech technology community has not developed algorithms and data structures that are optimally receptive for the incorporation of phonetic knowledge. Yet, as will be argued in this chapter, both fields have much to gain from working together more closely.

This chapter focuses on the value of phonetics for speech technology. After some general remarks about the two fields of research, we describe

Address for Correspondence:
Center for Spoken Language Understanding, OGI School of Science & Technology, Oregon Health & Science University

W. J. Barry and W. A. van Dommelen, The Integration of Phonetic Knowledge in Speech Technology, 149–166.

the linguistic, including phonetic, knowledge incorporated in certain text-to-speech synthesis (TTS) systems, with the goal of showing the *diversity* of such knowledge. Next, we argue that linguistic knowledge may play an important role in making TTS systems *domain-independent.* Finally, we give some examples of phonetics research problems whose resolution could be of direct benefit for TTS. Although this chapter is solely concerned with TTS, it is hoped that our remarks are also relevant for other speech technologies.

1. PHONETIC AND SPEECH TECHNOLOGY COMMUNITIES

Phonetics is a science concerned with finding acoustic, articulatory, and perceptual regularities in human speech. Phonetics has a broad reach, ranging from concrete phenomenological descriptions of the sound systems of various languages to abstract theoretical accounts. The key products of phonetics consist of knowledge dissemination and applications for speech and hearing diagnosis and remediation, second language teaching, and several other important contributions; but phoneticians typically do not construct software systems. The science of phonetics does not exist to serve the needs of speech technology, but obviously speech technology has to some degree made use of phonetic knowledge. Phoneticians' use of speech technology, on the other hand, has been largely limited to software tools for speech analysis.

It is important to distinguish between phonetics as currently practiced and phonetics-in-principle; the same distinctions can be made for speech technology. By necessity, of the near-infinitely many questions that would fall within the realm of phonetics, only a subset has been addressed. Given the separation between the two communities, it would be more or less a coincidence if this subset were to be precisely the subset of most relevance for speech technology.

Speech technology is a branch of engineering concerned with creating algorithms for processing or generating speech. Its key products are the dissemination of algorithm descriptions, mathematical results concerning these algorithms, and (typically software) implementations. Speech technology is not necessarily applied science, because its questions have opened up entirely new areas of basic science in statistics and algorithm research. But these areas, if anything, are rather remote from phonetics.

Although both areas are concerned with speech, based on anecdotal evidence we conclude that the communities associated with them are largely separate in terms of education, journals, and conferences. Even if a conference explicitly caters to both communities, usually sessions are

defined that largely attract people from only one community. Vocabularies differ, with the majority of speech technologists not knowing what a coronal consonant is and the majority of phoneticians not knowing what beam search is. As a historical note, it is of interest to realize that before the advent of digital methods speech technology was based on analog electrical circuits, which were explicitly linked to models of the human speech production system (Fant, 1960; Flanagan, 1972). This was naturally accompanied by a closer association between speech technology and phonetics. Once digitization set in, these links became less central in speech technology, and currently one can read speech technology papers that only tangentially refer to the fact that the input signal consists of speech.

In summary, phonetics and speech technology are fields of research that ask different questions about speech, use different vocabularies, pursue different goals, and are conducted by two largely separate scientific communities. However, the reasons for this situation have probably more to do with the very sociological factors that are critical for the creation and cohesion of scientific communities (Kuhn, 1996) than with any fundamental logical or scientific obstacles. Whether either field is ready for the type of ''scientific revolution'' as described by Kuhn remains to be seen, because, like the communities themselves, scientific revolutions are more a socio<!?show –[?tjl=20mm]$9#[?tjl]>logical than a purely intellectual phenomenon. Yet, some such revolution may be necessary for closer collaboration to take place.

2. TEXT-TO-SPEECH SYNTHESIS

2.1. Brief Description

We briefly describe TTS (Allen et al., 1987b; Keller et al., 2001; Sproat, 1997). Most TTS systems involve three stages. In the first (text analysis), symbolic representations are computed from text. These representations usually involve phoneme labels, prosodic tags (e.g., for word stress, sentence accent, or phrase breaks), and optionally also part-of-speech tags and parse trees. Target prosody components compute from these representations timing information (typically in the form of phoneme durations) and intonational information (typically in the form of a fundamental frequency curve). Finally, signal processing components search a speech corpus for appropriate fragments of recorded speech (acoustic units), concatenate these units, and optionally modify them so that the output speech exhibits the pre-computed target prosody.

Several variants on this basic architecture exist. For example, some systems (MITalk, DecTalk, Eloquent) do not use a speech corpus, but instead create speech from scratch by computing acoustic parameter trajectories via rules. These systems are often called *rule based* (which is unfortunate, because systems discussed below also use rules, be it for different TTS components).

In *standard concatenative systems* (e.g., the systems from Lucent, Elan Informatique, SVOX), the acoustic units consist of a set of di- and triphones that have been excised from a larger speech corpus and stored in a small corpus usually in the form of a single data table. These n-phones are unique in the sense that only one token exists of each.

In other systems (usually called "*corpus based systems*") the larger corpus is itself searched at run time, thus providing the system with multiple tokens to select from. In such systems, the target prosody may be used as a *selection criterion* instead of as a specification of how the acoustic units must be *modified* by the signal processing algorithms. It is also possible to build systems where a speech corpus is searched not on the basis of symbolic labels computed by text analysis, but on the basis of the textual input itself; in this type of system, prosody computation is skipped altogether. Recorded voice announcement systems are an extreme version of this type. However, whether a system has unique tokens or not – and hence must perform some type of search – is orthogonal to whether it performs signal modification.

We can put these systems in a larger perspective by observing some trends or dimensions. A first trend is that, as we move from rule based systems to traditional concatenative systems to corpus based systems, we see less emphasis on "knowledge" and more emphasis on data and intensive computation. Second, while the rule based systems and traditional concatenative systems were unabashedly intended to be fully general-purpose and able to handle any input text, corpus based systems are often optimized for a specific subdomain – even though that is not always admitted.

Underlying these trends are two facts. The first fact is very basic, and is simply that the hardware constraints that systems developed in the 80's and before had to contend with were quite forbidding. Hard disks were quite small (< 100 Mbytes for mid-sized computers in the mid-80's) and expensive. Corpus driven systems often require one GByte or more for the speech corpus alone, not including pronunciation dictionaries and other data files. Also, the search algorithms used by some corpus driven systems are quite compute-intensive, and would not have worked on these earlier computers.

2.2. Combinatorics, Generalization from Seen to Unseen Types, and Domain Independence

The second fact goes much deeper, and has to do with the fundamental challenges caused by the *combinatorics of language.* One of the central points made in this chapter is that a key reason for why speech technology should be eager to incorporate linguistic knowledge is that linguistics can play a key role in meeting these challenges.

It is well-known (van Santen, 1997; Moebius, 2001; Baaijen, 2000) that a relatively open textual domain (i.e., a domain that does not have severe restrictions on its vocabulary or grammar) contains an extremely large number of combinations of phone sequences and prosodic contexts. This is so, even if one restricts phone sequences to relatively short lengths (e.g., triphones) and uses a coarse characterization of prosodic context. Not surprisingly, the frequency distribution of these combinations is quite uneven, with some combinations occurring quite often and the overwhelming majority occurring rarely. However, the total probability mass of rare combinations is large enough that even in a small body of text (say, a sentence) something rare is bound to happen: ''Rare things happen frequently.'' To make things more complicated, frequency distributions of these combinations vary across different types of text. This situation appears to be a general property of distributions of almost any linguistic ''unit'', such as, in addition to these combinations, words, trigrams, and sentence structures.

These combinatorial properties of language materials have direct implications for system design and test. If the number of *unit types* is limited, the training materials could in principle cover all these types and all a system would have to do is to handle new unit tokens of already seen types. Even if the training materials do not quite cover all types, as long as training and test materials are sufficiently similar (as is invariably the case in speech recognition), the generalization task is not terribly challenging. However, if the training materials cannot cover all types and the system must work in materials quite different from the training materials, then the system has to face the much more daunting task of facing unseen unit types.

A central claim of this chapter is that *linguistics may provide precisely the type of domain-independent knowledge that is needed to handle unseen types.*

To illustrate this point, consider two approaches to predicting phoneme duration from *linguistic control factors* such as phonemic identity,

word stress, sentence accent, and position in the syllable, word, and phrase. One of these uses simple equations ("sum-of-products models"; van Santen, 1992; van Santen, 1994). For N factors, the formalism is

$$\mathrm{DUR}(\mathbf{f}) = \sum_{i \in T} \prod_{j \in I_i} S_{i,j}(f_j). \tag{1}$$

Here, f_j is a value on the jth factor, and $\mathbf{f} = f_1 \cdots, f_n$; $S_{i,j}$ is a parameter for the ith *product term* for the jth factor; T and I_i are sets of integers (van Santen, 1993); and DUR(f) is the predicted duration for factorial combination **f**. To illustrate, for the multiplicative model: $T = \{1\}$ and $I_1 = \{1, \ldots, n\}$ (a single multiplicative term that involves all factors); for the additive model: $T = \{1, \ldots, n\}$ and $I_i = \{i\}$ (N terms, each containing exactly one factor).

These equations reflect broad generalities, such as

- Holding all else constant:
  - The vowel [i:] is longer than the vowel [e].
  - The same vowel is longer in stressed syllables than in unstressed syllables.
  - The same syllable is longer in phrase-final position than in phrase-medial position.
- Phonemes belonging to the same phonemic class (e.g., vowels, voiceless fricatives) are affected in roughly the same way by these contextual factors.

These sum-of-products models contain few parameters that can be estimated reliably from a relative small training corpus. Even if the training data do not contain an unstressed [ae] in a word-medial syllable in an unaccented word at the end of major phrase, these equations produce a credible predicted duration because the data may contain a [schwa] in precisely this context and the data may contain [ae] and [schwa] in some other shared context allowing the system to estimate the ratio or difference (depending on the equation) of their durations. In other words, based on simple and plausible assumptions, the equations can produce reasonable predictions for unseen types *by extrapolation* from seen types.

Consider, on the other hand, a system that uses Classification and Regression Trees (Riley, 1990; Breiman et al., 1984). The essence of CART is that predictions for unseen cases are not obtained via extrapolation from seen types but *by pooling* them with seen types. Thus, it is

possible that the [ae] may be pooled with the [schwa] in the relevant branch of the tree. This is a problem, because the [ae] is an intrinsically much longer sound than the [schwa]. The equation-based approach handles this by extrapolation: it may predict that if on average in matched contexts the [ae] is 70% longer than the [schwa], than the same will hold in the context where no data were available for the [ae]; thus, if the duration of the [schwa] in that context was 85 ms, then the predicted duration for the [ae] would be 145 ms–probably much more accurate than the 85 ms predicted by CART. This contrast was exactly what was found by Maghbouleh (1996) in a comparison of the two methods.

The key difference between the two approaches here is that CART generalizes via similarity and the equations based method via extrapolation. Of course, there are numerous generalization situations where similarity is an appropriate principle, but extrapolation is clearly the principle that is appropriate for segmental duration prediction.

3. THE DIVERSITY OF PHONETIC KNOWLEDGE AVAILABLE TO SPEECH TECHNOLOGY

3.1. The Bell Labs TTS System

The contribution phonetics makes or could make to speech technology comes in many flavors. Before discussing this in more general terms, we have a look at the situation in TTS. More narrowly even, we analyze the Bell Labs TTS system (Sproat, 1997). The point made in this section is not whether incorporation of such knowledge is helpful, but to investigate what it means for a system to incorporate knowledge by surveying what types of information are incorporated in the Bell Labs system. In this section, we broaden the discussion from phonetic knowledge to linguistic knowledge.

The Bell Labs TTS system is a standard concatenative system as far as its signal processing is concerned. However, its other components are a virtual repository of more than 20 years of research on text analysis and prosody by several dozen scientists. In fact, prior to its complete rewrite as a prelude to commercialization in the mid-90's, its primary goal was more that of being a test bed for research than a commercial TTS engine. This makes the system quite unusual, but also quite useful for our exploration of the *diversity* of knowledge that can be used by speech technology.

3.1.1. *Text Analysis*

Its Text Analysis component computes phonemes and prosodic tags using a mixture of knowledge based algorithms, including dictionary

lookup, parts-of-speech tagging, syllabification rules that are based on the sonority hierarchy, morphological analysis based on general linguistic theories about morphology, heuristic pronunciation rules, statistically trained algorithms for word segmentation (Chinese), homograph disambiguation, accent assignment, and phrase break assignment. In addition, of course, a key role is played by the very symbol sets [phonemes and tags] that are used. One should not underestimate the difficulty of specifying these sets, in particular for lesser studied languages. In addition, quite a bit of knowledge also enters the choice of data features used for the statistically trained methods. For example, Wang and Hirschberg (1992) used in their statistical method a clever mixture of features, including parts of speech, distance from and to punctuation, and lexical items. Selection of these features was based on psycholinguistic research.

Finally, knowledge also enters on a general, "architectural", level. The current version of the system is based on weighted finite state transducers (WFST's). Sproat et al. (1997) have created compilers that generate WFST's from language-specific information having a variety of forms such as rewrite rules, dictionaries, and even CART trees. However, the runtime engine is completely language independent and uses powerful general-purpose algorithms developed for WFST's. This system was created as a direct result of work on a great variety of languages, including Chinese, Japanese, and Russian, that each posed particular problems that were not readily solvable in standard architectures. For example, Chinese text does not have word boundaries, Japanese text uses more than one symbol set, and in Russian "text normalization" (e.g., for pronouncing the "%" sign) requires non-trivial linguistic analysis and hence cannot be handled via a pre-processor, as is commonly done. The point is that the use of WFST's and the compilation tools are based on an understanding of the profound differences that exist between languages, and can be said to reflect linguistic knowledge.

Whether or not one considers this knowledge as belonging to the domain of phonetics or to other areas of linguistics, the examples make the general point that knowledge is used in many ways, some explicit and other–as in the WFST case–implicit.

3.1.2. *Duration*

Of the prosodic components, the component for duration based on sum-of- products models was discussed in Section 2.2. Suffice it to say that these models incorporate knowledge by their very structure (i.e., the fact that these equations naturally reflect regularities such as the invariable

lengthening effect of word stress), by the definition of phoneme/context classes (e.g., separate models are constructed for such classes as unvoiced fricatives in codas, or intervocalic voiced stops), by the predictive factors used or not used (e.g., the factor of location within a foot is not used), and by the ways these factors are defined (e.g., position in a phrase is simply coded as initial, medial, final, instead of in terms of the precise number of words or syllables between the target segment and the phrase boundaries). The decisions about which specific model to use, how to define phoneme/context classes, and how to construct the features are based on phonetic research on segmental duration. In this research, the goal is not that of estimating parameters on a specific training data set, but that of answering general, presumably domain-independent, questions such as: Do these two factors interact (Most don't)? Does it matter what consonant a vowel is preceded by (Not much)? Do word-penultimate syllables behave differently from other word-medial syllables (Italian: Yes; Most other languages: No)? The answers to these questions form yet another type of linguistic knowledge. This knowledge is based on focused speech production studies whose goal it is to answer precisely these questions. It is assumed that these answers are domain- and speaker-independent, and hence can be incorporated in the system. We contrast this with the precise values of parameters, or which exact sum-of-products model fits the best. That information may be highly speaker- or domain-dependent.

As with the use of WFST's in text analysis, the duration system has the key architectural feature of being general enough to have a language-independent run time engine that uses external data tables representing language specific information.

The approach to duration modeling makes a clear distinction between what is based on knowledge (e.g., using sum-of-products models, which factors matter) and what is based on data (parameter estimates, exact structure of sum-of-products model). Knowledge is expected to be domain- and speaker-independent. The function of this knowledge is to define a class of models and specify constraints on these models. As we have seen in Section 2.2, this is critical for the ability of a system to handle unseen cases.

3.1.3. *Intonation*

The intonation component is based on the superpositional model according to which a pitch curve can be written as the sum of simpler component curves: a phrase curve associated with phrases (at times, multiple levels of phrasing are used, each with its own phrase curve), accent curves

associated with pitch accents, and segmental perturbation curves associated with individual phonetic segments. Admittedly, the superpositional concept is controversial (Ladd, 1996), but that is not the point. The point is that this concept is based on knowledge, including hypotheses about quasi-independent processes in the vocal chords (Fujisaki, 1988) and analyses of production data (van Santen and Hirschberg, 1994; van Santen and Möbius, 2000). Further choices in the implementation, such as on which factors parameters depend, are also based on knowledge in the form of special-purpose studies. For example, we found that accent peak location depends on the phoneme class of the coda consonants, not on the individual consonant identities (e.g., [p] vs. [k]) but also not merely on the voicing feature. Finally, also the intonation system has a language independent run time engine, with external data tables representing language specific information.

3.1.4. *Signal Processing*

Finally, although the signal processing component is concatenative and hence is not based on the detailed level of articulatory or acoustic modeling of, for example, MITalk (Allen et al., 1987a), it nevertheless can be said to use knowledge (Olive and Liberman, 1985; Olive, 1990). First, there are language-dependent facts (e.g., one needs some triphones because of strong coarticulatory effects: in American English consonant-vowel-[r]–but not in UK English; in Italian trilled [r] (Shih, 1996); aspirated voiced stops in Hindi; vowel-devoicing in Japanese). Second, there are several details in the signal processing operations that reflect knowledge. For example, the temporal compression/stretching operation stretches out primarily the central portion of vowels instead of the initial and final portions. This is based on studies on vowel lengthening by, e.g., Gay (1968). Third, further details are based on perceptual studies showing that certain distinctions need or do not need to be made (e.g., one cannot hear the difference between [s-t] + [t-o] and [s-p] + [t-o]; in other words, the [s] does not depend on the place of articulation of the subsequent stop.)

3.2. Types of Knowledge Represented in the Bell Labs System

The Bell Labs System incorporates many different types of knowledge:

- Results from phonetics speech production and perception studies that addressed general, domain-independent issues:
  - Which factors are good predictors for pitch accent assignment or phrasing?

  - What are the general properties of the joint effects of contextual factors on duration?
  - How are sounds lengthened?

- Architectural design decisions that are based on an understanding of what languages have in common and along which dimensions they differ:
  - What data structures are needed to handle text analysis for English, Chinese, Japanese, and Russian?
  - Which class of equations can capture directional invariance, assuming that this is a universal feature of spoken language?
  - Which class of equations can handle pitch movement at multiple time scales?

- Language dependent details:
  - Phone list.
  - Phonotactics.
  - Coarticulatory patterns.
  - Lengthening of word-penultimate syllables in Italian.
  - Pronunciation dictionaries.

- Parameterized mathematical models based on domain-independent regularities:
  - Sum-of-product models.

- Parameterized mathematical models based on physiological studies of speech production:
  - Intonation model.

It should be emphasized, however, that almost all components of this system are at least partially data driven. In this sense, it is fundamentally different from the earlier MITalk system (Allen et al., 1987a) in which every component consisted of manually constructed rules and manually adjusted parameter values.

4. VALUE OF INCORPORATING LINGUISTIC KNOWLEDGE

The previous section showed that many different types of knowledge can be incorporated in TTS systems. However, this does not prove that this knowledge provides any value. Does it?

Speech recognition has for years been the envy of TTS researchers, not only because of better funding but also because it seems that the process of creating a speech recognition system for a new language or application seems easier and more sophisticated than the process of creating a new TTS system. Essentially, this effort is perceived to consist of the following steps:

1. Decide on phoneme set.
2. Obtain pronunciation dictionary.
3. Collect and optionally tag textual data for language modeling.
4. Collect and optionally transcribe speech data for acoustic model training.
5. Train language and acoustic models.

Is something similar conceivable for TTS? Consider a *minimal knowledge* TTS *system*, which would be a system that contains just two subcomponents: Text analysis, mapping text onto a symbolic representation using some inductive learning engine, and a speech generation component that searches a tagged speech corpus. TTS construction would consist of similar steps:

1. Decide on internal symbolic representations, including phoneme label set and prosodic tags.
2. Obtain pronunciation dictionary.
3. Collect, transcribe, and tag textual data for training the text analysis subsystem.
4. Collect, label, and optionally tag speech data to form the acoustic inventory.
5. Train inductive learning system on the tagged text.

At least for open-domains, the odds seem to be against this scenario as a realistic goal. First, currently no minimal knowledge system exists whose open-domain performance is better than that of the best commercial systems. For example, even using the most sophisticated learning engines (Sproat, 2000), error rates in text- to-phone conversion are in excess of 10% (19.7% word error rate for Sproat), which is well above the 5% or better WER of the commercial systems.

Second, the amount of training data needed to learn how to pronounce the "%" sign in Russian may be prohibitive, even if it is in principle possible to learn this automatically.

Third, for commercial-grade TTS systems, the issue is how to most effectively reach 100% of Acceptable Performance Levels (APL, presumably well

below perfection). If one uses an inductive learning method to quickly reach 80% APL but then still needs months of manual labor to attain 100% APL, little is gained compared to a process that is completely manual but requires only 1 month to reach 100% APL. This puts a *user interface constraint* on inductive learning methods: They must be amenable to manual intervention if they cannot reach 100% APL. Current inductive learning engines generate data structures that are not suitable for manual fine tuning.

We propose that the most realistic goal is one that takes optimal advantage both of domain-independent information that can be provided by linguistics and of continued progress in inductive learning methods, data collection, and automated data annotation methods. To make this possible, we need TTS architectures constructed of components that can absorb and represent linguistic knowledge, and at the same time are maximally trainable. The Bell Labs system provides several examples of such components. Yet, it is also clear that still too many components of the Bell Labs system contain manually provided information, including some of the rules tables used by text analysis.

What research is required to reach this goal? First, analysis is needed on what the necessary sub-functions are of text-to-speech conversion, either in general or for a certain domain class. For example, do we really need part-of-speech tagging, parsing, or target pitch contour generation? This analysis has both a performance aspect (i.e., what do these sub-functions contribute to overall system performance?) and a more theoretical, linguistic aspect (what facts about spoken language necessitate this or that analysis?).

Second, for those sub-functions that are deemed necessary, we need to analyze how they can be designed to optimally incorporate linguistic knowledge and be trainable at the same time.

Third, where relevant, we need to design user interfaces for tools that allow linguistic experts to interactively enter or optimize data tables, e.g., for language- specific details.

Fourth, and perhaps most important, we need to specify for each sub-function what facts about spoken language we need to know (the next section will discuss some of these in detail).

It is critical to realize that each of these four types of research involves close collaboration between linguists and speech technologists.

5. SOME PHONETIC QUESTIONS THAT SPEECH TECHNOLOGY NEEDS TO BE ANSWERED

As one builds a speech technology system or creates tools used for its construction, it often happens that a decision has to be made that is

not backed up by known facts. For example, methods for automatic selection of acoustic units typically use a distance measure applied to cepstral parameters. There are no studies showing that this is a good idea. Many design decisions are of this nature, and more will become visible as linguists and speech technologists collaborate on the four research areas listed in the previous section.

Here is a brief list of research questions that need an answer–an answer that requires phonetic research:

Perception of spectral discontinuities of the type that occur in concatenative synthesis. Automatic unit selection methods, whether at run time or off-line, need a measure that successfully predicts speech quality resulting from concatenating specific units. Several attempts have been made to predict quality using spectral distance measures between the start and end frames of to-be-concatenated units, but with remarkably little success (Klabbers and Veldhuis, 1998; Wouters and Macon, 1998). Spectral discontinuities form an acoustic signal whose perception is poorly understood. It may involve complex auditory frequency/time interactions that are not properly reflected by these simple measures.

Perception of discontinuities in intonation contours. Natural F_0 contours are far from continuous: They are interrupted by voiceless sounds, creaks, and many other effects. Yet, we are able to detect discontinuities rather well when they are generated by TTS. What makes these discontinuities detectable? What signal processing modifications can render them non-detectable?

Understanding sub-segmental timing in speech production. TTS methods that modify the temporal structure of acoustic units usually stress or compress these units uniformly (an exception being the signal processing used by the Bell Labs system). Is this audibly sub-optimal? If so, what non-uniform methods should be used?

Mimicking vowel reduction. Concatenative TTS systems, when faced with the need to generate reduced vowels, have usually two options: Shorten a non-reduced vowel, or substituting a [schwa]. However, it seems plausible that a reduction continuum exists, in which case one need run-time modification algorithms (Wouters and Macon, 2002a; Wouters and Macon, 2002b).

Understanding variability in speech production–both inter- and intra-speaker variability. When a speaker is instructed to repeat a sentence without any changes, some variation occurs nevertheless. Mimicking such variation may be important to reduce perceived monotony of synthetic speech. Is this variation locally random, e.g. can it be mimicked

by randomly perturbing the F_0 contour and individual phoneme durations, or does this variation involve pitch and duration changes with a longer time scale? If so, how can we mathematically describe this?

Acoustic invariances of perceptually equivalent pitch contours. Listeners are sensitive to certain pitch modifications, but not to others that, by any simple physical measure, are larger (d'Imperio and House, 1997; Kohler, 1990). A better understanding of what constitutes a *just noticeable difference* for pitch contours is critical for pitch contour generation.

Multidimensional modeling of all acoustic prosodic features – F_0, local acceleration, spectral balance, loudness, etc. Prosodic cues tend to co-occur. For example, phrase boundaries may involve the simultaneous lowering of F_0, deceleration of spectral movement (e.g., formant movement), and various correlates of decreasing sub-glottal pressure (e.g., spectral balance and loudness.) In addition, speakers differ in terms of which cue they most prominently use. Capturing these phenomena requires a multi-dimensional approach in terms of both measurement, modeling, and prosody generation.

How to measure the impact of emotional speech on listeners. The ability to generate emotional speech is important. However, an important shortcoming of much research on emotional speech is the way the perception is measured. Typically, we ask the listener which of several emotions was portrayed. While high levels of correct recognition are certainly a necessary condition for the speech having an emotional *impact* on the listener, they are not sufficient. After all, when we synthesize in two modes, one with completely flat F_0 and the other with randomly agitated F_0, high depressed vs. angry recognition scores will be obtained, but this does not mean that the listener experienced the voice as truly having these emotions. We need more clever approaches, either indirect methods or physiological measures.

Phonology of intonation. Currently, the ToBI system is almost universally used to describe intonation at the phonological level. Over the years, many papers critical of this approach have been written, yet no general alternative has been proposed. Is ToBI ready for replacement? If so, with what?

6. CONCLUSIONS

Speech technology uses more linguistic knowledge and concepts, and a greater diversity, than is generally realized. We believe that speech technology could benefit substantially by incorporating even more knowledge. Some of this knowledge is already available, but there is a large list of

phonetics research questions that are currently not addressed by the phonetics community. At the same time, the incorporation of linguistic knowledge and concepts requires speech technology systems to have receptive architectures. Neither this research nor these architectures will materialize unless organizational and educational bridges are built between these two fields of research.

The recommendation is obvious and has been made by others (e.g., Moore 1995): we need to enhance growth of a "bridge field" in the form of *mathematical* or *computational phonetics.* In practice, this means (i) infusing phonetics education with more mathematics and computer science, and speech technology education with more phonetics; (ii) offering either special sub-tracks in these fields or even joint degrees in phonetics, electrical engineering, and computer science; (iii) organizing conferences with truly joint sessions instead of, as is currently the rule in both Eurospeech and the International Conference on Spoken Language Processing, joint conferences with separate sessions.

ACKNOWLEDGEMENT

This research has been supported by NSF Grant 0082718.

REFERENCES

Allen, J., Hunnicut, S., and Klatt, D. *From text to speech: The MITalk System.* Cambridge (UK): Cambridge University Press, 1987a.

Allen, J., Hunnicutt, S., and Klatt, D. *From text to speech: The MITalk system.* Cambridge: Cambridge University Press, 1987b.

Baaijen, D. *Word frequency distributions.* Dordrecht (The Netherlands): Kluwer, 2000.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. *Classification and regression trees.* Monterey (CA): Wadsworth & Brooks, 1984.

d'Imperio, M. and House, D. Perception of questions and statements in Neapolitan Italian. In: *Proceedings of the Fifth European Conference on Speech Communication and Technology*, Rhodes, 1997.

Fant, G. *Acoustic theory of speech production.* Mouton, The Hague, 1960.

Flanagan, J.L. *Speech analysis, synthesis and perception*, Volume 3 of *Kommunikation und Kybernetik in Einzeldarstellungen.* Springer, Berlin. 2. erw. Aufl.; 1. Aufl. 1965, 1972.

Fujisaki, H. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In: Fujimura (ed.), *Vocal physiology: voice production, mechanisms and functions.* New York: Raven, 1988.

Gay, T. Effect of speaking rate on diphthong formant movements. *Journal of the Acoustical Society of America* 44 (1968): 1570–1573.

Keller, E., Bailly, G., Monaghan, A., and Terken, J. *Improvements in Speech Synthesis.* John Wiley & Sons, 2001.

Klabbers, E. and Veldhuis, R. On the reduction of concatenation artifacts in diphone synthesis. In: *Proceedings ICSLP*, Sydney, Australia, 1998: 1983–1986

Kohler, K. Macro and micro F0 in the synthesis of intonation. In: J. Kingston and M. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech.* Cambridge: Cambridge University Press, 1990: 115–138.

Kuhn, T. *The Structure of Scientific Revolutions, 3rd Edition.* University of Chicago Press, 1996.

Ladd, D. *Intonational phonology.* Cambridge (UK): Cambridge University Press, 1996.

Maghbouleh, A. An empirical comparison of automatic decision tree and hand-configured linear models for vowel durations. In: *Proceedings of the Second Meeting of the ACL Special Interest Group in Computational Phonology.* Association for Computational Linguistics, 1996.

Moebius, B. Rare events and closed domains: Two delicate concepts in speech synthesis. In: *Fourth ESCAWorkshop on speech synthesis*, Scotland: Pitlochry, 2001: 41–46.

Moore, R. Computational phonetics. In: *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, 2 (1995): 68–71.

Olive, J. A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds. In: *Workshop on speech synthesis.* Autrans France: ESCA, 1990: 25–30.

Olive, J. and Liberman, M. Text to speech–an overview. *Journal of the Acoustic Society of America,* Suppl. 1 78(Fall) (1985): s6.

Riley, M. Tree–based modeling for speech synthesis. In: *Workshop on speech synthesis.* Autrans France: ESCA, 1990: 229–232.

Shih, C. Synthesis of trill. In: *Proceedings of the International Conference on Spoken Language Processing.* Philadelphia: ICSLP, 1996: 2223–2226.

Sproat, R. (ed.), *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.* Boston (MA): Kluwer, 1997.

Sproat, R. Corpus-based methods and hand-built methods. In: *Proceedings ICSLP*, Beijing, China, 2000.

Sproat, R., Möbius, B., Maeda, K., and Tzoukermann, E. Multilingual text analysis. In: R. Sproat (ed.), *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Chapter 3, Boston (MA): Kluwer, 1997: 31–87.

van Santen, J. Contextual effects on vowel duration. *Speech Communication* 11 (1992): 513–546.

van Santen, J. Analyzing N-way tables with sums-of-products models. *Journal of Mathematical Psychology*, 37(3) (1993): 327–371.

van Santen, J. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8 (1994): 95–128.

van Santen, J. Combinatorial issues in text-to-speech synthesis. In: *Proceedings Eurospeech–97*, Rhodos, Greece, 1997: 2511–2514.

van Santen, J. and Hirschberg, J. Segmental effects on timing and height of pitch contours. In: *Proceedings ICSLP '94*, 1994: 719–722.

van Santen, J. and Möbius, B. A quantitative model of F_o generation and alignment. In: A. Botinis (ed.), *Intonation: Analysis, Modelling and Technology.* Dordrecht: Kluwer, 2000: 269–288.

Wang, M. and Hirschberg, J. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6 (1992): 175–196.

Wouters, J. and Macon, M. Perceptual evaluation of distance measures for concatenative synthesis. In: *Proceedings ICSLP*, Sydney, Australia, 1998: 2747–2750.
Wouters, J. and Macon, M. Effects of prosodic factors on spectral dynamics. I. Analysis. *Journal of the Acoustical Society of America* 111(1) (2002a): 417–427.
Wouters, J. and Macon, M. Effects of prosodic factors on spectral dynamics. II. Synthesis. *Journal of the Acoustical Society of America* 111(1) (2002b): 428–438.

HELMER STRIK

IS PHONETIC KNOWLEDGE OF ANY USE FOR SPEECH TECHNOLOGY?

ABSTRACT. Although it has often been advocated that more phonetic knowledge should be incorporated in speech technology, the amount of phonetic knowledge used in speech technology has decreased over the years. In order to get a better understanding of why this is the case, some examples of attempts to transfer phonetic knowledge to speech technology are presented. These examples make clear that there are several reasons why this transfer is problematic: different approaches are used in the fields of phonetics and speech technology, phonetic knowledge is based on small amounts of 'lab speech' and therefore does not generalize to 'real speech', the knowledge is not complete, and the knowledge is not quantified in the right format.

KEYWORDS. Phonetic knowledge, speech technology, ASR, TTS

1. INTRODUCTION

''Is phonetic knowledge any use?'' was the title of the panel discussion that took place at Eurospeech 2001 on Friday September 7, 2001 in Aalborg. This panel discussion was the second part of the Eurospeech special event entitled ''Integration of Phonetic Knowledge in Speech Technology''. In this paper we will take a look at the integration of phonetic knowledge in speech technology. We start with some notes on the two terms: 'phonetic knowledge' and 'speech technology'.

Speech technology is a term that covers many fields, like speech coding, speech-to-speech translation, text-to-speech (TTS), concept-to-speech, speaker identification, speaker verification, speaker tracking, automatic speech recognition (ASR), speech understanding, etc. Of all these fields, only ASR and TTS are addressed in the current paper, while the main focus is on ASR.

Giving a short and clear definition of phonetic knowledge is not straightforward. In fact, what exactly constitutes phonetic knowledge

Address for Correspondence:
A^2 RT, Dept. of Language and Speech, University of Nijmegen, The Netherlands

W. J. Barry and W. A. van Dommelen, The Integration of Phonetic Knowledge in Speech Technology, 167–180.

has been the topic of many discussions. In the context of the current article, we will not attempt to establish what the exact nature of phonetic knowledge is (e.g. to specify what exactly is phonetic and what is phonological knowledge). We believe that more important questions concern the role of phonetic/linguistic knowledge in speech technology: to what extent is it used, should this increase or diminish, etc. Consequently, in the current paper, the focus is on phonetic knowledge in a broad sense, which sometimes may even mean more general linguistic knowledge.

It has often been advocated that more phonetic knowledge should be used in ASR and TTS (Stevens, 1960; Zue, 1983; Zue, 1991; Pols, 1999). However, in many ASR and TTS systems it is not straightforward how phonetic knowledge should be integrated into these systems. One way of doing this is by using articulatory(-based) features, which has been tried in various research projects. In general, the goal of integrating phonetic knowledge into ASR and TTS systems was to increase the performance of these systems. However, ASR and TTS have also been regarded as means to test existing phonetic knowledge, to find out whether gaps and/or errors in the existing phonetic knowledge were present, and where. Furthermore, it has been suggested that ASR and TTS should be (partly) integrated, because human speech production and perception are not independent (Stevens, 1960).

Although many seem to be in favor of integrating (more) phonetic knowledge in speech technology, in the last decades we have witnessed a decrease in the amount of phonetic knowledge used in ASR and TTS (e.g. Zue, 1983; Zue, 1991). However, it is certainly not the case that the use of phonetic/linguistic knowledge has been abandoned in current systems. For instance, ASR and TTS systems make use of the knowledge that speech consists of words, that these words do not occur in a random order, that these words are made up of syllables and phonemes, that these phonemes do not occur in a random order, and much more knowledge on speech production, acoustics, and perception. More specifically, when we develop ASR and TTS systems, we often make use of e.g. the phoneme inventory of a language, a lexicon, grapheme-to-phoneme conversion, phonetic transcriptions, segmentations, and phonetic features, which are often derived using knowledge about speech perception.

In Section 2 we will argue that phonetics and speech technology are essentially two different worlds, which hinders the transfer of phonetic knowledge to speech technology. Some examples of (not) using phonetic knowledge in speech technology are given in Section 3. First, a description is given of three examples of attempts to use phonetic knowledge in ASR

which were pursued in research carried out at our department. They are presented in chronological order. Two other examples are discussed at the end of Section 3. Finally, the discussion is presented in Section 4.

2. PHONETICS AND SPEECH TECHNOLOGY: TWO DIFFERENT WORLDS

Why has the amount of phonetic knowledge used in speech technology decreased over the years? An obvious answer would be: Because systems in which less phonetic knowledge is used perform better. For many people (researchers, developers, retailers and users) this is indeed the most important aspect of a system: it should perform well. Therefore, if a system that uses less phonetic knowledge performs better than one using more phonetic knowledge, the former is preferred. However, this answer does not provide any insight into why the transfer of phonetic knowledge to speech technology is so difficult.

Part of the answer is certainly related to the fact that phonetics and speech technology are essentially two different worlds. This should not be underestimated. At the universities of most countries, research and education in phonetics and speech technology are conducted by different people in different faculties, i.e. those of linguistics and engineering. Consequently, many differences exist between these two groups of researchers: they study different theories, acquire different practical skills, and use different jargons. To a large extent they even have different frames of reference, carry out experiments differently, etc.

These differences between the two worlds are certainly a problem, and hinder the transfer of knowledge to some extent. Interestingly, the situation in The Netherlands is quite different from that of most other countries. In some Dutch universities, research and education in phonetics and speech technology take place in the same faculty, i.e. the faculty of Arts. However, although the gap between the two worlds should thus be smaller in The Netherlands, the role of phonetic knowledge in speech technology is not noticeably larger than in other countries. So there must be other reasons that hinder the transfer of knowledge.

These reasons might be found in the different approaches used in phonetics and speech technology. Let us take a closer look at those differences. In order to make the differences clearer, a somewhat caricatured overview is presented of a classic phonetic versus a speech technology approach (see Table 1). Although in most cases the differences will not be so extreme, this comparison is useful to get an idea of the what hampers transfer of phonetic knowledge to speech technology.

Table 1. The classic phonetic vs. a speech technology approach.

Approach	Classic phonetic	Speech technology
condition	controlled	less controlled
setting	studio, lab	many places
sound quality	high	varied: noise, etc.
speech style	formal	informal, spontaneous
articulation	careful	varied: hypo- to hyperart.
database	small, balanced	large, less balanced
subjects	few	many
processing	manual	automatic
analysis	deterministic	statistical
features	formants, LPC, etc.	cepstra, (rasta-)PLP, etc.
approach	linguistic	information-theoretical
goal	knowledge, theory	applications

Table 1 is based on a table from a presentation I gave in Nijmegen in 1996 at a meeting of the 'Dutch Organization of Phonetic Sciences' (see http://fonsg3.let.uva.nl /FonetischeVereniging/). The presentation was entitled 'Two methods of speech research: The classic phonetic and the speech technological approach' (the original Dutch title was: ''Twee methodes van spraakonderzoek: klassiek fonetische & spraak-technologische'').

Some clarification is in order here. In a prototypical classic phonetic experiment, a factorial design is used to make proper statistical analysis possible. Preferably, all cells in the factorial design are filled with the same number of observations. Care is taken to control other (known) factors, to reduce their (disturbing) effect as much as possible. Therefore, high quality sound is often used in a controlled setting (a studio), instead of e.g. spontaneous speech in a train station. For instance, in investigating lexical stress, subjects are asked to carefully pronounce contrastive pairs like ''SUBject'' and ''subJECT'' in a very controlled way (see also Section 3.2).

With such (classic) phonetic experiments a great deal of phonetic knowledge has been acquired over the years. The question is whether this phonetic knowledge can be used in speech technology, and of course how.

3. USING PHONETIC KNOWLEDGE IN SPEECH TECHNOLOGY: SOME EXAMPLES

3.1. Duration Model

Within the European ESPRIT project POLYGLOT, an isolated word recognition (IWR) system that had been originally developed for Italian

(Billi et al., 1989), had to be localized to a number of other European languages, including Dutch. This system made use of some phonetic knowledge, among others a duration model. This duration model contained statistics on the duration of phonetic units, which essentially were classes of phones with similar properties (Strik and Konst, 1992). What was needed for the IWR system were the conditional probabilities of a certain duration given the class of phones: P(duration | class of phones).

In order to obtain this duration model for Dutch, we first had a look at the literature. We found that research on this topic had indeed been carried out (e.g. Nooteboom, 1972; Nooteboom and Slis, 1972; Koopmans van Beinum, 1980). Although part of the phonetic knowledge in these publications was quantitative, it was not sufficient to derive the duration model needed, mainly for the following two reasons. [1] The phonetic knowledge was not complete: data on vowels were present, but not on consonants. [2] The knowledge was not in the correct format: It was specified in terms of means (and, sometimes, standard deviations), while for the duration model a probability density function was needed. Since the required duration model could not be derived from existing phonetic knowledge, we decided to use a data-driven method to obtain it (Strik and Konst, 1992). Isolated words were recorded, labeled, segmented and on the basis of these data a duration model was calculated.

In Table 2 mean and standard deviation values of the durations of some short vowels are given. These values are compared to the measurements of Koopmans van Beinum (1980): mean and standard deviation values of five measurements of the duration of vowels in isolated monosyllabic words spoken by an untrained male speaker. Of the various conditions for which vowel durations were measured by Koopmans

Table 2. Duration of short vowels (SAMPA notation is used in this article).

	Strik and Konst		Koopmans van Beinum	
Phone	Mean	SD	Mean	SD
I	91	24	124	24
U	98	29	108	22
O	99	25	108	23
A	103	23	120	19
i	105	37	140	12
E	107	25	124	15
u	111	25	150	17
y	140	55	136	23

van Beinum (1980), this was the condition that most closely matched the isolated word condition of this ASR system. The average values found by Koopmans van Beinum (1980) are larger (except for /y/), which is not surprising since she only used monosyllabic words and the database in Strik and Konst (1992) contains both monosyllabic and polysyllabic words.

These findings make clear that, besides the two reasons already mentioned above in this section, there is another reason why it is problematic to transfer existing knowledge to a speech technology application: The existing phonetic knowledge is based on data that is not representative of the speech that will be used in the application. In this case: only monosyllabic words (in the phonetic experiment) versus monosyllabic and polysyllabic words (for the ASR). Furthermore, it is questionable whether the five measurements of a single male subject are representative of the whole population.

Although duration has been studied a great deal, and thus substantial knowledge about duration should be present, duration is hardly used in current ASR systems. However, it is often used in TTS systems, in which case also data-driven methods are used to derive the duration models (see e.g. Lopez and Hernandez, 1995; van Santen, Sproat, Olive, and Hirschberg, 1996).

3.2. Lexical Stress

Phonetic research has shown that there are systematic acoustic differences between (the vowels in) syllables with and without lexical stress (see e.g. van Bergem, 1993; Sluijter and van Heuven, 1996). It has been observed that stressed syllables have a longer duration, higher energy, less spectral tilt, and a different vowel quality (i.e. more like a full vowel than like a reduced vowel). Given these systematic differences, one would expect that this knowledge could be used to improve the performance of ASR systems. A couple of years ago, this issue was investigated at our department. The procedure that was followed is described below.

irst, different models for vowels in stressed and unstressed syllables were trained (Kuijk, Heuvel and Boves, 1996). The recognition results on independent test-sets showed no clear improvements in the performance of the ASR system. Nevertheless, the resulting models for the vowels in stressed and unstressed condition were different, since swapping the models (i.e. using models trained on stressed vowels to recognize unstressed ones, and vice versa) led to higher error rates (Kuijk, Heuvel and Boves, 1996).

Other attempts at making use of lexical stress in ASR have led to varying results. Adda-Decker and Adda (1992) found improvements

for a French corpus, but not for the American English DARPA-RM corpus. Hieronymus et al. (1992) reported a 65% reduction in word error rate and a 45% reduction in sentence error rate. More recently, Wang and Seneff (2001) obtained a small but significant relative improvement of 5.5% in word error rate.

In order to get a better understanding of why knowledge on lexical stress cannot easily be applied to obtain substantial improvements in the performance of ASR systems, a more detailed study was carried out (Kuijk and Boves, 1999). Measurements of various phonetic features were made for 5000 phonetically rich sentences from the Dutch POLYPHONE corpus. A comparison was made of the phonetic feature values in stressed and unstressed condition. For instance, the distributions of the durations of the vowels /9y/ and /a:/ are shown in Figure 1. These distributions clearly illustrate the two extremes that were observed in comparing the distributions of the phonetic features: from almost no difference to large differences. Significant differences were found in the majority of cases, reflecting that systematic acoustic differences are present. Such differences might be useful to classify vowels as either stressed or unstressed. This possibility was verified in a number of tests, using both raw and normalized phonetic features. The results for correct classification of stress varied from 57.16% to 76.05%, for the various vowels. In conclusion, although there are systematic and significant differences between vowels in stressed and unstressed syllables, the resulting classification scores are not very high.

In trying to understand these results one should keep in mind that even if the differences are significant the overlap can be considerable. This is the case for almost all distributions of the phonetic features in this experiment. Other (classic) phonetic experiments have generally yielded

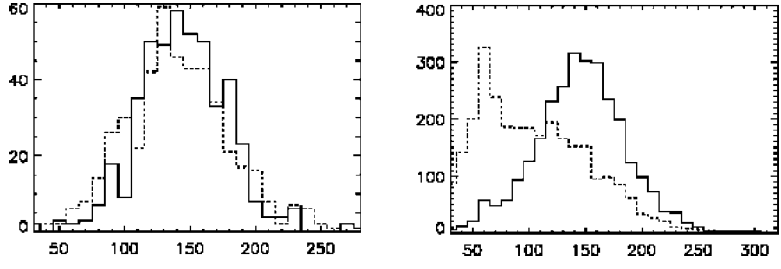

Figure 1. Distributions of the durations (in ms) of the vowels /9y/ and /a:/ (solid line: stressed condition, dotted line: unstressed condition).

smaller overlaps, because in these experiments the effects of other factors were reduced as much as possible by using a controlled setting: e.g. stress-minimal pairs (like "SUBject" versus "subJECT") were carefully pronounced in identical phonetic contexts. However, in real life the effects of other factors are present and cannot be ruled out. The consequences are that effects of lexical stress which are present in 'lab speech' are blurred by the effects of other factors in 'real speech'.

In this case the main reason why phonetic knowledge fails to improve ASR performance is that this knowledge is based on carefully controlled speech that is not representative of the speech encountered in everyday life. In addition, one should realize that knowledge about lexical stress is rather qualitative in nature: although in many publications on this topic measurement data are presented, it is obvious that there is no ready-made 'lexical stress model' that can be plugged directly into an ASR system.

3.3. Pronunciation Variation Modeling for ASR

A well-known problem in ASR is pronunciation variation. Various methods to model pronunciation variation at the lexical level have been investigated, in order to enhance the performance of ASR systems (for an overview see Strik and Cucchiarini, 1999; and Strik, 2001).

Since knowledge about pronunciation variation is available in the literature, it seems logical to employ this knowledge in ASR systems. In general, knowledge on pronunciation variation is qualitative and is often expressed in the form of rewrite rules. With these rewrite rules, pronunciation variants can be generated and subsequently added to the lexicon. In this way the performance of an ASR system can be improved (Kessens, Wester and Strik, 1999). This can, for instance, be observed in Figure 2 (taken from Kessens, 2002; and Kessens, Cucchiarini and Strik, 2003). The upper curve (labeled 'Lexicon') shows the word error rates (WERs) when pronunciation variants are added to the lexicon. When going from 1 to about 1.5 variants, the WER becomes lower. However, when more variants are added the WER goes up again and even reaches levels that are much higher than that of the baseline system. Probably this is because the confusability in the lexicon becomes too large if too many variants are added to the lexicon.

Somewhat better results can be obtained if the acoustic models are retrained (see curve 'HMMs' in Figure 2). The best results are obtained if the probabilities of the variants are taken into account in the language

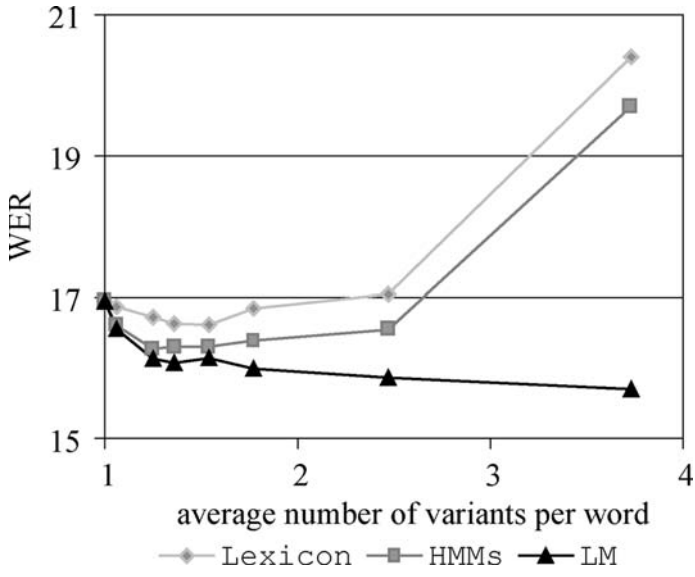

Figure 2. WERs for the different testing conditions.

model (LM) of the ASR (see curve 'LM' in Figure 2), but these probabilities are not readily available in the literature. A possibility then is to use a knowledge-based approach: start with the known rules, and calculate the probabilities of these rules, or the pronunciation variants generated with these rules, on the basis of a speech corpus. Such a knowledge-based approach has often been resorted to (see references in Strik and Cucchiarini, 1999; Strik, 2001). Although in this approach knowledge is used (i.e. rules), it is important to notice that the probabilities (which are essential) have to be derived from data, preferably substantial amounts of representative data.

Another possibility is to use a data-driven approach in which both the rules and their probabilities are derived from data (see references in Strik and Cucchiarini, 1999; Strik, 2001). The data-driven method generally comes up with known rules, which provide a description of the connected speech processes that are present in the speech corpus under investigation, plus many new rules which were not yet known (Wester, Kessens and Strik, 1998; Kessens, Wester and Strik, 2000; Kessens, Strik and Cucchiarini, 2000). Consequently, error rates obtained with data-driven approaches are usually lower (Kessens, Strik and Cucchiarini, 2000; Wester and Fosler-Lussier, 2000; Wester, 2002).

To sum up, linguistic knowledge can be used to model pronunciation variation for ASR. Simply using the knowledge as is (i.e. in the form of rewrite rules) can enhance the performance of an ASR system. However, even lower error rates can be obtained if probabilities of the rules (or variants) are derived from recorded and labeled data. And, if the data are available, one can probably best resort to data-driven methods, since they generally yield the best results (and in this way new rules can be learned). In this case the main obstacle to using existing knowledge is that the knowledge is not complete, and that it is not quantitative in nature.

3.4. Prosodic Models and Language Models

Besides the three examples taken from our own research, which were presented above, many more examples can be found in the literature, of which two are mentioned here.

The first example concerns prosodic models. Despite the enormous amount of phonetic/linguistic research on prosody that has been carried out, prosodic models are rarely used in ASR systems. Some reasons why this is the case are presented in Batliner et al. (2001). An important reason is that in most prosodic models too much emphasis is put on intonation (pitch, F0), and thus these models are not complete since prosody does not manifest itself in terms of F0 alone. In fact, F0 cannot even be varied in isolation without affecting other acoustic properties of the speech signal like spectral tilt and intensity (Strik, 1994).

The last example we want to mention is that of language models used in ASR. Generally, N-grams are used, which are simple stochastic models that can easily be integrated into ASR systems. Although syntax has been studied extensively, and many grammars have been proposed and developed over the years, so far (classic) linguistics has not provided a viable alternative to the N-grams (see e.g. Rosenberg, 2000). To a large extent this is due to the fact that this branch of linguistics has mainly been engaged with written language and not with spoken language. When we speak we often produce utterances that are not grammatically correct. And, to make things even more difficult, we also produce many disfluencies. Some studies have focused explicitly on spoken language, and tried to incorporate linguistically motivated language models in ASR systems (see the many references in Brill, Florian, Henderson, and Mangu, 1998). However, none of them succeeded in achieving substantial improvements over the N-gram.

4. DISCUSSION

What impedes the transfer of phonetic knowledge to speech technology? First of all, it is clear that in order to be used in speech technology, phonetic knowledge has to be incorporated into the computational framework of a speech technology system. There are several factors that make this incorporation problematic for much of the existing phonetic knowledge, which mainly has been obtained through controlled (classic) phonetic experiments. Some of these problems were illustrated in the examples in the previous section. To summarize, the main problems that emerged from the examples in the previous section are that the knowledge is based on small amounts of 'lab speech' and therefore does not generalize to 'more realistic speech', that the knowledge is not complete, and finally that it is not quantified at all or not quantified in the right format. In other words, phonetics does not provide ready-made quantitative models that can be plugged directly into a system.

These quantitative models can be derived on the basis of the large speech corpora that are available nowadays, with knowledge-based or data-driven methods, or combinations of these two types of methods. If the existing knowledge is not complete, as is often the case, then it is probably best to use data-driven approaches. Initial ideas about phonetic phenomena could come from (controlled) phonetic experiments. Subsequently, these ideas should be tested and quantified using large speech corpora. In this way knowledge can be acquired which can more easily be integrated in speech technology.

Of course, one could wonder whether more phonetic knowledge should be used in speech technology at all. A reason for doing so, which is often mentioned in this context, is that humans perform better than machines on many tasks. However, should an ASR system have ears and a basilar membrane, or should a TTS system have a larynx and a tongue (see also Hermansky, 1998)? No! We do not need to make replicas of (parts of) humans. Another extreme is not using phonetic knowledge at all. In this case, e.g., a speech corpus is seen as just a bunch of CDs, files or signals for which the word error rate or another error criterion should be minimized. It is obvious that to improve speech technology systems using phonetic/linguistic knowledge can be useful. A good example to illustrate this is knowledge about human auditory perception, which was applied to improve ASR performance (Hermansky, 1998). Nowadays, perception-based features (such as Mel, Bark or PLP) are used in most ASR systems.

Another reason why phonetic/linguistic knowledge could be useful is the following. Progress with current ASR and TTS techniques has been steady but slow. This could indicate that the ceiling of the performance for current techniques has almost been reached. Therefore, the best way to proceed is probably not to put only a lot of extra effort into fine-tuning the existing techniques, but instead to study some innovative approaches too. And although the complete solution cannot be found in current phonetic/linguistic knowledge, this knowledge should certainly be taken into consideration while searching for new techniques for better systems.

Speech production is a process that is constrained at various levels: acoustic, phonetic, phonological, lexical, syntactic, and semantic. Knowledge about these constraints could be of benefit to speech technology. To this end, these constraints have to be identified, described (in a certain formalism), and quantified in such a way that they can be incorporated in a complete computational framework. The best results in ASR so far have been obtained with a stochastic computational framework, so it is likely that the description and the quantification of the constraints should be of a stochastic nature. These constraints can be described at different levels (multiple tiers). Information missing on one level can then be derived from, or complemented with, information from other levels.

So far, the emphasis in ASR and TTS has been on word recognition and synthesis. Since speech is mainly used for communication, the focus of research should shift more towards understanding and expressing messages, i.e. speech-to-concept and concept-to-speech (see also Zue, 1991, and Furui, 2000). This does not only require phonetic knowledge, it also requires knowledge from many other disciplines. Between these worlds even more gaps will exist. For instance, psycholinguistic models often use the correct phone(me) sequences as input, and many natural language processing models take the correct word sequences as input. In ASR the correct phone(me) and word sequences are not readily available. Therefore, these models cannot be directly integrated with ASR systems. In order to integrate models from different disciplines, a lot of gaps still have to be bridged.

ACKNOWLEDGEMENTS

I would like to thank two anonymous reviewers and my colleagues (in alphabetical order) Loe Boves, Catia Cucchiarini, Henk van de Heuvel, Judith Kessens, David van Kuijk, Ambra Neri, Mirjam Wester and Febe de Wet for their comments on a previous version of this paper.

REFERENCES

Adda-Decker, M. and Adda, G. Experiments on stress-dependent phone modeling for continuous speech recognition. In: *Proceedings of ICASSP-92*, San Francisco, USA, 1992: 561–564.

Bergem, D.R. van. Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, 12 (1993): 1–23.

Batliner, A., Möbius, B., Möhler, G., Schweitzer, A., and Nöth, E. Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground. In: *Proceedings of Eurospeech-2001*, Aalborg, Denmark, 2001: 2285–2288.

Billi, R., Arman, G., Cericola D., Massia, G., Mollo, M., Tafini, F., Varese, G., and Vittorelli, V. A PC-based large vocabulary isolated word speech recognition system. In: *Proceedings of Eurospeech-89*, Paris, France, 1989: 157–160.

Brill, E., Florian, R., Henderson, J., and Mangu, L. Beyond N-Grams: Can Linguistic Sophistication Improve Language Modeling? In: *Proceedings of COLING/ACL 1998 Conference*, Montreal, Canada, 1998: 186–190.

Furui, S. Steps towards natural human-machine communication in the 21st century. In: *Proceedings of COST249 Workshop on Voice Operated Telecom Services*, Ghent, Belgium, 2000: 17–24.

Hermansky, H. Should recognizers have ears? *Speech Communication*, 25 (1998): 3–28.

Hieronymus, J.L., McKelvie, D., and McInness, F.R. Use of acoustic sentence level and lexical stress in HMM speech recognition. In: *Proceedings of ICASSP-92*, San Francisco, USA, 1992: 225–229.

Kessens, J.M. *Making a difference: On automatic transcription and modeling of Dutch pronunciation variation for automatic speech recognition*. Ph.D. dissertation, University of Nijmegen, the Netherlands, 2002.

Kessens, J.M., Cucchiarini, C., and Strik, H. A data-driven method for modeling pronunciation variation. *Speech Communication*, 40 (2003): 517–534.

Kessens, J.M., Strik, H., and Cucchiarini, C. A bottom-up method for obtaining information about pronunciation variation. In: *Proceedings of ICSLP-2000*, Beijing, China, 2000: 274–277.

Kessens, J.M., Wester, M., and Strik, H. Improving the Performance of a Dutch CSR by Modeling Within-word and Cross-word Pronunciation. *Speech Communication*, 29 (1999): 193–207.

Kessens, J.M., Wester, M., and Strik, H. Automatic Detection and Verification of Dutch Phonological Rules. In: *PHONUS5, Proceedings of the ''Workshop on Phonetics and Phonology in ASR''*, Saarbrücken, Germany, 2000: 117–128.

Koopmans van Beinum, F.J. *Vowel contrast reduction: an acoustic and perceptual study of Dutch vowels in various speech conditions*. Ph.D. Thesis, University of Amsterdam, The Netherlands, 1980.

Kuijk, D. van, Boves, L. Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication*, 27 (1999): 95–111.

Kuijk, D. van, Heuvel, H. van den, and Boves, L. Using lexical stress in continuous speech recognition for Dutch. *Proceedings ICSLP-96*, Philadelphia, USA, 1996: 1736–1739.

Lopez-Gonzalo, E. and Hernandez-Gomez, L.A. Automatic Data-Driven Prosodic for Text to Speech. In: *Proceedings Eurospeech-95*, Madrid, Spain, 1995: 585–588.

Nooteboom, S.G. *Production and perception of vowel duration: a study of durational properties of vowels in Dutch*. Ph.D. Thesis, University of Utrecht, The Netherlands, 1972.

Nooteboom, S.G. and Slis, I.H. The phonetic feature of vowel length in Dutch. *Language and Speech*, 15 (1972): 301–316.

Pols, L.C.W. Flexible, robust, and efficient human speech processing versus present-day speech technology. In: *Proceedings of ICPhS-99*, San Fransisco, USA, 1999: 9–16.

Rosenfeld, R. (2000). Two decades of Statistical Language Modeling: Where Do We Go From Here? In: *Proceedings of the IEEE*, 88(8), August 2000: 1270–1278.

Santen, J. van, Sproat R., Olive J., and Hirschberg, J. *Progress in speech synthesis*. Springer-Verlag, New York, 1996.

Sluijter, A.M.C. and Heuven, V.J. van. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100 (1996): 2471–2485.

Stevens, K.N. Toward a model for speech recognition. *Journal of the Acoustical Society of America*, 32 (1960): 47–55.

Strik, H. *Physiological control and behaviour of the voice source in the production of prosody*. Ph.D. dissertation, University of Nijmegen, the Netherlands, 1994.

Strik, H. Pronunciation adaptation at the lexical level. In: *Proceedings of the ITRW 'Adaptation Methods for Speech Recognition'*, Sophia-Antopolis, France, 2001: 123–130.

Strik, H. and Cucchiarini, C. Modeling pronunciation variation for ASR: a survey of the literature. *Speech Communication*, 29(1999): 225–246.

Strik, H. and Konst, E. A Duration Model for Phonetic Units in Isolated Dutch Words. In: *AFN-Proceedings*, Universiteit of Nijmegen, 15 (1992): 71–78.

Wang, C. and Seneff, S. Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain. In: *Proceedings of Eurospeech-2001*, Aalborg, Denmark, 2001: 2761–2764.

Wester, M. *Pronunciation variation modeling for Dutch automatic speech recognition*. Ph.D. Dissertation, University of Nijmegen, The Netherlands, 2002.

Wester, M. and Fosler-Lussier, E. A comparison of data-derived and knowlegde-based modeling of pronunciation variation. In: *Proceedings of ICSLP-2000*, Beijing, China, 2000: 270–273.

Wester, M., Kessens, J.M., and Strik, H. Modeling pronunciation variation for a Dutch CSR: testing three methods. In: *Proceedings of ICSLP-98*, Sydney, Australia, 1998: 2535–2538.

Zue, V. The Use of Phonetic Rules in Automatic Speech Recognition. *Speech Communication*, 2 (1983): 181–186.

Zue, V. From Signals to Symbols to Meaning: On Machine Understanding of Spoken Language. In: *Proceedings of the XIIth International Congress of Phonetic Sciences 1991*, Aix-en-Provence, France, 1991: 74–83.

Text, Speech and Language Technology

1. H. Bunt and M. Tomita (eds.): *Recent Advances in Parsing Technology.* 1996 ISBN 0-7923-4152-X
2. S. Young and G. Bloothooft (eds.): *Corpus-Based Methods in Language and Speech Processing.* 1997 ISBN 0-7923-4463-4
3. T. Dutoit: *An Introduction to Text-to-Speech Synthesis.* 1997 ISBN 0-7923-4498-7
4. L. Lebart, A. Salem and L. Berry: *Exploring Textual Data*. 1998 ISBN 0-7923-4840-0
5. J. Carson-Berndsen, *Time Map Phonology*. 1998 ISBN 0-7923-4883-4
6. P. Saint-Dizier (ed.): *Predicative Forms in Natural Language and in Lexical Knowledge Bases.* 1999 ISBN 0-7923-5499-0
7. T. Strzalkowski (ed.): *Natural Language Information Retrieval.* 1999 ISBN 0-7923-5685-3
8. J. Harrington and S. Cassiday: *Techniques in Speech Acoustics.* 1999 ISBN 0-7923-5731-0
9. H. van Halteren (ed.): *Syntactic Wordclass Tagging.* 1999 ISBN 0-7923-5896-1
10. E. Viegas (ed.): *Breadth and Depth of Semantic Lexicons.* 1999 ISBN 0-7923-6039-7
11. S. Armstrong, K. Church, P. Isabelle, S. Nanzi, E. Tzoukermann and D. Yarowsky (eds.): *Natural Language Processing Using Very Large Corpora.* 1999 ISBN 0-7923-6055-9
12. F. Van Eynde and D. Gibbon (eds.): *Lexicon Development for Speech and Language Processing*. 2000 ISBN 0-7923-6368-X; Pb: 07923-6369-8
13. J. Véronis (ed.): *Parallel Text Processing*. Alignment and Use of Translation Corpora. 2000 ISBN 0-7923-6546-1
14. M. Horne (ed.): *Prosody: Theory and Experiment*. Studies Presented to Gösta Bruce. 2000 ISBN 0-7923-6579-8
15. A. Botinis (ed.): *Intonation*. Analysis, Modelling and Technology. 2000 ISBN 0-7923-6605-0
16. H. Bunt and A. Nijholt (eds.): *Advances in Probabilistic and Other Parsing Technologies*. 2000 ISBN 0-7923-6616-6
17. J.-C. Junqua and G. van Noord (eds.): *Robustness in Languages and Speech Technology*. 2001 ISBN 0-7923-6790-1
18. R.H. Baayen: *Word Frequency Distributions*. 2001 ISBN 0-7923-7017-1
19. B. Granström, D. House and. I. Karlsson (eds.): *Multimodality in Language and Speech Systems*. 2002 ISBN 1-4020-0635-7
20. M. Carl and A. Way (eds.): *Recent Advances in Example-Based Machine Translation.* 2003 ISBN 1-4020-1400-7; Pb 1-4020-1401-5
21. A. Abeillé: *Treebanks.* Building and Using Parsed Corpora. 2003 ISBN 1-4020-1334-5; Pb 1-4020-1335-3
22. J. van Kuppevelt and R.W. Smith (ed.): *Current and New Directions in Discourse and Dialogue.* 2003 ISBN 1-4020-1614-X; Pb 1-4020-1615-8
23. H. Bunt, J. Carroll and G. Satta (eds.): *New Developments in Parsing Technology.* 2004 ISBN 1-4020-2293-X; Pb 1-4020-2294-8

24. G. Fant: *Speech Acoustics and Phonetics.* Selected Writings. 2004 ISBN 1-4020-2373-1; Pb 1-4020-2789-3
25. W.J. Barry and W.A. Van Dommelen (eds.): *The Integration of Phonetic Knowledge in Speech Technology.* 2005 ISBN 1-4020-2635-8; Pb 1-4020-2636-6
26. D. Dahl (ed.): *Practical Spoken Dialog Systems.* 2004 ISBN 1-4020-2674-9; Pb 1-4020-2675-7
27. O. Stock and M. Zancanaro (eds.): *Multimodal Intelligent Information Presentation.* 2005 ISBN 1-4020-3049-5; Pb 1-4020-3050-9
28. W. Minker, D. Bühler and L. Dybkjaer (eds.): *Spoken Multimodal Human-Computer Dialogue in Mobile Environments.* 2004 ISBN 1-4020-3073-8; Pb 1-4020-3074-6