T-Labs Series in Telecommunication Services

Florian Hinterleitner

Quality of Synthetic Speech

Perceptual Dimensions, Influencing Factors, and Instrumental Assessment

Springer

T-Labs Series in Telecommunication Services

More information about this series at http://www.springer.com/series/10013

Florian Hinterleitner

Quality of Synthetic Speech

Perceptual Dimensions, Influencing Factors, and Instrumental Assessment

Springer

Florian Hinterleitner
Quality and Usability Lab
Institute of Software Engineering
and Theoretical Computer Science,
Berlin Institute of Technology
Berlin
Germany

ISSN 2192-2810 ISSN 2192-2829 (electronic)
T-Labs Series in Telecommunication Services
ISBN 978-981-10-3733-7 ISBN 978-981-10-3734-4 (eBook)
DOI 10.1007/978-981-10-3734-4

Library of Congress Control Number: 2017930163

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Acknowledgements

First, I would like to thank my supervisor Prof. Dr.-Ing. Sebastian Möller for his support. Sebastian, you have been an excellent mentor, from the time I started to work on my Magister thesis, until I finished writing this PhD thesis. Thank you for all the contributions, ideas, and discussions. This would not have been possible without you!

Also, I would like to thank Prof. Dr.-Ing. Ulrich Heute. The discussions we had during the duration of the DFG project have been tremendously helpful. I always left our meetings with tons of new ideas and a checklist of solved problems. Moreover, I want to thank the third committee member, Prof. Simon King, for the productive collaboration over the past years. Your agreement on co-examining my thesis is highly appreciated.

Furthermore, I am especially grateful for the excellent team play with my colleague Dr.-Ing. Christoph Norrenbrock over the past years. All the discussions we had improved the quality of my work by miles. Thank you!

When I started working at the Quality & Usability Lab as a student over 8 years ago, I had no idea where this journey would lead me. During this time I worked with so many great colleagues of whom some turned into close friends. I want to thank Tilo Westermann for endless discussions about structure, phrasing, and the close race towards the finish line, which you unfortunately won. Thanks to Tobias Hirsch, Friedemann Köster, and Falk Schiffner for taking over my work on module descriptions, dealing with QISPOS chaos, and teaching *Speech Communication*, which made it possible for me to only focus on my thesis in the past couple of months. I want to thank Ina Wechsung for her lessons on statistics, Benjamin Weiss for proofreading my thesis and giving me helpful feedback, my foosball buddies Jens Ahrens, Benjamin Bähr, Justus Beyer, Klaus-Peter Engelbrecht, and Steffen Zander for distracting or not distracting me, and my *roomies* Julia Seebode, Niklas Kirschnick, Karim Helwani, and Florian Kretzschmar for tolerating my constant swearing when I was stuck in MATLAB or was about to be defeated by SPSS. I also want to thank the students I had the pleasure to supervise during their bachelor or master thesis and the students who worked with me on different

research projects. Special thanks to Iason Georgakopoulos for your tremendous support throughout the past two years. And lastly, working at this chair would have not been possible without the amazing assistance of Irene Hube-Achter and Yasmin Hillebrenner. Thank you for the great organizational support!

Special thanks to my family. I am incredibly grateful for all the support during my time in Berlin. And I also want to thank my grandpa for always giving me advice for life, whether I wanted to hear it or not, my close friends for balancing my life and taking my mind off work, and Leigh-Anne Robinson for supporting and understanding me during the past months.

Contents

Acronyms

AcC	Accent Command
ACCP	Acceptance
ACR	Absolute Category Rating
ANIQUE	Auditory Non-Intrusive Quality Estimation
ANOVA	ANalysis Of VAriance
AOD	Absence of Disturbances
BC	Blizzard Challenge
BF	Base Frequency
BUMP	BUMPiness
C	Calmness
CLID	CLuster IDentification Test
CLIN	Clink
CONT	Content
COPR	Comprehension Problems
CS	Continuous Scale
CTS	Concept-to-Speech
CV	Cross Validation
DI	Diphone
DIAL	Diagnostic Instrumental Assessment of Listening quality
DNN	Deep Neural Networks
DRT	Diagnostic Rhyme Test
DSTO	Distortions
DSTU	Disturbances
EMOT	Emotion
EP	Ending Point
FA	Factor Analysis
FAI	Fluency and Intelligibility
FD	Frequency Domain
FLUE	Fluency
FO	Formant

FS	Feature Selection
G2P	Grapheme-to-Phoneme
GUI	Graphical User Interface
HISS	Hiss
HMM	Hidden Markov Model
HTS	HMM-based Speech Synthesis System
INTE	Intelligibility
INTO	Intonation
IRR	Infinite Impulse Response
IRS	Intermediate Reference System
ITU	International Telecommunication Union
ITU-T	Telecommunication Standardization Sector of the International Telecommunication Union
KMO	Kaiser–Meyer–Olkin
LOFA	Level of Familiarity
LOTO	Leave-One-Test-Out
LP	Linear Prediction
LPC	Linear Predictive Coding
LSTE	Listening Effort
MANOVA	Multivariate ANalysis Of VAriance
Mary	Modular Architecture for Research in Synthesis
MaryTTS	Mary Text-to-Speech
MDS	Multidimensional Scaling
MFCC	Mel-Frequency Cepstral Coefficient
MOS	Mean Opinion Score
MRT	Modified Rhyme Test
NATU	Naturalness
NLP	Natural Language Processing
NOIS	Noise
NOV	Naturalness of Voice
OIMP	Overall Impression
PAF	Principal Axis Factor
PC	Pairwise Comparison
PCA	Principal Component Analysis
PESQ	Perceptual Evaluation of Speech Quality
PhC	Phrase Command
POLQA	Perceptual Objective Listening Quality Assessment
POLY	Polyphony
PPR	Physiological Perception Range
PQ	Prosodic Quality
PR	Perceptual Regularization
PSOLA	Pitch Synchronous Overlap and Add
RASP	Rasping Sound
RHYT	Rhythm
RMSE	Root Mean Square Error

RP	Regular Perception
RPM	Regular Perception Model
RPR	Regular Perception Range
SAMPA	Speech Assessment Methods Phonetic Alphabet
SD	Semantic Differential
SDS	Spoken Dialog System
SPAU	Speech Pauses
SPEE	Speed
SP	Starting Point
SR	Stepwise multiple linear Regression
ST	Sorting Task
STRE	Stress
SUS	Semantically Unpredictable Sentences
SVM	Support Vector Machine
SVR	Support Vector Regression
TD	Time Domain
TENS	Tension
TTS	Text-to-Speech
US	Unit Selection
VAD	Voice Activity Detection
VPLT	Voice Pleasantness
WB	Wide-Band
XML	Extensible Markup Language

Abstract

Text-to-Speech (TTS) synthesis, i.e., artificially produced speech, has finally attained a quality level that makes it possible to include it into ordinary services that are used by common people. With the increasing processing power of smartphones and the development of intelligent personal assistants like Siri, Cortana, and Google Now, synthetic speech started to affect even more people. Therefore, within the past couple of years, TTS has made its way from a geeky accessory to a normal part of everyday life.

Nonetheless, modern TTS systems still suffer from diverse quality constraints: frequent concatenations and temporal manipulations in diphone synthesis cause discontinuous speech, HMM synthesis can lead not only to natural sounding but also to very buzzy and muffled speech, and the quality of unit selection voices not only depends on the degree of the fit, but also on the appropriateness of the available speech units. Therefore, the resulting impairments all yield different perceptual impressions. Thus, the quality of synthetic speech is of multidimensional nature.

Therefore, research towards perceptual quality dimensions of synthetic speech is reviewed and two experiments towards perceptual quality are conducted. Their findings are compared with the state of the art and a set of five perceptual quality dimensions is derived. They are: (i) *naturalness of voice*, (ii) *prosodic quality*, (iii) *fluency and intelligibility*, (iv) *absence of disturbances*, and (v) *calmness*. Moreover, a test protocol is designed that recommends an experimental setup to assess these five dimensions.

In addition, several factors that influence these dimensions are analyzed. First, the findings of two studies show that the relevance of these dimensions shift depending on the use case (short messages readers vs. synthesized audiobooks). Second, a significant effect of a speaker's voice of a speech corpus is verified for all dimensions. And third, it is shown that the size of the speech corpus for unit selection voices significantly affects all dimensions.

Furthermore, different approaches towards instrumental quality assessment of synthetic speech are examined. Two linear regression models are developed and employed to estimate the quality of TTS signals. Even though they reach

correlations between estimated score and auditory rating of up to 0.74, they are outperformed by two more complex, non-linear approaches. One of these non-linear measures is utilized with the aim to improve the quality of MaryTTS unit selection voices. Even though this goal could not be achieved, the study highlights different approaches to further improve the prediction accuracy and therefore also the quality of the generated voice.

Chapter 1
Introduction

1.1 Motivation

The quality of Text-to-Speech (TTS) synthesis, i.e., artificially produced human speech, has increased remarkably over the past years. Until recently, when talking about synthetic speech, most people thought of robot-like voices like the one of Steven Hawking's formant synthesizer [1]. State-of-the-art systems, however, are finally able to generate speech that is close to human produced speech. With the popularity of smartphones and the development of intelligent personal assistants like Siri, Cortana, or Google Now, TTS has found its way into the lives of many people. Nonetheless, modern TTS systems still suffer from diverse quality constraints: depending on the type of system and synthesis method different distortions occur, e.g., diphone synthesizers usually sound very artificial due to their frequent concatenations, Hidden Markov Model (HMM) synthesis can lead to very natural sounding but often also muffled speech signals, and the quality of unit selection synthesizers depends to a high degree on how well the speech units fit to the utterance that is to be synthesized. Therefore, the quality of synthetic speech is of multidimensional nature. Thus, two different systems can score similar overall quality ratings in a listening test, while reaching contrasting scores on different quality dimensions. In addition to the TTS system itself, further factors influence the perceived overall quality or specific quality dimensions, e.g., the application the TTS system is designed for: due to the shorter exposure, a speech synthesizer will trigger a different perceptual impression when employed to announce weather information over the phone compared to a synthesized version of a novel. Furthermore, the quality of corpus based speech synthesizers also depends on the size of the speech corpus and speaker of the recordings.

During the development of a TTS system, regular evaluations are necessary to assess the progress. Depending on the quality aspect of interest, different kinds of auditory tests can be carried out: to check whether the system is able to carry information on a segmental or supra-segmental level, articulation and intelligibility tests can be carried out [2], comparison tests examine if a human listener is able to

F. Hinterleitner, *Quality of Synthetic Speech*, T-Labs Series
in Telecommunication Services, DOI 10.1007/978-981-10-3734-4_1

comprehend the content of the synthesized signal [3], and overall quality tests like the one recommended in ITU-T Rec. P.85 [4] are able to capture different quality aspects, e.g., *listening effort*, *comprehension problems*, *voice pleasantness*, and the *overall impression*. However, doubts have been cast on the suitability of various of the included items [5, 6]. Moreover, since this protocol was developed in 1995 when unit selection synthesis just started to become popular and long before HMM synthesis was developed, it is not clear whether the protocol is still able to capture all relevant quality aspects of state-of-the-art TTS systems.

Considering that auditory tests are extremely time consuming and cost intensive, constant auditory assessment of TTS systems in development is rarely feasible. Therefore, instrumental measures that estimate the quality perceived by a listener, without the requirement of inviting test participants for extensive listening tests, could greatly simplify the process of developing new TTS voices. Taking the TTS system development one step further by integrating such a measure into an existing TTS system could even improve the selection of speech units to be concatenated. This would then result in an increase in perceived overall quality of the final synthetic speech signal.

Considering that synthetic speech can be seen as distorted human speech, an application of existing quality measures for coded speech signals (i.e., signals transmitted through telephone networks) for the quality assessment of synthetic speech, seems to be a worthwhile attempt. Several measures that estimate the quality of coded signals are standardized by the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T) [7, 8] and have already been tested on synthetic speech [9–11] with mixed results. Moreover, few measures were developed to estimate the quality of specific TTS systems, e.g., quality prediction models for a unit selection synthesizer [12] and for an HMM voice [13]. Nonetheless, no standardized measures that are able to give analytic information (i.e., information on different perceptual quality dimensions) on a TTS signal exist today.

Therefore, this book seeks to answer the following research questions:

RQ1 *Which perceptual quality dimensions are relevant for state-of-the- art TTS systems?*

RQ2 *How should a listening test be designed in order to capture all rele- vant quality dimensions?*

RQ3 *Which factors in uence these perceptual quality dimensions?*

RQ4 *How can the quality of synthetic speech be assessed by an instrumental measure?*

RQ5 *Which requirements does an instrumental measure need to fulfill in order to be integrated into a TTS system?*

1.2 Outline

Chapter 2 presents the fundamentals of speech synthesis. The general setup of a speech synthesizer is introduced with a focus on different approaches to speech signal generation. Moreover, the open-source TTS synthesis platform Mary Text-to-Speech (MaryTTS) is described. MaryTTS will be the basis for applying instrumental assessment during the synthesis process, as outlined in Chap. 7.

Chapter 3 starts off with a definition of the term *quality* in the context of speech synthesis, and thereafter introduces a taxonomy for the quality assessment of synthetic speech. Furthermore, several auditory quality evaluation metrics are introduced and state-of-the-art instrumental quality measures for coded speech signals are displayed.

In Chap. 4, an overview of different studies on perceptual quality of TTS systems is given. Moreover, two extensive studies on perceptual quality dimensions are presented: in the first one, a semantic differential is developed, while the second one utilizes a direct comparison method for the evaluation of TTS stimuli.

In the following, Chap. 5 investigates factors influencing TTS quality, e.g., the application the system is used in, the voice of the recorded speaker, the size of the used speech corpus, and the utterance that is to be synthesized.

Chapter 6 employs different reference-based and reference-free approaches for the quality prediction of synthetic speech. Moreover, several new measures are developed and tested on the databases that were introduced in the previous chapters.

In Chap. 7 the benefit of the integration of a quality prediction model into a TTS system is investigated, difficulties that arise with this task are highlighted, and ways to resolve these problems are discussed.

Finally, Chap. 8 concludes the findings and gives an outlook on future work.

References

1. Hallahan WI (1995) DECtalk Software: Text-to-Speech Technology and Implementation. Digit Tech J 7:5–19
2. van Bezooijen R, van Heuven V (1997) Handbook of Standards and Resources for Spoken Language Systems, chapter Assessment of Synthesis Systems. Mouton de Gruyter, pp 481–563
3. Delogu C, Conte S, Sementina C (1998) Speech Communication, chapter Cognitive Factors in the Evaluation of Synthetic Speech. Elsevier, Amsterdam, pp 153–168
4. ITU-T Rec. P.85 (1994) A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices. International Telecommunication Union, Geneva
5. Viswanathan M, Viswanathan M (2005) Measuring Speech Quality for Text-to-Speech Systems: Development and Assessment of a Modified Mean Opinion Score (MOS) Scale. Comput Speech Lang 19:55–83
6. Sityaev D, Knill K, Burrows T (2006) Comparison of the ITU-T P.85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems. In: Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech), pp 1077–1080
7. ITU-T Rec. P.563 (2004) Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony. International Telecommunication Union, Geneva

8. ITU-T Rec. P.862 (2001) Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. International Telecommunication Union, Geneva
9. Cernak M, Rusko M (2005) An Evaluation of Synthetic Speech Using the PESQ Measure. In: Proceedings of Forum Acusticum 2005. Budapest, pp 2725–2728
10. Möller S, Heimansberg J (2006) Estimation of TTS Quality in Telephon Environments Using a Reference-free Quality Prediction Model. In: Proceedings of the 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, pp 56–60
11. ITU-T Contribution COM 12-180 (2008) Single-Ended Quality Estimation of Synthesized Speech: Analysis of the Rec. P.563 Internal Signal Processing. Deutsche Telekom AG (Authors: Möller S, Falk TH), ITU-T SG12 Meeting. Geneva, Switzerland
12. Chu M, Peng H (2001) An Objective Measure for Estimating MOS of Synthesized Speech. In: Proceedings of the 7th International Conference on Speech Communication and Technology (EUROSPEECH 2001), vol 3. Aalborg, Denmark, pp 2087–2090
13. Do C-T, Evrard M, Leman A, d'Alessandro C, Rilliard A, Crebouw JL (2014) Objective Evaluation of HMM-Based Speech Synthesis System Using Kullback-Leibler Divergence. In: Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech 2014). Singapore, pp 2952–2956

Chapter 2
Speech Synthesis

This chapter gives an introduction to speech synthesis. A general structure of TTS systems is introduced and the four main steps for producing a synthetic speech signal are explained. The main focus is put upon different methods for the speech signal generation, namely: parametric methods, concatenative speech synthesis, model-based synthesis approaches and hybrid models. Moreover, distortions that are specific for these systems are discussed. Finally, the open-source MaryTTS system is introduced.

2.1 Setup of a Speech Synthesizer

Different approaches towards synthetic speech have been developed over the years. So called *canned speech* systems use prerecorded phrases and play them back without, or only with very little changes. Such systems are therefore only employed in limited-domain systems, e.g., train station announcements. A more sophisticated approach is taken by Concept-to-Speech (CTS) systems. The main idea of CTS is the generation of waveforms directly from a linguistic description, i.e., without any specified input text. Although the idea of bypassing the difficulties of natural language processing seems promising, CTS systems have not managed to play an important role in speech synthesis.

Considering the limitations of *canned speech* and the shortcomings of CTS, this book exclusively concentrates on Text-to-Speech (TTS) synthesis. Even though there is a multitude of different approaches to TTS, there are similar steps every system has to go through in order to produce a synthetic speech signal. A general structure for a TTS system can be seen in Fig. 2.1. Such systems usually consist of four main steps: natural language processing, prosody generation, concatenation, and speech generation. These four steps are explained in the following sections.

F. Hinterleitner, *Quality of Synthetic Speech*, T-Labs Series
in Telecommunication Services, DOI 10.1007/978-981-10-3734-4_2

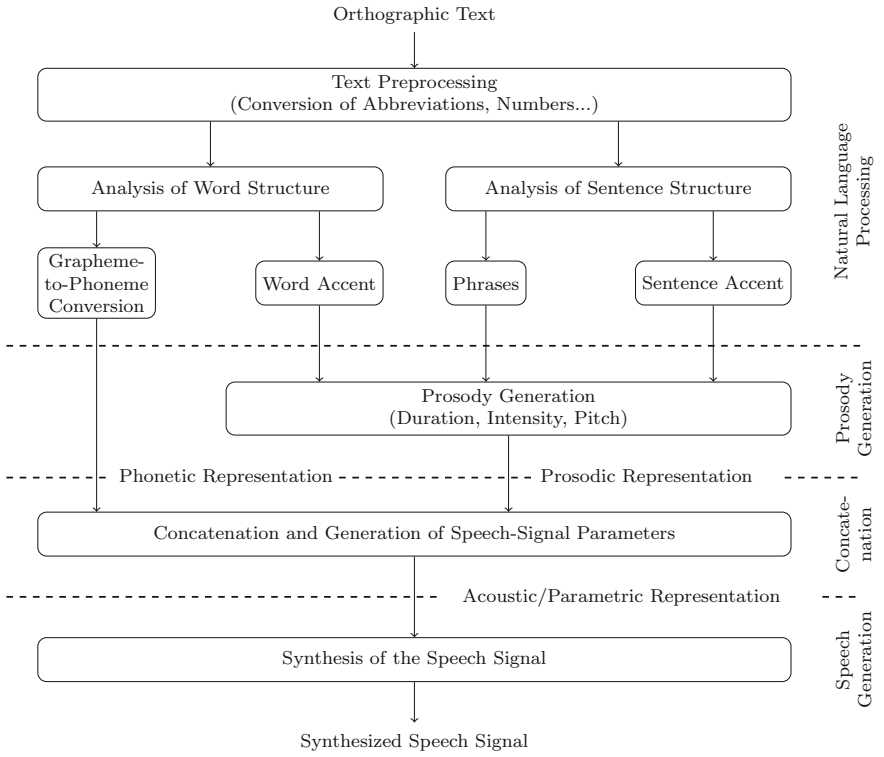


Fig. 2.1 General structure of TTS systems based on [1]

2.1.1 Natural Language Processing (NLP)

The first module of every TTS system is a text preprocessing unit. It analyzes the orthographic text and identifies special cases, e.g., abbreviations, numbers, foreign language terms, proper names etc. Those cases require special treatment, e.g., numbers have to be transformed into written text.

The preprocessing step is followed by an analysis of word and sentence structure. Therefore, on the one hand a morphologic analysis identifies word stems, prefixes, and suffixes which are important for the word stress. On the other hand, the sentence level structure is evaluated. This reveals sentence accents and information on phrases (i.e., groups of words). Word and sentence accents as well as identified phrases are needed in the following for the prosody generation.

Moreover, the morphologic output is used by the Grapheme-to-Phoneme (G2P) unit to transform the graphemes into their corresponding phonemes and thus create a phonetic representation of the orthographic text in a computer-readable format, e.g., using the (SAMPA) [1, 2].

2.1.2 Prosody Generation

The prosody generation unit uses information on word and sentence accents along with information on phrases to create the corresponding prosody of the orthographic text, i.e., duration, intensity and pitch.

To achieve this, procedures like the Fujisaki model [3, 4] can be applied. This model describes a pitch contour as a superposition of *phrase commands* and *accent commands* and an underlying *base frequency*. A detailed view on the Fujisaki model and features that can be derived from it is given in Sect. 6.2.2.1.

2.1.3 Concatenation and Generation of Speech-Signal Parameters

Based on the phonetic representation of the orthographic text generated by the G2P unit and the prosodic information, the concatenation unit creates a continuous sequence of signal parameters and/or articulation gestures.

Those first three units (NLP, prosody generation, and concatenation) solely depend on the given orthographic input text and are thus completely independent of the speaker of the speech signal to be synthesized.

2.1.4 Speech Signal Generation

This section introduces different methods for the speech signal generation within a TTS system based on the continuous sequence of acoustic units and/or signal parameters.

2.1.4.1 Parametric Speech Synthesis

In contrast to all other synthesis techniques that will be discussed in the following sections, parametric synthesizers do not use any prerecorded speech data, they generate the speech signal solely based on their input parameters. The basic principle behind this approach is the source-filter model of the human voice production system as shown in Fig. 2.2. According to this model, a speech signal can be generated from a voiced or an unvoiced excitation signal (or a combination of both) and parameters that describe the filter characteristics of the vocal tract, e.g., through parameters that reflect the shape of the pharynx and the mouth, the position of the tongue, and the rounding of the lips etc. Thus, these parameters specify a simple model of the vocal tract geometry. This model can then be used for an acoustic simulation of the human speech production. Such articulatory synthesizers [5] are still being developed,

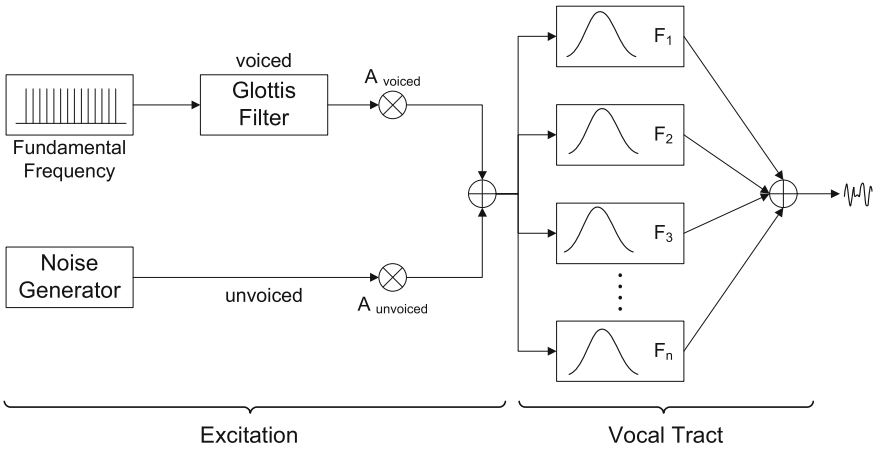


Fig. 2.2 Source-filter model of the human voice production

however, due to their inferior quality and their high computing time, they are mainly used for research.

An easier way to implement a parametric speech synthesizer is the realization of the vocal tract by formant filters.[1] These so called (FO) synthesizers are still being used, especially in systems with strong memory restrictions.

Formant Synthesizer

First attempts towards a synthesis based on formant filters have already been made in the early 1960s [6] and stayed popular long after that. The development of FO synthesis peaked in the late 80s/early 90s just before TTS research turned towards the concatenation of speech units. During this time, the Klatt synthesizer [7] was one of the state-of-the-art systems. Klatt's FO synthesis [8] (and FO synthesis in general) closely follows the source-filter model in which an impulse train is used to model voiced parts of a signal while white noise is the basis for the excitation of unvoiced sounds. Combining both excitation types adds breathiness to the generated speech signal and is used for the creation of fricatives. The filter itself can be realized through 2nd order IRR filters which represent the resonances of the vocal tract. In Klatt's FO synthesizer, a cascading of several sections (cascading model) is used to generate voiced sounds while adding several sections together (parallel model) is used to synthesize fricatives and stops.

While FO synthesizers are able to create "clean" sounding, intelligible speech, the achieved quality is far from that of natural speech. Reasons for this are the too simplistic models for both the excitation signal as well as the vocal tract [9]. Typical artifacts of FO synthesis are "metallic" voices that sound very artificial. Therefore, FO synthesizers cover the lower end of TTS quality of the synthesizers discussed in this chapter. Nonetheless, due to their parametric nature, FO synthesizers are easily

[1] Formants are peaks in the envelope of speech sounds and thus define the characteristics of a sound.

customizable, e.g., modifications of speech rate can be executed by simple parameter changes. This is an important feature, especially for visually impaired people who prefer acoustic cues rather than written text. By increasing the speech rate they are able to comprehend even faster [10]. Studies have shown that blind people are even able to comprehend ultra-fast synthetic speech at speaking rates above 17 syllables per second while a normal speech rate features only 3–5 syllables per second [11].

2.1.4.2 Concatenative Speech Synthesis

The development of concatenative speech synthesizers brought a great leap forward for the quality of TTS. By playing back prerecorded speech samples, the quality of synthesized speech was now theoretically able to equal that of natural speech. In reality, however, the quality of concatenative systems varies greatly: If the corpus contains units that are close to the text to be synthesized, the quality can be almost human-like. If that is not the case, severe distortions can occur. Thus, the inventory design is a crucial task. In most cases tailor-made speech corpora are recorded to fit the context of the systems.

The choice of unit size is one of the most important decisions for a TTS system. Finding the right balance between a preferably small footprint and a quality that is sufficient for the use case is a challenging task. At first sight it may seem that building a speech corpus based on the phonemes of a language, e.g., 42 phonemes for English, would cover every necessary sound. In practice, however, this would lead to major distortions at the transitions between the units. This is due to the coarticulation effect, i.e., the influence of a phone on its proceeding phone. Therefore, the smallest unit used in speech synthesis is the diphone, i.e., a unit that starts at the center of a phone and ends at the center of the following phone and thus covers the coarticulation. For the English language this leads to an approximate number of units of about 1500 diphones. Longer units, e.g., triphones, demisyllables, syllables etc., yield superior quality with the drawback of a much larger database. Building a TTS using, for example, a word inventory, would cause between 100 K and 1.5 M units [8].

But even with longer units synthesized speech still contains distortions that make it sound artificial. These distortions are due to two types of discontinuities [8]:

- *spectral discontinuities* occur when the formants of two aligned units do not match;
- *prosodic discontinuities* emerge from unfitting pitch curves at the concatenation point.

There are two different approaches to deal with these discontinuities:

- A technique called *Pitch Synchronous Overlap and Add (PSOLA)* manipulates the units in the time or spectral domain to correct these discontinuities.
- *Unit Selection (US) synthesis* uses an inventory that contains multiple instances of every unit with different pitch and formant curves and chooses the most appropriate one during concatenation.

Both approaches are discussed in the following sections.

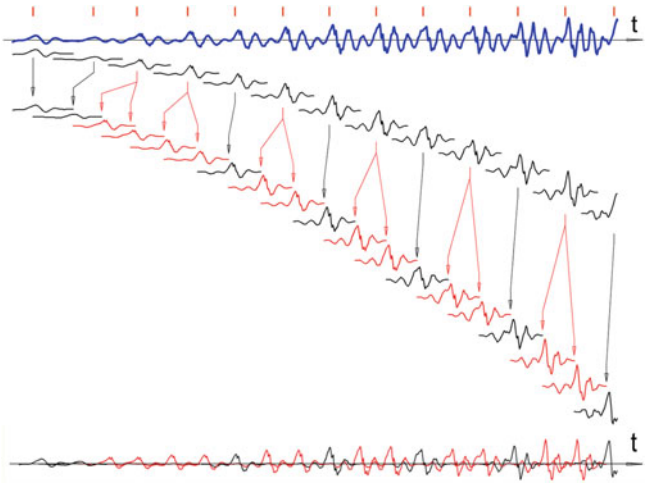


Fig. 2.3 Using PSOLA to increase the pitch of a signal by 50% [12]

Pitch Synchronous Overlap and Add (PSOLA)

In order to ensure a smooth transition, i.e., eliminating spectral and prosodic discontinuities, amplitude, duration and pitch of conjoined units have to be adjusted. The amplitude can easily be varied just by multiplying the waveform to the desired value. Applying PSOLA makes it possible to also change duration and pitch of each segment. In general, however, these modifications introduce distortions in each unit. To keep them as small as possible the most common PSOLA approach operates in the time domain (TD-PSOLA). This still causes degradations, but the gain of prosodically modified concatenated units usually is greater than the introduced distortions.

An example of how PSOLA works is shown in Fig. 2.3. Here the pitch gets increased by 50%. The upper part of the figure shows a diphone unit in blue with the red ticks marking its pitch cycles. A window function is applied which splits up the original diphone unit into smaller PSOLA units that cover two pitch cycles (waveforms in black). By doubling half of the PSOLA units (waveforms in red) while keeping the duration of the whole diphone unit constant the resulting diphone (waveform at the bottom of the figure) then features a pitch increased by 50%.

A similar technique is implemented in the MBROLA algorithm which has reached popularity due to its non-commercial availability. Since both approaches are very similar and are necessary for diphone concatenation they will both be referred to as (DI) systems.

Other, less common approaches apply PSOLA in the in the Frequency Domain (FD-PSOLA) or via Linear Prediction (LP-PSOLA). Given their high computational complexity, these approaches are rarely used.

Compared to the parametric approaches, PSOLA generates a decent quality which is moreover more independent from the context to be synthesized than the US

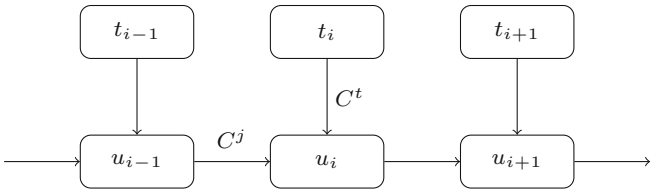


Fig. 2.4 Unit selection costs [15]

approach which will be described in the next section. Furthermore, PSOLA can operate on a reasonably small inventory. The downside of this technique is its sensitivity to accurate pitch marks. Moreover, changes in pitch are limited to ± 0.5 octaves and spectral discontinuities can only be corrected by the computationally more complex LP-PSOLA [12].

Unit Selection (US) Synthesizer

Even though PSOLA synthesis yields decent quality, the signal manipulations that come with this approach induce new distortions. In order to overcome the need for adjustments in speech units, Black et al. introduced the Unit Selection (US) approach for their synthesis platform CHATR [13, 14]. In contrast to PSOLA, a US synthesizer stores a huge database of speech units. Therefore, during the concatenation process a US system can choose from a large pool of candidate units. Each unit in the database has an attached feature vector that describes, e.g., phonetic and prosodic context (duration, power, and pitch) and acoustic join costs. This enables a synthesizer to pick units that guarantee smooth pitch and formant transitions. Hence, the main challenge for a US system is to select the most appropriate units.

The basis for the selection of units are the two cost functions [15, 16] shown in Fig. 2.4. Here, a unit from the database is denoted by u_i whereas a target unit is given by t_i. Thus the cost functions are defined as follows:

- *target costs* $C^t(u_i, t_i)$ are given as the difference between a database unit u_i and a target unit t_i which it should represent
- *join costs*[2] $C^c(u_{i-1}, u_i)$ estimate the quality of two joined units (u_{i-1} and u_i) and can be computed *offline*, before the actual synthesis takes place.

Thus, the process of unit selection means to find a balance between target and join costs without putting too much emphasis on one of them. An optimization of this problem leads to smooth transitions between all chosen units and thereby also to a minimum of discontinuities for pitch and formant curves of the respective speech database.

While a pure US system concatenates the units from the database without any signal manipulations, newer approaches shape both pitch and formant curves in

[2] Also known as *concatenation costs*.

order to ensure even smoother transitions. However, these manipulations occur far less often than in PSOLA synthesis.

The main advantage of the US approach is its perceptually high quality. However, to achieve high quality with a US system, the available units have to match the context to be synthesized. If this is not the case, glitches may occur and thus the quality will be severely degraded. Even a single glitch in an otherwise perfectly synthesized sentence will spoil the perceptual impression of a listener. Therefore, the quality of a US system strongly depends on the context and can vary drastically over time. Moreover, if there are no modifications to the units in the database, the synthesized speech is limited to the speaking style of the original recordings. More sophisticated approaches include a PSOLA algorithm to make slight changes to the units before concatenation in order to guarantee smoother transitions.

2.1.4.3 Statistical Parametric Speech Synthesis

The idea behind the US approach is to find the best fitting units and to connect them without any modifications of the signal so that no additional distortions are induced. Statistical parametric synthesis pursues a different objective which Black describes as *"generating the average of some set of similarly sounding speech segments"* [17], i.e., a natural speech database is converted into a statistical parametric model which can then be used for speech generation. In the following, the basic approach that is based on (HMMs) is introduced.

Hidden Markov Model (HMM) Synthesizer

HMM speech synthesis was first introduced by Tokuda for Japanese [18] and later adapted for the English language [19]. Figure 2.5 shows an overview of the two parts of an HMM synthesizer.

In the training part, excitation (e.g., $log\ F_0$) and spectral parameters (e.g., MFCC, their delta, and delta-delta coefficients) are extracted from a given database of natural speech, similar to a speech recognizer. These parameters are then used to train context-dependent HMMs. To appropriately realize the temporal structure of a speech signal, each HMM features state duration densities that are modeled by Gaussian distributions.

For the synthesis, the desired text is first converted into a context-based label sequence. According to this sequence a sentence HMM is constructed by concatenating the trained context-dependent HMMs. From this sentence HMM excitation and spectral parameters can be derived. In conjunction with the state durations those parameters are then used to synthesize the desired text.

This creates very smooth speech which does not feature the occasional glitches of US systems. Moreover, HMM synthesis guarantees robust quality. In addition, a system's footprint is very small and voice characteristics can be easily adjusted. However, when comparing HMM synthesizers to US systems, the overall quality

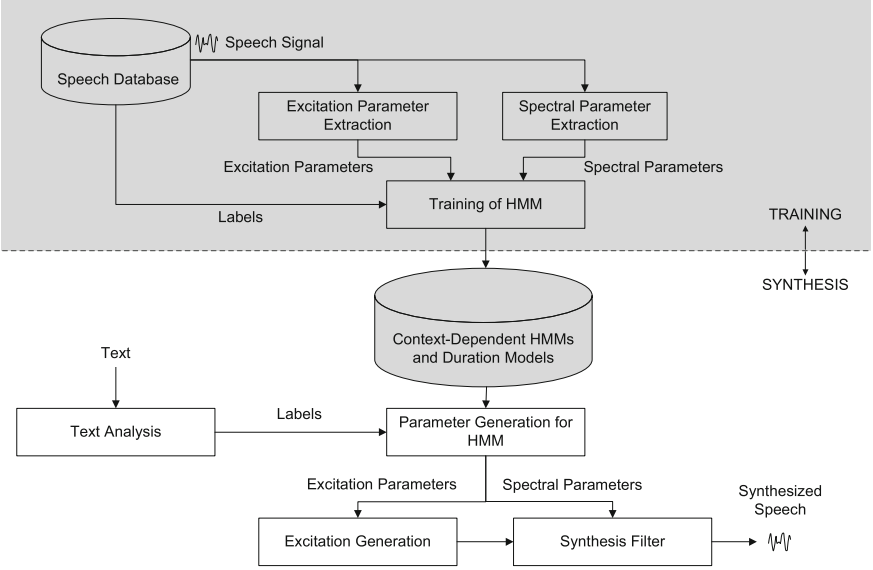


Fig. 2.5 Overview of a typical HMM based speech synthesis system [19]

still lags behind. This is mainly due to three factors: the used vocoder,[3] the modeling accuracy, and an over-smoothing of excitation and spectral parameters. These impairments are responsible for the characteristic buzzy and muffled sound of speech generated by HMM synthesizers [17, 20].

2.1.4.4 Hybrid Models

In an effort to combine the positive aspects of concatenative models with those of statistical parametric synthesizers, while bypassing their downsides, different approaches on hybrid models were developed.

In [21] a system is introduced that consists of a regular HMM in the training phase. For the actual speech generation part, however, these trained HMM are then used to select an optimal phone-sized unit sequence. Therefore, this model utilizes statistical criteria for the calculation of target and join costs.

In [22] an approach on multiform segment synthesis is proposed. The used segments are either *template segments* which consist of real waveforms or *model segments* which are abstractions of speech segments produced by an HMM. The target of this algorithm is to identify the best combination of model and template segments to synthesize a given text.

[3] A vocoder provides a parametric representation of a speech signal.

Table 2.1 Comparison of the main characteristics of different TTS approaches

	Formant Synthesis	PSOLA Synthesis	Unit-Selection Synthesis	HMM Synthesis
Sound Quality	poor	decent	very natural	very smooth and intelligible; not very natural
Variability of Quality	robust	robust	hit or miss*	robust
Perceptual Impression of Degradations	metallic, artificial voice	rough speech	sonic glitches, discontinuous speech	buzzy and muffled
Reasons for Degradations	simplistic model	frequent concatenations of speech units; modifications of PSOLA units	rough transitions between concatenated units; lack of adequate units in the database	vocoder; modeling accuracy; over-smoothing
Footprint	very small	small	large	small
Customizability	voice characteristics can be easily adjusted	fixed to speaker of database; changes in duration and pitch are feasible	fixed to speaker and speaking style of database	voice characteristics can be easily adjusted

Note: Hybrid systems are not included in this table due to a lack of available systems.
* strong dependency on corpus and the text to be synthesized

2.1.4.5 Advantages and Disadvantages of All Approaches

As mentioned in the previous section, different approaches on speech signal generation not only lead to different levels of quality but they also feature different perceptual impressions. An overview of the characteristics of these methods is given in Table 2.1.

2.2 The Mary Text-to-Speech System (MaryTTS)

This section presents an overview of the open-source TTS system MaryTTS [23]. MaryTTS is a Java-based speech synthesis platform that is able to generate diphone, unit selection, and HMM voices. It was developed in collaboration between the DFKI's Language Technology Lab[4] and the Institute of Phonetics[5] at the Saarland University. The basic architecture is shown in Fig. 2.6.

MaryTTS is able to process plain input text, SABLE text[6] and SSML[7] text. Initially, the input has to be converted intoMaryTTS's internal XML-based representation language (MaryXML). Plain text input can be directly converted, whereas SABLE and SSML have to pass through a parser before conversion. Then the tokenizer cuts the input text into tokens (i.e., words and punctuation marks) and the preprocessing unit performs a text normalization which converts numbers and abbreviations. Afterwards, a part-of-speech tagger adds word category information to each token.

From there on, two parallel branches exist: one where the prosody of the signal is modeled using GToBi,[8] a German version of ToBi (a set of conventions for transcribing the intonation and prosodic structure of speech). The second branch, first deals with inflection endings, i.e., correct endings are assigned to ordinals and abbreviations which were identified in the preprocessing, and second, pronunciation rules for each unit are looked up in a lexicon or generated by rule, if no lexicon entry exists.

Then the prosody information and the transcriptions of the input text are merged in the phonological processing module. Here the resulting phonological representation can be restructured on the basis of phonological rules. The generated information is again added to the MaryXML structure. The following module then transforms the symbolic information into the physical domain, i.e., GToBi and a set of duration rules add information on duration and pitch to the MaryXML structure.

---

[4]DFKI Language Technology Lab: http://www.dfki.de/lt/, last accessed 22.04.2016.

[5]Institute of Phonetics, Saarland University: http://www.coli.uni-saarland.de/groups/WB/Phonetics/, last accessed 22.04.2016.

[6]SABLE: https://www.cs.cmu.edu/~awb/festival_demos/sable.html, last accessed 21.04.2016.

[7]SSML: https://www.w3.org/TR/speech-synthesis/, last accessed 21.04.2016.

[8]GToBi: http://www.gtobi.uni-koeln.de/gm_gtobi_modell.html, last accessed 22.04.216.

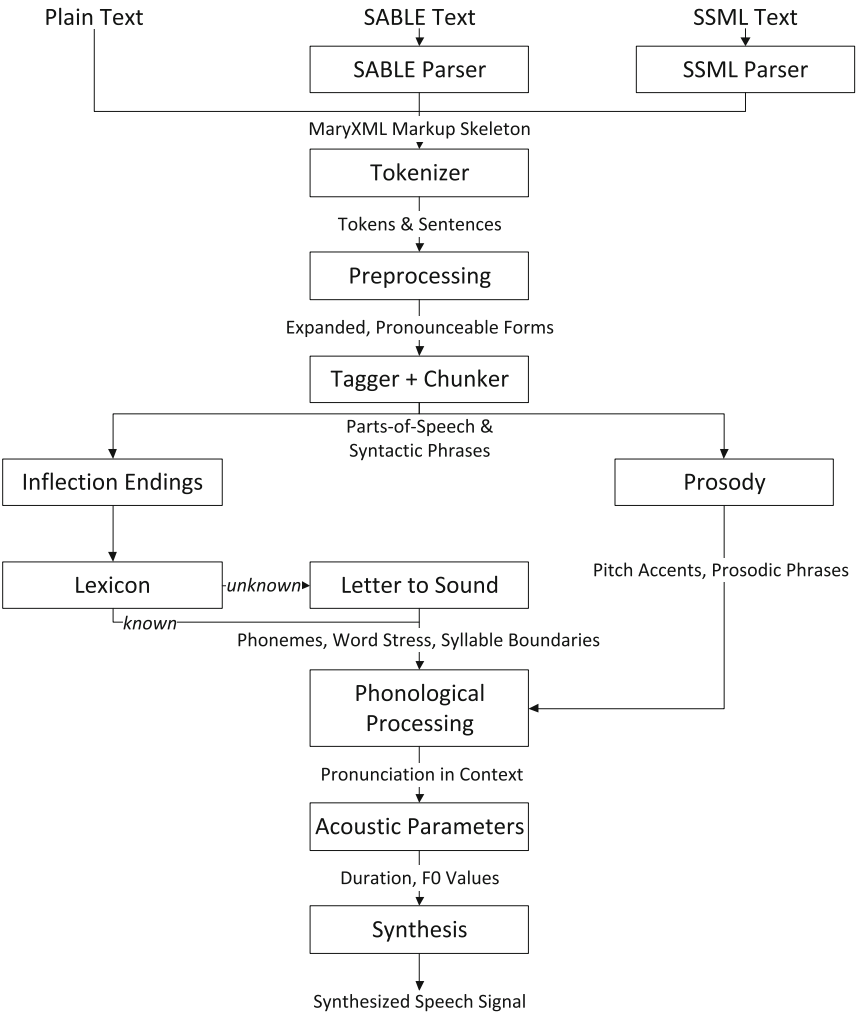


Fig. 2.6 Processing architecture within MaryTTS [23, 24]

Finally, the synthesis module utilizes the updated MaryXML information to generate a synthesized speech file. Therefore, an MBROLA based synthesizer can be applied to generate a diphone voice and more recent versions also allow the generation of unit selection as well as HMM voices.

References

1. Vary P, Heute U, Hess W (1998) Sprachsynthese. Digitale Sprachsignalverarbeitung. B.G. Teubner, Stuttgart, pp 465–497
2. Pfister B, Kaufmann T (2008) Sprachverarbeitung - Grundlagen und Methoden der Sprachsynthese und Spracherkennung. Springer, New York
3. Fujisaki H (1981) Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. Acoust Anal Physiol Interpret STL-QPSR 22:1–20
4. Fujisaki H (2004) Information, Prosody, and Modeling with Emphasis on Tonal Features of Speech. Speech Prosody 23:1–10
5. Flanagan JL, Ishizaka K, Shipley KL (1975) Synthesis of Speech from a Dynamic Model of the Vocal Cords and Vocal Tract. BELL Syst Tech J 54(3):485–506
6. Klatt DH (1987) Review of Text-To-Speech Conversion for English. J Acoust Soc Am 82:737–793
7. Klatt DH (1980) Software for a Cascade/Parallel Formant Synthesizer. J Acoust Soc Am 67:971–995
8. Huang X, Acero A, Hon H-W (2001) Spoken Language Processing - A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR, New Jersey
9. Taylor PA (2009) Text-To-Speech Synthesis. Cambridge University Press, Cambridge
10. Ley J (2016) Blind programmieren - Wenn der Computer schneller spricht, als ein Sehender hört. http://www.golem.de/news/blind-programmieren-wenn-der-computer-schneller-spricht-als-ein-sehender-hoert-1601-117767.html. Accessed 01 Nov 2016
11. Moos A, Trouvain J (2007) Comprehension of Ultra-Fast Speech - Blind versus Normally Hearing Persons. In: Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, pp 677–680
12. Institut für Kommunikationsforschung und Phonetik, Universität Bonn. Projekt MiLCA – Gesprochene Sprache. https://web.archive.org/web/20070613001637/, http://www.ikp.uni-bonn.de/dt/lehre/Milca/mmk/content/mmk_s322.xhtml. Accessed 10 Feb 2015
13. Black A, Taylor PA (1994) CHATR: A Generic Speech Synthesis System. In: Proceedings of the 15th International Conference on Computational Linguistics (COLING), Kyoto, Japan, pp 983–986
14. Campbell NW (1996) CHATR: A High-Definition Speech Re-Sequencing System. In: Proceedings of the 3rd ASA/ASJ Joint Meeting, Hawaii, USA, pp 1223–1228
15. Hunt AJ, Black AW (1996) Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In: Proceedings of the 21st International Conference on Acoustics, Speech and Signal Processing, Atlanta, Georgia, USA, pp 373–376
16. Campbell NW, Black A (1996) Progress in Speech Synthesis. Prosody and the Selection of Source Units for Concatenation Synthesis. Springer, New York, pp 279–291
17. Black A, Zen H, Tokuda K (2007) Statistical Parametric Speech Synthesis. In: Proceedings of the IEEE International Conference of Acoustics, Speech and Processing (ICASSP), Honolulu, Hawaii, USA, pp 1229–1232
18. Tokuda K, Yoshimura T, Masuko T, Kobayashi T, Kitamura T (2000) Speech Parameter Generation Algorithms for HMM-base Speech Synthesis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Instanbul, Turkey, pp 1315–1318
19. Tokuda K, Zen H, Black AW (2002) An HMM-Based Speech Synthesis System Applied to English. In: Proceedings of 2002 IEEE Speech Synthesis Workshop (SSW), Santa Monica, USA, pp 227–230
20. Zen H, Tokuda K, Black A (2009) Statistical Parametric Speech Synthesis. In: Speech Communication, vol 51. Elsevier Science Publishers BV, pp 1039–1064
21. Ling Z-H, Qin L, Lu H, Gao Y, Dai L-R, Wang R-H, Jiang Y, Zhao Z-W, Yang J-H, Chen J, Hu G-P (2007) The USTC and iFlytek Speech Synthesis Systems for Blizzard Challenge 2007. In: Proceedings of the 3rd Blizzard Workshop in conjunction with the 6th ISCA Workshop on Speech Synthesis (SSW), Bonn, Germany

22. Pollet V, Breen A (2008) Synthesis by Generation and Concatenation of Multiform Segments. In: Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2011), Brisbane, Australia, pp 1825–1828
23. Schröder M, Trouvain J (2001) The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. In: Proceedings of the 4th ISCA Workshop on Speech Synthesis
24. DFKI GmbH. Mary Text-to-Speech: Architecture Walkthrough, 20 April 2016. http://mary.dfki.de/documentation/module-architecture.html

Chapter 3
Auditory and Instrumental Quality Evaluation Metrics

This chapter starts off with a definition of *quality* and its related aspects followed by a detailed description of the quality perception and assessment process inside a listener. Moreover, it gives an overview of auditory quality assessment methods for synthesized speech. Therefore, a taxonomy of assessment tasks and techniques for TTS signals is introduced. A deeper insight into a variety of auditory tests for the evaluation of global quality as well as specific quality aspects is presented in Sect. 3.3.

Until this point, there are no standardized methods for the instrumental quality assessment of TTS signals, therefore, Sect. 3.4 introduces common approaches in the related field of instrumental quality assessment of distorted speech signals. Their benefit for the quality assessment of synthetic speech will be presented in Chap. 6.

3.1 What Is Perceptual Quality?

First, relevant definitions of quality and its related concepts are outlined, and secondly, the quality perception and quality formation process of a listener is examined.

Depending on the context it is used in, the definitions of the term *quality* vary. In general it specifies how good or bad something is. A definition for the field of speech perception, that will be used for the remainder of this book, is introduced by Jekosch [1]: **Quality** is the

> "*result from a perception and assessment process in which a subject compares the perceived features of an entity with the individual expectations, and/or appropriate requirements, and/or social demands*."

F. Hinterleitner, *Quality of Synthetic Speech*, T-Labs Series
in Telecommunication Services, DOI 10.1007/978-981-10-3734-4_3

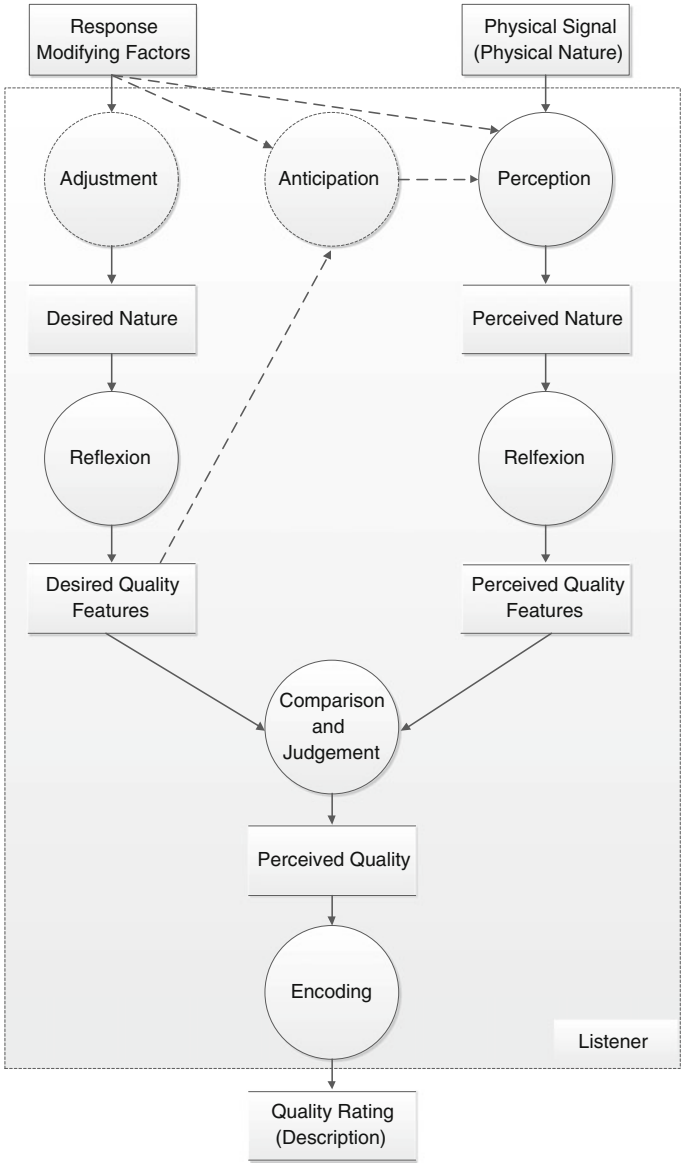


Fig. 3.1 Schematic description of a quality event [2, 3]. *Note circles* represent perceptual processes, two *parallel horizontal lines* represent storages for different types of representations, and boxes outside of the person represent input and output information. Moreover, *continuous lines* represent direct input to the perception process while *dashed lines* represent indirect input

In which **entity** is a

> "*material or immaterial object under observation*" [1]

and a **feature** is defined as a

> "*recognizable and nameable characteristic of an entity*" [1].

Therefore, a **quality feature** is

> "*a recognized and designated characteristic of an entity that is relevant to an entity's quality*". [1]

These quality features will be referred to as **perceptual quality dimensions** in the remainder of this book.

With these definitions in mind, the quality perception and assessment process of a listener can be described as in Fig. 3.1. Even though, different definitions of the terms *assessment* and *evaluation* can be found in the literature, both terms will be used synonymously in this book.

This process is triggered by a physical signal which is, in the context of this book, a speech sample generated by a speech synthesizer. This signal is perceived and the listener reflects upon this input. This step creates *perceived quality features*. In parallel a similar process is executed on the listener's experiences which leads to *desired quality features*.

A comparison and judgment between the desired and perceived quality features yields the perceived quality. However, these desired quality features do not only affect the comparison and judgment step, they can also influence the perception of the input signal through anticipation of the listener. The same effect can arise from response modifying factors that originate outside of the listener, e.g., background noise, situation and application-specific content, etc.

However, since the perceived quality only exists inside the listener, an encoding has to be executed in order to make the perceived quality visible. In the field of speech synthesis this encoding is performed via listening tests. Depending on the quality feature of interest, different types of listening tests can be applied. A deeper insight into several different test mechanics is discussed in the following sections. A more elaborated version of the quality perception and assessment process is given in [4].

3.2 Taxonomy for the Quality Assessment of Synthetic Speech

In Fig. 3.2 a modified version of the taxonomy of assessment tasks by van Bezooijen and van Heuven [5] is displayed. Hereby every path from the top to the bottom yields a meaningful specification of a test scenario. The four fundamental dichotomies are described in detail in the following subsections. Moreover, in the subsequent sections a more detailed view on auditory (Sect. 3.3) as well as instrumental measures (Sect. 3.4) is given.

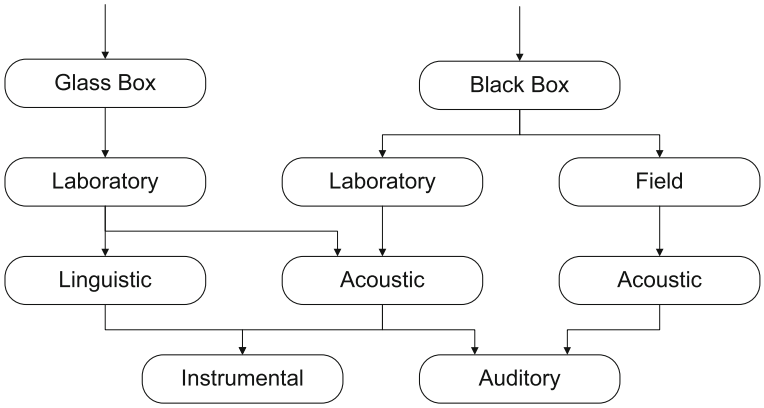


Fig. 3.2 Taxonomy for the evaluation of synthesized speech (modified version of [5])

3.2.1 Glass Box Versus Black Box

As introduced in Fig. 2.1 from Chap. 2.1, every TTS system consists of a multitude of different modules. Developers of such systems usually have total control over every module. Therefore, in a *glass box* approach it is possible to modify a certain module while keeping all others unchanged. Thus, you can, for example, check whether an updated G2P module outperforms the previous version.

In the case of commercial systems the user commonly has no control over certain modules. Hence the only part of the TTS system that can be evaluated in such a *black box* approach is the synthesized speech signal as a result of a sequence of all involved modules.

Therefore, *glass box* approaches are used in order to receive diagnostics on individual modules while *black box* approaches assess the performance or quality of the whole system.

3.2.2 Laboratory Versus Field Studies

Laboratory tests feature a highly controlled environment, e.g., listening to stimuli over head phones in a soundproof booth. This guarantees results that contain a very low level of noise and that are comparable between different studies and laboratories. The drawback that comes with *laboratory testing* is the artificial situation test participants find themselves in. A test setup that brings participants into a real life situation in which they would normally use the to-be-tested TTS system can make up for some of the introduced artificiality but in order to record data that actually reflects the real use of a TTS system a *field test* has to be conducted.

On the other hand, the lack of control in a *field test* over the test setup yields results that are not reproducible and can strongly differ between test participants, e.g., when evaluating a Spoken Dialog System (SDS) on a cell phone while walking along a road, the sound of passing cars can influence the intelligibility of one participant while another participant might use the SDS in a quiet environment without induced noise.

3.2.3 Linguistic Versus Acoustic

Figure 2.1 shows that the setup of a TTS system consists of four main parts: Natural Language Processing (NLP), prosody generation, concatenation, and speech generation. *Linguistic testing* involves evaluating units from the NLP part of a system, e.g., checking if numbers are processed correctly or if words that are not part of the local lexicon are transformed into a meaningful phonological representation. Besides the linguistic testing, the NLP unit can also be tested at the acoustic level by evaluating the generated speech signal. Therefore, only errors of the NLP unit that actually lead to audible impairments affect the evaluation. However, a detected error can not only be traced back to the NLP unit, it could also have its origin in the prosody generation, the concatenation, or the actual speech generation algorithm.

3.2.4 Auditory Versus Instrumental

The common way to assess the quality of TTS systems is to invite human listeners to evaluate a system based on their *auditory* impression. Even though, human listeners generate noisy data, due to, e.g., different preferences or hearing abilities, and listening tests are very time consuming and cost intensive, to date it is still the only way to reliably assess the quality of TTS systems. In order to overcome these restricting factors research has started to develop *instrumental* measures that can predict the quality a user would perceive when using this device.

3.3 Auditory Quality Evaluation Metrics

Judgment tests are the most common approach to auditory quality evaluation. Test participants are asked to give ratings on different attribute scales or compare two or more stimuli with each other. Scales can either refer to global aspects of the TTS signal such as the overall quality impression or to specific quality features of a system like the rhythm of the generated voice. Alternatively, a system can also be evaluated via *functional* tests which indirectly grade a system. Hereby, test participants do not use scales to rate, e.g., the intelligibility, but solve a task from which the intelligibility

can be derived. This bypasses user ratings which are often noisy due to, e.g., different preferences of every single participant or contrasting interpretations of rating scales.

3.3.1 Functional Tests[1]

The primary goal of speech is to serve the communication of information. Thus, a key requirement to synthetic speech is that it is intelligible, i.e., that the linguistic information can be discerned by the listener. However, intelligibility is commonly not enough, and synthetic speech - even if it was 100% intelligible - is not perceived as human-like, mostly due to a lack in naturalness, which largely determines the overall quality. In the following, functional tests to evaluate the intelligibility of a TTS signal are described.

3.3.1.1 Intelligibility on Word Level

Since consonants are more problematic to synthesize, the test material of intelligibility tests on a word level mainly focuses on them. In the following the simple Diagnostic Rhyme Test (DRT), its successor the Modified RhymeTest (MRT), and the more complex CLuster IDentification Test (CLID) are presented. See [5] for a detailed overview of these and further intelligibility tests.

Diagnostic Rhyme Test (DRT)

The DRT [7, 8] uses a fixed set of meaningful words to test for intelligibility of the initial consonant. The examined items are of the form CVC, i.e., an initial **C**onsonant followed by a medial **V**owel followed by a final **C**onsonant. One auditory stimulus and one word pair as answer option are presented at a time. The word pair consists of two words which differ only in the initial consonant, e.g., *dune* and *tune*. The listener marks which of those two words he thinks was presented. For each of six categories (i.e., voicing, nasality, etc.) specific word pairs are chosen. The intelligibility is expressed by the total error rate or the percentage of correct initial consonants.

Modified Rhyme Test (MRT)

The MRT [8] is an extension of the DRT which is able to test for initial as well as final consonant intelligibility. The test items consist of sets of six one-syllable words. Half of the set differs in initial while the other half differs in final consonant, e.g., *bus*, *bug*, *but*, *buff*, *bun*, and *buck*. The listener has to identify which of the six items in the list was presented. The intelligibility is given as initial and final consonant error rate, or as overall percentage of correct consonants.

[1] The content of this section has previously been published in a slightly different version in [6].

CLuster IDentication Test (CLID)

The previous two approaches are fast, reliable, easy to administer, and no training of the participants is required. However, the intelligibility may be overestimated since the participants can chose words from the presented categories; thus, there is a probability to select the right word by chance. In addition, the words presented in a set are meaningful, but not equally frequent in a language; thus, there is an inherent distortion of the participants' responses, which is due to their knowledge of the language. A more balanced approach in intelligibility testing which overcomes the limitations of rhyme tests is the CLID test [9]. On the basis of linguistic statistics gathered from speech databases containing monosyllables, an automatic word generator is used to create phonotactically correct monosyllables of the type $C^i V C^j$ (where i and j represent the number of initial and final consonants, respectively). These, mostly non-sense words, are evaluated in an open response test where participants have to accomplish a task like:

"Please write down what you have heard in such a way that another person would read it aloud in the same way as you heard it originally." [10]

This guarantees that the participants are not biased by any given response categories.

Subsequently, the recognition rates can be computed on word and on cluster level (initial, medial, and final consonant).

3.3.1.2 Intelligibility on Sentence Level

While intelligibility tests on word level lead to very diagnostic results, tests on sentence level are more similar to speech perception in normal communication situations.

Semantically Unpredictable Sentences (SUS)

In the most common test methodology, short semantically unpredictable sentences [11] are used, i.e., they do not occur in real life. The advantage of SUS results from the fact that, even though the syntax of each sentence is correct, the whole sentence does not make sense, thus the listeners can not rely on a semantic context. This increases the importance of the acoustic characteristics of the TTS signal. A SUS test uses five different syntactic structures, e.g., *subject - verb - object* could yield the sentence *"The strong way drank the day."* [12]. Ten sentences for each of the five categories are produced and assessed in random order in a listening test.

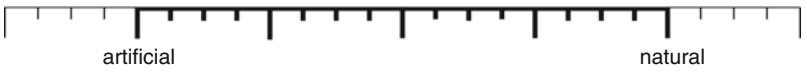


Fig. 3.3 Naturalness scale with separate scale and end points [14]

3.3.2 Judgment Tests[2]

While the previous section discussed evaluation methods focused on functional testing, i.e., intelligibility was measured by how well listeners correctly identify words and phrases, the current section addresses judgment tests. Since different TTS approaches lead to different distortions in the synthetic speech signal (see Table 2.1), the quality of TTS signals is generally of multidimensional nature. Therefore, the judgment tests highlighted in this section depict a multidimensional analysis of the characteristics of a TTS system. Here listeners are instructed to rate stimuli along a number of attribute scales determining specific quality features of a system or comparing two or more stimuli concerning one quality feature and thus yielding very analytic results [5]. The most simple way would be to ask listeners to rate, e.g., the naturalness of TTS stimuli, on a 5-point Absolute Category Rating (ACR) scale and to build a Mean Opinion Score (MOS) of all ratings. In order to make the task for the test participants easier, continuous scales with separate scale and end points can be used as proposed by Bodden and Jekosch [14] (see Fig. 3.3). This reduces the impact of two effects often noticed when subjects use rating scales: first, most subjects avoid to give ratings on the end of scales because they expect even better or worse stimuli to come. Moreover, scales with fixed end points make it hard for subjects to differentiate between stimuli of an either very good or very bad quality. All the studies presented in the Chaps. 4 and 5 assess quality via such continuous scales.

The two main approaches to analyzing perceptual quality dimensions with the help of human listeners are discussed in the following.

3.3.2.1 Semantic Differential (SD)

In a Semantic Differential (SD), pre-defined attribute scales featuring antonym pairs at the ends of each scale are used to measure the auditory impression of the listeners. This guarantees a direct relationship between the used attribute scales and the perceptual quality dimensions that can be derived from them. Therefore, the results are usually easy to interpret. On the downside, the ratings of the test participants are always limited to the set of presented scales. If a perceived quality feature can not be expressed by any of the presented scales, this information will be lost. Thus, it is crucial to carefully choose a set of scales for the listening test. To reduce the

[2]Parts of the content of this section have previously been published in a slightly different version in [13] and [6].

influence of the test designers to a minimum, a suitable set of scales can be developed through several pretests, i.e., the goal of a first pretest is to collect attributes and corresponding attribute scales which describe the auditory impression of the listeners; in a second pretest, this set of attribute scales can be reduced to a final selection of scales.

While individual scales can target very analytic quality features of a system, e.g., *disturbances* in the signal or the *pronunciation* of the speaker, there are also items that assess the general satisfaction of a user, e.g., *overall impression* or *acceptability* of the system. Such a standardized protocol is discussed in the following paragraph.

ITU-T Rec. P.85

A judgment test for the auditory assessment of TTS systems which is close to an SD was published in the ITU-T Rec. P.85 [15], *"A method for subjective performance assessment of the quality of speech voice output devices"* in 1995. In contrast to an SD the quality is rated on 5-point ACR scales with quality describing labels at every step. Even though this protocol is now over 20 years old and was thus developed far before current state-of-the-art TTS systems emerged and triggered a huge leap in the quality of synthesized speech, it is still the only standardized auditory assessment technique for TTS systems.

It is recommended to assess the quality of telecommunication services which provide synthetic speech output, be it via concatenating sentences, parts of sentences, or via TTS synthesis. A test according to this recommendation should include at least 5 different synthesis systems and at least one reference condition (e.g., natural speech corrupted with a degradation or a known synthesis system). The test is designed in a way that requires the listener to concentrate on the content of each message, i.e., before rating each stimulus on 5-point ACR scales the test participants have to answer several questions concerning the information contained in the stimulus. Thus, the messages should contain a fixed part that is specific to the addressed use case and a variable part that differs between stimuli. The duration of the stimuli should be between 10 and 30 s. A possible test phrase in a mail order shopping scenario would be:

Mr. Zimmerman, you have ordered running shoes, color: white, size: 41, price: 61€. They will be delivered to you in 9 days.

The name, the shoe color, it's size, the price and the delivery date are variable parts that can be requested from the listener.

Each stimulus is presented twice consecutively. After the first presentation the test participants answer questions on the information contained in the stimulus, and after the second presentation they judge the quality of the presented stimulus on different rating scales. Each listening test consists of 3 sessions: a training session, where the test participants should get used to the test procedure, the environment, and get an impression of the quality range of the stimuli in the test, and two main sessions that either use scales concerning the intelligibility or the quality of the synthesizers. The intelligibility scales test the *listening effort*, *comprehension problems* and the

articulation while the quality scales assess *pronunciation*, *speaking rate* and *voice pleasantness*. Furthermore, each session also includes the scales *overall impression* and *acceptance*.

An extension of P.85 towards the assessment of synthesized audiobooks was proposed in [16]. Therefore, scales that are relevant when synthesizing the content of books were included, e.g., scales that assess *intonation* and *emotion*. The development of these additional attribute scales will be presented in Sect. 5.1.1.

Given the complexity of this test, the question arises why an evaluation via P.85 should be preferred over a simpler intelligibility test or a MOS test addressing overall quality. These three methods were compared in [17]. The results showed that high intelligibility ratings do not necessarily come together with high ratings on the naturalness and overall impression scales. Moreover, the best ranked synthesizer in the naturalness test did not get the best rating on the overall quality scale. Thus, while a simple MOS naturalness test can give a basic overview of the quality of synthesizers, the P.85 yields far more fine grained information about the performance of a system.

However, this evaluation protocol has also been heavily criticized. In [18] the authors suggested extensive modifications:

- natural speech reference stimuli should not be included because they affect the mean ratings of TTS systems and thus tend to diminish the differences between them
- items that assess *naturalness*, *audio flow*, and *ease of listening* should be included
- the item *speaking rate* should be modified so that one end of the scale represents an optimal speed while the other end indicates extremely slow or extremely fast speech

One of the main points that has already been addressed in [17] is the fact that many of the recommended scales are highly correlated. Thus, some of the scales mainly measure the same perceptual impression, when they should actually cover all perceptual quality features of synthetic speech. These features are not necessarily orthogonal but certainly exhibit smaller correlations than the scales from P.85.

Given these points of criticism and the fact that all state-of-the-art synthesis techniques (unit selection, HMM, and hybrid synthesis) have been developed far after the publication of this protocol, the need for a method that captures all artifacts that come along with current TTS systems arises. The development of such an SD protocol is presented in Sect. 4.2. A similar approach was used to generate a test protocol for the auditory evaluation of synthesized audiobooks. The approach as well as the resulting protocol are displayed in Sect. 5.1.1.

3.3.2.2 Pairwise Comparison (PC) and Multidimensional Scaling (MDS)

In comparison to an SD, the Pairwise Comparison (PC) approach with subsequent Multidimensional Scaling (MDS) is solely based on the perceptual impression of the listener and not on any given rating scales. Participants are instructed to rate the

similarity of one quality feature of pairs of speech signals, e.g., similarity in naturalness. Therefore, every stimulus in a set of n stimuli has to be compared to all remaining $n-1$ stimuli. The outcome is a matrix that represents the similarity between all stimuli [19]. Via an MDS algorithm, the dimensionality of this matrix can then be reduced until the solution is interpretable but still represents the observed stimulus space. However, since a complete comparison of all stimuli leads to $\frac{n(n-1)}{2}$ comparisons and a listening-test duration of several hours per subject, this approach is hardly deployable with larger sets of objects. For these cases, Tsogo [20] proposed a Sorting Task (ST). Here, subjects are instructed to build groups of stimuli that are similar to each other while being different from the stimuli in other groups. This yields one incidence matrix per subject from which a similarity matrix can be derived that can be further processed as described above. Even though the MDS approach has the advantage that the participant's ratings are not influenced by given rating scales, its major drawback is the interpretability of the resulting dimensions. MDS dimensions as such give no indication on their interpretation, thus, additional knowledge about the nature of the stimuli has to be obtained, e.g., via expert listening, rating scales or measures derived from the synthesis system.

An extensive study on perceptual quality dimensions using the ST and a subsequent MDS on a dataset of various different TTS systems is presented in Sect. 4.3.

3.4 Instrumental Quality Evaluation Metrics

Given the time, effort and costs auditory listening tests implicate, the need for instrumental measures arises. This is especially true for developers who need to evaluate their systems on a regular basis. Unfortunately, to date there are no standardized measures that were developed for the quality assessment of synthesized speech. However, several measures exist that are able to predict the quality of encoded speech signals. Therefore, the distortions these measure evaluate do not stem from the concatenation of speech units or the generation through a parametric synthesizer but rather they are introduced through the transmission of signals over a telephone channel.

Given the fact that in both cases distortions in speech signals need to be assessed, similar instrumental measures could be used. Therefore, this section gives an overview of instrumental measures that are used for the quality prediction of distorted speech signals. Two main approaches exist: firstly, reference based-measures use an undistorted source signal as a comparison standard while reference-free measures are used if such an undistorted signal is not available. Some of these measures have already been applied on synthetic speech. The achieved results will be discussed in Sect. 6.

3.4.1 Reference-Based Measures[3]

Several reference-based models have been developed to predict distortions in natural speech introduced by transmission channels of telephone networks. Therefore, they use a clean reference signal and evaluate the perceptual distance to the distorted signal. In the following 3 different measures are introduced.

3.4.1.1 ITU-T Rec. P.862 (PESQ)

The ITU-T Rec. P.862 [22] for speech quality assessment of narrow-band telephone networks was released in 2001. With P.862.2 (WB-PESQ) [23] Perceptual Evaluation of Speech Quality (PESQ) was extended to wide-band speech signals. Figure 3.4 shows an overview of the PESQ model (the wide-band version basically follows the same stages).

Firstly, several preprocessing steps take place: level alignment, IRS [24] filtering mimicking the transfer characteristics of traditional handset telephones, voice activity detection, followed by a time alignment between the original speech signal and the degraded signal. Secondly, both signals are converted via a perceptual model. Therefore, the signals are transformed into internal representations which emulate the ones of the human auditory system. This step comprises spectrum computation, band integration, frequency and gain compensation, as well as loudness compression. In the final step a distance between the 2 perceptually transformed signals is computed. This includes disturbance, masking and asymmetry computation, frequency and time averaging, and results in a final PESQ quality score. Subsequently, a mapping function transfers the results onto a MOS scale.

For WB-PESQ the IRS receive filter was replaced with a bandpass filter in the range 200–8000 Hz and the mapping function was adjusted.

3.4.1.2 DIAL

Due to the ongoing advancements in telephone networks the ITU-T started a standardization project to come up with a new reference-based measure. As one of the contestants the Diagnostic Instrumental Assessment of Listening quality (DIAL) model [25, 26] was proposed. Its framework is presented in Fig. 3.5.

The preprocessing consists of active speech level normalization, voice activity detection, and the time alignment from the PESQ model. The main model combines 3 building blocks: the *core model* estimates non-linear degradations introduced by speech processing systems. The *dimension estimators* introduced by Wältermann [27] cover the perceptual dimensions directness/frequency content (DFC), noisiness (N), loudness (L) and continuity (C) of speech signals. Finally, the *cognitive model*

[3] Parts of the content of this section have previously been published in a slightly different version in [21].

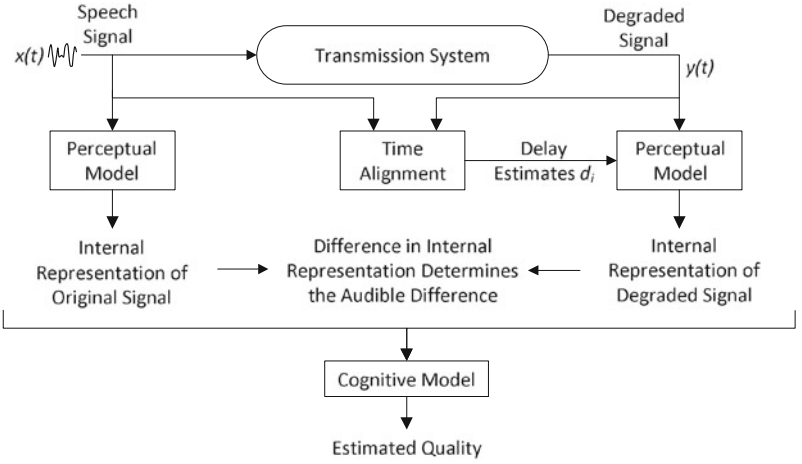


Fig. 3.4 Overview of the PESQ model

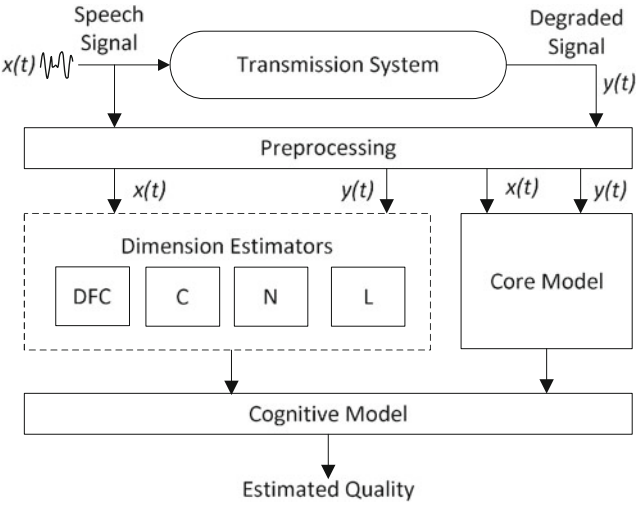


Fig. 3.5 Overview of the DIAL model

uses the so far computed score for each dimension to simulate the cognitive process of a human listener. Besides narrow-band (300–3400 Hz) and wide-band (50–7000 Hz) signals, the DIAL model can also be used in a super-wide-band (50–14000 Hz) context.

3.4.1.3 ITU-T Rec. P.863 (POLQA)

The Perceptual Objective Listening Quality Assessment (POLQA) model [28] was released in 2011 and defines an algorithm for speech quality assessment of state-of-the-art telephony systems. Being the successor of PESQ, its framework is very similar to the one in Fig. 3.4. Moreover, it allows quality estimation of super-wide-band speech signals and the assessment of networks and codecs that introduce time warping.

Firstly, a time alignment takes place. Therefore both signals are split up into small chunks. The delay between each chunk of the reference signal and the distorted signal is calculated and used to adjust the sampling rate of the degraded signal. Subsequently, both signals are transformed into an internal representation of the human auditory system similar to the PESQ model. Additionally, low levels of noise in the reference signal which might lead to the impression of a signal of minor quality are eliminated. This represents the idealization process that subjects usually go through during their quality judgment. Then the cognitive model uses 6 indicators (frequency response, noise, room reverberation and 3 indicators that describe the internal difference in the time-pitch-loudness domain) to compute a final estimated MOS score.

PESQ and POLQA have already been tested for synthetic speech impaired by coding and packet loss [29–31]. Moreover, PESQ has been tested on narrow-band TTS signals [32]. Its impressing performance on single-word TTS signals lead to the approach of verifying these results and comparing the performance of PESQ with the performance of the two other state-of-the-art algorithms DIAL and POLQA. Even though all of these algorithms are designed for a different domain their performance on synthetic speech can give an impression of what is possible in the area of quality estimation of TTS systems. The performance of PESQ, DIAL, and POLQA on synthetic speech will be discussed in detail in Chap. 6.

3.4.2 Reference-Free Measures

In most cases, especially when evaluating TTS systems, a “clean reference”, i.e., an undistorted natural reference signal, is not existent. Therefore, reference-free methods have to be applied. In general, such a measure uses the given signal to estimate an undistorted reference signal. This signal can then be used as a reference, just like in the reference-based approaches described in the previous section. The following section discusses a reference-free approach that is currently used for the instrumental assessment for distorted speech signals.

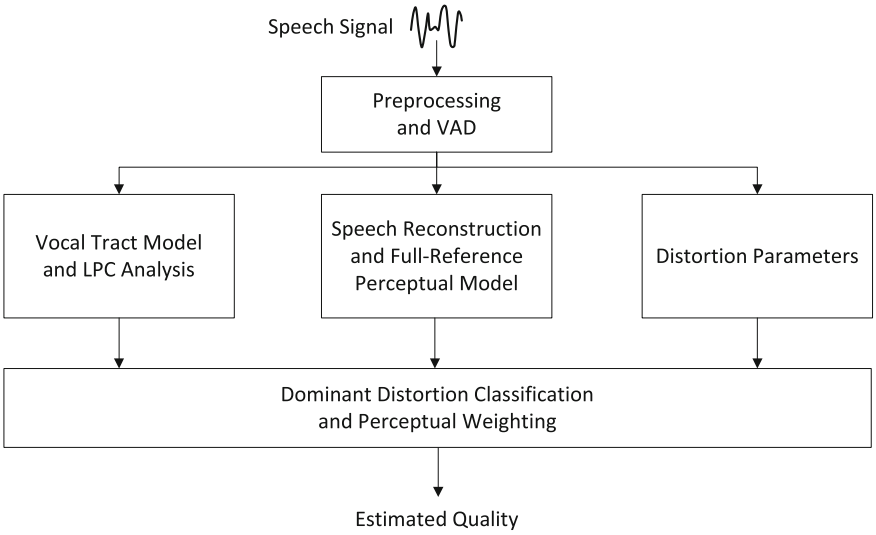


Fig. 3.6 Overview of the P.563 model

3.4.2.1 ITU-T Rec. P.563

In 2004 the ITU-T introduced the Rec. P.563 [33] as a standard for the quality assessment of narrow-band telephony applications. An overview of its structure is given in Fig. 3.6. Its main parts are described in the following:

- *Preprocessing*
  In the preprocessing step of P.563 the input signal is normalized to $-26\,\mathrm{dBov}$. Subsequently two additional versions of the signal are created to cover the characteristics of different types of terminals, e.g., hands-free phones and mobile phones. Moreover, a Voice Activity Detection (VAD) is performed to omit unvoiced parts of the signal.

- *Pitch Synchronous Vocal Tract Model and LPC Analysis*
  Within the vocal tract model, P.563 tries to simulate the human speech production system. Therefore, a series of tubes with varying diameters model the human vocal tract. Given the fact that the inertance of the human vocal tract prohibits abrupt changes, a distortion in the signal is detected if sudden changes in tube diameter occur.

- *Speech Reconstruction and Full-Reference Perceptual Model*
  A speech reconstruction module generates a quasi-clean speech signal from the distorted input signal. From there on a reference-based module, similar to PESQ and POLQA, is used to estimate the quality of the degraded input signal.

- *Distortion-Specific Parameters*
  This unit detects distortions which are specific for certain degradations, e.g., noise, temporal clipping, robotization.

- *Dominant Distortion Classification and Perceptual Weighting*
  Since human listeners mainly focus on the most prominent distortion when multiple distortions are present at the same time, the distortion-specific parameters from the previous step are then weighted linearly according to their influence on speech quality.

Finally, an estimated quality value is computed. A more detailed description of P.563 and its internal parameters can be found in [34].

The P.563 algorithm, together with the similar approaches ANIQUE [35], ANIQUE+ [36] and an algorithm presented by the company Psytechnics [37, 38], have already been tested on synthetic speech [39–42]. The achieved results as well as further research in this area is presented in Chap. 6.

References

1. Jekosch U (2005) Voice and Speech Quality Perception - Assessment and Evaluation. Springer, New York
2. Jekosch U (2004) Basic Concepts and Terms of "Quality", Reconsidered in the Context of Productsound Quality. Acta Acustic United Acust 90(6):999–1006
3. Raake A (2004) Assessment and Parametric Modelling of Speech Quality in Voice-Over-IP Networks. Ph D. Thesis, Ruhr-Universitt, Bochum
4. Raake A, Egger S (2014) Quality of Experience - Advanced Concepts, Applications and Methods. In: Quality and Quality of Experience, Springer, Switzerland, pp 11–34
5. van Bezooijen R, van Heuven V (1997) Assessment of Synthesis Systems. In: Handbook of Standards and Resources for Spoken Language Systems, Mouton de Gruyter, Berlin, pp 481–563
6. Hinterleitner F, Norrenbrock C, Möller S, Heute U (2014) Text-To-Speech Synthesis. In: Quality of Experience - Advanced Concepts, Applications and Methods, Springer, Berlin, pp 179–194
7. Voiers WD (1977) Diagnostic Evaluation of Speech Intelligibility. In: Hawley ME (ed) Speech Intelligibility and Speaker Recognition, vol 2 of Benchmark papers in Acoustics. Dowden, Hutchinson, and Ross, Stroudburg, pp 374–384
8. ASA S3.2-2009 (2009) American National Standard Method for Measuring the Intelligibility of Speech over Communication Systems. American National Standards of the Acoustical Society of America, Washington
9. Jekosch U (1992) The Cluster-Identification Test. In: 2nd International Conference on Spoken Language Processing (ICSLP), Banff, Alberta, Canada, pp 205–208
10. Kraft V, Portele T (1995) Quality Evaluation of Five German Speech Synthesis Systems. Acta Acust 3:351–365
11. Benot C, Grice M, Hazan V (1996) The SUS Test: A Method for the Assessment of Text-To-Speech Synthesis Intelligibility Using Semantically Unpredictable Sentences. Speech Commun 18(4):381–392
12. Jekosch U (1993) Speech Quality Assessment and Evaluation. In: Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech). Berlin, Germany, pp 1387–1394

13. Hinterleitner F, Norrenbrock C, Möller S (2013) Is Intelligibility Still the Main Problem? A Review of Perceptual Quality Dimensions of Synthetic Speech. In: Proceedings of the 8th ISCA Speech Synthesis Workshop (SSW 2013), pp 167–171
14. Bodden M, Jekosch U (1996) Entwicklung und Durchfhrung von Tests mit Versuchspersonen zur Verifizierung von Modellen zur Berechnung der Sprachbertragungsqualitt. Technical report, Institut fr Kommunikationsakustik, Ruhr-Universitt, Bochum
15. ITU-T Rec. P.85 (1994) A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices. International Telecommunication Union, Geneva
16. ITU-T Contribution COM 12-37 (2013) Proposal for an Appendix to Rec. P.85 of the Evaluation of Speech Output for Audiobook Reading Tasks. Deutsche Telekom AG (Authors: Möller S, Hinterleitner F), ITU-T SG12 Meeting, Geneva, Switzerland, 19 - 28 March 2013
17. Sityaev D, Knill K, Burrows T (2006) Comparison of the ITU-T P.85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems. In: Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech), pp 1077–1080
18. Viswanathan M, Viswanathan M (2005) Measuring Speech Quality for Text-to-Speech Systems: Development and Assessment of a Modified Mean Opinion Score (MOS) Scale. Comput Speech Lang 19:55–83
19. Borg I, Groenen P (2005) Modern Multidimensional Scaling - Theory and Applications, 2nd edn. Springer Series in Statistics, New York
20. Tsogo L, Masson MH, Bardot A (2000) Multidimensional Scaling Methods for Many-Objects Sets: A Review. Multivar Behav Res 35:307–319
21. Hinterleitner F, Zabel S, Möller S, Leutelt L, Norrenbrock C (2011) Predicting the Quality of Synthesized Speech Using Reference-Based Prediction Measures. In: Proceedings of the 22th Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Aachen, Germany, pp 99–106
22. ITU-T Rec. P.862 (2001) Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. International Telecommunication Union, Geneva
23. ITU-T Rec. P.862.2 (2007) Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs. International Telecommunication Union, Geneva
24. ITU-T Rec. P.48 (1993) Specification for an Intermediate Reference System. International Telecommunication Union, Geneva
25. Côté N, Koehl V, Gautier-Turbin V, Raake A, Möller S (2010) An Intrusive Super-Wideband Speech Quality Model: DIAL. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech), Makuhari, Japan
26. Côté N (2011) Integral and Diagnostic Intrusive Prediction of Speech Quality. T-Labs Series in Telecommunication Services, Springer, Berlin
27. Wltermann M, Scholz K, Raake A, Heute U, Möller S (2006) Underlying quality dimensions of modern telephone connections. In: Proceedings of the 7th Annual Conference of the ISCA (Interspeech 2006). International Speech Communication Association (ISCA)
28. ITU-T Rec. P.863 (2011) Perceptual Objective Listening Quality Assessment (POLQA). International Telecommunication Union, Geneva
29. ITU-T Contribution COM 12-99 (2010) Packet Loss and Coding Impact on Quality of Synthesized Speech Predicted by PESQ and P.563 Models. Ministry of Transport, Construction and Regional development of the Slovak Republic (Author: Počta P), ITU-T SG12 Meeting, Geneva, Switzerland, April 2010
30. ITU-T Contribution COM 12-220 (2011) Predicting the Quality of Synthesized and Natural Speech Impaired by Coding Using P.862 and P.863 Models. Ministry of Transport, Construction and Regional development of the Slovak Republic (Author: Počta P), ITU-T SG12 Meeting, Geneva, Switzerland, October 2011
31. ITU-T Contribution COM 12-221 (2011) Predicting the Quality of Synthesized and Natural Speech Impaired by Packet Loss Using P.862 and P.863 Models. Ministry of Transport, Construction and Regional development of the Slovak Republic (Author: Počta P), ITU-T SG12 Meeting, Geneva, Switzerland, October 2011

32. Cernak M, Rusko M (2005) An Evaluation of Synthetic Speech Using the PESQ Measure. Proceedings of the Forum Acusticum 2005, Budapest, pp 2725–2728
33. ITU-T Rec. P.563 (2004) Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony. International Telecommunication Union, Geneva
34. Malfait L, Berger J, Kastner M (2006) P.563 - The ITU-T Standard for Single-Ended Speech Quality Assessment. IEEE Trans Audio Speech Lang Process 14:1924–1934
35. Kim D-S (2005) ANIQUE: An Auditory Model for Single-Ended Speech Quality Estimation. IEEE Trans Speech Audio Process 13(5):821–831
36. Kim D-S, Tarraf A (2006) Enhanced Perceptual Model for Non-Intrusive Speech Quality Assessment. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France, pp 829–832
37. Gray P, Massara RE, Hollier MP (1998) Constraint-Based Pitch-Cycle Identification Using a Hybrid Temporal Spectral Method. In: Proceedings of the 105th Audio Engineering Society Convention
38. Gray P, Hollier MP, Massara RE (2000) Non-Intrusive Speech Quality Assessment Using Vocal Tract Models. IEE Proc Vision Image Signal Process 147:493–501
39. Heimansberg J (2006) Instrumentelle Schätzung der Qualität synthetischer Sprache. Bachelorarbeit, Institut für Kommunikationsakustik, Ruhr-Universität Bochum
40. Möller S, Heimansberg J (2006) Estimation of TTS Quality in Telephon Environments Using a Reference-free Quality Prediction Model. In: Proceedings of the 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, pp 56–60
41. ITU-T Contribution COM 12-180 (2008) Single-Ended Quality Estimation of Synthesized Speech: Analysis of the Rec. P.563 Internal Signal Processing. Deutsche Telekom AG (Authors: Möller S, Falk TH), ITU-T SG12 Meeting, Geneva, Switzerland, May 2008
42. Möller S, Kim D-S, Malfait L (2008) Estimating the Quality of Synthesized and Natural Speech Transmitted Through Telephone Networks Using Single-ended Prediction Models. Acta Acust United Acust 94:21–31

Chapter 4
Perceptual Quality Dimensions

Chapter 2 introduced a general structure of TTS systems and highlighted different approaches towards signal generation. Although nowadays TTS systems produce speech signals that sound very humanlike, they still feature degradations. Considering that these degradations are system specific, i.e., degradations of a PSOLA synthesizer evoke a different perceptual impression than the degradations of an HMM synthesizer, the quality of TTS systems is of multidimensional nature.

In the following, Chap. 3 introduced several auditory evaluation metrics that are able to assess specific quality features. Several studies have employed these auditory methods to derive perceptual quality dimensions. However, they were all restricted in one way or the other, e.g., they only analyzed one or very few systems, they employed signals of a very short duration, or only a limited number of attribute scales was utilized in the experiment. Thus, there is no unified picture of perceptual quality dimensions of state-of-the-art TTS systems. Therefore, this chapter seeks to answer RQ1 and RQ2:

> *Which perceptual quality dimensions are relevant for state-of-the-art TTS systems?*
>
> *How should a listening test be designed in order to capture all relevant quality dimensions?*

The chapter begins with a review of several studies concerning the quality of TTS systems followed by a discussion of the restrictions of these studies. Afterwards, two studies on two extensive German TTS databases are presented. In the first study a Semantic Differential (SD) is developed that covers all perceptual quality features of state-of-the-art TTS systems, while the second study is based on a direct comparison of TTS stimuli without the help of any attribute scales. Following that, the results are brought in line with the findings from Sect. 4.1, and five universal quality dimensions, as well as a test protocol for identifying them in an efficient way, are derived.

F. Hinterleitner, *Quality of Synthetic Speech*, T-Labs Series
in Telecommunication Services, DOI 10.1007/978-981-10-3734-4_4

4.1 State-of-the-Art Perceptual Quality Dimensions[1]

This section gives an overview of studies on perceptual quality of TTS that have been conducted within the past 20 years as well as an interpretation of the resulting perceptual quality dimensions. Moreover, the restrictions of the test setups and their influence on the resulting quality dimensions is highlighted. Each study will be later on referred to by its abbreviation introduced in brackets in the title.

4.1.1 Study: Kraft and Portele (Kraft1995)

In 1995, Kraft and Portele [2] evaluated five German speaking TTS systems in an auditory listening test. The database consisted of stimuli produced by two formant synthesizers (male voices) and three concatenative diphone/demisyllable synthesizers (two female voices, one male voice). The 44 subjects were instructed to rate the stimuli on eight presented ACR scales with five to six categories. Six familiar and unfamiliar passages were synthesized with a total duration of about 100 words. A subsequent Principal Component Analysis (PCA) with Promax rotation revealed two factors which were connected to (i) *prosodic and long term attributes* and to (ii) *segmental attributes*. Even though the first dimension was linked to prosody, it also comprises attribute scales that are specific to the voice of the systems, such as *naturalness* and *pleasantness*.

4.1.2 Study: Mayo et al. I (Mayo2005)

In a pilot study Mayo et al. [3] unveiled the perceptual quality dimensions of the Festival synthesizer [4]. In this study, eight sentences from the TIMIT database [5] were chosen and synthesized with an English speaking female voice. The stimulus duration varied from 1.9 to 4.1 s. Eight native speakers of English which were all experienced with listening to synthetic speech took part in a PC test. They were instructed to rate whether the two presented stimuli were *similar* or *different* in terms of naturalness. The responses were compiled into a dissimilarity matrix which was then processed via an MDS analysis. The resulting dimensions were interpreted through visual and auditory analysis of the configuration of the stimulus space. The first dimension represents (i) *prosodic cues* which reflect the appropriateness of duration and intonation. The second dimension is linked to (ii) *segmental and unit-level cues*. It describes the appropriateness of units selected for synthesis as well as the number of selected units.

[1] Parts of the content of this section have previously been published in a slightly different version in [1].

4.1.3 Study: Viswanathan and Viswanathan (Vis2005)

To test the reliability and validity of the test method proposed in the ITU-T Rec. P.85 [6] which was introduced in Sect. 3.3.2.1, Viswanathan et al. [7] conducted a series of five consecutive listening tests. In the final study, stimuli produced by five English speaking TTS systems were evaluated on nine 5-point ACR scales. Additionally, participants were instructed to also rate the *overall quality* and the *acceptability* of the systems. The investigated systems used either phones or sub-phone units for concatenative synthesis. The synthesizers included algorithmic variations for pitch and duration generation. The stimuli were rated by 128 naïve test participants. A Factor Analysis (FA) revealed two factors: Dimension 1 is related to the extent to which speech is similar to natural human speech and was thus labeled (i) *naturalness*; Dimension 2 describes how well the content of the signal can be understood, hence it can be assigned to the (ii) *intelligibility* of the signal.

4.1.4 Study: Seget (Seget2007)

Seget [8] conducted a study using speech material from six German TTS systems. The stimuli were created by diphone-based synthesizers using the PSOLA technique and unit-selection systems. A total of 10 speech samples have been generated per TTS system, half for male speakers and half for female ones. The synthesized speech samples have an average duration of 12 s and consist of two utterances separated by a silence interval of approximately 2 s. All stimuli were bandpass filtered according to ITU-T Rec. G.712 [9]. The listening test closely followed the ITU-T Rec. P.85 [6]. Thus, besides the rating of the stimuli on eight ACR scales, the 17 test participants were also given a parallel task. As suggested in P.85, the listening test also included natural speech reference files. A subsequent Principal Axis Factor (PAF) analysis with subsequent Promax rotation revealed two dimensions. The first dimension consists of scales concerning the naturalness of the synthesized voice as well as prosodic attributes of the signal. The second dimension comprises scales that cover the fluency and intelligibility of the signal. Thus, dimension 1 was labeled (i) *naturalness and prosody* while dimension 2 was named (ii) *intelligibility*.

4.1.5 Study: Hinterleitner (Hint2010)

In 2010 Hinterleitner [10] evaluated six German TTS systems in a listening test that followed closely the requirements of the P.85 test protocol. The TTS systems featured female and male voices of PSOLA and US synthesizers. Moreover, one female and one male voice of a natural speaker were included in the test. The five synthesized utterances per system exhibit an average duration of 6 s. All files were preprocessed

according to the ITU-T Rec. G.711 [11] and bandpass filtered according to ITU-T Rec. G.712 [9]. Twenty-five subjects (12 female, 13 male) were invited to rate the stimuli on four rating scales (*overall impression*, *voice pleasantness*, *naturalness*, *listening effort*) while also mastering a parallel task. Prior to the listening test every subject completed a training phase which helped them get used to the test setup and the quality range of the presented TTS stimuli. Since the quality was only evaluated on four rating scales no subsequent FA was performed.

4.1.6 Study: Mayo et al. II (Mayo2011)

Mayo et al. [12] pursued the investigations described in Sect. 4.1.2. Twenty-four sentences from the TIMIT [5] corpus were selected and synthesized with an English speaking female voice by unit selection TTS system Festival [4]. The average duration of the stimuli was 2.7 s. Thirty participants took part in a PC test, where they were instructed to rate the similarity of a pair of stimuli in terms of *naturalness*. Two types of acoustic analyses were carried out: the automatic analysis consisted of measures that were computed by Festival during the synthesis process (e.g., target and join costs) and measures that were derived from those features (e.g., total cost, target costs of different types of diphones); the manual analysis included comparisons with natural speech files (e.g., number of transcription/pronunciation errors per synthetic utterance).

A subsequent MDS analysis yielded three dimensions. Through visual, auditory and cluster analysis these dimensions could be linked to (i) *overall join quality/quantity*, (ii) *join distribution and detectability*, and (iii) *unit appropriateness and prosody*. By the chosen wording the first two dimensions seem to be connected to segmental attributes that concern the *fluency* and the *intelligibility* of the speech signal, while the third dimension represents global characteristics that describe the *prosodic quality* of the signal.

4.1.7 Restrictions of Discussed Studies

This section outlines the similarities and differences of the studies presented in the previous sections and their impact on the resulting quality dimensions. An overview of all relevant characteristics of the experimental setups can be seen in Table 4.1 and is discussed in the following.

One of the main restrictions of all presented studies is their limitation to very few synthesis techniques. Strictly speaking, no more than two different TTS techniques are covered in each study. This limits the spectrum of featured degradations and thus also the resulting quality dimensions.

Moreover, resulting quality dimensions depend on the different types of synthesizers that were part of the dataset, thus synthesizer-specific characteristics, e.g., the noise of HMM-synthesizers or the sonic glitches of concatenative systems, can nat-

Table 4.1 Comparison of the main characteristics of the different test setups of the 6 presented studies

Study Language	Kraft1995 German	Mayo2005 English	Vis2005 English	Seget2007 German	Hint2010 German	Mayo2011 English
Synthesizer Type:						
Formant	✓					
Diphone (PSOLA/MBROLA)	✓		✓	✓	✓	
Unit Selection		✓		✓	✓	✓
HMM						
Number of Systems	5	1	5	6	6	1
Number of Configurations*	5	1	5	12	12	1
Stimuli per Configuration	6	8	9	5	5	24
Quality Assessment via	ACR	PC	ACR	ACR	ACR	PC
Number of Scales	8	-	9	7	4	-
Natural Reference	no	no	no	yes	yes	no
Duration of Stimuli	100 words	1.9-4.1 s	20-25 words	12 s	6 s	2.7 s
f/m Voices	both	female	both	both	both	female
Separate Sessions for f/m	no	-	no	no	yes	-
Preprocessing	-	-	-	G.712	G.711, G.712	-
Sampling Rate	n/k	n/k	n/k	8 kHz	8 kHz	n/k

* A configuration denotes a specific combination of one voice/one speech corpus and one synthesis system.

urally only be assessed if these types of systems are part of the study. Accordingly, studies that only feature formant synthesizers and diphone based PSOLA systems, e.g., as in Kraft1995, are most likely to lead to different dimensions than studies that only assess unit-selection synthesizers, e.g., as in Mayo2005 and Mayo2011.

Furthermore, none of the studies covers the very popular technique of HMM synthesis. Therefore, a perceptual quality dimension that would cover the noise HMM systems produce can not emerge from those studies.

Besides the included TTS systems the studies that use an SD (Kraft1995, Vis2005, Seget2007, and Hint2010) are also restricted to the rating scales presented to the test participants. Firstly, these studies only feature a small number of presented rating scales. Thus, none of those 4 studies lead to more than two quality dimensions (in the case of Hint2010 no FA was performed due to only four presented scales). Secondly, most of the studies (Vis2005, Seget2007, Hint2010) are based on the partially outdated evaluating protocol P.85 from 1995. Therefore, degradations that are distinctive for modern-day TTS systems might not have been captured.

Moreover, two studies are also crucially influenced by the inclusion of natural reference stimuli. Even though TTS systems nowadays can sound very natural, in most cases a quality gap between synthesized speech and natural speakers can be easily discerned. Thereby, test participants tend to give lower ratings to TTS stimuli if a natural reference speaker is part of the test. As a consequence, the range on a rating scale that is normally used for TTS systems shrinks. Therefore, quality aspects that are relevant when only listening to synthetic speech might get suppressed due to the small range that is available for TTS systems.

In addition, the evaluation of a system can also be compromised due to too short TTS samples, e.g., the studies Mayo2005 and Mayo2011 use stimuli with an average duration of below 4 s. These very short utterances might affect a rating of more global attributes, e.g., the prosody of a system.

Moreover, the studies Kraft1995, Vis2005, and Seget2007 presented female and male TTS stimuli in the same session. Thus, test participants with a preference for speakers of one gender might have been entrapped to give lower ratings to stimuli of a speaker of the opposite sex due to a direct comparison. This effect can be softened when female and male stimuli are presented in different sessions.

The studies Seget2007 and Hint2010 investigated the use of TTS in telephone services. Thus all stimuli were G.712 bandpass filtered and in case of Hint2010 also G.711 coded. Therefore, coding distortions were induced that are not generated by a synthesis system.

Those two studies were also the only ones that provided information concerning the sampling rate (8 kHz). This raises the question whether distortions that affect higher frequencies (above 8 kHz) could be perceived at all in any of the presented studies.

Considering all the differences of the presented studies it is not surprising that the resulting quality dimensions are not consistent. With the goal to reveal a set of universal perceptual quality dimensions of state-of-the-art TTS systems, two extensive studies were conducted. Their setups and results are presented in the following two sections.

4.2 Semantic Differential and Factor Analysis[2]

In Sect. 3.3.2 two different approaches for multidimensional analysis of TTS systems were introduced. In the following a Semantic Differential (SD) protocol is developed and subsequently tested on an extensive database of synthesized speech.

Since an SD uses pre-defined attribute scales to measure the auditory impression of the listeners a direct relation between the used attribute scales and the derived quality dimensions is guaranteed. On the downside, due to the given set of scales, this approach cannot guarantee that all relevant perceptual dimensions are actually solicited from the test participants. To reduce the influence of the test designers to a minimum, a suitable set of scales has to be developed through several pretests. In pretest 1 attributes describing the auditory impression of the listeners are collected. These terms are converted into scales and presented in a second pretest. An analysis of the data of the second pretest leads to a final selection of scales which are presented in the final SD experiment. On the basis of these attribute ratings, perceptual quality dimensions can be derived with the help of a factor analysis. The development of a protocol for TTS quality assessment will be described in the following section while a statistical analysis of the gathered data and the discovered perceptual quality dimensions are discussed in Sect. 4.2.2.

4.2.1 Experimental Setup

This section gives an overview of the database of speech synthesizers collected for the listening tests. Moreover, it describes the approach used to gain a relevant set of attribute scales that describe the perceptual space of TTS systems in a more or less complete way.

4.2.1.1 Test Database

Ten German sentences from the EUROM.1 corpus [14] were chosen as source material. Utterances containing place names, proper names, or words from a foreign language were excluded as they often differ from German pronunciation rules which are likely to cause problems for speech synthesizers. To avoid user fatigue but still guarantee a valid impression of the occurring distortions, the sentences were shortened to a length of about 10 s each.

To capture a broad variety of distortions synthetic speech files generated from 14/15 different German speaking TTS systems for female/male speakers, for some

[2]Parts of the content of this section have previously been published in a slightly different version in [13].

of them with up to six different voices. Thus, data from 35/28 different configurations[3] (female/male) could be produced. Besides the synthetic speech files the database also contains stimuli from 4/4 amateur (female/male) and 4/4 professional (female/male) natural speakers. Even though the goal of these tests was to develop a protocol for synthetic speech, natural speakers were included in the first pretest in order to obtain quality describing attributes of "ideal" speech. All speech files were downsampled to 16 kHz and level normalized to −26 dBov using the speech-level meter [15].

The database that was used for the main test contains speech material synthesized by the following TTS systems[4] (the synthesis type and the number of female/male voices for each synthesizer are given in brackets):
Acapela Infovox3 (US, 1/1), AT and T Natural Voice (US, 1/1), atip Proser (DI, 1/1), BOSS (US, 1/0), Cepstral Voices (US, 1/1), Cereproc CereVoice (US, 1/1), DRESS (DI, 1/1), Loquendo TTS (US, 1/1), MARY bits (US, 1/1), MARY hmm-bits (HMM, 1/1), MARY MBROLA (DI, 1/1), NextUp Talker (DI, 1/1), NextUp TextAloud3 (FO, 0/1), Nuance RealSpeak (US, 2/1), SVOX (US, 1/1), SyRUB (DI, 0/1).

4.2.1.2 Pretest 1

The objective of pretest 1 was to collect a broad basis of attributes describing auditory features of synthetic speech. Therefore audio files from 12/13 different TTS systems with female/male voices plus two different natural speakers per gender were presented. Twelve expert listeners (4 female, 8 male) from the Quality and Usability Lab of the TU Berlin took part in the test. The stimuli were presented in a quiet conference room environment via headphones (AKG K601) and a high-quality sound device (Roland Edirol UA-25) in randomized order. Two sessions were conducted, one with female and one with male voices, with a break of 5 min in between. Every TTS system was covered with two stimuli. The listeners were instructed to write down nouns, adjectives and antonym pairs describing their auditory impression. Furthermore, they were asked to give an intensity rating for each attribute on a scale ranging from 1 to 10.

The listening test resulted in 2179 collected terms out of which 296 unique descriptions were found. These attributes were condensed into 44 scales. Attribute scales that mainly rate quality features concerning individual voice character and accent and those that rate the same perceived quality features were omitted. The remaining scales were weighted by frequency of occurrence, and the 28 most named ones were chosen for pretest 2.

[3] A configuration denotes a specific combination of one voice/one speech corpus and one synthesis system.

[4] While the same TTS systems were used in pretest 1 and 2, the configurations deviate slightly from what was used in the main test.

4.2.1.3 Pretest 2

To narrow down the set of attribute scales from pretest 1 to a manageable number, a second pretest was conducted. Here, audio files from 19/20 different configurations of TTS systems with female/male voices were presented. Nine expert listeners (3 female, 6 male) from the Quality and Usability Lab of TU Berlin and 13 naïve listeners (8 female, 5 male, mainly students from the TU Berlin) took part in the test. All naïve listeners were paid for their participation.

The purpose of this test was to find the set of attribute scales that was most suitable for the final SD experiment and thus describes the quality stimulus space most precisely. Therefore, the subjects were instructed to only use scales that were most relevant for their auditory impression. The stimuli were presented in randomized order in two sessions, one with female voices and one with male voices, with a 5 min break in between. The stimuli were presented via headphones (AKG K601) and a high-quality sound device (Roland Edirol UA-25) in a sound proof booth.

To narrow down the number of attribute scales, *unnatural melody versus natural melody* which correlated highly ($R > .60$) with the other scales that rate naturalness, and thus measures similar features, was omitted before the main test. Moreover, scales that were used rather rarely were dropped.

In order to gain a first impression of the perceptual space, a PCA with Varimax rotation was performed on the remaining items and three factors were extracted. Subsequently, all items with high loadings on multiple factors and items with communalities $< .45$ were discarded. This led to a set of 16 attribute scales which is presented in Table 4.2.

Table 4.2 Set of attribute scales that was used in the main SD experiment

Abbr.	Label	Attribute Scale
BUMP	Bumpiness	bumpy vs. not bumpy
CLIN	Clink	clinking vs. not clinking
DSTO	Distortions	distorted vs. undistorted
DSTU	Disturbances	undisturbed vs. disturbed
FLUE	Fluency	interrupted vs. fluent
HISS	Hiss	hissing vs. not hissing
INTE	Intelligibility	unintelligible vs. intelligible
NATU	Naturalness	artificial vs. natural
NOIS	Noise	noisy vs. not noisy
POLY	Polyphony	several voices vs. one voice
RASP	Rasping Sound	raspy vs. not raspy
RHYT	Rhythm	unnatural rhythm vs. natural rhythm
SPEE	Speed	fast vs. slow
STRE	Stress	unnatural stress vs. natural stress
TENS	Tension	tense vs. calm
VPLT	Voice Pleasantness	unpleasant vs. pleasant
OIMP	Overall Impression	bad vs. excellent

Note labels have been translated from German wordings

Table 4.3 Mix of synthesis techniques present in the main experiment

Synthesis Type	Female	Male
Formant (FO)	0	1
Diphone (DI) (PSOLA/MBROLA)	4	5
Unit Selection (US)	10	8
HMM	1	1

4.2.1.4 Main Test

For the main test a set of 15 different synthesizer configurations per gender was chosen. The mix of different synthesis techniques can be seen in Table 4.3. For each configuration two different stimuli were presented. The test was split in three parts: since the subjects were not familiar with the quality as well as degradations of TTS signals, a training with three different stimuli covering the whole quality range from the TTS signal database was conducted. In the second and third part female and male stimuli were presented or vice versa, with a 5 min break in between.

Thirty naïve subjects aged 21 to 65 (15 female, 15 male, μ=27.9, σ=7.9) took part in the test. Most of them were students from the local university. None of them had any known hearing disabilities. All subjects were paid for their participation. The stimuli were presented via headphones (AKG K601) and a high-quality sound device (Roland Edirol UA-25) in a sound proof booth.

After listening to each stimulus the participants had to rate the Overall Impression (OIMP) of the signal on a continuous scale ranging from *bad* to *excellent*. Subsequently, a quality estimate for the attribute scales determined in pretest 2 had to be adjusted via a slide presented on the test GUI. The rating on the OIMP scale was separated from the ratings on the remaining scales in order to guarantee that the test participants first thought about their impression of the system as a whole and after that about degradations of specific quality features.

4.2.2 Statistical Analysis

The following section describes the Factor Analysis (FA) that was used to come up with an interpretable perceptual space. Moreover, the resulting quality dimensions are analyzed.

Prior to the execution of any statistical methods the ratings of all test participants were screened for plausibility. Therefore, box plots featuring the ratings on all dimensions were created for all stimuli. One participant had more than 5% outlier ratings and was thus excluded from the study.

4.2.2.1 Factor Analysis

A first analysis of the data collected in the main test showed that the item SPEE almost always exclusively loaded on a single factor. Thus, all other items created the remaining dimensions. To split these dimensions up in order to enable a more detailed view of the perceptual space the item SPEE was discarded from further analysis.

A PAF analysis of the remaining 15 items revealed three factors. Separate PAFs for female and male stimuli showed a similar factor structure, thus one analysis over the whole dataset seemed sufficient. The three factors account for 61.47% of the total variance. This value could not be increased significantly by extracting more than three factors. Residuals were computed between the observed and reproduced correlations: 5 (4%) were non redundant with absolute values greater than .05.

It was assumed that quality dimensions are not independent of each other, i.e., an impairment in one dimension can also affect the quality of other dimensions to a certain degree. Therefore, an oblique rotation (Promax rotation with $\kappa = 4$), which leads to correlated factors, was performed subsequently. The value of the accounted variance after rotation will not be analyzed because of a massive overestimation due to correlated factors.

Table 4.4 Factor pattern matrix. *Note* For better readability values below .20 are suppressed

SCALE	FACTOR LOADINGS		
	1	2	3
Stress	.93		
Naturalness	.90		
Rhythm	.89		
Voice Pleasantness	.77		
Tension	.66		
Bumpiness	.58		
Distortions	.45	.32	
Hiss		.75	
Noise		.65	
Rasping Sound		.59	
Disturbances		.49	.27
Clink	.27	.38	
Polyphony			.79
Intelligibility	.21		.57
Fluency	.29		.50
EIGENVALUES	6.38	1.88	.96
% OF VARIANCE	42.5	12.5	6.4
CRONBACH'S α	.90	.69	.74

Table 4.5 Factor correlation matrix

FACTOR	1	2	3
1	1.00	.45	.73
2	.45	1.00	.52
3	.73	.52	1.00

The FA resulted in the factor pattern matrix shown in Table 4.4. For clarity, values below .20 were suppressed. Due to the oblique Promax rotation the resulting factors are not orthogonal. The correlations among the three factors can be seen in Table 4.5.

4.2.2.2 Resulting Quality Dimensions

In order to obtain a meaningful interpretation of the quality dimensions, items with high cross-loadings, meaning similar loadings on multiple factors,[5] have to be excluded before interpretation. In this case this applies to the items DSTO and CLIN.

Factor 1 is highly correlated with the items STRE, NATU and RHYT, thus it represents the ***naturalness*** of the TTS signal. The items with high loadings on factor 2 (HISS, NOIS, RASP) are all related to ***disturbances*** in the signal. Factor 3 seems to reflect ***temporal distortions***, e.g., concatenation artifacts which occur in US synthesis. The effect of the item POLY which contributes the most to this dimension can be witnessed when, e.g., two units with a slightly different speed get connected. This creates the impression of two different voices speaking at the same time.

Figure 4.1 shows a graphical representation of Table 4.4. It has to be stated that a TTS signal with high values on all three dimensions is perceived as *very natural*, *not disturbed*, and *not temporally distorted*. The two items with high cross-loadings (DSTO, CLIN) in the factor pattern matrix do also stand out here. Both only reach very low values on all three dimensions, thus they do not account much for any of these. Furthermore, the item BUMP is not only correlated with *naturalness* but also with the dimension *temporal distortions* ($R = .59$). This is hardly surprising since temporal distortions can be perceived as bumps in a speech signal.

Moreover, as an effect of the oblique rotation, it has to be stated that all factors are correlated. Especially factor 1 and factor 3 show a very high correlation. This means that a very natural sounding TTS system will most likely be bound to the impression of a fluent, intelligible speaker.

In order to get a clearer view on the quality features of each single system a mapping of the stimuli in the perceptual space is presented in Fig. 4.2. Figure 4.2a displays the values for the different systems for dimensions 1 (*naturalness*) and 2 (*disturbances*), in which the subscripted character (f/m) represents the speaker gender; FO systems are marked with cyan asterisk, US synthesizers with green dots, DI synthesizers with blue squares, and HMM synthesizers with red diamonds.

[5] $|\text{loading factor A - loading factor B}| < .20$.

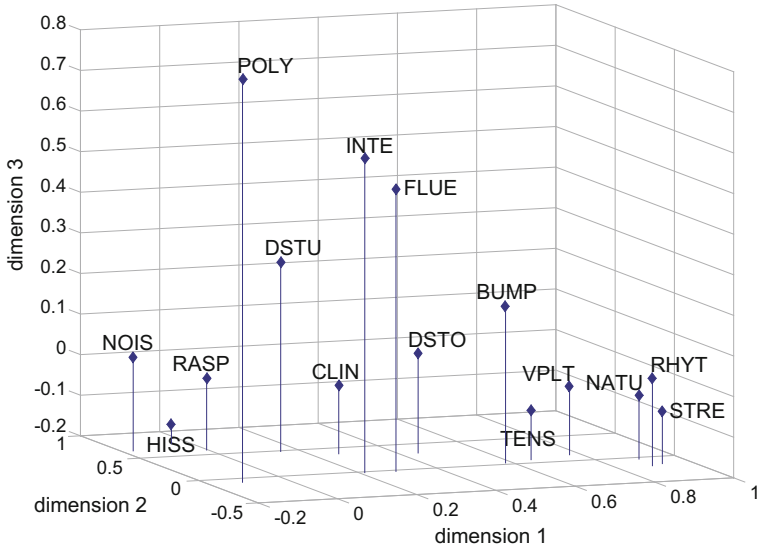
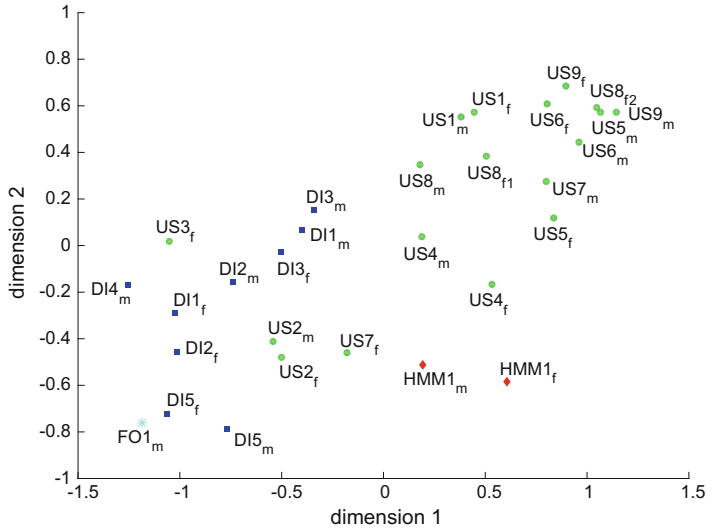


Fig. 4.1 Mapping of the quality scales in three dimensional perceptual space

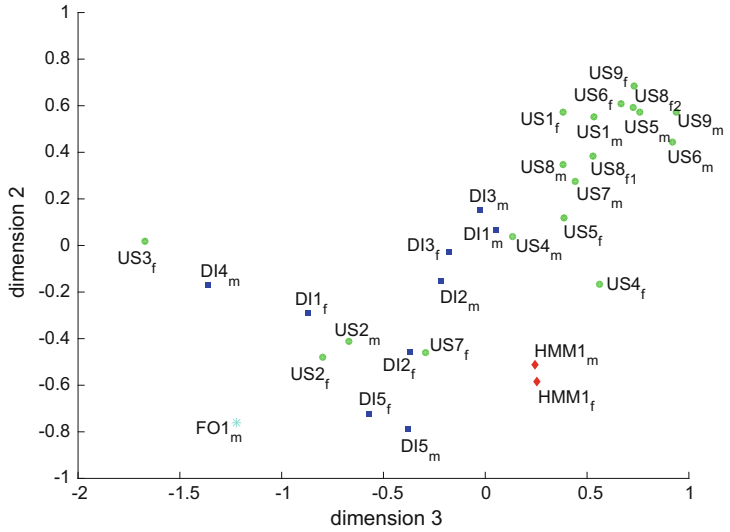
Independent of the speaker's gender most synthesizers build clusters, e.g., US1, US2, US6. It is striking, however, that some systems obviously do not stick together: US7$_f$ and US7$_m$ for instance are clearly separated with differing values for all three dimensions. This indicates that the speech material used for the female US7 system features disturbances that could not be found in the corresponding male data. Moreover, it can be stated that some systems reach high values on one axis while only low ratings on the other, e.g., HMM1 is perceived as natural but also disturbed. Besides, systems with equal values in the disturbances dimension, for instance US3$_f$, US4$_m$, and US5$_f$ can be perceived as natural sounding as well as artificial.

Figure 4.2b shows the perceptual space of dimension 3 (*temporal distortions*) and 2 (*disturbances*). Again, the clustering effect can be observed. However, both figures reveal that undisturbed stimuli always get rated as natural as well as not temporally distorted. In contrast, natural and not temporally distorted synthesizers can also be perceived as disturbed (e.g., HMM1).

An analysis of the correlations of the extracted dimensions with the rating on the OIMP scale can be seen in Table 4.6. Therefore, the *naturalness* dimension accounts the most for the overall impression rating.



(a) Mean factor values per synthesizer for dimensions 1 (naturalness) and 2 (disturbances).



(b) Mean factor values per synthesizer for dimensions 3 (temporal distortions) and 2 (disturbances).

Fig. 4.2 Mapping of the stimuli in the perceptual space

Table 4.6 Correlations between the three dimensions and the OIMP

DIMENSION	1	2	3
OIMP	.81	.47	.56

4.3 Sorting Task and Multidimensional Scaling[6]

The previous section showed how to create a set of attribute scales that can be used in an SD to assess the quality of TTS synthesizers. This study clearly led to a better understanding of perceptual quality of synthetic speech, considering that TTS stimuli generated by all popular speech synthesis methods (FO synthesizers, DI concatenation, US synthesis, HMM synthesis) were evaluated on scales that most likely cover all common TTS artifacts. Therefore, the results are relevant for the development and optimization of modern TTS systems. Moreover, the study yielded three quality dimensions that were easily interpretable and generally intelligible. Nevertheless, this method also implies some drawbacks:

1. During the two pretests the attributes were mainly solicited from audio and speech experts as test participants. This fact guaranteed that all relevant artifacts are captured and reproduced in the derived rating scales. However, the disadvantage of this approach is that experts also perceive degradations that are not as relevant for the quality impression of normal listeners. Though naïve listeners might be able to discern stimuli with respect to specific degradations (i.e., on corresponding attribute scales), this score does not necessarily affect their quality impression to a large degree. Hence, through the process of a factor analysis this can lead to dimensions that only have a minor influence on the overall quality.
2. Even though the intention was to keep the developed scale labels as simple as possible, one cannot be certain that the naïve listeners of the main test understood all scales in the same way.
3. Furthermore, methods that assess quality on global scales always limit the test participants' rating to the presented scales.
   Therefore, it is crucial to investigate if those three quality dimensions could also be derived through a listening test that is not based on given rating scales, but rather on the unrestricted perceptual quality impression of the listeners itself.

To overcome these drawbacks, this section presents further research on the above mentioned dimensions. In the following, a set of TTS stimuli that was used in the listening test is described. Furthermore, the principle of Multidimensional Scaling (MDS) and the Sorting Task (ST) that was chosen for the listening test are explained. The results of the experiments are statistically analyzed and an interpretation of the resulting perceptual dimensions is given.

[6]Parts of the content of this section have previously been published in a slightly different version in [16].

Table 4.7 Mix of synthesis techniques present in the ST experiment

SYNTHESIS TYPE	FEMALE	MALE
Formant (FO)	2	5
Diphone (DI) (PSOLA/MBROLA)	9	10
Unit Selection (US)	17	10
HMM	2	2

4.3.1 Experimental Setup

In the following, the TTS database is introduced, the principle of MDS is presented, and the listening test following the ST method is described. Moreover, to help with the interpretation of the obtained perceptual dimensions, a post test using the attribute scales developed in Sect. 4.2 is conducted.

4.3.1.1 Test Database

To guarantee a fair comparison of the test stimuli, one German sentence[7] with a synthesized duration of approximately 5 s was chosen for the listening test. Twenty different TTS systems were selected (some of them with different voices) and used to synthesize this utterance. All in all, 16 female systems in 30 different configurations[8] and 19 male systems in 27 different configurations were used. The TTS systems that were used to generate the stimuli are the following ones (the synthesis type and the number of female/male voices for each synthesizer are marked in brackets):
Acapela Infovox3 (US, 2/1), AT and T Natural Voices (US, 1/1), atip Proser (DI, 2/1), BOSS (US, 2/0), Cepstral Voices (US, 1/1), Cereproc CereVoice (US, 1/1), DRESS (DI, 4/4), ESpeak (FO, 0/1), Fonix Speech FonixTalk (FO, 2/2), IVONA (US, 1/1), Loquendo TTS (US, 1/1), MARY bits (US, 2/2), MARY hmm-bits (HMM, 2/2), MARY MBROLA (DI, 2/3), Meridian Orpheus (FO, 0/1), NextUp Talker (DI, 1/1), NextUp TextAloud3 (FO, 0/1), Nuance RealSpeak (US, 4/1), SVOX (US, 2/1), SyRUB (DI, 0/1).

Thus, the database contains stimuli generated by FO synthesizers, Diphone (DI) synthesizers, Unit Selection (US) synthesizers, and one HMM-synthesizer (HMM) (Table 4.7).

The generated speech files were downsampled to 16 kHz (if necessary) and level normalized to −26 dBov using the speech-level meter [15].

[7] German: *Letzte Nacht habe ich die Haustür geöffnet um die Katze nach draußen zu lassen.* English translation: *Last night I opened the front door to let the cat out.*.

[8] A configuration denotes a specific combination of one voice/one speech corpus and one synthesis system.

4.3.1.2 Main Test

This section describes the main principle of MDS as well as a method suitable for large object sets. To simplify the interpretation of the dimensions extracted by the MDS algorithm, a post-test was carried out in which all stimuli were analyzed on the 16 attribute scales that were developed in Sect. 4.2.

Pairwise Comparison and Multidimensional Scaling

The main idea of multidimensional scaling is to identify orthogonal perceptual dimensions without prior knowledge about the nature of the stimuli, by asking test participants to scale the dissimilarities between pairs of stimuli. The dissimilarities between stimuli can then be transformed into a stimulus space in which the between-point distances correspond to the dissimilarities between stimuli.

Dissimilarities are usually derived in listening tests in which each stimulus in a set of n stimuli is compared with all other $n-1$ stimuli. Subjects rate the similarity of two stimuli on a scale with the end points *very similar* and *not similar at all*. The outcome is a matrix that represents the similarity between all stimuli [17].

Via an MDS algorithm, the dimensionality of this matrix can be reduced until the solution is interpretable but still represents the observed stimulus distances. As a badness-of-fit measure for the MDS representation, Kruskal [18] introduced the Stress function (low Stress values indicate a better fit).

The downside of PC tests is that a complete comparison of all stimuli leads to $\frac{n(n-1)}{2}$ comparisons. With large sets of objects the amount of comparisons reaches a level that is not suitable for the assessment in listening tests. With the database described in Sect. 4.3.1.1 this would yield 435/351 comparisons for female/male stimuli and a test duration per subject of over two hours. Therefore, a method is introduced in the following paragraph to derive dissimilarity matrices without a full PC test.

Sorting Task

Tsogo [19] proposed to use an ST when dealing with large object sets. Subjects are instructed to build groups of stimuli that are similar to each other while being different from stimuli in other groups. This results in one $n \times n$ incidence matrix per subject containing zeros and ones representing unsimilar and similar objects. Adding the matrices of all test participants together yields one similarity matrix from which one can easily derive a dissimilarity matrix as an input for the MDS algorithm.

To avoid that test participants sort stimuli with respect to gender, the test consisted of two sessions, one with female and one with male stimuli. The stimuli had to be sorted in up to eight groups with a minimum of two stimuli per group. Fourty naïve subjects aged between 19 and 37 (20 female, 20 male, $\mu = 25.6$, $\sigma = 3.25$) took part. All of them were native German speakers, none had any known hearing disabilities. All test participants were paid for their participation. The stimuli were presented via headphones (AKG K601) and a high-quality sound device (RME Hammerfall DSP Multiface II) in a soundproof booth.

4.3.1.3 Post-Test

MDS dimensions as such give no indication on their interpretation unless the stimuli are analyzed along the identified dimensions via expert listening or an additional auditory test. Thus, an interpretation is often a vague and moreover a highly subjective task. Therefore, a post-test in which all stimuli were rated on the 16 attribute scales from Sect. 4.2 was conducted.

The stimuli were presented in two groups (one with female, one with male stimuli) in randomized order. Five expert listeners from the Quality and Usability Lab of the TU Berlin and seven naïve subjects aged between 23 and 31 (5 female, 7 male, $\mu = 27$, $\sigma = 2.64$) took part in the post-test. All of them were native German speakers without any known hearing disabilities. The stimuli were presented via headphones (Sennheiser HD 485) and a high-quality sound device (Roland Edirol UA-25) in a quiet listening environment.

4.3.2 Statistical Analysis

In the following an MDS is applied on the data gathered from the PC test and the ratings on the attribute scales are used to interpret the resulting quality dimensions.

4.3.2.1 Multidimensional Scaling

Via a non-metric MDS [20] three dimensions were extracted for both female and male stimuli. The statistical fit parameter Stress1 for the female/male solution reached values of 0.07/0.06 and was thus far below the Stress1 values for random data as reported in [21]. To maximize the variance in each dimension the stimulus space was Varimax-rotated.

In order to ensure a meaningful interpretation, the Pearson correlation coefficient R between the factor scores and the median values of the ratings on the 16 attribute scales from the post-test was computed. Among others, dimension 1 correlated highly with the items VPLT ($|R| \geq .80$) and INTE ($|R| \geq .80$), dimension 2 correlated highly with RHYT ($|R| \geq .80$) and FLUE ($|R| \geq .70$), and dimension 3 reached the highest correlation with SPEE ($|R| \geq .50$). Dimensions 1 and 2 both also correlated highly with the item NATU ($|R| \geq .80$). But since this scale does not help to discern between both dimensions it was not taken into account for the following optimization.

In the next step the point configurations for both female and male stimuli were further rotated in a way which maximized the correlation between each dimension and the item(s) mentioned in the previous paragraph. The correlations between the optimized rotated dimensions and the items are shown in Table 4.8.

As can be seen, dimension 1 not only correlates highly with VPLT and INTE but also with the items NATU and STRE for female and male data. The correlation with OIMP is the highest for this dimension. Dimension 2 achieves high correlations

Table 4.8 Pearson correlation between rotated factor scores and attribute scale ratings

SCALES	DIMENSION 1		DIMENSION 2		DIMENSION 3	
	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE
bumpiness	−.61		−.81	−.79		
clink	−.67	−.60	−.56		.62	
distortions	−.72	−.77	−.77	−.70		
disturbances	−.63	−.69	−.77	−.69		
fluency	.59	.53	**.88**	**.78**	−.53	
hiss	−.54					
intelligibility	**.83**	**.87**	.73	.64		
naturalness	.85	.85	.77	.88		
noise pleasantness	**.88**	**.87**	.78	.78		
polyphony	−.64	−.63	−.77	−.58	.54	
rasping sound	−.56	−.63				
rhythm	.80	.76	**.86**	**.84**	−.55	
speed					**.54**	**.65**
stress	.84	.74	.76	.83	−.53	
tension	−.71	−.51	−.59	−.59	.68	.54
overall impression	.90	.89	.76	.76		

Note for better readability correlations with $|R| < .50$ are suppressed; the correlations for the selected scales are in bold

with RHYT and FLUE but also with BUMP and NATU. The correlations between dimension 3 and the attribute scales are much lower than for the other two dimensions. However, it is the only dimension that correlates with the item SPEE ($|R| \geq .50$). Moreover, it correlates with the item TENS and the female data achieves a correlation of ($R = .62$) with CLIN.

4.3.2.2 Resulting Quality Dimensions

The 2D-mapping of the stimuli in the perceptual space of dimensions 1 and 2 and dimensions 1 and 3 can be seen in Figs. 4.3 and 4.4. The two Fig. 4.3a, b display the perceptual space for the female and male stimuli in the perceptual space of dimensions 1 and 2 while the Fig. 4.4a, b show the results for the female and male stimuli in the perceptual space of dimensions 1 and 3. Subscripted indices indicate different voices; FO systems are marked with cyan asterisk, US synthesizers with green dots, DI synthesizers with blue squares, and HMM synthesizers with red diamonds.

For the interpretation of the extracted dimensions all stimuli were sorted according to their value in each dimension. The auditory impression of the sorted stimuli along with the correlations from Table 4.8 served as an indication for the interpretation of the dimensions. The voices of high-ranked stimuli in dimension 1 sounded very human-like even if the speech was somehow distorted. These stimuli can be

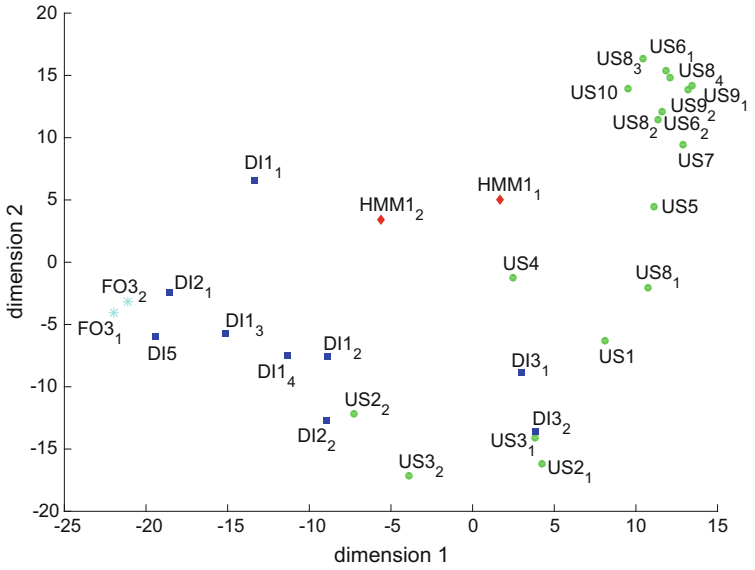
described as voices with personality and charisma; they evoke the impression of listening to a real human being. Thus, this dimension was labeled **naturalness of voice**. The auditory impression for stimuli sorted along dimension 2 confirms the high positive correlations with RHYT and FLUE and the high negative correlation with BUMP. Stimuli with low values in this dimension lacked natural sounding prosody while they often also cause the impression of a stuttering speaker. Therefore, this dimension can be associated with **temporal distortions** (low values indicate severe temporal distortions). Stimuli with high values in dimension 3 were slowly speaking and relaxed while voices with low values sounded more stressed and restless. This impression is confirmed by the correlations from Table 4.8. Thus dimension 3 describes the **calmness** of the voice.

A closer look at Fig. 4.3 reveals a clustering effect: most US systems build a cluster in the upper right corner, most DI synthesizers are in bottom center of the figure and FO synthesizers can be found at the far left of each figure. The only systems that achieve high values in dimension 2 are US synthesizers. However, not all US stimuli that sound very human-like also reach high values in dimension 2 ($US2_1$, $US2_2$, $US3_1$, and $US3_2$ in Fig. 4.3a and $US2_1$ in Fig. 4.3b).

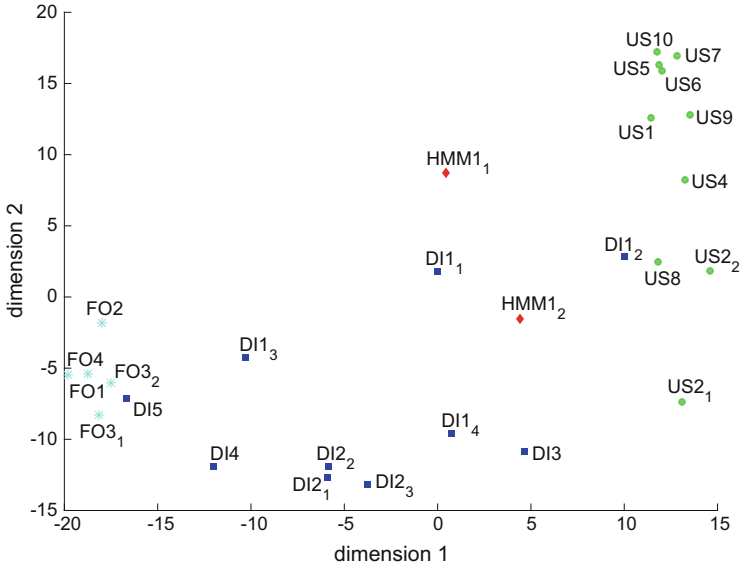
Dimension 2 on the other hand is linked to prosody and concatenation artifacts. Thus, especially DI synthesizers that connect lots of small speech units should score low values in this dimension. In this context it is interesting to see that the stimuli in Fig. 4.3a with the lowest values in dimension 2 are also US synthesizers ($US2_1$, $US3_2$). However, these are two non commercial, scientific systems that apparently have major issues with the pitch contour at the junctions between units. Strikingly, the stimulus $US8_1$ in Fig. 4.3b achieves one of the highest ratings of male stimuli in dimension 1, still its OIMP rating was only 2.5 (on the MOS scale which ranges from 1 to 5). The main reason for that is the low score in the dimension *temporal distortions*. Thus, the voice sounds human-like while the quality is still mediocre.

In Fig. 4.4b the clustering effect for the male stimuli can be seen even more clearly. The US systems sound very human-like and show similar speech rates, the DI systems more or less span across the whole range of dimension 1 with rather low values in dimension 3, FO synthesizers build a cluster on the far left of the figure, and the two HMM voices can be found in the top center of the figure. The highest values in dimension 3 and therefore the system with the lowest perceived speech rate is the HMM synthesizer for female as well as for male voices. After listening to all stimuli the auditory impression of system DI5 stood out. Even though it is a DI synthesizer it sounded very metallic and artificial. This could point to a system that uses coded speech units. Thus, the proximity of DI5 and the FO systems especially in the Figs. 4.3b and 4.4b is not surprising.

Furthermore, it can be stated that most of the stimuli that were produced by the same TTS system build clusters regardless of the voice's gender, e.g., DI2, HMM1, US9. Nonetheless, some of the stimuli of one system score very differently in one dimension (US8 in Fig. 4.3a, DI1 in Fig. 4.3b). Moreover, while high values in the dimensions 1 and 2 indicate a better overall impression, Fig. 4.4a and 4.4b denote that the highest-ranked stimuli on the scale OIMP ($US9_1$, $US9_2$, US7 in Fig. 4.4a

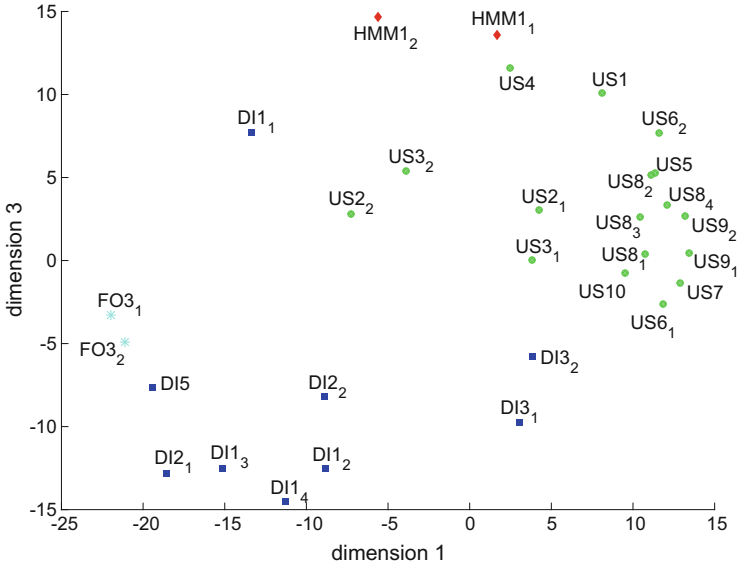


(a) Mapping of the female stimuli in the perceptual space.

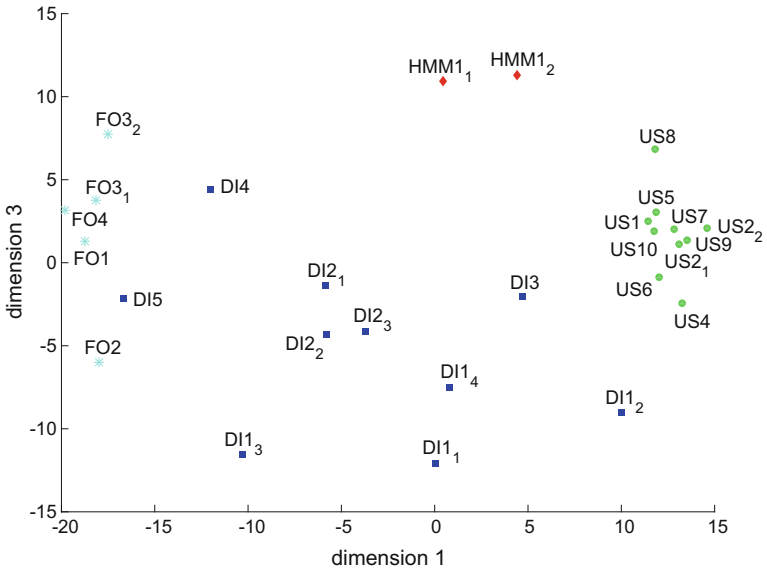


(b) Mapping of the male stimuli in the perceptual space.

Fig. 4.3 Mapping of the stimuli in the perceptual space of dimensions 1 and 2



(a) Mapping of the female stimuli in the perceptual space.



(b) Mapping of the male stimuli in the perceptual space.

Fig. 4.4 Mapping of the stimuli in the perceptual space of dimensions 1 and 3

Table 4.9 Correlations between the three dimensions and the OIMP

DIMENSION	1	2	3
OIMP	.89	.76	.33

and $US2_2$, US9, US4 in Fig. 4.4b) show medium values in dimension 3. Therefore, a medium speech rate seems to achieve the best listening impression.

An analysis of the correlations of the extracted dimensions with the rating on the OIMP scale can be seen in Table 4.9. Therefore, the *naturalness of voice* dimension accounts the most for the overall impression rating.

4.4 Summary of the SD/FA and ST/MDS Studies[9]

The studies presented in Sect. 4.1 all featured different restrictions, e.g., they all included one or only very few synthesis techniques, none contained any HMM systems, they mostly used attribute scales that were developed in the mid 90s, some used a natural reference speaker etc. This ultimately lead to diverse sets of perceptual quality dimensions.

Thus, in order to come up with a set of universal perceptual quality dimensions one had to eliminate these restrictions. This was done in the studies that were discussed in Sects. 4.2 and 4.3. Tabel 4.10 sums up the characteristics of these studies.

As can be seen, all relevant synthesis techniques were covered in both tests. Therefore, most likely both databases contained all relevant quality degradations that are present in state-of-the-art TTS systems. By applying two different auditory quality evaluation metrics, one based on attribute scales specifically developed for the quality assessment of TTS systems and one that was solely based on the perceptual impression of the listeners, it is safe to assume that all quality features that are relevant for TTS systems were present in those studies.

On first sight, however, the results from Sects. 4.2 and 4.3 differ slightly. The study SD/FA yielded the dimensions *naturalness*, *temporal distortions*, and *disturbances* while ST/MDS lead to the dimensions *naturalness of voice*, *temporal distortions*, and *calmness*.

[9] Parts of the content of this section have previously been published in a slightly different version in [16].

Table 4.10 Comparison of the main characteristics of the different test setups of studies SD/FA and ST/MDS

	SD/FA	ST/MDS
Year	2011	2012
Language	German	German
Synthesizer type:		
Formant	✓	✓
Diphone (PSOLA/MBROLA)	✓	✓
Unit Selection	✓	✓
HMM	✓	✓
Number of systems	16	20
Number of configurations	30	57
Stimuli per configuration	2	1
Quality assessment via	CS	ST/CS
Number of scales	16	16
Natural reference	no	no
Length of stimuli	10 s	5 s
f/m voices	both	both
Separate sessions for f/m	yes	yes
Preprocessing	-	-
Sampling rate	16 kHz*	16 kHz*

* Only few systems generated synthesized files of only 8 kHz.

Nonetheless, a closer look reveals that the dimension *naturalness of voice* specifies the broad dimension *naturalness* from SD/FA, that consisted of scales assessing the naturalness as well as the prosody of a system. The results show that the quality dimension *naturalness of voice* apparently covers how human-like the synthetic voice sounds. This dimension gives an answer to the question: does the listener have the impression that this TTS signal has been produced by a human being or does it sound like it was produced by a computer?

Even though the dimensions *temporal distortions* of both tests carry the same name, they represent slightly different perceptual impressions. While the dimension from the SD/FA study exclusively evaluates temporal distortions that originate from concatenation artifacts from DI or US synthesizers the dimension from the ST/MDS study also covers long-term prosodic characteristics of speech signals. Therefore, it does not only account for local concatenation errors but also for the prosody of the whole synthesized utterance.

Furthermore, the *calmness* dimension of ST/MDS corresponds to the dimension *speed* that was also detected previously in SD/FA but seemed to be of minor importance. Apparently, the impression of fast speech also evokes a feeling of stress and restlessness while a slow speaker sounds relaxed and in some cases even a bit drowsy.

A wide difference can be noticed concerning the dimension *disturbances*. While this dimension was highly relevant in the SD/FA study, the ST/MDS experiment did not capture a perceptual importance of disturbances for the participants of this study. Though listeners could clearly distinguish, e.g., the grade of noise and hiss in the signal through the attribute scales presented in the post-test, this can not be stated for the ST/MDS experiment. This effect seems to be due to the nature of most TTS signals: even though the quality improved dramatically over the years, there are still major issues that catch the attention of listeners. Those issues mainly concern naturalness and temporal distortions and become so important that minor problems like disturbances are masked by the degradations in the first two dimensions. Moreover, listeners are used to impairments like noise and hiss through, e.g., coding and transmission artifacts in cell phone or IP-based communication and might thus be more tolerant with respect to disturbances in the signal. The SD/FA clearly yielded very analytic information while the ST/MDS resulted in dimensions that are completely unaffected (e.g., by attribute scales) and thus represent the relevant dimensions that were perceived by the listeners in this test.

The analysis of the results with respect to the types of TTS systems (FO, DI, US, HMM synthesis) brought interesting insights concerning, e.g., concatenation techniques and speech rate. From a system-developer's point of view it would have also been interesting to analyze the results with respect to the inventory size or the amount of training data. But, unfortunately, this information is unknown especially for all of the commercial systems.

4.5 Universal Perceptual Quality Dimensions[10]

The differences in all presented studies concerning the quality assessment methods, the included synthesizer types, the different stimulus durations and many more yielded ambiguous results. In the following, a comparative overview of the perceptual quality dimensions resulted from the studies in Sects. 4.1, 4.2, and 4.3 is given. Moreover, it will be shown that these dimensions can be linked to five universal perceptual quality dimensions of synthetic speech, which are:

- **Naturalness of Voice (NOV)**
- **Prosodic Quality (PQ)**
- **Fluency and Intelligibility (FAI)**
- **Absence of Disturbances (AOD)**
- **Calmness (C)**

Table 4.11 shows the perceptual quality dimensions of each study and the universal quality dimensions they can be linked to, along with attributes that are relevant for the corresponding dimension.

[10]Parts of the content of this section have previously been published in a slightly different version in [1].

Table 4.11 Perceptual quality dimensions of synthetic speech

Dimensions	Relevant Scales	Kraft1995	Mayo2005	Vis2005	Seget2007	Mayo2011	SD/FA	ST/MDS
Naturalness of Voice	*naturalness* *voice pleasantness*	Prosody		Naturalness	Naturalness and Prosody		Naturalness	Naturalness of Voice
Prosodic Quality	*stress* *rhythm* *prosody* *intonation*		Prosodic Cues	Intelligibility		Unit Appropriatness and Prosody		Temporal Distortions
Fluency and Intelligibility	*fluency* *intelligibility* *bumpiness* *polyphony*	Segmental	Segmental or Unit Level Cues		Intelligibility	Overall Join Quality/Quantity Join Distribution and Detectability	Temporal Distortions	
Absence of Disturbances	*hissing* *noise* *rasping* *disturbances*						Disturbances	
Calmness	*speed* *tension*						Speed	Calmness

4.5.1 Naturalness of Voice

As can be seen in Table 4.11, the dimension NOV is part of the outcome of most studies. High values in this dimensions indicate a voice that sounds natural, a voice where listeners can imagine that it was produced by a human being, a voice with character. The exceptions are the two MDS experiments (Mayo2005 and Mayo 2011). However, this can be explained considering the stimuli from those tests: they were all generated from the same voice by the Festival synthesizer. Thus, none of them differed in voice characteristics. Therefore, this dimension was not relevant for the test participants at that point in time.

4.5.2 Prosodic Quality

Due to the overlap of the first two dimensions in some studies (Kraft1995, Seget2007, SD/FA) and the bonding with the FAI dimension in others (Vis2005, ST/MDS), the second dimension seems to be a bit more vague. Mayo2005 and Mayo2011 show that this dimension can indeed be regarded as an independent dimension. It represents the prosody of the synthesized utterance, thus high values in PQ imply natural prosody while low values occur when rhythm or stress of the generated voice deviate from that of a regular human speaker.

4.5.3 Fluency and Intelligibility

The third prominent dimension covers fluency and intelligibility and it can be found in all studies. This dimension captures segmental artifacts that are characteristic for synthesizers that concatenate smaller units like PSOLA synthesizers or diphone-based US systems. This can create the impression of a stuttering speaker or even two speakers speaking at the same time. The MDS study Mayo2011 shows that this dimension can be further split up, at least for US synthesizers. On the contrary, the overlap of PQ and FAI in the ST/MDS experiment shows that these two dimensions are not always easy to distinguish for naïve listeners.

4.5.4 Absence of Disturbances

The dimension AOD could only be retrieved from the extensive experiments in SD/FA. This is most likely due to the fact that the presented scales were developed with the help of speech and audio experts who are able to focus on various types of degradations. Even though the test participants could clearly distinguish, e.g.,

the grade of noise and hiss in the signal, these degradations were obviously less important to them than issues concerning the naturalness of the voice or the prosody of the signal. Nonetheless, this dimension can be useful to assess, e.g., the quality of HMM synthesizers or systems that concatenate coded speech units which can produce noisy speech signals.

4.5.5 Calmness

Finally, the dimension C was found in SD/FA and ST/MDS. It is associated with the rate of speaking. However, a high rate of speaking not only evokes the impression of a person speaking fast, the generated voices also sounded very tense, stressed, and restless while slowly speaking voices generally generated the impression of a relaxed speaker. This dimension, however, appears to be less important since most of the speech synthesizers run at a similar speech rate. Nonetheless, when assessing the quality of fast synthesizers, like they are deployed in reading devices for the blind [22], this quality aspect can play a crucial role.

These five dimensions measure all relevant perceptual features state-of-the-art TTS systems produce. Nonetheless, this does not mean that they are necessarily relevant in each study (see Table 4.11). Depending, e.g., on the kind of quality assessment method or the synthesizer types under test, some dimensions might overlap, correlate highly with one another, or might not be relevant at all, while in another test scenario they are strongly needed to distinguish the quality of two TTS systems.

4.5.6 Instructions for TTS Quality Assessment

In order to reliably assess the aforementioned quality dimensions, this section presents a test protocol which features attribute scales for each dimension and questions that further explain the perceptual concept behind each dimension.

Table 4.11 not only shows the perceptual dimensions that resulted from each study, but also attribute scales that are relevant for the assessment of each dimension. The presented scales were part of several studies and featured high factor loadings in their respective factor analyses. The following list links two of those attribute scales to each dimension and depicts questions that help a test participant to better understand these scales.

- **NOV:**
  attribute scales:
  - voice pleasantness: *unpleasant voice versus pleasant voice*
  - naturalness: *artificial versus natural*

Does the speaker have an unpleasant or a pleasant voice?
Does the voice sound artificial or natural?
Do you have the impression of listening to a human being?

- **PQ:**
  attribute scales:
  - stress: *unnatural stress versus natural stress*
  - rhythm: *unnatural rhythm versus natural rhythm*

  Are the pronounced words emphasized naturally?
  Does the intonation sound natural to you?

- **FAI:**
  attribute scales:
  - intelligibility: *unintelligible versus intelligible*
  - fluency: *interrupted versus fluent*

  How easy is it to understand the speaker?
  Do you have problems to comprehend some of the words?
  Is the speaker stuttering or does the voice sound fluent?

- **AOD:**
  attribute scale:
  - disturbances: *disturbed versus undisturbed*
  - noise: *noisy versus not noisy*

  Do you recognize any disturbances in the signal, e.g., noise, hissing, clinking, or rasping sounds?

- **C:**
  attribute scales:
  - speed: *slow versus fast*
  - tension: *tense versus calm*

  Does the speaker speak slowly or fast?
  Does the speaker sound relaxed or restless?

The listening test should be designed according to the instructions below:

- *Speech material*:
  In order for the test participants to be able to observe all quality features, a minimum stimulus duration of 5 s is recommended. Utterances containing place names, proper names, or words from a foreign language should be excluded as they often differ from pronunciation rules of the language under test, which is therefore likely to cause problems for speech synthesizers.

- *Attribute scales*:
  To simplify the task of differentiating between stimuli of an either very good or very bad quality, scales should feature separate scale and end points as shown in Fig. 3.3. If a scale covering the *overall impression* of the listeners is included in the test, it should be rated prior to the remaining scales, in order to guarantee that the test participants first think about their impression of the system as a whole and after that about degradations of specific quality features.
- *Test procedure*:
  The listening test should take part in a quiet listening environment, preferably in a sound proof booth. Each test should start with a short training session in which each participant has the possibility to get used to the setup, e.g., the utilized GUI. This training session should consist of high and low quality stimuli so that the participants get an idea of the quality range that will be presented in the main test. To avoid listener fatigue, participants should take a short break if the experiment lasts longer than 30 min. Female and male voices should be rated in separate sessions in order to avoid any gender specific preferences. And finally, to avoid any impact with regard to the order of the scales or the order of the stimuli, both the sequence of scales and the playlist of the stimuli should be randomized between subjects.

4.6 Summary

Two studies towards perceptual quality dimensions of synthetic speech were presented in this chapter. Comparing the findings of these studies with the outcome of several studies that have been conducted in the past, lead to a set of five perceptual quality dimensions that can be seen as universal for state-of-the-art TTS systems. They were named: **Naturalness of Voice (NOV)**, **Prosodic Quality (PQ)**, **Fluency and Intelligibility (FAI)**, **Absence of Disturbances (AOD)**, and **Calmness (C)**. Moreover, a test protocol was designed that is able to capture a listener's impression of each of the dimensions.

References

1. Hinterleitner F, Norrenbrock C, Möller S (2013) Is Intelligibility Still the Main Problem? A Review of Perceptual Quality Dimensions of Synthetic Speech. In: Proceedings of the 8th ISCA Speech Synthesis Workshop (SSW 2013), pp 167–171
2. Kraft V, Portele T (1995) Quality Evaluation of Five German Speech Synthesis Systems. Acta Acust 3:351–365
3. Mayo C, Clark RAJ, King S (2005) Multidimensional Scaling of Listener Responses to Synthetic Speech. In: Proceedings of the 6th Annual Conference of the International Speech Communication Association (Interspeech), pp 1725–1728

4. The Centre for Speech Technology Research, The University of Edinburgh. The Festival Speech Synthesis System. http://www.cstr.ed.ac.uk/projects/festival/. Accessed 08 Jan 2016
5. Garofolo JS (1988) Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonteic Continuous Speech Database. National Institute of Standards and Technology (NIST), Gaithersburgh, MD
6. ITU-T Rec. P.85, (1994) A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices. International Telecommunication Union, Geneva
7. Viswanathan M, Viswanathan M (2005) Measuring Speech Quality for Text-to-Speech Systems: Development and Assessment of a Modified Mean Opinion Score (MOS) Scale. Comput Speech Lang 19:55–83
8. Seget K (2007) Untersuchungen zur auditiven Qualität von Sprachsyntheseverfahren (Study of Perceptual Quality of Text-to-Speech Systems). Diplomarbeit, Lehrstuhl für Netzwerk- und Systemtheorie, Christian-Albrechts-Universität Kiel
9. ITU-T Rec. G.712, (2001) Transmission Performance Characteristics of Pulse Code Modulation Channels. International Telecommunication Union, Geneva
10. Hinterleitner F (2010) Vorhersage der Qualität synthetischer Sprache mittels eines signalbasierten Maßes. Magisterarbeit, Quality and Usability Lab, TU Berlin
11. ITU-T Rec. G.711, (1993) Pulse Code Modulation (PCM) of Voice Frequencies. International Telecommunication Union, Geneva
12. Mayo C, Clark RAJ, King S (2011) Listeners' Weighting of Acoustic Cues to Synthetic Speech Naturalness: A Multidimensional Scaling Analysis. Speech Commun 53:311–326
13. Hinterleitner F, Möller S, Norrenbrock C, Heute U (2011) Perceptual Quality Dimensions of Text-to-Speech Systems. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011), pp 2177–2180
14. Chan D, Fourcin A, Gibbon D, Grandstrom B, Huckvale M, Kokkonakis G, Kvale K, Lamel L, Lindberg B, Moreno A, Mouropoulos J, Senia F, Trancoso I, Veld C, Zeiliger J (1995) EUROM-A Spoken Language Resource for the EU. In: Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH 1995), pp 867–870
15. ITU-T Rec. P.56, (1993) Objective Measurement of Active Speech Level. International Telecommunication Union, Geneva
16. Hinterleitner F, Norrenbrock C, Möller S, Heute U (2012) What Makes this Voice Sound so Bad? A Multidimensional Analysis of State-of-the-Art Text-to-Speech Systems. In: Proceedings of the 2012 IEEE Workshop on Spoken Language Technology (SLT 2012), pp 240–245
17. Borg I, Groenen P (2005) Modern Multidimensional Scaling - Theory and Applications, 2nd edn. Springer Series in Statistics, New York
18. Kruskal J, Wish M (1978) Multidimensional Scaling. In: Quantitative Applications in the Social Sciences, vol 07–11. Sage
19. Tsogo L, Masson MH, Bardot A (2000) Multidimensional Scaling Methods for Many-objects Sets: A Review. Multivar Behav Res 35:307–319
20. Seber GAF (1984) Multivariate Observations. Wiley, New York
21. Sturrock K, Rocha J (2000) A Multidimensional Scaling Stress Evaluation Table. Field Methods 12(1):49–60
22. Moos A, Trouvain J (2007) Comprehension of Ultra-fast Speech - Blind vs. "Normally Hearing" Persons. In: Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, pp 677–680

Chapter 5
Influencing Factors on Perceptual Quality

The previous chapter reviewed (a) the results of several studies concerning perceptual quality dimensions of TTS, (b) described original work related to this domain, and (c) introduced a set of five universal quality dimensions. This chapter pursues research in this domain by discussing factors influencing these dimensions. Thus, this chapter addresses RQ3:

Which factors influence these perceptual quality dimensions?

Until now, the TTS stimuli under test featured a duration of a couple of seconds which is typical for applications such as email and short message readers as well as smart-home systems. Since the quality of TTS systems has noticeably increased over the past years, using speech synthesizers to read books has suddenly become a feasible task. For this application, however, quality features different from the ones in the use case of short message readers can suddenly come into focus, or at least the weighting of the known features can be altered. Therefore, a test protocol for the auditory quality assessment of synthesized audiobooks is developed in Sect. 5.1.1 and tested in Sect. 5.1.2.

A further important influencing factor on data-driven systems is the voice of the speaker a TTS system is based on. How the perceived quality is affected by a speakers voice, and which aspects of a voice are relevant for a TTS system, is examined in Sect. 5.2.

Moreover, the quality of popular data-driven TTS approaches like US and HMM synthesis also depends on the size of the natural speech database and the training database, respectively. To which extent the size of the speech corpus affects a US system is shown in Sect. 5.3 by creating new voices for the MaryTTS synthesizer.

Thus, the studies presented in this chapter examine (i) the influence of the *application* a synthesizer is used in (TTS in audiobook reading tasks vs. TTS for short messages), (ii) the importance of the *voice of the speaker*, and (iii) the impact of the *size of the speech corpus*. The first factor is chosen due to recent research on this topic in the Blizzard Challenge and the second and third factor were selected because

F. Hinterleitner, *Quality of Synthetic Speech*, T-Labs Series
in Telecommunication Services, DOI 10.1007/978-981-10-3734-4_5

of the general popularity of corpus-based speech synthesis. However, other factors, e.g., the gender of the speaker, selection strategies of unit selection synthesizers, or the compilation of a speech database for voice building, are also worth investigating and should thus be part of future research.

5.1 Influence of the Application[1]

Since 2005, the Blizzard Challenge (BC) gathers developers of TTS systems to compare techniques in building corpus-based speech synthesizers. The fact that all participants get the same speech corpus to build their systems on assures a comparability between all synthesizers. In 2011 the BC has announced a special task concerning audiobooks read by TTS systems for 2012. With this new application area, quality aspects of TTS like listening effort, the ability for emotional speech, or the placing of speech pauses in a way that supports the comprehension of the text, get more important.

Depending on which aspect of the system is to be evaluated, different types of listening tests are recommended: articulation and intelligibility experiments [2] test whether the TTS signal is able to carry information on a segmental or supra-segmental level; comprehension tests [3] show if the listener can discern the content; and overall quality tests [4] capture different global quality aspects and dimensions [5]. However, none of these methods is specialized in measuring quality aspects like the ones mentioned in the previous paragraph. Therefore, new ways of evaluating the quality of TTS systems have to be designed.

Thus, this section addresses RQ3. More precisely, the following questions will be answered:

- Which perceptual dimensions are relevant for synthesized audiobooks?
- How should a listening test be designed to capture these perceptual dimensions?

Therefore, this section presents the development and application of an evaluation protocol for the subjective assessment of TTS in audiobook reading tasks. In Sect. 5.1.1, an overview of the experimental setup of the pretest including the used speech material, the TTS stimuli, the design of the experiment, the test procedure, the analysis of the test data and a discussion of the results is given. Finally, Sect. 5.1.2 presents the outcome of the main test during the BC 2012.

5.1.1 Pretest

This section presents the development of an auditory quality assessment protocol for quality evaluation of TTS in audiobook reading tasks.

[1] Parts of the content of this section have previously been published in a slightly different version in [1].

Table 5.1 Selected books for the audiobook pretest

ID	Text Type	Author	Book
1	Long sentences	Sven Regener	*Der kleine Bruder**
2	Direct speech, incomplete sentences	Douglas Adams	*The Hitchhiker's Guide to the Galaxy*
3	Higher level of lexis, complex sentence structure	Charles Dickens	*The Adventures of Oliver Twist*
4	Poetic, picturesque	Antoine Saint-Exupéry	*Wind, Sand and Stars*
5	Direct speech, basic language	Tommy Jaud	*Resturlaub**
6	Action, short sentences	Thomas Harris	*Hannibal*
7	Children's book	Astrid Lindgren	*Pippi Longstocking*
8	Thriller	Ken Follett	*Code to Zero*

* no English translation available

5.1.1.1 Experimental Setup

Test Database

Eight passages from German issues of the books in Table 5.1 were chosen as material for the listening test. These paragraphs were selected with the objective to cover a wide variety of writing styles and book categories including thrillers, funny books, action-packed passages, books for children, books with very long sentences, and passages containing almost only direct speech.

In order to ensure a high quality listening experience, two German unit selection voices of the TTS systems CereProc CereVoice (CV, female: Gudrun, male: Alex) and the IVONA (IV, female: Marlene, male: Hans) were selected for the listening test. Each of the systems was used to synthesize the same passages from the books listed in Table 5.1. TTS systems that had problems synthesizing English names had to be manually adjusted to ensure a normal pronunciation. Additionally, four samples were synthesized by the IV Marlene voice and subsequently manually optimized (IVO). The optimization included an adjustment of wrong articulated words and an improvement of pauses between sentences and paragraphs. The mean length of all stimuli was 54.7 s with an average of 138 words.

Attribute Scales

Since this study was conducted prior to the design of the test protocol introduced in Sect. 4.5.6, an inclusion of the highlighted attribute scales was not an option. Therefore, attribute scales from the ITU-T Rec. P.85 (see Sect. 3.3.2.1) were selected for the evaluation of the generated stimuli. The chosen items were modified in order to adjust them to the task of assessing audiobook stimuli. Furthermore, attribute

scales were added to assess specific quality features that should be considered when evaluating TTS audiobooks.

The additional items were selected based on the review of current literature. Prosodic elements like communicative, structuring, aesthetic, and emotional aspects can be seen as the most important factors for reading and interpreting books [6], thus most of the selected scales are focused on prosodic evaluation (STRE, COPR, SPAU, INTO, EMOT). The 11 items are discussed in the following:

- *Overall Impression (OIMP)*:
  This scale evaluates the overall quality of the synthesized signal from *bad* to *excellent*.
- *Voice Pleasantness (VPLT)*:
  Measures the degree of voice pleasantness from *very unpleasant* to *very pleasant*.
- *Stress (STRE)*:
  Unnatural stress and accentuations are often perceived as very annoying and thus also have a great influence on the text comprehension [6]. Therefore, the item covering anomalies in pronunciation from the ITU-T Rec. P.85 was included with slight modifications.
- *Listening Effort (LSTE)*:
  Describes the effort a listener is required to make when listening to this voice over a longer period of time.
- *Comprehension Problems (COPR)*:
  This item captures any comprehension problems that might occur due to badly synthesized speech.
- *Acceptance (ACCP)*:
  The binary acceptance item from the ITU-T Rec. P.85 questionnaire was modified into a continuous rating scale.
- *Speech Pauses (SPAU)*:
  The SPAU item evaluates if punctuation marks (e.g., periods, commas, question marks, exclamation marks, colons, etc.) have been converted into appropriate speech pauses between words, sentences, and paragraphs in a way that supports the comprehension of the text [7].
- *Intonation (INTO)*:
  This item captures if the produced pitch curve fits to the type of sentence, e.g., the pitch of interrogative sentences usually increases at the end of a sentence whereas the pitch of declarative sentences decreases [6, 7].
- *Emotion (EMOT)*:
  Variation of emotion is achieved by variations of sound pressure, intonation, speech pauses and volume [6]. To ensure an authentic reading experience, the voice should reflect the atmosphere of the scene and the moods of the characters [8]. This is captured by the EMOT item.

Furthermore, the following two attribute scales were included to test whether they influence the test participants's judgment on the other items:

- *Content (CONT)*:
  This item measures if the participants liked the content of the synthesized passage on a scale from *no, not at all* to *yes, very much*.
- *Level of Familiarity (LOFA)*:
  LOFA captures if the participants knew the book prior to the listening test.

Test Procedure

Twenty-five naïve subjects aged 19–32 years (13 female, 12 male, $\mu = 25.3$, $\sigma = 3.1$) took part in the test. All of them were native German speakers. None of them suffered from any known hearing problems or dyslexia. All subjects were paid for their participation. The stimuli were presented via headphones (AKG K601) and a high-quality sound device (Roland Edirol UA-25) in a soundproof booth. The listening test was designed within subjects, i.e., all participants listened to all stimuli.

The subjects were instructed to first rate the OIMP of the stimulus on a continuous rating scale. Subsequently, quality estimates for the other nine scales had to be given via a slider presented on the GUI. This was done with the objective to first think about the overall quality and after that about specific quality features of the presented stimuli. In the end, the test participants had to rate if they knew the presented passage prior to the listening test. To avoid any impact with regard to the order, the sequence of scales (except for OIMP, CONT and LOFA) was randomized between subjects. To make themselves familiar with the test procedure, all subjects first had to pass a training phase with two stimuli that were not included in the main test. The main test consisted of two blocks with 18 stimuli and a 5 min break in between.

After the test, boxplots featuring the ratings on all dimensions were created for all stimuli. Two test participants had more than 5% outlier ratings and were thus excluded from the study.

5.1.1.2 Statistical Analysis

This work has been previously published in [1, 9] in a slightly different version. Further analysis on the data is presented by placing the focus of the statistical analysis on the derived quality dimensions. Therefore, the influence of *text type* and *TTS system* on the extracted factors is examined.

Factor Analysis

To get an impression of the perceptual space that was captured in the listening test, a PCA was carried out. All items except for LOFA, which only had nominal scale level, and OIMP, since it comprises the information from the other scales, were included.

Three factors were extracted which account for 71.8% of the total variance. Subsequently, an oblique rotation (Promax rotation with $\kappa = 4$) was performed in order to

Table 5.2 Factor pattern matrix. *Note* For better readability, values below .20 are suppressed

Scale	Factor loadings		
	1	2	3
Voice Pleasantness	.96		
Acceptance	.89		
Listening Effort	.82		
Comprehension Problems	.44	.43	
Intonation		.95	
Speech Pauses		.83	
Emotion	.24	.61	
Stress	.30	.55	
Content			.98
Eigenvalues	4.56	0.98	0.92
% of variance	50.1	10.8	10.3
Cronbach's α	.86	.79	–

Table 5.3 Factor correlation matrix

Factor	1	2	3
1	1.00	.62	.13
2	.62	1.00	.14
3	.13	.14	1.00

obtain interpretable factors since correlated dimensions were assumend. The resulting factor pattern matrix can be seen in Table 5.2.

Due to the oblique rotation method the factors are no longer orthogonal. The correlations between the three factors can be seen in Table 5.3.

To ensure a meaningful interpretation of the perceptual space, items with high cross-loadings[2] will not be taken into account. In this case this only applies to COPR.

Factor 1 includes the items VPLT, LSTE, and ACCP. Thus, it covers the perceptual construct that is related to the **listening pleasure** the TTS systems achieve. With the high loading of the items INTO and SPAU, the second dimension seems to reflect the **prosody** of the signal. Moreover, the two other scales that account for factor 2 express emotion and stress. Factor 3 only consists of the item CONT thus it represents the **content appreciation**.

[2] $|\text{loading factor A} - \text{loading factor B}| < .20$.

Resulting Quality Dimensions

The mapping of the stimuli in the perceptual space of dimension 1 and 2 is displayed in Fig. 5.1, in which the subscripted characters (f/m) represent the speaker gender and the subscripted numbers (1–8) the text ID from Table 5.1. As it can be seen the stimuli form one cluster for each of the synthesis systems. The stimulus IV_{f3} is the only outlier. Its values for the dimension 1 and 2 are far below the mean value for the IV system. This impression is confirmed by the ratings this stimulus achieved in the listening test: it scored lowest on INTO and VPLT while getting the highest values in COPR and LSTE of all female IV stimuli.

It is surprising that two stimuli of the manually optimized version of the female voice of IVONA (IVO_{f4}, IVO_{f7}) are inferior in both dimensions to the stimuli synthesized by the original TTS system. However, IVO_{f8} performed better than IV_{f8} and even achieved the best dimension 1 value of all stimuli. This shows that simply by adjusting wrong articulated words and improving the pause lengths between sentences and paragraphs even the quality of good synthesizers can be improved, but also worsened.

In addition, it is noticeable that most stimuli of the same *text type* and the same *system* reach similar prosody values (e.g., CV_{f7} and CV_{m7}, IV_{f5} and IV_{m5}). However, for a given system the listening pleasure of the male stimuli is always superior to the female voices. This accounts for the data from the CV as well as the IV synthesizer

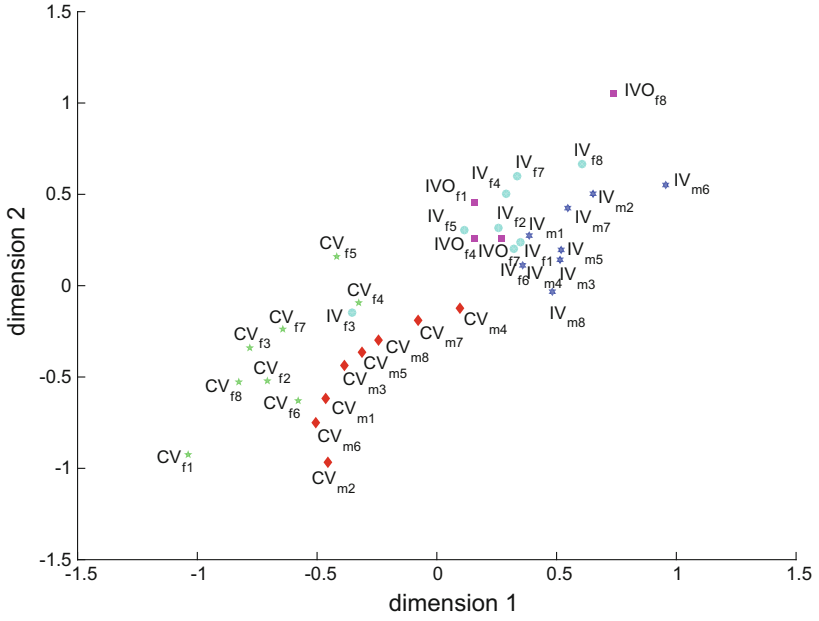


Fig. 5.1 Factor loadings for the dimensions 1 and 2

Table 5.4 Linear model of predictors for overall impression. *Note* $R^2 = .71$

	B	SE B	β	p
CONSTANT	3.36	0.02		$p \leq .001$
FACTOR 1	0.61	0.03	.59	$p \leq .001$
FACTOR 2	0.37	0.03	.35	$p \leq .001$
FACTOR 3	−0.04	0.02	−.04	$p = .029$

(except for text 8). Furthermore, it can be stated that both synthesizers have difficulties with different kinds of texts, e.g., the stimuli of *text type* 4 are one of the best rated stimuli of synthesizer CV whereas IV_{m4} is one of the worst rated of IV.

These findings already show that the quality of the synthesized speech signal mostly depends on the *system* and not on the *text type*, even though the quality of the tested synthesizers varies highly between *text types*. Further statistical analysis on this effect is presented in the following sections.

Scatter plots of dimension 3 in comparison to dimension 1 or 2 showed no interesting clusters.

Importance of the Factors for the Overall Quality

To unveil the importance of each dimension for the experienced overall quality, a linear regression was performed with the dimensions as predictors and OIMP as outcome variable. The result can be seen in Table 5.4. The table lists the three factors, their beta values (B), their standard errors (SE B), and their standardized values (β).

The model explains 71% of the total variance. Moreover, the table shows that Factor 1 contributes most to the overall impression ($\beta = .59$). Factor 2 reaches a β value of .35. Factor 3 that only consists of the item CONT only reaches a β value of $-.04$ and is thus irrelevant for the perceived overall quality.

Influence of *Text Type* and the TTS *System*

In order to get deeper insight into the data, a MANOVA was performed. It further examines how the dependent variables *text type* and *system* (CV, IV and IVO with male and female voices) influence the three dimensions. Using Pillai's trace, a significant effect of *text type* ($V = 0.10$, $F(21, 2379) = 3.6$, $p \leq .001$) and *system* ($V = 0.28$, $F(12, 2379) = 20.5$, $p \leq .001$) was found while no combined effect (*text type*system*) could be discovered ($V = 0.10$, $F(72, 2379) = 1.1$, $p = .249$).

The main effect of the independent variable *text type* on the dimensions is displayed in Table 5.5.

Therefore, the *text type* does not have a significant effect on the *listening pleasure* but on the *prosody* of the synthesized signal. Moreover, the *text type* also has a highly

Table 5.5 Effect of *text type* on the three dimensions

	F	p
Dimension 1	1.933	$p = .062$
Dimension 2	2.238	$p = .029$
Dimension 3	7.649	$p \leq .001$

Table 5.6 Effect of *system* on the three dimensions

	F	p
Dimension 1	55.159	$p \leq .001$
Dimension 2	36.756	$p \leq .001$
Dimension 3	0.308	$p = .873$

significant effect on dimension 3. This is not at all surprising since this dimension captures whether a test participant liked the content of the presented stimuli or not.

The effect of the TTS *system* is presented in Table 5.6.

The system has a highly significant effect on dimensions 1 and 2 while it does not have an effect on dimension 3.

Influence of Familiarity of Texts

To analyze whether the test participants were biased if they knew some of the books prior to the listening test, the data was split in one set of ratings that were given by subjects that did not know the synthesized passages and one set with ratings by subjects that knew the passages. The mean values of each stimulus from each dataset were compared and no significant differences could be found. Therefore, the LOFA of the text does not introduce any bias on the ratings of other scales.

5.1.1.3 Discussion and Suggestions for the Blizzard Challenge

A PCA on the gathered data from the listening test revealed three dimensions: one that covers the listening pleasure, one that comprises prosodic impressions, and one that represents whether a test participant likes the content of the presented stimulus.

A MANOVA revealed that the *system* (combination of TTS system and male/female voice) has a significant influence on the dimensions 1 and 2 while it does not have an influence on dimension 3. Moreover, *text type* has a significant influence on the dimensions 2 and 3.

Therefore, the following suggestions were forwarded to the Blizzard Challenge organizers:

- the *system* has a highly significant effect on the dimensions 1 and 2 which were shown to be most relevant for the overall impression. Therefore, items that capture these dimensions should be included in a questionnaire.
- the *text type* has a significant (but minor, when compared to the *system*) influence on the values of dimension 2, thus the use of a variety of texts of different categories and with different writing styles was proposed.
- the *Level of Familiarity* does not have a significant influence on the ratings on other scales. Therefore, this item can be excluded from further tests.
- the item *Comprehension Problems* has high cross-loadings on the dimensions 1 and 2 and should thus be dropped since it does not help to discern between the 2 dimensions.
- the item *Content* represents the 3rd dimension from the factor analysis. Since the *system* does not have a significant effect on it and it is also not relevant for the overall impression this item should not be included in further tests.

5.1.2 Main Test[3]

Considering the findings from the previous section, the organizers of the Blizzard Challenge chose the experimental setup for the audiobook task which is described in the following section.

5.1.2.1 Experimental Setup

This section shows an overview of the books that were synthesized and presented in the listening test. Moreover, the attribute scales that were used to evaluate the stimuli are presented and the test procedure of the online listening test is described.

Test Database

Due to the large number of stimuli that needed to be evaluated for the BC, the listening test was carried out online. Therefore, it was open to the public and the texts that were synthesized were limited to copyright-free and out-of-copyright sources.[4] Books were selected with the idea to cover a wide variety of writing styles and book categories as proposed in Sect. 5.1.1.3. The 13 chosen passages featured a mean stimulus duration of 44.5 s. The full list of authors and books can be seen in Table 5.7.

[3]The content of this section has previously been published in a slightly different version in [10].

[4]Many Books http://manybooks.net/, Project Gutenberg http://www.gutenberg.org.

Table 5.7 Selected books for the audiobook task of the BC

Author	Book
Jane Austen	Emma
Charles Dickens	Oliver Twist
Arthur Conan Doyle	The Hound of the Baskervilles
Alexandre Dumas	The Three Musketeers
Jerome K. Jerome	Three Men in a Boat
Franz Kafka	The Trial
Edgar Allan Poe	The Fall of the House of Usher
Mary Shelley	Frankenstein or the Modern Prometheus
Mark Twain	Alonzo Fitz
Mark Twain	Those Extraordinary Twins
Jules Verne	Twenty Thousand Leagues Under the Sea
H.G. Wells	Time Machine
P.G. Wodehouse	My Man Jeeves

Test Procedure

The online listening test consisted of nine sections of which two were dedicated for the evaluation of audiobook stimuli. One of the audiobook sections also included a natural reference voice while the other one did not. Both sections consisted of 10 synthetic stimuli from 10 different TTS systems. All in all 230 audiobook stimuli were evaluated during the listening test. The test was designed between subjects, therefore each stimulus was rated by at least 20 test participants.

Taking into account the findings from the pretest in Sect. 5.1.1 the following seven attribute scales were chosen for the online evaluation:

- Overall Impression (OIMP)
- Voice Pleasantness (VPLT)
- Speech Pauses (SPAU)
- Stress (STRE)
- Intonation (INTO)
- Emotion (EMOT)
- Listening Effort (LSTE)

The scores on each scale were given on a continuous slider. Just like in the pretest, attribute scales with separate scale and end points were used as proposed by Bodden and Jekosch [11]. A screenshot of the GUI from the online listening test can be seen in [10].

5.1.2.2 Statistical Analysis

Factor Analysis

A PAF analysis was conducted on the seven presented items. The KMO measure verified the sampling adequacy for the analysis, KMO = .90, and all KMO values for the individual items were >.86, which is well above the acceptable limit of .50. Bartletts's test of sphericity, $\chi^2(21) = 21459.16$, indicated that correlations between the items were sufficiently large for a PAF. An initial analysis was run to obtain the eigenvalue for each component in the data. Only one item had an eigenvalue over Kaiser's criterion of 1. The scree plot also indicated to retain one component. However, given the very large sample size (4809), the easy interpretability after an extraction of two factors, and the fact that a 2-factor model allows a deeper insight into the human perception of TTS signals, two components were retained in the final analysis. These two factors together explained 75.61% of the total variance.

Since correlated quality dimensions were assumed, an oblique rotation method (Promax rotation with $\kappa = 4$) was chosen. The resulting factor pattern matrix can be seen in Table 5.8. For clarity, values below .40 are suppressed. Due to the oblique rotation the factors are not orthogonal, rather they correlate with .78.

Resulting Quality Dimensions

To achieve a meaningful interpretation of the resulting quality dimensions, items with high cross-loadings, meaning high loadings on both factors, were omitted before the interpretation. This only applies to the item LSTE.

Given the high loadings of the items INTO and STRE and the medium loading of SPAU on factor 1, this dimension is clearly linked to the **prosody** of the speech signal.

Table 5.8 Factor pattern matrix. *Note* For better readability, values below .40 are suppressed

Scale	Factor loadings	
	1	2
Intonation	.86	
Stress	.76	
Emotion	.60	
Speech Pauses	.54	
Voice Pleasantness		.90
Overall Impression		.87
Listening Effort	.44	.47
Eigenvalues	4.67	0.62
% of variance	66.8	8.9
Cronbach's α	.84	.91

The item EMOT which assesses the ability of TTS systems to synthesize appropriate emotions, i.e., through prosodic modulations, also contributes to this interpretation.

Dimension 2 correlates highly with the items VPLT and OIMP. This indicates that dimension 2 is related to the voice of the speaker as well as the naturalness of the signal and the overall experience. Thus, this dimension can be associated with the **listening pleasure**. The affiliation of the item OIMP indicates that this dimension is in fact more important for the quality impression of the listener than dimension 1.

5.1.2.3 Discussion

Comparing the results from this study to the outcome of the pretest in Sect. 5.1.1 reveals major similarities. The assignment of the items to the quality dimensions is nearly the same as in the previous study. Thus, both dimensions represent about the same perceptual quality impression of the user as before. Nevertheless, there are some differences: first of all, the item LSTE which highly correlated with the dimension *listening pleasure* before, has very high cross-loadings in the current study and thus did not help to discern between the factors.

Interestingly, the importance of both dimensions shifted in the current study. While in Sect. 5.1.1 the dimension *listening pleasure* explained most of the variance in the data and was thus the first factor, the order of importance is reversed in the current study. Nonetheless, since the item OIMP correlates highly with factor 2 here, this dimension is clearly the most relevant when it comes to the overall quality impression of the user.

5.1.3 Conclusions

The development and testing of an evaluation protocol for synthesized audiobooks has been described in this section. A factor analysis revealed two main dimensions, namely *listening pleasure* and *prosody*. Furthermore, the pretest led to a third dimension which was labeled *content appreciation*. However, since this dimension has no significant importance for the overall quality it can be disregarded. With high correlations of VPLT and OIMP with the dimension *listening pleasure* it is save to assume that it can be associated with the universal quality dimension NOV. High correlations of INTO and STRE with dimension prosody suggest a relation between this dimension and the PQ.

Even though these two dimensions are most relevant to the overall quality of a synthesized book, other dimensions of minor perceptual importance should not be ruled out. Due to the long duration of the stimuli it is possible that the impression of listeners is mainly influenced by those two dominant perceptual constructs, nonetheless a dimension that, e.g., classifies concatenation artifacts, may exist here as well. This could be tested via a PC test/ST with subsequent MDS. However, performing a

direct auditory comparison of stimuli of such a long duration is an almost impossible task for test participants.

Furthermore, it could be revealed that the familiarity of the presented books has no significant influence on the quality ratings. The synthesized passage itself however influences the prosodic dimension. Therefore, when assessing the quality of TTS in synthesized audiobook tasks, a wide variety of different text types should be used.

5.2 Influence of a Speakers Voice[5]

In order to present a comprehensive overview of TTS quality, most of the studies mentioned in Chap. 4 evaluated numerous synthesizers. Therefore, it is not possible to draw a conclusion on the effect of the speaker of a speech corpus (i.e., the "voice" of the synthesizer) on the quality of a TTS system. To measure such an effect, stimuli of different voices generated by one TTS system would have to be evaluated. In this section a study is presented which used two synthesizers, each with five different voices, to unveil the impact of a speakers voice on the quality of the synthesized speech samples perceived by a user.

In addition to the speaker's voice there is always an influence of the size of the speech corpus on the quality of the synthesizer (more on this topic will be discussed in Sect. 5.3). In this study equal corpora sizes within the TTS systems under test are assumed.

Thus, this section addresses RQ3 by answering the following questions:

- What is the influence of the voice of the speech material on the quality of a TTS system?
- Which perceptual quality dimensions are affected by a speaker's voice?

The remainder of this section is organized as follows: in Sect. 5.2.1 the experimental setup is described. This includes the TTS systems that were used in the test, the speech material that was synthesized, the attribute scales that were used to rate the stimuli, and a description of the test procedure. Section 5.2.2 outlines the statistical analyses that were conducted to extract perceptual dimensions and to determine the influence of a speaker's voice on the perceived quality of a TTS system, and Sect. 5.2.3 concludes the presented findings.

5.2.1 Experimental Setup

This section gives an overview of the TTS systems, the speech material, and the rating scales that were used in the experiment. Moreover, it describes the test procedure.

[5]The content of this section has previously been published in a slightly different version in [12].

Table 5.9 Selected utterances used in the listening test

Utterance-ID	Utterance
1	*It's my turn to feed the baby again tonight. I hope she's not off her food. Then there's the bath and getting her ready for bed.*
2	*My sister is terrified of the dark. She absolutely refuses to go out alone at night. She wants someone to go with her all the time.*
3	*I'm in a mood for something light and entertaining. There's sure to be some old American musical or other. They certainly don't make them like that anymore nowadays.*
4	*There seems to have been some mistake. I ordered a teddy bear from the catalog and was billed for an electric lawnmower. And I don't even have a garden.*
5	*Singing is an expression of deep-felt emotion. It can indicate extreme happiness and deep sorrow. Strangely, though, different nations tend to one or the other end of the spectrum.*

5.2.1.1 Test Database

In order to be able to make a valid statement about the influence of a speaker's voice on the quality of a TTS system the synthesizers IVONA (IV) and ACAPELA (AC) were selected. These two unit selection systems were chosen because of their high quality and the amount of different voices. three female (IV: f1, f2, f3; AC: f4, f5, f6) and two male (IV: m1, m2; AC: m3, m4) American English voices were chosen in both cases to synthesize the five utterances that are introduced in the following paragraph. The stimuli used for the listening test had an average duration of 9 s.

Five English passages from the EUROM.1 corpus [13] were selected as material for the listening test. Utterances containing place names, proper names, or words from a foreign language were excluded as they often differ from English pronunciation rules, this is likely to cause problems for speech synthesizers. The five passages used in the listening test are shown in Table 5.9.

5.2.1.2 Attribute Scales

For this listening test, the following two attribute scales for each of the perceptual quality dimensions introduced in Sect. 4.5 were selected:

- **Naturalness of Voice (NOV)**: naturalness, voice pleasantness
- **Prosodic Quality (PQ)**: stress, rhythm
- **Fluency and Intelligibility (FAI)**: fluency, intelligibility
- **Absence of Disturbances (AOD)**: noise, disturbances
- **Calmness (C)**: speed, tension

Moreover, the following scales which have proven valuable when evaluating the likeability of a voice [14] were included:

- Darkness: dark versus bright
- Pitch: low versus high
- Pitch Variation: monotonous versus varied
- Pressure: pressed versus lax
- Resonance: toneless versus sonorous
- Likeability: unlikeable versus likeable

Additionally, the Overall Impression (OIMP) was assessed on a scale ranging from *bad* to *excellent*.

5.2.1.3 Test Procedure

Twenty-four naïve subjects (10 female, 14 male) aged between 19 and 63 ($\mu = 30.1$, $\sigma = 4.2$) took part in the test. All of them were native English speakers though not everyone had an American English background. All subjects were paid for their participation. The stimuli were presented via headphones (AKG K601) and a high-quality sound device (Roland Edirol UA-25) in a sound proof booth. The test was designed within subjects, i.e., all participants rated all stimuli.

The test participants were instructed to first rate the *overall impression* of the stimulus on a continuous rating scale. Subsequently, quality estimates for the other 16 scales had to be given via a slider presented on a GUI. To avoid any impact with regard to the order of the attribute scales or the order of the stimuli, both the sequence of scales and the playlist of the stimuli were randomized between subjects. To make themselves familiar with the test procedure, all test participants first had to pass a training phase with three stimuli. The main test consisted of two blocks with female and male stimuli and a 5 min break in between.

After the test, boxplots featuring the ratings on all dimensions were created for all stimuli. Two test participants had more than 5% outlier ratings and were thus excluded from the study.

5.2.2 Statistical Analysis

To unveil the inherent perceptual quality dimensions, a factor analysis was performed. In a further step, a linear regression analysis was conducted, in order to investigate how the obtained quality dimensions contribute to the perceived overall quality. Finally, a MANOVA was calculated to explore the influence of the speaker's voice on the overall quality as well as on the perceptual quality dimensions.

5.2.2.1 Factor Analysis

A PAF analysis on the 16 attribute scales was carried out and resulted in four perceptual dimensions. The Kaiser–Meyer–Olkin (KMO) measure verified the sampling adequacy for the analysis, KMO = .87, and all KMO values for individual items were above .62, which is well above the acceptable limit of .5. The four factors account for 65.6% of the total variance in the dataset. Residuals were computed between the observed and reproduced correlations: 9 (7%) were nonredundant residuals with absolute values greater than 0.05. A subsequent Varimax rotation lead to the rotated factor matrix shown in Table 5.10.

A reliability analysis was performed and reached good values (Table 5.10) for all factors (all Cronbach's $\alpha \geq .7$) except for factor 4 ($\alpha = .65$).

The two-tailed Pearson correlation coefficient R between each factor score and the MOS of the overall impression scale was computed: factor 1 reaches a very strong correlation ($R = .71$) while the other factors only feature a moderate (factor 2: $R = .34$, factor 3: $R = .31$) or weak positive relationship (factor 4: $R = .21$). All correlations are significant on the .01 level.

Table 5.10 Rotated factor matrix

Scale	Factor loadings			
	1	2	3	4
Stress (PQ)	.82			
Rhythm (PQ)	.81			
Naturalness (NOV)	.76		.32	
Voice Pleasantness (NOV)	.60	.46		
Fluency (FAI)	.60			
Likeability	.53	.39	.38	
Intelligibility (FAI)	.41			
Pressure		.80		
Tension (C)		.75		
Speed (C)		.54		
Disturbances (AOD)	.37	.51		
Noise (AOD)		.48		
Resonance			.76	
Pitch Variation			.63	
Pitch				.69
Darkness				.65
Eigenvalues	5.6	2.4	1.4	1.0
% of variance	35.4	15.1	8.8	6.3
Cronbach's α	.89	.77	.71	.65

Note For better readability values below .30 are suppressed. The abbreviations in brackets indicate the corresponding perceptual quality dimensions as introduced in Sect. 4.5

Table 5.11 Linear model of predictors for overall impression

	B	SE B	β	p
CONSTANT	3.73	0.02		$p \leq .001$
FACTOR 1	0.97	0.03	.65	$p \leq .001$
FACTOR 2	0.44	0.03	.29	$p \leq .001$
FACTOR 3	0.35	0.03	.22	$p \leq .001$
FACTOR 4	0.32	0.03	.19	$p \leq .001$

Note $R^2 = .66$

5.2.2.2 Resulting Quality Dimensions

A closer look at the rotated factor matrix in Table 5.10 reveals that the three most important perceptual dimensions for synthesized speech [15] all coincide with the first factor. Hence, the first factor represents the NOV, the PQ, and FAI of the TTS systems. Therefore, the strong correlation between factor 1 and the OIMP scale is plausible. Factor 2 features the scales that were highlighted as relevant for the perceptual dimensions AOD and C [15] plus the scale *pressure* which is not surprising since it is similar to the scale *tension* ($R = .70$). Factor 4 displays the scales *pitch* and *darkness* and is thus related to the *pitch* and the *spectral center of gravity* of the speaker. Factor 3 features the scales *resonance* and *pitch variation*. The interpretation of this factor seemed difficult, thus the stimuli were sorted by their factor 3 scores and experts listened to them. As a result this factor is assumed not to be related to prosodic features, but instead it is related to spectral characteristics of the TTS signals. Still, the interpretation of this factor is vague and should be part of future research.

5.2.2.3 Importance of the Factors for the Overall Quality

To unveil the importance of each resulting factor for the experienced overall quality, a linear regression was performed with the factors as predictors and the overall impression as outcome variable. The results can be seen in Table 5.11. The table lists the four factors, their beta values (B), their standard errors (SE B), and their standardized values (β).

The model explains 66% of the total variance. Moreover, the table shows that Factor 1 contributes most to the overall impression ($\beta = .65$) while the other factors reach β values between .19 and .29.

5.2.2.4 Influence of a Speaker's Voice

An initial MANOVA examined the influence of the independent variable *speaker* of the TTS voice on the OIMP scale and the four factors from the factor analysis. To

Table 5.12 p-values of the post-hoc Scheffé test for the IVONA data

		f1	f2	f3	m1	m2
f1	MOS	–	.421	**.000**	**.000**	.735
	FACTOR 1	–	.681	**.000**	**.019**	.984
	FACTOR 2	–	**.031**	.493	.075	**.000**
	FACTOR 3	–	1.000	**.011**	.942	**.000**
	FACTOR 4	–	**.000**	.147	.988	.940
f2	MOS	.421	–	**.000**	**.005**	.989
	FACTOR 1	.681	–	**.000**	.448	.938
	FACTOR 2	**.031**	–	.731	.998	**.000**
	FACTOR 3	1.000	–	**.013**	.953	**.000**
	FACTOR 4	**.000**	–	**.000**	**.000**	**.000**
f3	MOS	**.000**	**.000**	–	**.034**	**.000**
	FACTOR 1	**.000**	**.000**	–	.054	**.000**
	FACTOR 2	.493	.731	–	.884	**.000**
	FACTOR 3	**.011**	**.013**	–	.111	.290
	FACTOR 4	.147	**.000**	–	.385	.562
m1	MOS	**.000**	**.005**	**.034**	–	**.001**
	FACTOR 1	**.019**	.448	.054	–	.095
	FACTOR 2	.075	.998	.884	–	**.000**
	FACTOR 3	.942	.953	.111	–	**.000**
	FACTOR 4	.988	**.000**	.385	–	.999
m2	MOS	.735	.989	**.000**	**.001**	–
	FACTOR 1	.984	.938	**.000**	.095	–
	FACTOR 2	**.000**	**.000**	**.000**	**.000**	–
	FACTOR 3	**.000**	**.000**	.290	**.000**	–
	FACTOR 4	.940	**.000**	.562	.999	–

Note p-values less or equal to .05 are marked bold

avoid any influence of the speech synthesizers on the results, the dataset was split up by TTS system and separate MANOVAs were performed for the data of AC and IV systems.

Using Pillai's trace, a significant effect of *speaker* for the IV dataset[6] and for the AC dataset[7] was found. The results of the post-hoc Scheffé tests for the IV and AC dataset can be seen in Tables 5.12 and 5.13, respectively.

Both tables highlight that some voices differ from all other voices concerning the overall quality, e.g., Table 5.12 depicts that voice f3 and m1 differ from all other voices of the IV dataset. The same applies for voice f5 in the AC dataset. Accordingly, the data indicates that the voice of the speech corpus has a significant effect on the perceived overall quality of a TTS system.

[6]effect of *speaker* for the IV dataset: ($V = 0.45$, $F(15, 1494) = 17.7$, $p \leq .001$).

[7]effect of *speaker* for the AC dataset: ($V = 0.66$, $F(15, 1494) = 28.1$, $p \leq .001$).

Table 5.13 p-values of the post-hoc Scheffé test for the ACAPELA data

		f4	f5	f6	m3	m4
f4	MOS	–	**.000**	.431	.263	.088
	FACTOR 1	–	**.000**	.515	.086	**.018**
	FACTOR 2	–	**.000**	.123	**.030**	**.000**
	FACTOR 3	–	.353	1.000	.748	**.037**
	FACTOR 4	–	**.000**	**.000**	**.000**	**.000**
f5	MOS	**.000**	–	**.000**	**.000**	**.000**
	FACTOR 1	**.000**	–	**.000**	**.000**	**.000**
	FACTOR 2	**.000**	–	**.000**	**.000**	.160
	FACTOR 3	.353	–	.292	.973	.875
	FACTOR 4	**.000**	–	.978	.052	**.000**
f6	MOS	.431	**.000**	–	.998	.938
	FACTOR 1	.515	**.000**	–	.890	.601
	FACTOR 2	.123	**.000**	–	.987	**.000**
	FACTOR 3	1.000	.292	–	.681	**.026**
	FACTOR 4	**.000**	.978	–	.217	**.000**
m3	MOS	.263	**.000**	.998	–	.989
	FACTOR 1	.086	**.000**	.890	–	.986
	FACTOR 2	**.030**	**.000**	.987	–	**.000**
	FACTOR 3	.748	.973	.681	–	.510
	FACTOR 4	**.000**	.052	.217	–	**.000**
m4	MOS	.088	**.000**	.938	.989	–
	FACTOR 1	**.018**	**.000**	.601	.986	–
	FACTOR 2	**.000**	.160	**.000**	**.000**	–
	FACTOR 3	**.037**	.875	**.026**	.510	–
	FACTOR 4	**.000**	**.000**	**.000**	**.000**	–

Note p-values less or equal to .05 are marked bold

Moreover, there are some voices that differ from the others with regard to one of the factors, e.g., in the IV dataset f2 differs from all other voices concerning factor 4 and m2 differs from all other voices concerning factor 2. In the AC dataset the voice f5 sticks out: it differs from all other voices concerning OIMP, factor 1 and from all but voice m4 concerning factor 2. In addition, voice f4 and voice m4 both differ from the other voices concerning factor 4.

In conclusion, significant differences between a speaker and all other voices of the corresponding dataset could be found: this applies for OIMP (IV and AC), factor 1 (AC), factor 2 (IV), and factor 4 (IV and AC).

These differences are visualized in Fig. 5.2 for the OIMP scale. The boxplot shows the variance in overall impression for all voices of the systems AC (red) and IV (blue) for all five utterances. As can be seen, the overall quality variance within a system is much higher then between both systems. Therefore, selecting a voice for a TTS system is crucial because it tremendously influences the perceived quality impression.

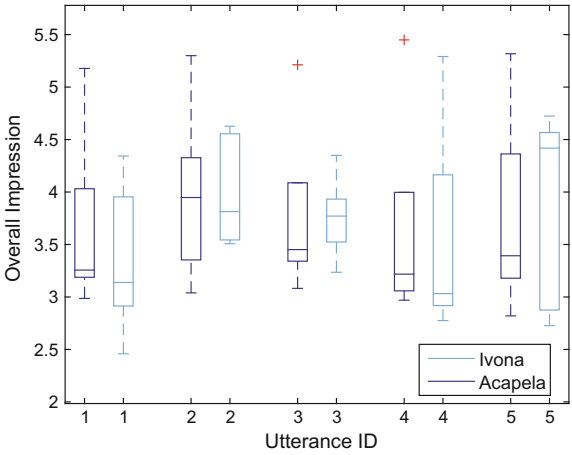


Fig. 5.2 Boxplot of the mean overall impression ratings for the five synthesized utterances

Within the scope of the MANOVAs, an effect of the test material on the overall quality for the IV systems was also discovered. However, this is a minor effect compared to the influence of the speaker's voice. Further research on the influence of the material is presented in Sect. 5.3.

5.2.3 Conclusions

As shown in Table 5.10, the factor analysis revealed four factors for this dataset. A comparison with the five universal perceptual quality dimensions unveils that these five dimensions are distributed among factor 1 and 2. However, it has to be taken into consideration that this database consists of two state-of-the-art unit selection synthesizers with a variety of different voices. Therefore, it is plausible that some of the presented attribute scales contribute to the same factor whereas they would define different perceptual constructs if they were used to assess, e.g., databases with TTS stimuli produced by systems of different quality and/or of different synthesizer type (e.g., PSOLA synthesizer, US systems, and HMM synthesizer) like it was shown in Sect. 4.5.

Moreover, some of the scales that were introduced through previous likeability studies built new factors. Factor 4 can be associated with the *pitch* and the *spectral center of gravity* of the signals whereas the meaning of factor 3 could not be clarified completely. Furthermore, the scale *likeability* features very high cross-loadings between factors 1, 2, and 3. Thus, it does not contribute to discriminate between those perceptual constructs and should be omitted in future research. In contrast, the scale *pressure* supports the impression of the dimension C and can be seen as a substitute for one of the scales associated with this dimension.

When analyzing the influence of the different voices on the overall quality and the extracted factors of the TTS systems IV and AC, a significant effect was found for both datasets. Moreover, it could be shown that some of the voices differ from the others of the same system concerning the overall impression and the factors 1, 2, and 4.

The interpretation of factor 3 seemed difficult and even expert listeners could not fully resolve the problem. Further research will have to focus on this perceptual sensation.

5.3 Influence of Corpus Size and Utterance[8]

Nowadays, the most common approaches for TTS are unit selection and HMM synthesis. Both methods are data-driven, thus they build a voice out of a database of prerecorded speech. Therefore, the size of such a speech corpus also has a significant impact on the quality of the synthesized speech signal. The extent to which the corpus size influences the perceptual quality will be investigated in this section by creating different unit selection voices with the MaryTTS system (see Sect. 2.2). Moreover, the impact of the synthesized utterance on the five universal quality dimensions will also be examined.

Thus, this section addresses RQ3 by seeking answers to the following questions:

- Does an increase in corpus size necessarily lead to an increase in quality?
- How strong is the influence of the corpus size on the five perceptual quality dimensions?
- Which quality dimensions are affected by the content of the synthesized utterance?

In the following section the experimental setup is described. This includes the generation of different unit selection voices with MaryTTS and the listening test procedure itself. In Sect. 5.3.2 different statistical analyses are performed and the influence of the *corpus size*, the *utterance*, and their interaction effect are investigated. Finally, Sect. 5.3.3 concludes the findings.

5.3.1 Experimental Setup

This section gives an overview of the generated TTS voices, the speech material, and the rating scales that were used in the experiment. Moreover, it describes the test procedure.

[8]Parts of the content of this section have been submitted for publishing in a slightly different version [16].

5.3.1.1 Test Database

TTS Systems

A German speech corpus spoken by a professional male speaker was used as a basis for the TTS voices. The full speech corpus consists of files containing names, addresses, names of places and countries, abbreviations, directions, numbers, and short sentences mainly containing news content.

In order to unveil the influence of the size of the speech corpus on different perceptual quality dimensions, voices based on different subsets of the full speech corpus A were created. Thus, B is a subset of A, C is a subset of B, etc. The subsets were selected in a way that guaranteed an equal distribution of content (names, addresses, numbers, etc.) across all subsets. The duration of each corpus can be seen in Table 5.14.

Due to the prerequisites of MaryTTS the high quality recordings had to be downsampled to 16 kHz sampling rate. Based on these corpora six unit selection voices were created with the MaryTTS [17] system, using MAUS [18, 19] to align text and audio.

Even though the voices all feature the same speaker, they will be referred to as Voices A–F in the following.

Speech Material

Table 5.15 shows all utterances that were synthesized for the listening test. Utterances containing place names, proper names, or words from a foreign language were excluded as they often differ from German pronunciation rules and are thus likely to cause problems for speech synthesizers. Five shortened German passages from the EUROM.1 spoken language resource database [13] were selected (Utterances 01–05). Moreover, five utterances (Utterances 06–10) that were part of the recordings of corpus A but not of the five subsets B to F were chosen. Therefore, since Utterances 06 to 10 were part of the speech corpus of Voice A, the synthetic speech generated by Voice A for these utterances equals natural speech.

Table 5.14 Corpus sizes for voices A to F

Corpus	Duration (hh:mm)
A	07:49
B	06:31
C	04:20
D	02:10
E	00:52
F	00:11

Table 5.15 Selected utterances used in the listening test

Utterance-ID	Utterance
00	*Was soll ich heute abend nur essen? Ich habe noch einen Eintopf in der Tiefkühltruhe.*
01	*Letzte Nacht habe ich die Haustür geöffnet, um die Katze nach draußen zu lassen. Plötzlich hörte ich, wie die Tür hinter mir zufiel.*
02	*Von hier aus gibt es zu meinem Haus auch eine Abkürzung über den Hügel. Die meisten Leute erzählen, dass es dort spukt.*
03	*Kannst du mir sagen, was heute abend im Fernsehen kommt? Ich hätte Lust auf etwas Leichtes und Amüsantes.*
04	*Das Singen ist ein Ausdruck von tiefgreifenden Gefühlen. Mal kann es extremes Glück, mal tiefe Trauer ausdrücken.*
05	*Ich kann auf langen Schiffsreisen einfach nicht schlafen. Dazu sind die Sitze viel zu ungemütlich.*
06	*Der Reinerlös dieses Festes soll für eine dringend notwendige Erweiterung der Kindergartenküche verwendet werden.*
07	*Entsprechend der zuvor vom Parlament verabschiedeten Verfassungsänderung, können nur Abgeordnete Minister werden.*
08	*Brandenburgischen Zollfahndern ist ein Schlag gegen den international organisierten Zigarettenschmuggel gelungen.*
09	*Nach Auflösung von vereinzelten Nebelfeldern Wechsel von Sonne und Wolken.*
10	*Zum Sommeranfang sorgt ein Hoch für warmes Wetter in Mitteleuropa.*

Furthermore, an additional utterance (Utterance 00) from the EUROM.1 spoken language resource database was synthesized for the training session participants had to take prior to the listening test.

While the utterances from the EUROM.1 spoken language resource database feature short stories from everyday life, the Utterances 06–10 comprise news content. The average duration of all test stimuli is around 8 s.

5.3.1.2 Attribute Scales

The objective was to assess the test participant's perceptual impression on several different perceptual dimensions. Therefore, several of the attribute scales that were introduced in Sect. 4.5 were selected. Two attribute scales were selected for each dimension. The perceptual quality dimensions and their associated attribute scales are:

- **Naturalness of Voice (NOV)**: naturalness, voice pleasantness
- **Prosodic Quality (PQ)**: stress, rhythm
- **Fluency and Intelligibility (FAI)**: fluency, intelligibility
- **Absence of Disturbances (AOD)**: noise, disturbances
- **Calmness (C)**: speed, tension

5.3.1.3 Test Procedure

Thrity naïve participants (13 female, 17 male) aged 18 to 35 ($\mu = 25.6$, $\sigma = 4.1$) were invited to take part in the listening test. All of them were native German speakers and were paid for their participation. The stimuli were presented via headphones (AKG K601) and a high-quality sound device (Roland Edirol UA-25) in a soundproof booth. The test was designed within subjects, i.e., all participants rated all stimuli.

Stimuli were presented within a Matlab GUI, where ratings could be given on a continuous scale ranging from 0 to 100. To avoid any impact with regard to the order of the scales or the order of the stimuli, both the sequence of scales and the playlist of the stimuli were randomized between subjects.

The test consisted of three parts. At first, all participants started off with a training phase where they could get used to the user interface and the quality range of the presented stimuli. Three stimuli generated by the Voices A, D, and F were rated in this session. In case any problems occurred during the training, every participant had the chance to ask questions after this first part. Subsequently, the main test took place. Part 2 and 3 consisted of 30 stimuli each, interrupted by a 5 min break to avoid listener fatigue.

5.3.2 Statistical Analysis

This section shows the results of the statistical analysis of the gathered data. Prior to the execution of any statistical methods the ratings of all test participants were screened for plausibility. Therefore, boxplots featuring the ratings on all dimensions were created for all stimuli. Three test participants had more than 10% outlier ratings and were thus excluded from the study.

5.3.2.1 Overview of Gathered Data

For the following statistical analysis, the values of the five perceptual dimensions were computed by calculating the mean value of the items associated with each dimension. A first overview of the data from the listening test is given in the boxplots in Fig. 5.3 for the dimension NOV. Each box represents all scores in the NOV dimension for the Voices A–F. Moreover, the boxes are grouped by the utterance (01–05: dark blue boxes, 06–10: light blue boxes).[9]

There are three things that stand out:

1. Utterances 06–10 synthesized by Voice A reach by far the best rating in the NOV dimension. This effect was expected since these stimuli represent a natural voice. Thus, these ratings can be seen as an upper quality threshold.

[9] Similar results occur for the other four dimensions.

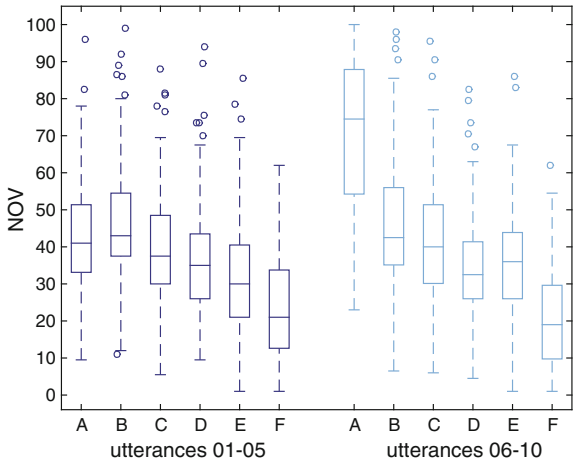


Fig. 5.3 Boxplot of the NOV scores for all corpus sizes (A–F) grouped by utterance-ID

2. Even though Voice A equals natural speech for the Utterances 06–10, there were still single ratings below 30. One reason for this is probably the neutral speaking style of the professional speaker during the recordings of the speech corpus. Furthermore, due to the downsampling to 16 kHz, which was a prerequisite of MaryTTS, all spectral information in the speech signals above 8 kHz were lost. This will also have affected the perceived voice pleasantness.
3. The highest median value for the Utterances 01–05 is obtained by Voice B. This however is a surprising result since larger corpora were expected to lead to better voice quality.

5.3.2.2 Influence of Corpus Size and Utterance

In order to analyze the influence of all *corpus sizes* (A–F), the Utterances 06–10 had to be excluded to avoid any bias due to the naturalness of Voice A on these sentences (see Fig. 5.3). Therefore, the following statistical analyses are performed on the data of the Utterances 01–05.

Five separate ANOVAs[10] with each of the five perceptual dimensions as dependent variable, the *corpus size* (A-F) and the *utterance* (01–05) as fixed factor, and the *test participants* as random factor were performed.

Influence of Corpus Size

There were significant effects of *corpus size* on all five perceptual dimensions. The results are shown in Table 5.16. The effect of *corpus size* on all five dimensions is

[10]The assumptions for the execution of ANOVA have been met.

Table 5.16 Influence of *corpus size* and *utterance* on the five perceptual quality dimensions

Dimension	Corpus Size	Utterance	test subject	Corpus Size * Utterance
NOV	$F(5, 754) = 71.4$	$F(4, 754) = 10.8$	$F(26, 754) = 18.9$	$F(20, 754) = 5.6$
PQ	$F(5, 754) = 24.5$	$F(4, 754) = 28.0$	$F(26, 754) = 14.3$	$F(20, 754) = 4.5$
FAI	$F(5, 754) = 157.7$	$F(4, 754) = 28.9$	$F(26, 754) = 10.0$	$F(20, 754) = 9.2$
AOD	$F(5, 754) = 60.1$	$F(4, 754) = 4.5$	$F(26, 754) = 18.0$	$F(20, 754) = 6.4$
C	$F(5, 754) = 43.6$	–	$F(26, 754) = 11.1$	$F(20, 754) = 3.7$

Note $p < .001$ for all reported F-ratios

also reflected in the bar graph in Fig. 5.4. To unveil significant differences between the *corpus sizes* post-hoc tests (REGWQ) were performed for all five ANOVAs.

As can be seen in Fig. 5.4, similar patterns occur for all five dimensions: the larger the corpus of the corresponding voice, the higher the rating. This, however, is not true for Voice A. Despite Voice A being the one with the largest speech corpus, Voice B reaches the highest score in all but the PQ dimension. The only dimension in which this difference is significant is FAI. Therefore, the most fluent speech was not produced by the voice with the largest corpus but by a voice that was built on a subset of this corpus. This is an outcome that was not excepted. Therefore, target and join costs for Voices A and B were compared for several of the utterances that were rated better when synthesized by Voice B. However, since the corpus influences the computation of the join costs of a voice in MaryTTS, diphones which are part of Voice A as well as Voice B feature different join costs depending in which voice they are used. Therefore, a comparison between the costs of different voices is not permissible. Thus, the source of this effect remains unclear. It seems, however, as if the unit selection algorithm of MaryTTS chooses unit that can lead to a decrease in quality when the speech corpus gets increased above a certain point. It is unknown if the unit selection algorithm of MaryTTS was ever used on a speech corpus of over

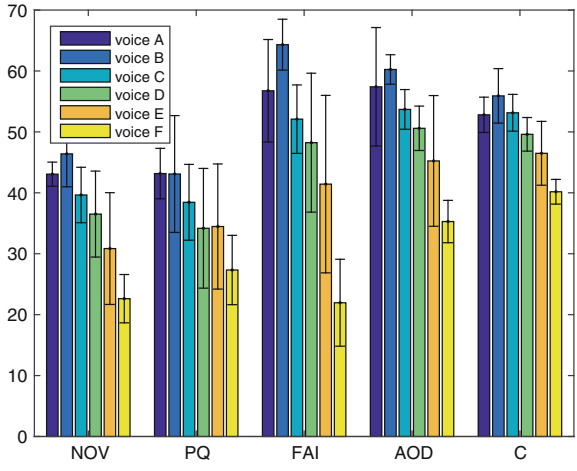


Fig. 5.4 Bar chart of mean scores of all voices on the 5 dimensions (error bars illustrate the standard deviation)

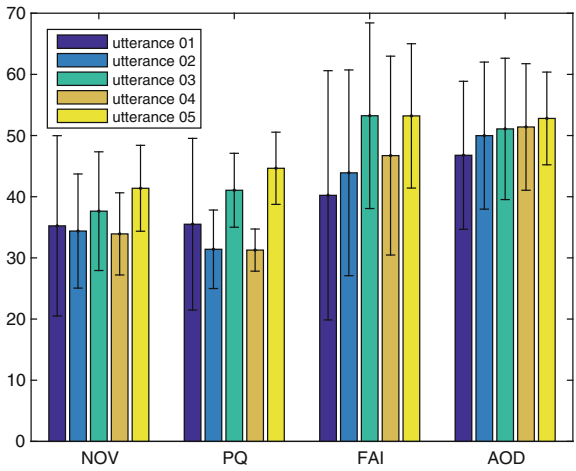


Fig. 5.5 Bar chart of mean scores of all utterances on the dimensions NOV, PQ, FAI, and AOD (error bars illustrate the standard deviation)

seven hours; e.g., the German BITS corpus [20] which was also used to create unit selection voices with MaryTTS [17] is only around 3:40 hours per speaker.

Furthermore, at the lower end of the quality range, Voice F was found to be rated significantly lower than all other voices on the five dimensions.

When comparing the range of ratings of the quality dimensions, the smallest variance could be found for the PQ dimension. Here three groups that differ significantly from each other were found: A&B, C&E, and D&E. The biggest variance is achieved for FAI, leading to significant differences between A, B, the group C&D, E and F. Thus, the *corpus size* has only a limited influence on the prosody while highly affecting the fluency and intelligibility of the voices.

Influence of Utterance

There were significant effects of *utterance* on all but the calmness dimension (see Table 5.16). This effect is also shown in the bar graph in Fig. 5.5.

Utterance 05 sticks out on the positive end of all dimensions. It got significantly better ratings than all other utterances in the NOV dimension and together with Utterance 03 it gets the best ratings in PQ and FAI. Therefore, these two utterances seem to be an easier task than the other utterances.

Furthermore, while Utterances 01, 02, and 04 do not differ significantly concerning the NOV dimension, Utterances 02 and 04 got rated significantly lower in the PQ dimension compared to Utterance 01. However, when looking at FAI, Utterances 02 and 04 perform significantly better than Utterance 01.

Thus, the *utterance* that is to be synthesized can impact multiple perceptual quality dimensions (see Utterances 03 and 05) but it can also boost the performance in one dimension while decreasing the performance in another (see Utterance 01 vs Utterances 02 and 04).

Even though there was a significant effect on AOD no interesting patterns could be found. For the C dimension, no significant effect of *utterance* was detected.

Interaction Effect

Moreover, an interaction effect of *corpus size*utterance* was detected on all of the perceptual quality dimensions. This interaction effect is exemplarily shown for the FAI dimension in the line plot in Fig. 5.6 (similar results occur for the other four dimensions).

A trend as the one found for Utterance 03 was expected, i.e., a larger corpus triggers an increase in the quality dimensions (here fluency). But here a larger corpus does not necessarily help to increase the quality (see the change from corpus E to D or B to A for Utterance 02) and in certain cases a larger corpus can also lead to a dramatic decrease in quality (see change from corpus B to A for Utterance 01). Moreover, for Utterance 05 a strong increase in fluency can be seen when increasing the corpus from F to E and from then on the fluency fluctuates around a certain level.

These results show that even though a larger corpus possibly offers units with smaller target and join costs, finding these units is not an easy task.

Additionally, there was also a significant effect of the *test subjects* on all five dimensions. This indicates different preferences between test subjects. Future research on user characteristics can lead to more information on this effect.

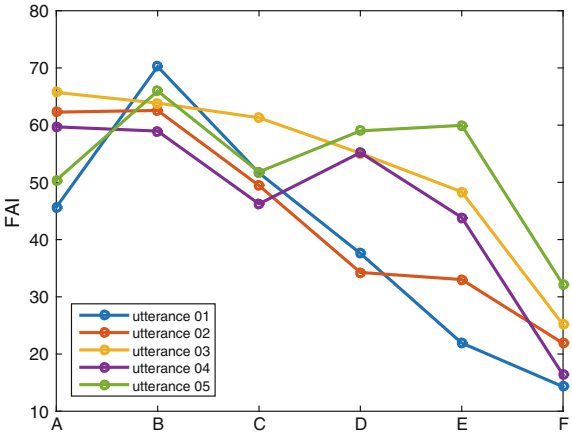


Fig. 5.6 Line plot of mean FAI scores for utterances 01–05 for all voices

5.3.3 Conclusions

Through statistical analyses, the effect of the *size of the speech corpus*, the *utterances*, and the interaction effect between both of them on different perceptual quality dimensions could be verified.

The *size of the speech corpus* affects all dimensions with the highest significant effect on FAI. Surprisingly, for this dimension the system with the second largest speech corpus achieved significantly higher ratings than all other systems. This may indicate a threshold in the unit selection algorithm of MaryTTS from which on a system does not further improve its quality when increasing its database. The exact source of this effect remains unclear but will be further investigated. The smallest effect of the size of the *speech corpus* could be determined for the PQ dimension.

Secondly, an effect of the synthesized *utterance* on all but the dimension C could be detected. Some utterances were an easier task for the TTS system than others, which lead to significantly higher scores in the first four dimensions. On the other hand, the synthesized *utterance* can affect dimensions differently, i.e., this can lead to higher scores in one dimension while achieving lower scores in another.

Moreover, an interaction effect *corpus size*utterance* was found for all dimensions. It was expected that the synthesized quality of an utterance increases continuously with increasing corpus size. However, this was not true for all of the synthesized utterances: some utterances reached their peak in quality at a small corpus size and fluctuated around this value when further increasing the corpus size, while the quality of others increased linearly until their peak at the second largest corpus size.

These results show that the unit selection process of a TTS system is a very sensitive task. Even though a larger corpus theoretically holds units that are able to create superior synthetic speech, selecting the right units is crucial. In order to further improve this process, a quality predictor could be included. Such a predictor could provide quality estimates for the concatenation of different candidate units with the units that already have been selected. Therefore, by selecting the candidate unit that yields the highest score, the predictor can help to optimize the unit selection process. Different approaches on instrumental quality prediction for TTS signals are presented in the following chapter.

5.4 Summary

This chapter analyzed different influencing factors on the perceptual quality of TTS systems. First, an evaluation protocol for synthesized audiobooks was designed and tested during the BC 2011. The outcome highlights the importance of the two dimensions *listening pleasure* and *prosody* for this use case. Moreover, a study examined the influence of the *voice of a speaker* for corpus-based TTS systems. A significant effect on the *overall impression* and three of the four extracted factors was found. Finally, the influence of the *corpus size* and the *utterance* to be synthesized was

examined by creating several unit selection voices in MaryTTS. A significant effect for *corpus size* was found on all five universal perceptual quality dimensions, while the *utterance* only affected four of them.

References

1. Hinterleitner F, Neitzel G, Möller S, Norrenbrock C (2011) An Evaluation Protocol for the Subjective Assessment of Text-to-Speech in Audiobook Reading Tasks. In: Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA), Florence, Turin, Italy
2. van Bezooijen R, van Heuven V (1997) Assessment of Synthesis Systems. In: Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, pp 481–563
3. Delogu C, Conte S, Sementina C (1998) Cognitive Factors in the Evaluation of Synthetic Speech. In: Speech Communication. Elsevier, pp 153–168
4. ITU-T Rec. P.85 (1994) A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices. International Telecommunication Union, Geneva
5. Hinterleitner F, Möller S, Norrenbrock C, Heute U (2011) Perceptual Quality Dimensions of Text-to-Speech Systems. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011), pp. 2177–2180
6. Rautenberg U, Schnickmann T (2007) Die Stimme im Hörbuch: Literaturverlust oder Sinnlichkeitsgewinn? In: Das Hörbuch - Stimme und Inszenierung. Harrassowitz Verlag, Wiesbaden, pp 21–54
7. Häusermann J, Janz-Peschke K, Rühr S (2010) Das Hörbuch - Medium, Geschichte, Formen. UVK Verlags-Gesellschaft, Konstanz
8. Burkey M (2007) Sounds Good To Me: Listening to Audiobooks with Critical Ear. http://audiobooker.booklistonline.com. Accessed 22 July 2011
9. Neitzel G (2011) Entwicklung eines Evaluationsverfahrens zur Bestimmung der Qualität synthetisch erzeugter Hörbücher. Bachelor Thesis, Quality and Usability Lab, TU Berlin
10. Hinterleitner F, Norrenbrock C, Möller S (2013) Perceptual Quality Dimensions of Text-To-Speech Systems in Audiobook Reading Tasks. In: Proceedings of the 24th Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Bielefeld, Germany, pp 44–49
11. Bodden M, Jekosch U (1996) Entwicklung und Durchführung von Tests mit Versuchspersonen zur Verifizierung von Modellen zur Berechnung der Sprachübertragungsqualität. Technical report, Institut für Kommunikationsakustik, Ruhr-Universität, Bochum
12. Hinterleitner F, Manolaina C, Möller S (2014) Influence of a Voice on the Quality of Synthesized Speech. In: Proceedings of the 6th International Workshop on Quality of Multimedia Experience (QoMEX 2014), pp 95–100
13. Chan D, Fourcin A, Gibbon D, Grandstrom B, Huckvale M, Kokkonakis G, Kvale K, Lamel L, Lindberg B, Moreno A, Mouropoulos J, Senia F, Trancoso I, Veld C, Zeiliger J (1995) EUROM-A Spoken Language Resource for the EU. In: Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH 1995), pp 867–870
14. Weiss B, Burkhardt F (2010) Voice Attributes Affecting Likability Perception. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010), pp 1934–1937
15. Hinterleitner F, Norrenbrock C, Möller S (2013) Is Intelligibility Still the Main Problem? A Review of Perceptual Quality Dimensions Of Synthetic Speech. In: Proceedings of the 8th ISCA Speech Synthesis Workshop (SSW 2013), pp 167–171
16. Hinterleitner F, Weiss B, Möller S (2016) Influence of Corpus Size and Content on the Perceptual Quality of a Unit Selection MaryTTS Voice. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016) (submitted)

17. Schröder M, Hunecke A (2007) Creating German Unit Selection Voices for the MARY TTS Platform from the BITS Corpora. In: Proceedings of the 6th ISCA Speech Synthesis Workshop (SSW), pp 95–100
18. Schiel F (1999) Automatic Phonetic Transcription of Non-Prompted Speech. In: Proceedings of the 14th International Congress Of Phonetic Sciences (ICPhS 1999), pp 607–610
19. Reichel UD (2012) PermA and Balloon: Tools for String Alignment and Text Processing. In: Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012), pp 1874–1877
20. Ellbogen T, Schiel F, Steffen A (2004) The BITS Speech Synthesis Corpus for German. In: Proceedings of LREC, Lisbon, Portugal, pp 2091–2094

Chapter 6
Instrumental Quality Assessment

During the development of new TTS systems, a continuous auditory quality assessment is desirable in order to see whether recent changes positively affect the overall quality of the system. Unfortunately, such auditory assessments are cost-intensive as well as time consuming. Therefore, a reliable instrumental quality measure could support the development of high quality TTS systems. As introduced in Sect. 3.4, several instrumental measures exist that are able to estimate the quality of speech signals distorted by coding and transmission artifacts. Considering synthetic speech being some kind of distorted natural speech, applying those measures to synthetic speech seems admissible. Two different kinds of measures exist: reference-based measures that use a *clean* reference signal, i.e., a signal without distortions, to compute a perceptual distance between the clean and the distorted signal, and reference-free measures that are able to estimate the quality solely based on the *distorted* signal. Some of these measures have already been tested on synthetic speech signals. Therefore, this Chapter seeks to answer RQ4:

> *How can the quality of synthetic speech be assessed by an instrumental measure?*

To be more specific, this chapter will give answers to the following questions:

- Are existing reference-based instrumental measuress able to estimate the quality of TTS signals?
- Which approaches can be employed in order to create a reference-free instrumental measure for TTS signals?

Therefore, this chapter presents results of the performance of standardized reference-based measures on synthetic speech. Moreover, recent approaches towards the reference-free instrumental quality prediction of synthetic speech are discussed and compared.

F. Hinterleitner, *Quality of Synthetic Speech*, T-Labs Series
in Telecommunication Services, DOI 10.1007/978-981-10-3734-4_6

6.1 Reference-Based Measures

Reference-based measures are widely popular in the domain of telephone coded speech signals. Therefore, several different measures are commonly used to evaluate speech signals transmitted through telephone networks, among them WB-PESQ, DIAL, and POLQA which were introduced in Sect. 3.4.1. While in this domain clean, undistorted signals, i.e., speech signals before a transmission through a telephone network, are usually always at hand, this is only very rarely the case when it comes to synthetic speech. In this context a *clean* signal would be a natural speech signal spoken by the speaker of the speech corpus of the TTS system. In the following sections, the three previously described models will be tested on different TTS databases that also contain a *clean* reference signal.

6.1.1 State of the Art

In very few studies existing reference-based measures have been used to evaluate the quality of TTS signals. Počta investigated the influence of packet loss on synthetic speech quality, measured with PESQ [1], PESQ, and POLQA [2] as well as the influence of coding measured with PESQ and POLQA [3]. While PESQ and POLQA were able to predict the quality of transmitted synthetic speech to a certain degree [2], his findings also showed a higher vulnerability of synthetic speech to packet loss impairments than naturally-produced speech [1]. However, the goal of these studies was not to investigate the quality of synthetic speech per se, but to find out about the influence of transmission artifacts on TTS signals. Therefore, the reference signal in this test setup was the "clean" TTS signal before it was distorted.

In contrast, Cernak and Rusko [4] utilized the PESQ measure to estimate TTS quality by comparing synthetic speech signals to natural speech recordings. While the predicted MOS only reached values slightly above 1, the correlations between auditory and predicted MOS were surprisingly high.

The outstanding results from Cernak and Rusko [4] lead to further research in this direction. The findings are presented in the following section.

The most recent study [5] in this field described the development of a reference-based measure for the quality prediction of synthetic speech. The model was trained and tested on data from the Blizzard Challenges (BCs) 2008 to 2013 and reached a correlation of up to .60 and .84 per file and system respectively.

6.1.2 Quality Prediction[1]

The three reference-based instrumental quality measures Wide-Band (WB)-PESQ, DIAL, and POLQA, which were introduced in Sect. 3.4.1, will be used to assess TTS

[1]The content of this section has previously been published in a slightly different version in [6].

signals which were evaluated during several BCs. In the following section the TTS databases are presented and the prediction results are discussed.

6.1.2.1 Blizzard Challenge Databases 2008–2010

The Blizzard Challenge (BC) is an annual contest for developers of TTS systems. Participants in the English part of the challenges 2008–2010 were provided with a speech corpus of the University of Edinburgh (*full corpus*) as well as two different subcorpora (*ARCTIC corpus* and *small corpus*[2]). These corpora were used to build data-driven TTS voices. A set of test sentences was released to the contestants, who were asked to submit synthesized versions within a limited time interval. An online listening test was conducted to evaluate *naturalness*, *intelligibility* and the degree of *similarity* to the original speaker. Each test also included natural reference stimuli of the same speaker the TTS systems were built on. These natural reference stimuli were not made available to the developers, therefore, the synthesized files and their corresponding reference are completely independent of each other.

In the following, the challenges of the years 2008, 2009, and 2010 are described.

- Blizzard Challenge 2008
  The data of the BC 2008 [7] consists of 18 speech synthesis systems, 1 natural speaker, and 2 systems from participants from previous challenges (a Festival-based system from CSTR[3] and the HTS[4] system from the Blizzard Challenge 2005). In an attempt to calibrate the results from year to year, the latter systems were used as benchmarking systems. For every synthesizer, 42 files were evaluated during the listening tests.
- Blizzard Challenge 2009
  The 2009 database [8] consists of 14 speech synthesis systems, 1 natural speaker, and 3 benchmarks systems (the 2 systems used during the BC 2008 and the HTS system from the BC 2007). Thirty-eight files generated by each system were judged during the evaluation phase.
- Blizzard Challenge 2010
  In 2010, 18 TTS developers took part in the challenge [9]. This database consists of TTS voices built on the *ARCTIC corpus* (14 participants, 1 natural reference, and 3 benchmark systems from previous challenges) and of synthetic speech samples that were built on the *rjs* speaker provided by Phonetic Arts (15 participants, 1 natural speaker, and 2 benchmark systems). Both databases consist of 36 files per synthesizer.

[2]The voices built on the *ARCTIC* and *small corpus* in the years 2008 and 2009 were unavailable. Thus, the research in these years concentrates on the TTS data on the basis of the *full corpus*.

[3]The Centre for Speech Technology Research: http://www.cstr.ed.ac.uk/, last accessed 22.04.2016.

[4]HMM-based Speech Synthesis System: http://hts.sp.nitech.ac.jp/, last accessed 22.04.2016.

Quality Evaluation

The listening tests were carried out online, using a design developed for BC 2007. Various listener types were employed, spanning from volunteers recruited via the challenge's participants, mailing lists, blogs for speech experts, and paid undergraduates. Since the results of all listeners were used during the evaluation, there will be no further differentiation. In the BC 2008, 438 listeners finished the whole test procedure whereas 365 completed the test in 2009 and 363 in 2010. The listener gender was anonymized, thus gender-related aspects could not be analyzed. The tests consisted of different sections where listeners had to rate differences in similarity, naturalness, and intelligibility. The evaluated files from these sections consisted of sentences from the genres *news* and *novel* and were sampled at 16 kHz.

6.1.2.2 Results and Discussion

The three reference-based instrumental quality measures WB-PESQ, DIAL, and POLQA were used to assess the quality of the TTS signals presented in the previous section. With the objective to estimate the overall quality of the participating systems, the current study focused on the MOS of the given naturalness ratings.

The corresponding natural speech files were used as reference signals in the estimation process. To evaluate the accuracy of the predicted results, Pearson's correlation coefficient R between the predicted and the auditory MOS averaged per system was computed. The results can be seen in Table 6.1.

None of the algorithms achieved satisfying results. Only for database BC 2008 DIAL and POLQA reached correlations above .40. Taking a look at Table 6.2 shows that the mean predicted MOS value is usually between 1 and 2 with a variance δ^2 just above 0. Obviously all three algorithms detect major distortions in all tested TTS systems which leads to constant low MOS values with no distinction between good and bad sounding TTS.

Usually, WB-PESQ, DIAL, and POLQA are used to evaluate audio material of a duration of at least 8–9 s. Because most of the databases consist of TTS files with a duration of 2–3 s, this seems to be one cause for the very low MOS values. Hence, groups of three TTS files of the same system from the database BC 2008 were concatenated and used as input for WB-PESQ and DIAL. The resulting scores

Table 6.1 Correlations between predicted and auditory MOS scores (per synthesizer)

DATABASE	WB-PESQ	DIAL	POLQA
BC 2008	.17	.49	.46
BC 2009	.19	.25	.33
BC 2010 rjs	.02	−.15	.35
BC 2010 arctic	.21	−.14	−.08

Table 6.2 Mean values and variances of the predicted MOS

DATABASE	WB-PESQ		DIAL		POLQA	
	$\overline{MOS}$	δ^2	$\overline{MOS}$	δ^2	$\overline{MOS}$	δ^2
BC 2008	1.38	0.22	1.97	0.63	1.22	0.06
BC 2009	1.32	0.17	2.09	0.03	1.08	0.02
BC 2010 rjs	1.11	0.00	1.91	0.03	1.34	0.09
BC 2010 arctic	1.13	0.03	2.00	0.03	1.20	0.05

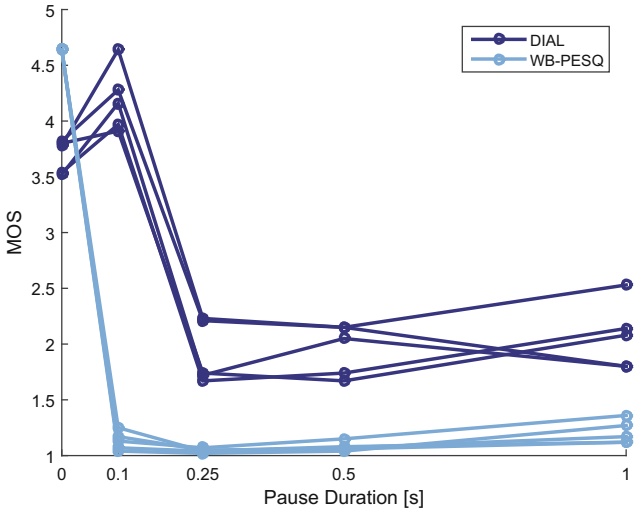


Fig. 6.1 Predicted MOS in relation to the duration of inserted pauses

were averaged per system and Pearson's Correlation Coefficient R was computed. The results showed little improvement for DIAL ($R = .59$) but none for WB-PESQ ($R = -.18$). Moreover, $\overline{MOS}$ as well as variance remained on a very low level.

One of the reasons for the very low MOS values might have been failures in the time alignment between the natural speech file and the TTS signal. Compared to natural speech, TTS systems often produce signals that comprise parts that are lengthened or shortened. This makes a time alignment more challenging than between a natural speech signal and a telephone-network-coded one. To simulate these TTS distortions, five to six small pauses with a duration of 0.1–1 s were inserted into four natural speech files. These files were tested with WB-PESQ and DIAL with the original natural speech files as reference signals. Figure 6.1 shows the resulting MOS.

The quality predicted by WB-PESQ for the natural speech files without any modifications is around 4.5. After the insertion of pauses (no matter which duration) the value drops below 1.5. DIAL estimates values between 3.5 and 4 for the original speech samples. For inserted pauses longer than 0.1 s the MOS predicted by DIAL

decreases to values between 1.5 and 2.5. Surprisingly, the natural speech files with inserted pauses of 0.1 s duration get a slightly better rating than the original files with DIAL.

The achieved correlations lag far behind the results achieved by Cernak and Rusko (as presented in Sect. 6.1.1). However, their approach differed in the point that they used only word long TTS samples which makes the time alignment between the natural speech and the TTS signals much easier. Hence, five samples from the BC 2008 database for five TTS systems were selected, one word per sample was cut, and the WB-PESQ scores were computed. However, the results could not be improved: $\overline{MOS}$ of 1.11, a variance of 0.03.

6.1.3 Conclusions

The correlations between the predicted and the auditory MOS were disappointing throughout all databases. The best correlations were achieved by POLQA on database BC 2008 ($R = .46$). Of course, it has to be noted that all of the tested predictors were used out of their original intended domain, therefore the achieved correlations do not contradict the good results attained for telephone-transmitted speech.

One of the reasons for the overall low predicted MOS seems to be an inaccurate time alignment between the TTS samples and the natural speech files. This is due to the non-linear distortions introduced by the TTS algorithms. For further studies, a dynamic time warping (and in case of POLQA a more extensive time warping) could be used as a preprocessing step to ensure exact time alignment of extremely temporally stretched or compressed signals. For now it has to be stated that standardized instrumental quality evaluation measures, like WB-PESQ, DIAL, or POLQA, can not be used to estimate the quality of synthetic speech when an actual natural reference signal is present. Since in most cases such a natural reference signal is not present, the development of reference-free measures is far more important.

6.2 Reference-Free Measures

Several reference-free predictors have been introduced in Sect. 3.4.2. The following state of the art section gives an overview of the performances of several of these reference-free predictors and introduces two recent approaches towards the instrumental quality measurement of synthetic speech. Moreover, two linear regression approaches are presented in detail. One is based on features that can be derived from the Fujisaki model, and the other one employs feature selection to minimize a large-scale feature set to few features that still cover most of the information.

6.2.1 State of the Art

Besides the reference-based measures utilized in the previous section, there are several reference-free measures that estimate the quality of speech signals transmitted through telephone networks. Among them, there is the ITU-T Rec. P.563 [10, 11] which was introduced in Sect. 3.4.2. Similar approaches are ANIQUE/ANIQUE+ [12, 13] and an algorithm developed by the company Psytechnics [14, 15]. Möller et al. [16] and Heimansberg [17] applied the P.563 model on three different databases of synthetic and natural speech, both degraded by telephone channels. While on some databases the correlation between auditory and estimated MOS reached an adequate level, on others the measure could not help to estimate the quality. When estimating the quality of mixed databases (synthetic as well as natural speech), P.563 did not achieve satisfying results. Moreover, Möller et al. [18] also tested ANIQUE+ and the algorithm developed by Psytechnics on the three aforementioned databases and concluded that the reference-free measures under test mainly predict the effect of the transmission channel and not the actual quality of a TTS system.

With the objective to investigate the low correlations of P.563, Möller and Falk [19–21] examined the internal parameters from which P.563 computes the final MOS estimate. They found that various parameters attained higher correlations with the auditory MOS than the final MOS estimate. The importance of the parameters, however, differed between female and male TTS signals. These findings suggested to develop separate quality prediction models for female and male synthesizers.

Designing an instrumental measure for the exact purpose of quality estimation of TTS signals has already been proposed in the early 90s [22]. Mariniak suggested to extract features of natural speakers and of synthetic speech signals and to compute a spectral distance between them. The deviation of the synthetic speech signals from human produced speech indicates the quality of the TTS system. This topic was revisited by Falk et al. [23]. They developed a measure which evaluates the perceptual distance between features extracted from TTS signals and a reference HMM which is trained on the behavior of natural speech. The obtained normalized log-likelihood indicates the similarity between synthetic and natural speech. This approach was tested on the *Seget2007* database (see Sect. 4.1.4) and reached correlations with auditory assessed quality features of up to .81 and .83 for female and male data, respectively [20, 23]. The results were compared to the P.563 measure and outperformed it for every quality feature.

The positive results of single P.563 features and the HMM predictor described in the previous paragraphs lead to further investigations in these directions [24, 25]. Therefore, two linear regression models were developed, one based on the P.563 features and one based on a set of approx. 1500 general speech features which are usually applied for the classification of speech metadata, such as emotion, gender and age. Given these large sets of parameters, a sequential feature selection had to be employed before the actual regression model was trained. Therefore, first features with $|R| \leq .25$ were omitted, then a PCA with Varimax rotation was performed to come up with a small set of factors which still holds most of the variance of the original

features. The factors derived from the P.563 features and the general speech features were the input for the training of four linear regression models (one per gender for each P.563 features and general speech features). These two models together with the HMM predictor were tested on three databases: *Seget2007*, *Hint2010*, and a smaller database which contains synthetic as well as natural speech degraded by telephone networks. While the HMM predictor attained a sufficient accuracy on two databases, its performance was disappointing especially considering female data. The two other models reached correlations between .64 and .90 for all databases. Even though the results were very positive, the high correlations have to be handled with caution considering that the models were not cross-validated.

In the following two very recent models will be described in more detail.

6.2.1.1 Support Vector Regression (SVR)

Norrenbrock developed an SVR quality prediction model [26] which is based upon the principle of Support Vector Machines (SVMs) [27]. The SVR model utilizes extracted features, e.g., MFCCs, for a mapping on a target quality feature, e.g., the MOS of a given overall impression rating.

By employing a kernel function, an SVM is able to model nonlinear relations, i.e., the input is mapped into a high-dimensional feature space and then a linear model can be constructed. The training data is approximated with limited precision ϵ and a model is constructed by minimizing this error.

Norrenbrock uses a special case of SVR called ν-SVR [28, 29], where ϵ is adaptively determined through a fixed constant ν. A radial basis function is employed as kernel type and features and auditory ratings are scaled to [0, 1] for training and testing. A supervised feature selection choses features with a minimum correlation of $|R| \geq .4$ for model evaluation.

6.2.1.2 Regular Perception Model (RPM)

Based on the theory of regular perception, Norrenbrock developed an RPM for TTS quality prediction [30]. The RPM makes use of the theory of Regular Perception (RP), introduced by Norrenbrock [30], which compares the values of extracted features of synthetic speech signals with the regular feature values for natural human speech.

The basic principle of regular perception is illustrated in Fig. 6.2. The figure shows a perceivable physical property, a Physiological Perception Range (PPR) as well as a Regular Perception Range (RPR). In the context of human speech, the PPR is restricted by the limitations of the human auditory system (i.e., the bandwidth between approximately 20 Hz and 20 kHz) and the RPR defines the range in which a specific attribute regularly lies, e.g., the range of the pitch of a human female speaker (approximately 200–250 Hz). If a specific property exceeds the upper threshold of the RPR or falls below the lower threshold, a listener perceives a deviation from the

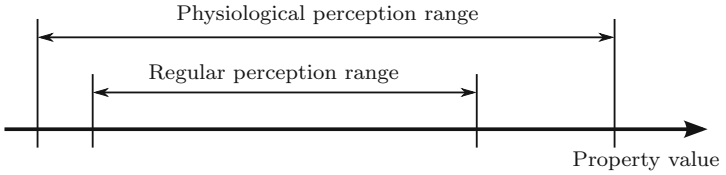


Fig. 6.2 Illustration of the perceptual ranges [30]

expected nature of this property, e.g., a female speaker with a pitch of 160 Hz evokes the impression of an unusually dark voice.

In a more general sense this means that features (e.g., Mel-Frequency Cepstral Coefficients (MFCCs)) that are utilized for quality estimation are first extracted from a database of natural reference speakers. The extracted values for each feature then span the RPR for each feature. Thus, if a feature that is extracted from a synthetic speech signal falls outside of the RPR of this feature for natural speech, it may carry information for the listener that the speech signal is not natural, and potentially also of degraded quality.

The Perceptual Regularization (PR) process for time-invariant features leads to a binary value $\{0, 1\}$ for each so called quality element while time-variant features are converted to values between 0 and 1, where a 0 indicates no degradation and every value greater than 0 implies a deviation from the norm. The derived quality elements are then utilized for the actual regular perception model.

The RPM then only selects positively correlating quality elements for model training. A quality estimate is computed according to the following logistic mapping:

$$\varphi(\tilde{y}) = \frac{1}{1 + exp(a\tilde{y} + b)} \tag{6.1}$$

were $\tilde{y}$ denotes the unmapped average and the parameters a and b are optimized using a nonlinear least-squares method.

Therefore, the *"RPM represents a gradual display to which extent the quality-relevant properties of the signal under test fall into their corresponding RPRs"* [30].

6.2.1.3 Results

This section presents quality prediction results achieved by the SVR model and the RPM [31], both developed and trained by Norrenbrock. Therefore, the models were trained on prosodic and MFCC features. The employed Cross Validation (CV) approach is introduced and the quality estimation accuracy is discussed.

Databases and Features

Both approaches were tested on the databases *SD/FA*, *ST/MDS*, and *Seget2007* which were introduced in Chap. 4. A comparative overview is given in the Tables 4.1 and 4.10.

As a basis for the quality prediction, several features are extracted from the signal. First, prosodic information is gathered. These properties cover (i) intonation properties on the level of voiced segments [32, 33], (ii) voice-source properties which estimate the irregularity of successive pitch periods with respect to length (jitter) and energy (shimmer) [34], and (iii) formal rhythm-associated properties based on vocalic and intervocalic durations [32]. Moreover, MFCCs are derived from the signal [30], i.e., 12th order MFCCs are evaluated for active speech frames with a window length of 25 ms and a frame shift of 12.5 ms. Furthermore, their delta and delta-delta values are included.

Cross Validation Setup

An inter-test CV is performed. Therefore, the model is trained on two databases and tested on the remaining one (this partitioning process is repeated in every possible combination). This CV setup will be referred to as Leave-One-Test-Out (LOTO).

Figure 6.3 depicts the CV scenario that is used for model training. The feature matrix is denoted by $\mathbf{X}$ while the target vector $\mathbf{y}$ contains the auditory ratings. For the k-th CV partitioning, the model training comprises the steps in the gray area in the upper part of the figure: feature normalization, supervised feature selection, and model training. The scaling information of the feature normalization, the indices of the selected features, and the model parameter vector, parametrize the estimation of the test data and are passed from the feature normalization, feature selection, and model training steps to their corresponding steps for the creation of the test model in the lower part of the figure. A comparison between the estimate $\hat{\mathbf{y}}_{\text{test}}$ and the true

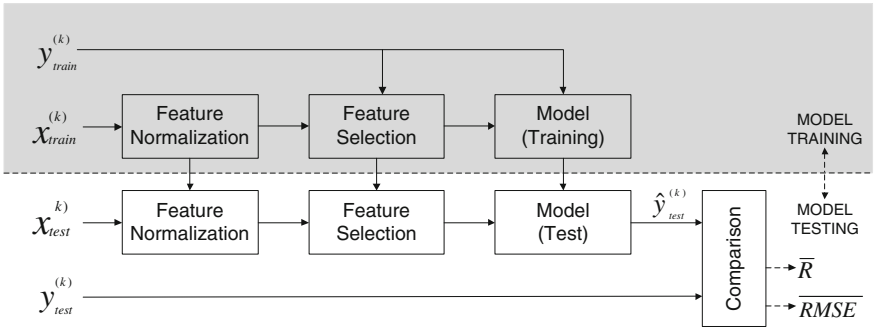


Fig. 6.3 Strict CV setup with feature selection [26]

Table 6.3 Performance of the SVR model and the RPM in a LOTO CV scenario [31]

DATABASE	GENDER	MODEL	NOV		PQ		FAI	
			$\overline{R}$	$\overline{\text{RMSE}}$	$\overline{R}$	$\overline{\text{RMSE}}$	$\overline{R}$	$\overline{\text{RMSE}}$
SD/FA, ST/MDS, Seget2007	FEMALE	SVR	.66	0.61	.53	0.68	.38	0.52
		RPM	.85	0.49	.85	0.54	.77	0.39
	MALE	SVR	.75	0.56	.68	0.61	.55	0.50
		RPM	.90	0.44	.82	0.43	.79	0.34

auditory ratings $\mathbf{y}_{\text{test}}$ over all k then yields the average Pearson correlation $\overline{R}$ and the average $\overline{\text{RMSE}}$.

The same setup will be employed in the following sections for random 3-fold CV [27] for intra-test model validation, i.e., models are trained on $^2/_3$ of the data of a given database, and tested on the remaining $^1/_3$ of that database. To ensure a reliable correlation value for one database, training and testing are repeated (e.g., 500 random CV partitionings) and the correlation values are averaged over all partitionings.

Results and Discussion

Both models were trained to estimate the scores of the following three perceptual quality dimensions: Naturalness of voice (NOV), Prosodic Quality (PQ), and Fluency and Intelligibility (FAI). As a measure of accuracy Pearson's correlation coefficient R was computed between the estimated and auditory assessed score for each quality dimension. Moreover, the Root Mean Square Error (RMSE) is reported. The results are shown in Table 6.3.

The RPM outperforms the SVR for each gender and in every dimension.

While the prediction accuracy of both models in the dimensions NOV and FAI is better on the male data, for the PQ dimension this effect is reversed for the RPM.

Considering the overall performance over all target dimensions the RPM is clearly the superior model. Further information on the performance of these and other models in different CV scenarios and trained on different features is presented in detail by Norrenbrock [31].

6.2.2 Linear Regression Models

This section presents the development of two new linear regression models. One is based on a set of prosodic features computed from parameters that can be extracted via the Fujisaki model, and one approach employs a three step feature selection to reduce a set of approx. 1500 general speech features to a set of a few that still cover most of the information. These models are trained on databases that were introduced

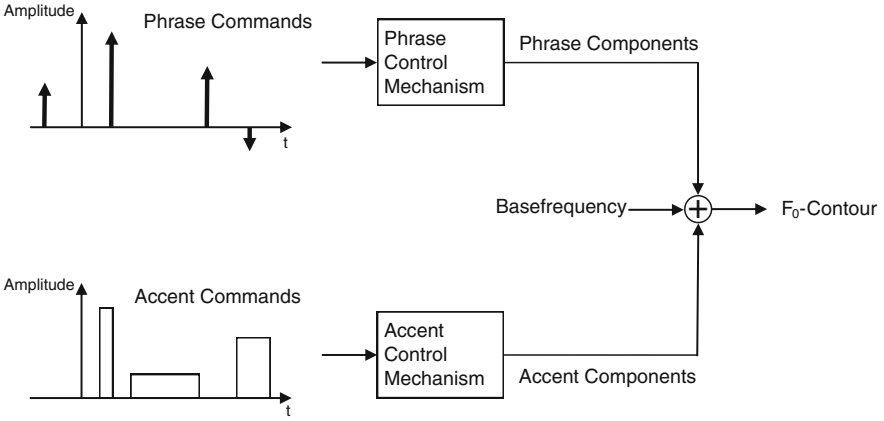


Fig. 6.4 Fujisaki model for the generation of F0 contours [36]

in Chap. 4 and the achieved results are compaired with the performance of the SVR model and the RPM.

6.2.2.1 Linear Regression Based on Fujisaki Features[5]

In the following, a brief overview of the Fujisaki model is presented as well as a detailed description of the developed features derived from the extracted parameters. Moreover, a linear regression model based on the extracted features is constructed.

Fujisaki Model

The F0-contour of speech signals contributes important non-linguistic information like *naturalness* and the current *emotion* of the speaker. Generally, such contours are characterized by a decline from onset towards the end of an utterance. During word accent, the F0-contour is superposed by local intonation humps.

The Fujisaki model [36] follows this principle by describing an F0-contour as a superposition of Phrase Commands (PhCs) and Accent Commands (AcCs) and an underlying Basefrequency (BF). The concept of this model can be seen in Fig. 6.4.

PhCs consist of several Starting Points (SPs), each of them with a specific amplitude, thus they describe a set of impulses. PhCs amplitudes as well as the onset time for the first PhC of a signal can have a negative sign. AcCs consist of SPs and Ending Points (EPs) that describe a set of stepwise functions. The time within one pair of SPs and EPs represents an accented block. In comparison to PhCs, all AcCs amplitudes

[5]Parts of the content of this section have previously been published in a slightly different version in [35].

and their onset times are always positive. The BF describes the minimum value of the logarithmized F0-contour throughout the signal.

The PhCs and AcCs are the input for two critically-damped second-order linear systems to these commands (*phrase control mechanism* and *accent control mechanism*, respectively). The PhCs and AcCs are assumed to be smoothed by the low-pass characteristics of their respective control mechanisms. The output of those control mechanisms (the *phrase components* and *accent components*) and the BF are then joined to form the pitch curve of an utterance. Thus, this model reduces the complexity of a pitch contour to a minimal set of three parameters (PhCs, AcCs, and BF) that still capture the main aspects of the pitch contour.

Fujisaki Features

Keeping in mind that the prosody of synthetic speech is one of the most relevant aspects when it comes to the impression of naturalness, utilizing a model that describes the pitch contour of speech signals as the basis for a quality predictor seems to be of great help. Therefore, the Fujisaki model implemented by Mixdorff especially for the use in German [37] was used to extract the above-mentioned parameters for all TTS files from the German TTS databases mentioned in the following section.

Fourty-seven statistical features based on the extracted Fujisaki parameters were computed. They comprise mean, minimum, maximum values as well as the variances of the extracted parameters. Moreover, several features based on the quantity of increasing/decreasing (in relation to the previous command) PhC/AcC segments in a signal were computed. All features can be derived from Eqs. (6.2)–(6.11) by combining the terms in curly brackets in every possible way, e.g., one of the features from Eq. (6.11) is called *maximum of distances between AcC SPs*.

$$\left\{\begin{matrix} mean \\ minimum \\ maximum \\ variance \end{matrix}\right\} of\ distances\ between \left\{\begin{matrix} \text{PhC SP}s \\ \text{AcC SP}s \\ \text{AcC EP}s \\ \text{AcC SP}s\ and\ following\ \text{EP}s \\ \text{AcC EP}s\ and\ following\ \text{SP}s \end{matrix}\right\} \tag{6.2}$$

$$\left\{\begin{matrix} mean \\ minimum \\ maximum \\ variance \end{matrix}\right\} of \left\{\begin{matrix} \text{PhC } amplitudes \\ \text{AcC } amplitudes \end{matrix}\right\} \tag{6.3}$$

$$number\ of \left\{\begin{matrix} increasing \\ decreasing \end{matrix}\right\} \left\{\begin{matrix} \text{PhC}s \\ \text{AcC}s \end{matrix}\right\} normalized\ by\ length\ of\ the\ signal \tag{6.4}$$

$$relative\ position\ of \left\{\begin{matrix} minimum \\ maximum \end{matrix}\right\} \left\{\begin{matrix} \text{PhC } amplitude \\ \text{AcC } amplitude \end{matrix}\right\} \tag{6.5}$$

$$\left\{\begin{matrix} minimum \\ maximum \\ sum\ of\ all \end{matrix}\right\} \text{AcC } block(s) \tag{6.6}$$

$$sum\ of\ all\ AcC\ blocks\ normalized\ by \begin{Bmatrix} maximum\ amplitude \\ maximum\ \text{AcC}\ block \end{Bmatrix} \tag{6.7}$$

$$ratio\ of\ increasing\ and\ decreasing \begin{Bmatrix} \text{PhC}s \\ \text{AcC}s \end{Bmatrix} \tag{6.8}$$

$$quantity\ of \begin{Bmatrix} \text{PhC}s \\ \text{AcC}s \end{Bmatrix} normalized\ by\ length\ of\ the\ signal \tag{6.9}$$

$$basefrequency \tag{6.10}$$

$$sum\ of\ PhC\ amplitudes\ normalized\ by\ maximum\ amplitude \tag{6.11}$$

Databases

The quality prediction models that are developed in the next paragraph were trained on the following TTS databases:

- *Seget2007* (see Sect. 4.1)
- *Hint2010* (see Sect. 4.1)
- *SD/FA* (see Sect. 4.2)
- *ST/MDS* (see Sect. 4.3)

Even though all databases were generated via similar test procedures, there are differences that have to be taken into account during the following steps and the interpretation of the results.

- *Seget2007* and *Hint2010* used ACR and computed a MOS for the overall quality scale ranging form 1 to 5, while *SD/FA* and *ST/MDS* used continuous scales with a range of 1–7.
- *Seget2007* and *Hint2010* employed natural speakers as reference stimuli. This leads to a compression of the range in which TTS stimuli are rated.
- All databases consist of partially different TTS systems. Hence, the ratings in one database always also depend on the range of quality of TTS systems in it.
- The mean duration of stimuli varies between databases from 5 to 12 s.

A comparative overview of all databases with all relevant information can be seen in the Tables 4.1 and 4.10.

Prediction Models

The aim was to develop one quality predictor based on the presented Fujisaki features. The *overall impression* ratings were chosen as target prediction values, since they were the only quality feature that was assessed in all of the four studies displayed in the previous section. Due to the different ranges of ratings the SD/FA and ST/MDS

Table 6.4 Results of stepwise multiple linear regression analysis for female voices. $R^2 = .26$

FEATURE	B	SE B	β
constant	2.073	0.539	
num dec AcC norm length	0.902	0.248	.304 ***
basefrequency	−0.008	0.003	−.212 *
min dist AcC EP SP	6.131	2.099	.248 **
mean dist PhC SP	0.300	0.107	.234 **

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$
Note see text for explanation of the features

scores had to be transformed to the standard MOS scale range (1–5) before all ratings could be merged.

As learned from previous research [25] the prediction efficiency of most features varies highly between genders. Hence, one stepwise multiple linear regression analysis was conducted for each gender. The auditory MOS of all four databases were used as response variable, while the 47 Fujisaki features described in Sect. 6.2.2.1 were used as predictors.

For both genders one significant model could be created. Table 6.4 lists the selected features for the female model, its beta values (B), their standard errors (SE B), and their standardized values (β). The four features denote the number of decaying AcCs normed by the length of the speech signal (*num dec AcC norm length*), the basefrequency of the signal (*basefrequency*), the minimum distance between EPs and the following SP of the AcCs in a signal (*min dist AcC EP SP*), and the mean distance between PhC SPs (*mean dist PhC SP*). Even though the RMSE for the female predictor is fairly low ($RMSE_f = 0.52$), the model only accounts for 26% of the variability in the outcome.

The male model (Table 6.5) consists of five predictors. These features denote the mean distance between PhC SPs (*mean dist PhC SP*), the quantity of AcCs normed by the length of the speech signal (*quantity AcC norm length*), the mean amplitude in the AcCs in a signal (*mean AcC amp*), the maximum distance between the EP and the following SP of the AcCs in a signal (*max dist AcC EP SP*), and the sum of all

Table 6.5 Results of stepwise multiple linear regression analysis for male voices. $R^2 = .39$

FEATURE	B	SE B	β
constant	1.135	0.407	
mean dist PhC SP	−0.354	0.089	−.334 ***
quantity AcC norm length	0.790	0.171	.404 ***
mean AcC amp	4.380	1.030	.512 ***
max dist AcC EP SP	0.274	0.072	.428 ***
sum AcC blocks	−0.519	0.208	−.302 *

$^{*}p < .05$. $^{***}p < .001$
Note see text for explanation of the features

Table 6.6 Pearson Correlation between predicted MOS and auditory MOS for each database

DATABASE	FEMALE		MALE	
	R	RMSE	R	RMSE
Seget2007	.48**	0.44	.58**	0.48
Hint2010	.05	0.61	.48*	0.55
SD/FA	.66**	0.65	.60**	0.62
ST/MDS	.61**	0.79	.75**	0.70

* $p < .05$. ** $p < .01$

AcC blocks in a signal (*sum AcC blocks*). The RMSE for this model is on the same level as the RMSE for the female predictor ($RMSE_m = 0.48$), however, the male model accounts for 39% of the variablity in the outcome.

Taking a look at both models reveals that the feature *mean dist PhC SP* is the only item that shows up in the female as well as the male predictor.

To test for over-fitting effects a leave-one-out cross-validation was conducted. The R^2 values for both models could be confirmed. The RMSE showed a minor increase for both the female ($RMSE = 0.55$) and the male ($RMSE = 0.51$) predictor. Thus, both models can be accounted to be stable.

Results and Discussion

Both models were used to compute quality estimates for the MOS for all available TTS files. As a measure of accuracy, Pearson's correlation coefficient R between predicted MOS and auditory MOS per database and gender and the RMSE is reported. The achieved correlations can be seen in Table 6.6.

The results for the female stimuli show a strong correlation for the two more complex databases (*SD/FA* and *ST/MDS*) and a medium correlation for *Seget2007*. For the female speech files from *Hint2010* no significant correlation could be achieved. The results for the male files are mostly superior to those of the female databases: for *Seget2007*, *SD/FA*, and *ST/MDS* strong correlations could be achieved while *Hint2010* still reaches $R = .48$.

When comparing the results across databases the correlations for *SD/FA* and *ST/MDS* stand out with $R \geq .60$. The lowest correlations for female and male data were achieved on the *Hint2010* database.

For most databases both predictors achieved strong correlations even though the four databases differ from each other in many ways (see Sect. 6.2.2.1). When comparing the results for *Seget2007* and *Hint2010* with the results for *SD/FA* and *ST/MDS* it is striking that these correlations are on a lower level. The cause could be that the first two databases contain natural speech stimuli while the others do not. This circumstance might well have led to a relatively lower rating of the TTS stimuli in those databases during the listening tests than if there would not have been any natural references in them. The same accounts for the various different TTS systems in

all four databases. The combinations in each of them also influence the ratings of the stimuli. Moreover, the average duration of the TTS signals differs strongly between databases. Until now, a minimal required signal duration for a reasonable prediction accuracy has not been determined. But it was observed that previous prediction models showed problems with shorter signals of about 3–4 s [38, 39].

Furthermore, the R^2 values for both models do not account for more than 40% of the variability in the outcome. Thus, to reliably estimate the quality of TTS systems, additional features will be necessary.

6.2.2.2 Feature Selection and Stepwise Regression for Large-Scale Feature Sets[6]

The previous section presented the development of a linear regression model on a set of 47 extracted features. This simple process, however, can not be adopted to large-scale feature sets. In order to conduct a linear regression analysis in such a case, first a feature selection algorithm has to be executed to reduce the amount of features. This section presents the development of a linear regression model based on a set of approx. 1500 general speech features and compares its performance with the SVR approach which was introduced in Sect. 6.2.1.1.

General Speech Features and Databases

As a basis for the quality prediction approach that will be developed, the feature extraction algorithm described in [40] will be used. The extracted features provide a broad variety of information on vocal expression patterns that are useful when classifying human emotions. As depicted in [25], the inherent information is also suitable when analyzing the quality of synthetic speech.

In the first step, the following low-level audio descriptors are extracted: pitch, loudness, MFCCs, spectrals, formants, and intensity. Subsequently, a statistic unit derives moments, extrema, linear regression coefficients, and ranges of the respective acoustic contours. Among others, this yields features like: pitch range, maximum value of the second formant, the mean of a Mel frequency cepstral coefficient, and the maximum change in spectral flux.

The feature extraction was applied on the databases *SD/FA* and *ST/MDS* (see Sects. 4.2 and 4.3) and 1495 features per file were extracted and used as an input for the quality prediction approach. Similarities and differences of these databases that are relevant for the discussion of the results are summarized in Sect. 4.4.

[6]The content of this section has previously been published in a slightly different version in [26].

General Cross-Validation Setup

Different CV schemes are used in order to investigate the extent to which the broad feature pool can be exploited for quality prediction. Due to the high number of available features, special care is advisable in order to avoid overfitting. First, random 3-fold CV [27] is used for intra-test model validation, i.e., models are trained on $^2/_3$ of the data of a given database, and tested on the remaining $^1/_3$ of that database. And second, a Leave-One-Test-Out (LOTO) CV is performed. Therein, the model is trained using one database and tested on the other (and vice versa). Thus, the results of different auditory tests are compared on the basis of model generalization. The CV setup introduced in Sect. 6.2.1.3 is employed for both CV schemes (see Fig. 6.3).

Three-Step Feature Selection and Stepwise Regression

This section introduces a feature selection approach which is able to cope with extensive feature sets. Due to the differences between female and male speech and the experience that the importance of features for quality prediction heavily depends on the speaker gender of the stimuli [23], separate quality prediction models for each gender were developed.

The main idea of this approach is to reduce a huge feature set via a three-step Feature Selection (FS) to a small subset (fewer than 10) while keeping enough relevant information to be able to build an effective quality prediction model via a Stepwise multiple linear Regression (SR). The FS-SR approach is based on the algorithm described in [41] and can be seen in Fig. 6.5.

In a first step the *Relief* algorithm [42] is applied to omit irrelevant features by a relevance ranking of all i features in the feature set. *Relief* finds the nearest *hit* and the nearest *miss* for each feature x_i in the set, i.e., another sample of the same class and another sample of a different class, respectively. Subsequently, it adjusts the relevance value r of each feature x_i according to (6.12):

$$r = (x_i - \text{near-hit}_i)^2 + (x_i - \text{near-miss}_i)^2 \tag{6.12}$$

The so-called *Relief-F* algorithm is an adaptation of *Relief* for the handling of noisy, incomplete, as well as multi-class data sets [43]. This approach utilizes the Matlab implementation of *Relief-F* to retain the 12.5% most relevant features, i.e., 187 features, for further processing.

In a second step redundancy is removed from this feature subset by applying the k-means algorithm [44] to cluster features into groups of similar features. Ten feature clusters are build and the feature with the highest *Relief-F* relevance value is selected as a representative of each cluster. Since the k-means cluster solution is strongly dependent on the randomly chosen starting points of the cluster centers, each execution of k-means yields slightly different clusters and representatives. Therefore, the occurrence of representatives throughout 2000 executions of k-means is counted

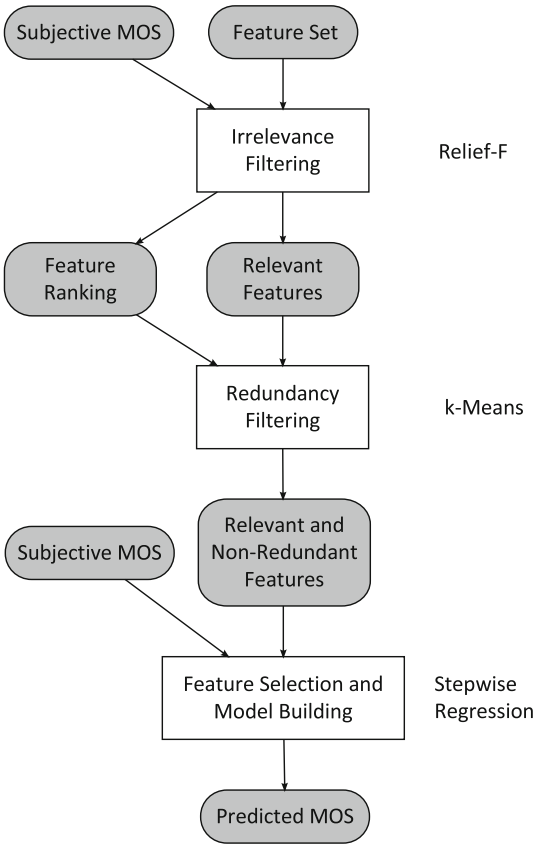


Fig. 6.5 Three-step feature selection and model building

and all features with an occurrence rate above 30% are selected. This reduces the number of features below 15 throughout all databases.

In a third and final step, a stepwise multiple linear regression is applied to select relevant features from the previous subset and to build a prediction model.

Thus, the steps Relief-F, k-means, and the stepwise multiple linear regression from Fig. 6.5 are all part of the *feature selection* blocks in Fig. 6.3. Moreover, the stepwise multiple linear regression also develops the final prediction model as shown in the blocks *model (training)* and *model (test)*. The feature normalization step is employed prior to the Relief-F algorithm and is covered by a z-score normalization of all features.

Results and Discussion

For comparing purposes, the SVR model which was introduced in Sect. 6.2.1 was trained on the same data. Both models were built considering the model assessment

Table 6.7 Performance of quality prediction models on the test sets (3-fold CV). The figures are averaged over 500 random CV partitionings

DATABASE	MODEL	MALE		FEMALE	
		$\overline{R}$	$\overline{\text{RMSE}}$	$\overline{R}$	$\overline{\text{RMSE}}$
SD/FA	FS-SR	**.52**	0.76	**.48**	0.78
	SVR	**.72**	0.57	**.63**	0.65
ST/MDS	FS-SR	**.49**	0.77	**.35**	0.94
	SVR	**.75**	0.68	**.57**	0.83

Table 6.8 Performance of quality prediction models on the test sets (LOTO CV)

DATABASE	MODEL	MALE		FEMALE	
		$\overline{R}$	$\overline{\text{RMSE}}$	$\overline{R}$	$\overline{\text{RMSE}}$
SD/FA, ST/MDS	FS-SR	**.74**	0.64	**.43**	0.88
	SVR	**.80**	0.55	**.61**	0.73

techniques presented in Sect. 6.2.2.2 (3-fold CV with 500 random partitionings and LOTO CV). The results can be seen in the Tables 6.7 and 6.8.

As can be seen from Table 6.7, the average correlation between auditory and predicted MOS for the 3-fold CV varies between .35 and .75. The error range is 0.57–0.94, where the MOS ratings have been scaled to the common ACR scale [1, 5] beforehand. Note that good predictive performance of the model is indicated by high correlations and low error values.

Moreover, this table shows that the SVR approach outperforms the FS-SR in the 3-fold CV scenario. When comparing the results across speaker gender, both models achieve a higher prediction accuracy for the male data.

Turning to the results of the LOTO CV in Table 6.8, the correlations vary between .43 and .80 while the error range is 0.55–0.88. Compared to the results of the 3-fold CV, the quality prediction task with LOTO CV is more ambitious because the models are trained on one database and tested on the other. Strikingly, the correlations for both models with LOTO CV on the male data exceed the averaged correlations that are achieved via a 3-fold CV. This effect applies especially for the FS-SR approach with an average 3-fold correlation of .51 compared to a correlation of .74 in the LOTO CV. A possible cause lies in the nature of the databases. As mentioned in Sect. 4.4. both databases contain a very similar set of TTS systems and are thus not completely independent. Moreover, the models built during the 3-fold CV are always trained on only $^2/_3$ of one database. This implicates that models in some of the 3-fold CV loops are built on a training set that contains no stimuli of synthesizer A while the test set contains only stimuli of synthesizer A (in extreme cases this can happen with entire synthesizer types, e.g., no US, DI, or HMM-synthesizer). This can lead to very low correlations in some 3-fold CV loops which affect the average correlation of the model to a degree that occasionally makes the 3-fold CV inferior to the LOTO

CV. Moreover, since the models in the LOTO CV are trained on the whole database they are theoretically more stable than the models from the 3-fold CV and as a result they can lead to better correlations on other databases.

Furthermore, it is noticeable that, compared to the 3-fold CV, the performance gap between FS-SR and SVR for the male data is considerably smaller. Additionally, the prediction accuracy of both models is substantially better for the male voices than for the female.

When comparing the number of features that were used for model building, a big difference between both models is noticable: while the FS-SR approach uses 4 features independent of CV method, the SVR models consist of over 200 features. Furthermore, the number of features for the SVR models depends on the type of CV: a 3-fold CV uses more features than a LOTO CV.

This points out that one reason for the superior performance of the SVR approach lies in the number of selected features. With between 55 and 69 times the number of features of the FS-SR models, the SVR models are far more comprehensive. A second advantage of the SVR approach is the ability to model nonlinear relationships while the FS-SR approach relies on the linear procedure of a stepwise regression. Thus, these two facts explain most of the different prediction accuracies as seen in the Tables 6.7 and 6.8.

6.2.3 Conclusions

Two different instrumental reference-free quality prediction models have been introduced in this section. First, the Fujisaki model implemented by Mixdorff [37] was used to extract parameters for four German TTS databases with over 200 samples. From these parameters, 47 statistical features were derived. One stepwise multiple linear regression analysis with these features as predictors and the auditory MOS as response variable for female and male data was conducted. Two stable models could be constructed depending on four features for the female data and five features for the male data. Both models have proven stable with only minor changes in R^2 in a leave-one-out cross-validation.

Pearson's correlation coefficient between the predicted MOS and the auditory MOS for each database and gender was computed. With the exception of the female data from *Hint2010* correlations between .48 and .75 could be reached.

Second, a feature extraction algorithm was used to extract 1495 speech features on two subjectively evaluated TTS databases. A three-step feature selection was employed to reduce the large set of features to a small group of features that still cover most of the relevant information. A stepwise multiple linear regression analysis was performed to come up with an estimated quality score.

Two different cross validation techniques were applied: a 3-fold CV, which leads to a high intra-test model validation, and LOTO CV, where the model is validated through a second database and thus leads to more generalizable models.

For comparison purposes, the SVR model, which was introduced in Sect. 6.2.1.1, was trained on the same databases.

The correlations achieved by the FS-SR model are as high as .52 in the 3-fold CV case, while the results for LOTO CV show a correlation of up to .74. In both cases the performance on the male data is superior to the performance on the female data.

However, the more complex SVR models outperform the FS-SR in every case. While for the 3-fold case the performance of SVR is clearly superior, the gap for the male data in the LOTO CV case is considerably smaller. When comparing these results to the results of the SVR model and the RPM which were presented in Sect. 6.2.1 the clearly most reliable measure is still the RPM.

The superior performance of both, the SVR model and the RPM, originates from their ability to model non-linear relationships and the utilization of a far larger feature set for model building. Whether this performance is good enough to integrate such a measure into an existing TTS system to aid during the unit selection process will be explored in the following chapter.

6.3 Summary

In this chapter different reference-based and reference-free instrumental quality estimation measures were presented. Several state-of-the-art instrumental models for the quality assessment of distorted natural speech were tested on synthetic speech. Neither the reference-based nor the reference-free models are helpful for the case of synthetic speech. In the reference-based case, the main reason for this seems to be major problems to find a correct temporal alignment between reference and synthetic speech file. Moreover, two reference-free linear regression models were developed, which achieved reasonable correlations on some databases. However, especially for female TTS files, both approaches were outperformed by the more complex SVR model and the RPM, which are able to model nonlinear relations.

References

1. ITU-T Contribution COM 12-99 (2010) Packet Loss and Coding Impact on Quality of Synthesized Speech Predicted by PESQ and P.563 Models. Ministry of Transport, Construction and Regional development of the Slovak Republic (Author: Počta, P.), ITU-T SG12 Meeting, Geneva, Switzerland
2. ITU-T Contribution COM 12-221 (2011) Predicting the Quality of Synthesized and Natural Speech Impaired by Packet Loss Using P.862 and P.863 Models. Ministry of Transport, Construction and Regional development of the Slovak Republic (Author: Počta, P.), ITU-T SG12 Meeting, Geneva, Switzerland
3. ITU-T Contribution COM 12-220 (2011) Predicting the Quality of Synthesized and Natural Speech Impaired by Coding Using P.862 and P.863 Models. Ministry of Transport, Construction and Regional development of the Slovak Republic (Author: Počta, P.), ITU-T SG12 Meeting, Geneva, Switzerland
4. Cernak M, Rusko M (2005) An Evaluation of Synthetic Speech Using the PESQ Measure. In: Proceedings of Forum Acusticum 2005, Budapest, pp 2725–2728

5. Latacz L, Verhelst W (2015) Double-Ended Prediction of the Naturalness Ratings of the Blizzard Challenge 2008–2013. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015), Dresden, Germany, pp 3486–3490
6. Hinterleitner, F, Zabel S, Möller S, Leutelt L, Norrenbrock C (2011) Predicting the Quality of Synthesized Speech Using Reference-Based Prediction Measures. In: Proceedings of the 22th Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Aachen, Germany, pp 99–106
7. Karaiskos V, King S, Clark RAJ, Mayo C (2008) The Blizzard Challenge 2008. In: Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA)
8. King S, Karaiskos V (2009) The Blizzard Challenge 2009. In: Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA)
9. King S, Karaiskos V (2011) The Blizzard Challenge 2010. In: Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA)
10. ITU-T Rec. P.563 (2004) Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony. International Telecommunication Union, Geneva
11. Malfait L, Berger J, Kastner M (2006) P.563 - The ITU-T Standard for Single-Ended Speech Quality Assessment. IEEE Trans Audio Speech Lang Process 14:1924–1934
12. Kim D-S (2005) ANIQUE: An Auditory Model for Single-Ended Speech Quality Estimation. IEEE Trans Speech Audio Process 13(5):821–831
13. Kim D-S, Tarraf A (2006) Enhanced Perceptual Model for Non-Intrusive Speech Quality Assessment. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France, pp 829–832
14. Gray P, Massara RE, Hollier MP (1998) Constraint-Based Pitch-Cycle Identification Using a Hybrid Temporal Spectral Method. In: Proceedings of the 105th Audio Engineering Society Convention
15. Gray P, Hollier MP, Massara RE (2000) Non-Intrusive Speech Quality Assessment Using Vocal Tract Models. IEE Proc Vis Image Signal Process 147:493–501
16. Möller S, Heimansberg J (2006) Estimation of TTS Quality in Telephon Environments Using a Reference-free Quality Prediction Model. In: Proceedings of 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, pp 56–60
17. Heimansberg J (2006) Instrumentelle Schätzung der Qualität synthetischer Sprache. Bachelorarbeit, Institut für Kommunikationsakustik, Ruhr-Universität Bochum
18. Möller S, Kim D-S, Malfait L (2008) Estimating the Quality of Synthesized and Natural Speech Transmitted Through Telephone Networks Using Single-ended Prediction Models. Acta Acust United Acust 94:21–31
19. ITU-T Contribution COM 12-180 (2008) Single-Ended Quality Estimation of Synthesized Speech: Analysis of the Rec. P.563 Internal Signal Processing. Deutsche Telekom AG (Authors: Möller, S. and Falk, T.H.), ITU-T SG12 Meeting, Geneva, Switzerland
20. Möller S, Falk TH (2009) Quality Prediction for Synthesized Speech: Comparison of Approaches. In: Proceedings of NAG/DAGA 2009 International Conference on Acoustics, pp 1168–1171
21. Falk TH, Möller S, Karaiskos V, King S (2008) Improving Instrumental Quality Prediction Performance for the Blizzard Challenge. In: Proceedings of Blizzard Challenge Workshop
22. Mariniak A (1993) A global framework for the assessment of synthetic speech without subjects. In: Porceedings of 3rd European Conference on Speech Processing and Technology (Eurospeech 1993), pp 1683–1686
23. Falk TH, Möller S (2008) Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems. IEEE Signal Process Lett 15:781–784
24. Hinterleitner F (2010) Vorhersage der Qualität synthetischer Sprache mittels eines signalbasierten Maßes. Magisterarbeit, Quality and Usability Lab, TU Berlin
25. Möller S, Hinterleitner F, Falk TH, Polzehl T (2010) Comparison of Approaches for Instrumentally Predicting the Quality of Text-to-Speech Systems. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010), pp 1325–1328

26. Hinterleitner F, Norrenbrock C, Möller S, Heute U (2013) Predicting the Quality of Text-To-Speech Systems from a Large-Scale Feature Set. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), pp 383–387
27. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. Springer, Heidelberg
28. Schölkopf B, Smola AJ, Williamson RC, Bartlett PL (2000) New Support Vector Algorithms. Neural Comput 12:1207–1245
29. Chang C-C, Lin C-J (2011) LIBSVM : A Library for Support Vector Machines. ACM Trans Intell Syst Technol 2(3):27:1–27:27
30. Norrenbrock C, Hinterleitner F, Heute U, Möller S (2015) Quality Prediction of Synthesized Speech Based on Perceptual Quality Dimensions. Speech Commun 66:17–35
31. Norrenbrock C (2014) Instrumental Quality Estimation for Synthesized Speech Signals. Ph.D. thesis, Christian-Albrechts-Universität zu Kiel
32. Norrenbrock C, Hinterleitner F, Heute U, Möller S (2012) Instrumental Assessment of Prosodic Quality for Text-To-Speech Signals. IEEE Signal Process Lett 19:255–258
33. Norrenbrock C, Hinterleitner F, Heute U, Möller S (2012) Quality Analysis of Macroprosodic F0 Dynamics in Text-To-Speech Signals. In: Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012), Portland, USA, pp 454–457
34. Norrenbrock C, Heute U, Hinterleitner F, Möller S (2011) Aperiodicity Analysis for Quality Estimation of Text-to-Speech Signals. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011), pp 2193–2196
35. Hinterleitner F, Norrenbrock C, Möller S (2012) On the Use of Fujisaki Parameters for the Quality Prediction of Synthetic Speech. In: Proceedings of the 23th Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Cottbus, (Germany), pp 112–119
36. Fujisaki H (1981) Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. Acoustical Analysis and Physiological Interpretations. STL-QPSR 22:1–20
37. Mixdorff H (2016) MANUAL for the FujiParaEditor - An Interactive Tool for Extracting Fujisaki Model Parameters. http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html. Accessed 15 May 2016
38. Hinterleitner F, Möller S, Falk TH, Polzehl T (2010) Comparison of Approaches for Instrumentally Predicting the Quality of Text-To-Speech Systems: Data from Blizzard Challenges 2008 and 2009. In: Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA)
39. Hinterleitner F, Möller S, Norrenbrock C, Heute U (2011) Vergleich unterschiedlicher Ansätze zur instrumentellen Vorhersage der Qualität von Text-to-Speech Systemen: Daten der Blizzard Challenge 2010. In: Proceedings DAGA 2011
40. Minker W, Lee GG, Mariani J, Nakamura S (2010) Salient Features for Anger Recognition in German and English IVR Portals. In: Spoken Dialogue Systems Technology and Design. Springer, Heidelberg
41. Bins J, Draper BA (2001) Feature Selection from Huge Feature Sets. In: Proceedings of the 8th IEEE International Conference on Computer Vision 2001 (ICCV 2001)
42. Kira K, Rendell LA (1992) The Feature Selection Problem: Traditional Methods and a New Algorithm. In: Proceedings of the 10th National Conference on Machine Intelligence, pp 129–134
43. Kononenko I, Simec E, Robnik-Sikonja M (1997) Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. Appl Intell 7:39–55
44. MacQueen J (1967) Some Methods for Classification and Analysis of Multivariate Observations. In: LeCam LM, Neyman J (eds) Proceedings of the 5th Berkley Symposium on Mathematics, Statistics and Probability. University of California Press, Berkley, pp 281–297

Chapter 7
Requirements for the Integration of an Instrumental Quality Measure into a Concatenative TTS System

In the preceding chapter, several approaches towards the instrumental quality assessment for TTS signals were displayed and correlations as high as .90 could be achieved by the cross validated RPM. The ultimate aim for the development of such models, however, is their integration into a TTS system, in order to further increase its quality. If the measure is able to estimate the perceived quality of a listener with a sufficient accuracy, it could be employed, e.g., for the selection of speech units.

This chapter explores the capabilities of the RPM as outlined in Sect. 6.2.1.2 for the quality improvement of unit selection voices created with MaryTTS. Therefore, the unit selection approach of MaryTTS is described and methods for the generation of multiple versions of the same utterance are introduced (Sect. 7.2). The RPM can then be used to choose the best of these alternatives. A listening test is conducted to examine whether some of the generated alternative versions actually feature a superior quality compared to the original MaryTTS output (Sects. 7.3 and 7.4). And lastly, different RPMs are applied to estimate the quality of the synthesized speech signals (Sects. 7.5 and 7.6) and the achieved results are used to specify requirements for the integration of instrumental quality measures (Sect. 7.7). Thus, this chapter addresses research question RQ5:

> *Which requirements does an instrumental measure need to fulfill in order to be integrated into a TTS system?*

More precisely, the following questions will be answered:

- How can the MaryTTS unit selection synthesizer be utilized to create multiple versions for the same input text?
- Is it possible to create alternative versions with a higher quality than the original MaryTTS output?
- Is the RPM able to automatically chose the superior version?

F. Hinterleitner, *Quality of Synthetic Speech*, T-Labs Series
in Telecommunication Services, DOI 10.1007/978-981-10-3734-4_7

7.1 Regular Perception Model (RPM)

Of all instrumental quality measures under test (see Sect. 6.2), the RPM achieved the highest accuracy, and was therefore chosen as a quality prediction measure for the upcoming task. In this section the training of the RPM and its attained quality prediction accuracy are described. All models that are trained in this section will be referred to as RPM1 later on.

7.1.1 Model Training

When the model from Sect. 6.2.1.1 was trained, its task was to predict the quality of TTS systems which were built upon recordings of different speakers. Thus, in order for the employed Regular Perception Ranges (RPRs) to be most comprehensive, they were extracted from various natural speakers. Therefore, a set of about 6 min duration of different German speakers from the Kiel Corpus [1] was selected for the RPR identification. From now on, this RPR will be referred to as *"a"*.

In the current case, however, the task of the RPM is to predict the quality of only one system that employs one specific speaker. Thus, RPRs were also extracted from a set of approximately 6 min duration, from the recordings that were used for the voice creation. This RPR will be referred to as *"b"*. For comparing purposes, separate models were trained on *a* and *b*.

The already subjectively evaluated database of MaryTTS unit selection voices from Sect. 5.3 was chosen for model training. An overview of the characteristics of this data set is given in Table 7.1. Besides featuring unit selection voices of different corpus sizes, and thus TTS signals of a wide range of quality, the unit selection voices that will be developed in the following section also employ the same speech corpus. Therefore, this database seemed to be an ideal training set for the instrumental measure.

In order for RPM1 to give a comprehensive impression of the quality of the estimated TTS signals, separate RPMs were trained on all of the five perceptual quality dimensions (NOV, PQ, FAI, AOD, and C).

Following the CV setup introduced in Sect. 6.2.1.3, a 3-fold CV with 500 random CV partitionings was employed for model training. All in all, 10 models were trained, five employing *a* and five emplyoing *b*.

7.1.2 Results

The achieved averaged correlations $\overline{R}$ between estimated and subjectively evaluated values of each quality dimension for the database from Sect. 5.3 are shown in Table 7.2. Moreover, absolute and relative differences between the models trained

Table 7.1 Characteristics of the data used for RPM1 training

Language	German
Synthesizer type:	
Formant	
Diphone (PSOLA/MBROLA)	
Unit Selection	✓
HMM	
Number of systems	1
Number of configurations*	6
Stimuli per configuration	10**
Quality assessment via	CS
Number of scales	10
Natural reference	no
Length of stimuli	8 s
f/m voices	male
Separate sessions for f/m	-
Preprocessing	-
Sampling rate	16 kHz

* A configuration denotes a specific combination of one voice/one speech corpus and one synthesis system.
** For one configuration only five TTS stimuli were available.

Table 7.2 RPM1 performance on the training data (Table 7.1)

	RPM1*a*		RPM1*b*		Absolute Difference		Relative Difference	
	$\overline{R}$	$\overline{\text{RMSE}}$	$\overline{R}$	$\overline{\text{RMSE}}$	$\overline{R}$	$\overline{\text{RMSE}}$	$\overline{R}$	$\overline{\text{RMSE}}$
NOV	.75	6.88	.82	6.06	0.08	−0.83	10.26%	−12.00%
PQ	.79	6.21	.84	5.55	0.05	−0.66	6.48%	−10.70%
FAI	.68	12.28	.78	10.82	0.09	−1.46	13.90%	−11.90%
AOD	.73	7.50	.81	6.64	0.08	−0.86	10.52%	−11.42%
C	.65	4.79	.74	4.33	0.08	−0.46	12.53%	−9.59%
Mean value	.72	7.53	.80	6.68	0.08	−0.85	10.73%	−11.12%

Note RPM1*a* indicates RPM1 trained employing *a* and RPM1*b* indicates RPM1 trained employing *b*

on *a* and *b* are given. For the interpretation of the corresponding $\overline{\text{RMSE}}s$ it has to be stated that the dimension values range from 0 to 100.

RPM1 achieves a mean correlation over all dimensions of .72 and .80 for RPM1*a* and RPM1*b*, respectively. Thus, an average increase of over 10% can be stated for the model, when trained on the RPR that was extracted from the speaker of the actual TTS voice. In fact, an increase in prediction accuracy can be noted for every single dimension. Therefore, RPM1 performs better if the RPR of the extracted features can be computed upon natural speech data that is spoken by the same speaker as the to be estimated TTS data.

To check if the performance can be further improved when the data for RPR computation increases, a set of about one hour of speech material, from the corpus the voice is built upon, was selected for RPR identification (*"c"*). Another RPM was trained (3-fold CV with 500 random CV partitionings) utilizing *c*, but no further improvements could be noted.

Overall, when comparing the achieved average correlations with the ones of the RPM from Table 6.3, a slightly lower correlation for NOV and FAI on the present data set was found. This however, could have been caused by the different CV method, which influences the correlations, as shown in Sect. 6.2.2.2.

7.2 Unit Selection Voice Creation in MaryTTS

This section illustrates the unit selection voice creation with MaryTTS. Therefore, the employed natural speech database is described, the utterances that were synthesized are introduced, and different approaches for the generation of multiple versions of the same utterance within MaryTTS are presented.

7.2.1 Test Database

The current study extends research that was presented in Sect. 5.3 and aims at further improving unit selection voices generated by MaryTTS. Therefore, the same recordings of a professional German speaker serve as a basis for the creation of two unit selection voices.

7.2.1.1 TTS Systems

This study aimes at creating one high quality and one low quality voice and at improving both of them by employing the RPMs trained in the previous section. Considering the findings of Sect. 5.3.2, two different subsets were chosen for voice building: a corpus size with an overall duration of 6:31 h and one with 2:10 h, corresponding to voice B and D (see Table 5.14), respectively. Both speech corpora consist of files containing names, addresses, names of places and countries, abbreviations, directions, numbers, and short sentences mainly containing news content.

Again, MAUS [2, 3] was used to align text and audio. All recordings were downsampled to 16 kHz sampling rate due to the prerequisites of MaryTTS.

7.2.1.2 Speech Material

In comparison to the study presented in Sect. 5.3, this time, shorter utterances were selected for synthesis. This is due to two factors: first, the synthesized files should be assessed in a PC test which requires stimuli of a short duration in order to be feasible for the test participants. Secondly, due to the procedure of generating alternative versions of each utterance, which will be explained in the following section, degradations in the signal (and thus a decline in quality) are more likely for stimuli of a longer duration.

Utterances were again chosen from the EUROM.1 spoken language resource database [4]. In order for RPM1 to be completely independent from its training data, only utterances that were not part of the study in Sect. 5.3 were selected. Utterances that exceeded a synthesized duration of 5 s were shortened to their main clauses. All 95 utterances were synthesized with the voices B and D; the average duration of all created speech files is 3.5 s.

7.2.2 Generation of Alternative Versions

Up to now, this section described the generated MaryTTS unit selection voices and the utterances that are to be synthesized. In order for the RPM1 to help to improve the quality of the generated speech, methods for the generation of multiple versions of the same utterances have to be developed. RPM1 can then be applied to chose the one with the highest estimated quality.

Before these approaches are introduced, the process of unit selection within MaryTTS will be illustrated. The unit selection procedure of MaryTTS is based on diphone units. Thus, the word *"ball"*, which transcribes to *bQl* in SAMPA notation, consists of the following diphones: *-b*, *b-Q*, *Q-l*, *l-*.[1]

An illustration of the unit selection process for the generation of the word *"ball"* can be seen in Fig. 7.1. Here, the algorithm has to chose between *i* candidates for the diphone *-b*, *j* candidates for the diphone *b-Q*, etc. Thus, considering that there were 10 candidates available for each diphone, this would lead to 10^4 possible paths just for the word *"ball"*. In order to not track every single path until the end, MaryTTS only keeps track of the last two diphones. Thus, when looking for candidates for the diphone *Q-l*, the decision for the diphone *-b* gets fixed and only paths from this fixed diphone will be considered from there on.

The decision for one candidate over the others is based upon the costs for these units. As introduced in Sect. 2.1.4.2, a unit selection synthesizer choses a unit with the objective to minimize target and join costs. By default, MaryTTS choses the candidate with the lowest costs. Thus, alternatives can be created by not selecting the candidate with the lowest costs, but the one with the second, third, or forth lowest cost. This leads to the generation of three alternative versions.

[1] *-b* and *l-* mark the starting and ending sounds of the word *bQl*.

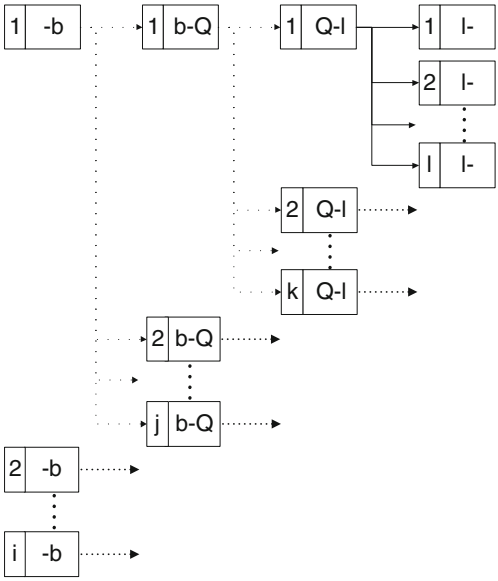


Fig. 7.1 Unit selection for the word *bQl*

The default weights for these costs in MaryTTS are: *target costs* = 0.7 and *join costs* = 0.3. Therefore, by emphasizing either target or join costs, different signals can be generated. In order to generate signals that clearly differ from the default MaryTTS output, two alternatives were created by weighting target or join costs with a factor of 0.99.

Thus, MaryTTS was assigned to generate the following six versions for each utterance:

- default output (target costs = 0.7, join costs = 0.3)
- second best path (target costs = 0.7, join costs = 0.3)
- third best path (target costs = 0.7, join costs = 0.3)
- fourth best path (target costs = 0.7, join costs = 0.3)
- best path (target costs = 0.99, join costs = 0.01)
- best path (target costs = 0.01, join costs = 0.99)

After the generation of all six versions of each utterance, RPM1 was applied to estimate the quality in each dimension. However, expert listening (four expert listeners from the Quality & Usability Lab of the TU Berlin) showed that RPM1 is not able to reliably select the best version of a set of six files.

Considering that in most cases the alternative would actually degrade the quality, due to not choosing the best path or unbalancing the ratio between target and join costs, it is not clear whether any of the alternatives feature a superior perceptual quality. To find out whether this is the case, 36 alternatives that exhibited improvements to expert listeners (six expert listeners from the Quality & Usability Lab of the TU Berlin) and their corresponding default version, were selected for a listening test which is described in the next section.

7.3 Experimental Setup

This section depicts the auditory assessment of original and alternative versions through pairwise comparison testing and a semantic differential.

7.3.1 Pairwise Comparison (PC)

In the PC test, participants were instructed to compare the original version of a stimulus with its alternative. By directly comparing two perceptual impressions, even smaller deviations in quality were expected to be observable compared to the SD task.

Therefore, pairs of stimuli (X and Y), an original and its corresponding alternative, were presented within a Matlab GUI. Then the participants had to decide on one of the six categories shown in Fig. 7.2.

Equally distributed values (ranging from −2.5 to 2.5) were assigned to each category.

7.3.2 Semantic Differential (SD)

The objective of the assessment via an SD was to identify perceptual differences in multiple dimensions between the original and its corresponding alternative version. In order to get a comprehensive impression of the listeners perception, the following quality dimensions were captured through the assigned attribute scales as introduced in Sect. 4.5:

- **Naturalness of Voice (NOV):** voice pleasantness
- **Prosodic Quality (PQ):** stress
- **Fluency and Intelligibility (FAI):** intelligibility
- **Absence of Disturbances (AOD):** disturbances

Dimension C was not included in this study due to its low importance, especially for the evaluation of files that all stem from the same TTS system. Moreover, a scale covering the *overall impression* (OIMP) of the listeners was included.

X sounds much better than Y	*X sounds better than Y*	*X sounds slightly better than Y*	*Y sounds slightly better than X*	*Y sounds better than X*	*Y sounds much better than X*

Fig. 7.2 The six categories from the PC task

7.3.3 Test Procedure

Thrity-two naïve participants (15 female, 17 male) aged 20 to 33 ($\mu = 26.47\,\sigma = 3.15$) were invited to take part in the listening test. All of them were native German speakers and were paid for their participation. The stimuli were presented via headphones (AKG K601) and a high-quality sound device (Roland Edirol UA-25) in a sound proof booth. The test was designed within subjects, i.e., all participants rated all stimuli.

The listening test consisted of three parts. At first, each participant was instructed to complete the PC task. The other two parts consisted of rating the stimuli from part 1 on the attribute scales introduced in Sect. 7.3.2. Stimuli were presented within a Matlab GUI, where ratings could be given on a continuous scale ranging from 0 to 100. To avoid any impact with regard to the order of the scales or the order of the stimuli, both the sequence of scales and the playlist of the stimuli were randomized between subjects.

To avoid listener fatigue, a 5 min break had to be taken after every part.

7.4 Statistical Analysis

In this section, the data gathered in the listening test is statistically analyzed with the objective to find out whether the selected alternative versions were perceived as being superior, compared to the original MaryTTS version.

Prior to the execution of any statistical methods, the ratings of all test participants were screened for plausibility. Therefore, separate boxplots, featuring the ratings from the PC and the SD task, were created for all stimuli. The PC ratings of three test participants had more than 10% outliers (4 or more outliers out of 36 given ratings) and were thus excluded from the study. Moreover, the data of three participants from the SD task had to be excluded due to more than 5% outliers ratings (18 or more outliers out of 360 given ratings).

7.4.1 PC Data

To check whether the selected alternative version was favored over the original, t-tests for each of the comparisons were computed. The results are shown in the Tables A.1 and A.2 for the files synthesized by voice B and D, respectively. Twenty-four of the thirty-six alternatives were perceived as significantly better than the original.

Table 7.3 Significant improvements detected in PC and SD

	SIGNIFICANT IMPROVEMENTS	
	ABSOLUTE	RELATIVE
PC	24	66.7%
OIMP	24	66.7%
NOV	19	52.8%
PQ	18	50.0%
FAI	24	66.7%
AOD	21	58.3%

7.4.2 SD Data

In order to gain insight into the improvements in each dimension, t-tests for the ratings on each attribute scale for all files were conducted. The detailed results are shown in the Tables A.3–A.38. Twenty-four alternatives achieved significantly higher scores for the *overall impression* as well as the FAI dimension.

7.4.3 Discussion

A summary of the analysis of both, the results of the PC and the SD, is shown in Table 7.3. Two surprising results were found: firstly, even though the PC was expected to be a simpler task for the participants due to the direct comparison of two stimuli, an equal number of improved stimuli were found in the SD for the scales OIMP and FAI. And secondly, although most of the significant improvements that were found for the PC data, overlap with the findings for the scales OIMP and FAI, this is not always the case. For example, for file 4 a significant improvement could be verified in the PC task (Table A.1), while no significant improvement was found for any of the attribute scales (Table A.6). The same effect was found in the opposite direction: file 23 was perceived as superior in the OIMP and the dimensions NOV, PQ, and FAI (Table A.25), while no improvement was found in the PC task (Table A.1).

7.5 Quality Prediction

This section examines the quality prediction performance for RPM1 and trains a new model (RPM2) on the data from the listening test conducted earlier in this chapter.

Table 7.4 Mean estimated and subjective scores

	MEAN SCORE ESTIMATED BY RPM1*b*	MEAN SUBJECTIVE SCORE
NOV	30.07	45.18
PQ	31.81	42.64
FAI	36.40	59.55
AOD	27.76	42.05

7.5.1 RPM1 Performance

The previously trained RPM1 was employed to estimate the quality of the files from the SD task of Sect. 7.3. However, RPM1 only achieved correlations between .06 and .09 for all assessed dimensions. This poor performance is also reflected in a severe underestimation of the scores in all dimensions, as shown in Table 7.4. These results also explain why RPM1 was not able to help during the selection process of alternatives in Sect. 7.2.2.

As previous studies have shown [5], quality estimation of TTS files with a short duration (<5 s) is a very challenging task. Therefore, one reason for the poor performance could be the difference in duration between the files RPM1 has been trained on (duration: 8 s) and the files that were assessed in the listening test (duration: 3.5 s).

7.5.2 RPM2 Training

In order to master this task, a new model (RPM2) was trained on the data from this listening test (see Table 7.5). Again, different models for the RPRs *a* and *b* were trained and a 3-fold CV with 500 random partitionings was employed for model training. Table 7.6 lists the correlation $\overline{R}$ and the $\overline{\text{RMSE}}$ averaged over all CV partitionings. Just as for RPM1 (see Table 7.2), an increase in prediction accuracy can be stated for the models employing *b*. This time, however, the average increase, compared to the models utilizing *a*, is only around 8%. Also, in comparison to the performance of RPM1 on the data from Table 7.2, a decrease of the averaged Pearson correlation coefficient of .09 and .12 for the models RPM2*a* and RPM2*b*, respectively, could be noted. This is, as stated before, most likely due to the very short duration of the estimated TTS files.

7.5.3 RPM2 Performance

To test the performance of the newly trained RPM2*b*, a set containing partly unseen data was compiled. This set comprises:

Table 7.5 Characteristics of the data used for RPM2 training

Language	German
Synthesizer type:	
Formant	
Diphone (PSOLA/MBROLA)	
Unit Selection	✓
HMM	
Number of systems	1
Number of configurations*	2
Stimuli per configuration	17/19
Quality assessment via	CS
Number of scales	5
Natural reference	no
Length of stimuli	3.5 s
f/m voices	male
Separate sessions for f/m	-
Preprocessing	-
Sampling rate	16 kHz

* A configuration denotes a specific combination of one voice/one speech corpus and one synthesis system.

Table 7.6 RPM2 performance on the training data (Sect. 7.3.2)

	RPM2*a*		RPM2*b*		Absolute Difference		Relative Difference	
	$\overline{R}$	$\overline{\text{RMSE}}$	$\overline{R}$	$\overline{\text{RMSE}}$	$\overline{R}$	$\overline{\text{RMSE}}$	$\overline{R}$	$\overline{\text{RMSE}}$
OIMP	.66	7.15	.65	7.15	0.00	0.00	−0.69%	0.03%
NOV	.64	6.40	.72	5.77	0.08	−0.62	13.24%	−9.76%
PQ	.56	9.00	.63	8.41	0.07	−0.58	12.66%	−6.49%
FAI	.66	7.90	.71	7.48	0.04	−0.42	6.59%	−5.31%
AOD	.63	7.84	.69	7.36	0.06	−0.48	8.81%	−6.14%
Mean value	.63	7.65	.68	7.23	0.05	−0.42	8.12%	−5.54%

Note RPM2*a* indicates RPM2 trained employing *a* and RPM2*b* indicates RPM2 trained employing *b*

- the alternatives with significantly improved quality, as identified by the SD listening test (see Table 7.3),
- their corresponding original MaryTTS output,
- and the other four alternatives that were not part of the test.

RPM2*b* was then used to estimate the OIMP and the scores of the four dimensions of each of the six versions of an utterance (the original and the five alternatives), with the objective to find out, whether RPM2*b* would select the alternative from the listening test over the other files. Table 7.7 highlights how often RPM2*b* rated the

Table 7.7 RPM2*b* selection success

	OIMP	NOV	PQ	FAI	AOD
RPM2*b* SELECTION SUCCESS	71%	68%	56%	71%	76%

alternative version from the test as the best version from the set of six files. Thus, RPM2*b* achieves a selection success of over 70% for the OIMP and the dimensions FAI and AOD.

However, since there are no subjective ratings available for the four alternatives that were not part of the test, it has to be stated that the alternative version from the listening test is not necessarily the best of all alternatives. Therefore, even files that RPM2*b* preferred over the original and the alternative form the test, could feature a higher perceptual quality than the original. Thus, the scores from Table 7.7 might actually be higher.

7.6 Automatic Selection of Alternative Versions

This section examines the use of RPM2*b* on completely unseen data and investigates whether RPM2*b* can be employed to improve MaryTTS.

Therefore, 200 utterances from the Kiel Corpus [1] (*Berlin* and *Marburg* sentences) were selected for synthesis by voices B and D. MaryTTS was used to generate the original output as well as the five alternatives described in Sect. 7.2.2. The average duration of all generated TTS files is 2 s.

Then, RPM2*b* estimated the OIMP and the scores of the first four dimensions. Due to the low average correlation of .68 (see Table 7.6), only files that reached ratings, 20% above the original in three of the five estimated values, were considered as improved output. This happened in 80 out of the 400 cases (200 utterances synthesized by voice B and D).

Subsequently, three expert listeners from the Quality & Usability Lab of the TU Berlin evaluated the selected alternatives in a PC task to check for perceptual improvements. While some of the alternatives reached a superior quality, the majority featured partly critical distortions that degraded the perceived quality. Therefore, even though some improvements could be achieved in the present case, the severe degradations that are part of some of the selected files, prevent the application of RPM2*b* as a reliable quality measure for TTS data of short duration.

7.7 Potential Improvements

When looking back at the performance of the RPM on different databases (see Tables 6.3 and 7.2) an average correlation of over .80 can be observed. However, when the model is applied to TTS files of a shorter duration (3.5 s), the correlations drop significantly (see Table 7.6). Nonetheless, it was assumed that a model that is trained on data of shorter duration, would still be able to estimate the quality of shorter files (2 s) with a sufficient accuracy. The results from the previous section, however, show that this is apparently not the case.

Therefore, in a next step, the RPM should be tested on files of around 10 s duration. Until now, there were two factors that prevented further research in this direction:

1. To detect significant differences between original and alternative, a PC task, which required short stimuli, was part of the conducted experiment (Sect. 7.3).
2. The approach for generating alternative versions is more likely to lead to degradations for longer stimuli (Sect. 7.2.2).

The analysis of the data gathered by the SD (Table 7.3) showed that some of the attribute scales are able to detect the same amount of significant differences, between original and alternative, as the PC task. Thus, assessing the quality of original and alternative only via attribute scales is now a real option.

Therefore, only the second limiting factor still holds. To overcome it, new methods for the generation of alternatives, which are less likely to lead to degradations, have to be established. Unit selection synthesis, however, can only be influenced in a limited number of ways, i.e., by changing the emphasis of join and target costs and by choosing different paths during unit selection. However, both techniques were exploited in this chapter. Therefore, other synthesis methods like HMM synthesis, which allow an easy adjustment of voice characteristics, should be explored in the future.

7.8 Summary

In this chapter, the RPM was applied to improve the quality of MaryTTS unit selection voices. First, methods for the creation of multiple versions of one utterance in MaryTTS were introduced. A listening test showed that some of the generated alternatives feature a higher quality than the default output of MaryTTS. The RPM was trained on this data and achieved correlations of around .68 over all dimensions. Nonetheless, the RPM was not able to accurately estimate the quality of a different set of TTS files. This is, however, probably attributed to the fact that the generated stimuli only featured a duration of around 2 s. As observed beforehand, TTS data of such a short duration complicates the prediction task immensely.

The main findings of this chapter are:

- The accuracy of the RPM can be increased, when the RPRs are computed based upon the natural speaker, the to be estimated TTS files are generated from.
- A small set of around 6 min of natural speech is sufficient for RPR computation.
- The introduced methods for the generation of alternative versions within MaryTTS can lead to an improved quality.
- The RPM is not able to predict the quality of short files (duration < 4 s) with a sufficient accuracy.

References

1. Institut für Phonetik und digitale Sprachverarbeitung, Christian-Albrechts-Universität Kiel. The Kiel Corpus of Read Speech, vol I. http://www.ipds.uni-kiel.de/publikationen/kcrsp.de.html. Accessed 15 May 2016
2. Schiel F (1999) Automatic Phonetic Transcription of Non-prompted Speech. In: Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS 1999), pp 607–610
3. Reichel UD (2012) PermA and Balloon: Tools for String Alignment and Text Processing. In: Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012), pp 1874–1877
4. Chan D, Fourcin A, Gibbon D, Grandstrom B, Huckvale M, Kokkonakis G, Kvale K, Lamel L, Lindberg B, Moreno A, Mouropoulos J, Senia F, Trancoso I, Veld C, Zeiliger J (1995) EUROM-A Spoken Language Resource for the EU. In: Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH 1995), pp 867–870
5. Hinterleitner F, Möller S, Falk TH, Polzehl T (2010) Comparison of Approaches for Instrumentally Predicting the Quality of Text-to-Speech Systems: Data from Blizzard Challenges 2008 and 2009. In: Proceedings of the Blizzard Challenge Workshop, International Speech Communication Association (ISCA)

Chapter 8
Conclusions and Future Work

This chapter summarizes the findings of this book, answers the research questions introduced in Chap. 1, and gives an outlook on future work.

8.1 Summary

The research presented in this book focused on the perceptual quality of synthetic speech. The constant improvement of TTS quality and the popularity of smartphones and their intelligent personal assistants like Siri, Cortana, and Google Now, made synthetic speech accessible to a multitude of people.

The theoretical background for this book is provided in the Chaps. 2 and 3. First, Chap. 2 gave an introduction into speech synthesis. A general setup of a TTS system was introduced and different approaches towards speech signal generation were discussed, e.g., parametric, concatenative, and statistical parametric speech synthesis. The resulting perceptual impression of these systems was examined and reasons for degradations in the generated speech signals were highlighted. Characteristic degradations like *noise*, for example, induce different perceptual impressions. As a result, the quality of synthetic speech is of multidimensional nature. Furthermore, the open source speech synthesis system MaryTTS was introduced.

In Chap. 3, the concept of perceptual quality was introduced and relating terms were specified. A taxonomy for the quality assessment of synthetic speech highlighted the four fundamental dichotomies: glass box versus black box, laboratory versus field studies, linguistic versus acoustic testing, and instrumental versus. auditory assessment. Furthermore, functional tests were introduced that are able to assess intelligibility on word and sentence level and a section concerning judgment tests proposed two different approaches towards multidimensional quality assessment: a Semantic Differential (SD) measures quality via pre-defined attribute scales and a Pairwise Comparison (PC) test evaluates the perceptual distance between pairs of stimuli. Additionally, different instrumental measures that are used for quality estimation of natural speech degraded by telephone channels were introduced.

F. Hinterleitner, *Quality of Synthetic Speech*, T-Labs Series
in Telecommunication Services, DOI 10.1007/978-981-10-3734-4_8

In the following, Chap. 4 gave an overview of state-of-the-art studies concerning perceptual quality dimensions of synthetic speech. Moreover, two experiments were conducted that derived perceptual quality dimensions via an SD and FA and an ST and MDS. An overview of all results was given and five perceptual quality dimensions were derived that can be seen as universal. Subsequently, an evaluation protocol was constructed that links two attribute scales to each of these dimensions. Subsequently, Chap. 5 presented research towards factors influencing these quality dimensions: (i) the perceptual quality of TTS audiobooks was examined, (ii) the influence of speaker's voice was investigated, and (iii) the influence of the size of the speech corpus was analyzed. In the following, Chap. 4 gave an overview of state-of-the-art studies concerning perceptual quality dimensions of synthetic speech. Moreover, two experiments were conducted that derived perceptual quality dimensions via an SD and FA and an ST and MDS. An overview of all results was given and five perceptual quality dimensions were derived that can be seen as universal. Subsequently, an evaluation protocol was constructed that links two attribute scales to each of these dimensions. Subsequently, Chap. 5 presented research towards factors influencing these quality dimensions: (i) the perceptual quality of TTS audiobooks was examined, (ii) the influence of speakers voice was investigated, and (iii) the influence of the size of the speech corpus was analyzed.

Exploring the possibilities of instrumental quality assessment, Chap. 6 first examined the use of reference-based quality measures that are commonly used for quality estimation of telephone coded speech signals and then covered reference-free approaches. More precisely: linear regression models were developed and their performance was compared with the performance of nonlinear state-of-the-art models.

Finally, Chap. 7 presented research towards the integration of the aforementioned models into a TTS system with the objective to improve its quality. Different methods for the generation of multiple versions of the same utterance were introduced, a listening test was conducted to prove that alternative versions can feature a superior quality compared to the default output, and the RPM was employed to automatically select alternative versions. The study concluded with suggestions towards potential improvements of this approach.

8.2 Conclusions

This section concludes the findings of this book by answering the research questions posed in Chap. 1.

RQ1 *Which perceptual quality dimensions are relevant for state-of-the-art TTS systems?*

The studies presented in Chap. 4 revealed the following five perceptual quality dimensions:

- **Naturalness of Voice (NOV)** indicates whether a listener imagines the voice being produced by a human or a machine.
- **Prosodic Quality (PQ)** assesses the prosody of TTS systems.
- **Fluency and Intelligibility (FAI)** captures segmental artifacts that can occur when concatenating speech units.
- **Absence of Disturbances (AOD)** measures the amount of disturbances, e.g., noise and hiss, in the signal.
- **Calmness (C)** quantifies whether the voice sounds relaxed and calm or rather stressed and restless.

It has to be noted that these dimensions were derived from TTS stimuli of a short duration (<15 s) and are therefore primarily significant for applications like short message readers, information systems, or smart-home assistants. Moreover, the dimensions can also only account for the quality of speech generated by FO, DI, US, or HMM systems.

RQ2 *How should a listening test be designed in order to capture all relevant quality dimensions?*

The findings of Chap. 4 led to the creation of a test protocol in Sect. 4.5.6 that is able to assess all five quality dimensions. An SD was constructed that links two attribute scales to each of the dimensions:

- **NOV**:
  - Voice pleasantness: unpleasant voice versus pleasant voice
  - Naturalness: artificial versus natural
- **PQ**:
  - Stress: unnatural stress versus natural stress
  - Rhythm: unnatural rhythm versus natural rhythm
- **FAI**:
  - Intelligibility: unintelligible versus intelligible
  - Fluency: interrupted versus fluent
- **AOD**:
  - Disturbances: disturbed versus undisturbed
  - Noise: noisy versus not noisy
- **C**:
  - Speed: slow versus fast
  - Tension: tense versus calm

Moreover, in order for the test participants to be able to observe all quality features, a minimum stimulus duration of 5 s was recommended. Furthermore, to simplify

the task of differentiating between stimuli of an either very good or very bad quality, scales should feature separate scale and end points as shown in Fig. 3.3.

RQ3 *Which factors influence these perceptual quality dimensions?*

Chapter 5 explored the influence of three factors on the perceptual quality:

- The **application** the TTS system is used in:

  When synthesizing passages of a longer duration, e.g., audiobooks, the importance of the quality dimensions shift. A test protocol was developed for the quality assessment of TTS in audiobook listening tasks that covers the following two dimensions:

  – The *listening pleasure* comprises the voice pleasantness and the overall impression of the listener.
  – The dimension *prosody* is linked to intonation, stress, emotion and speech pauses.

  A factor covering the *content appreciation* could also be identified, but was excluded due to its lack of influence on the *overall impression*.

  It has to be noted that these dimensions do not contradict the findings for RQ1. The resulting set of only two dimensions can be caused by the different application the TTS systems are used in (stimuli with a duration of around 50 s were assessed) and might also be due to the fact that the employed TTS systems were of a high quality, e.g., no FO systems were included.
- The **voice of the speaker** that was used to record the speech corpus:

  A factor analysis revealed the following four factors:

  – Factor 1 comprises the dimensions NOV, PQ, and FAI.
  – Factor 2 could be linked to the dimensions AOD and C.
  – Factor 3 gathers scales that assess *resonance* and *pitch variation*.
  – Factor 4 consists of the attribute scales *pitch* and *darkness*.

  Therefore, factor 4 could be linked to the pitch and the spectral center of gravity of the speaker. The interpretation of factor 3, however, could not be clarified completely.

  Furthermore, it could be shown that some of the voices under test differed significantly concerning the *overall impression* and the factors 1, 2, and 4. Thus, the voice of a speaker of a corpus-based TTS system greatly influences the perceived quality of the system. Therefore, the selection of a speaker is a crucial task.

- The **size of the speech corpus** and the to be synthesized **utterance**:
  - A significant effect of the *size of the speech corpus* was found on all five dimensions. The highest effect was found for the dimension FAI and the lowest for PQ. Surprisingly, the voice with the second largest speech corpus reached the highest rating in most dimensions. However, the cause of this effect could not be resolved completely.
  - A significant effect of the to be synthesized *utterance* was found on the first four dimensions. Thus, some utterances were an easier synthesis task than others. Moreover, the synthesized utterance can affect dimensions differently, i.e., this can lead to higher scores in one dimension while achieving lower scores in another.
  - A significant interaction effect *corpus size*utterance* was found for all five dimensions. It could be shown that the quality in all dimensions does not always incrementally increase with corpus size, e.g., some utterances reach their highest quality rating at lower corpus sizes and the quality of some even decreased towards larger corpus sizes.

RQ4 *How can the quality of synthetic speech be assessed by an instrumental measure?*

Chapter 6 examined the application and development of reference-free and reference-based instrumental quality measures.

- The existing **reference-based measures** PESQ, DIAL, and POLQA were employed for quality estimation of TTS signals. However, none of these models were able to estimate the quality of synthetic speech. The main reason seems to be an inaccurate temporal alignment between reference and TTS signal.
- Multiple **reference-free measures** were employed to estimate TTS quality:
  - A linear regression model, based on features derived from the Fujisaki model, was developed and led to average correlations of .45 and .60 for female and male files, respectively.
  - A three step feature selection was used to select relevant and non-redundant features out of a large scale feature set for model building. A linear regression model was trained and achieved correlations of .43 and .74 in a LOTO CV setup for female and male files, respectively.
  - A more precise estimation can be attained by employing nonlinear models. Especially the RPM achieved promising correlations of up to .90.

RQ5 *Which requirements does an instrumental measure need to fulfill in order to be integrated into a TTS system?*

This RQ was addressed in Chap. 7.

- Methods for the generation of multiple versions of one utterance for MaryTTS unit selection voices were introduced.
- A listening test has shown that some alternatives reach a significantly higher quality than the default MaryTTS output.
- The accuracy of the RPM could be increased by up to 10% by employing natural speech data for RPR computation that was spoken by the same speaker as the to be estimated unit selection voice.
- However, the estimated quality values that were provided by the RPM could not help to select alternatives that feature a superior quality. This seems to be mainly attributed to difficulties in estimating the quality of TTS data of a short duration (<4 s).
- Exploring the application of HMM synthesis for a simpler creation of alternative versions was suggested as one way to generate high quality alternatives with a duration of over 4 s.

8.3 Future Work

This section presents future work in the three main research areas that were addressed in this book.

8.3.1 Perceptual Quality Dimensions

While the first three dimensions (NOV, PQ, and FAI) appeared in most of the examined studies, the dimensions AOD and C could only be derived from the two experiments presented in Chap. 4. However, a significant impact on the overall impression could be identified in both cases. Therefore, a next step should concentrate on examining the influence of these dimensions on other TTS databases.

Furthermore, the studies towards perceptual quality dimensions of synthetic speech presented in this book focused on German and English TTS voices and therefore only covered Germanic languages. The importance of the derived dimensions, however, may shift in other languages of the Indo-European language family and will most likely change for tone languages like Mandarin or Thai. Thus, in a next step, research should be expanded towards other language families.

8.3.2 Influencing Factors

The currently examined use case of TTS in audiobook reading tasks revealed two perceptual quality dimensions. The deviation from the results of studies examining the application of TTS in, e.g., short message readers, is not surprising, however,

research could investigate whether a third dimension that also affects the overall impression, can be derived.

Moreover, recent research just started to explore the use of TTS in kids audiobooks. Given the simpler content of audiobooks targeted at an age group between 5 and 10 years and the intended vivid speaking style of the narrator, this does not only require new approaches towards quality assessment due to, e.g., limited reading skills and therefore a limited ability to read a questionnaire, but might also result in different perceptual dimensions.

The study on the influence of a speaker's voice on the perceptual quality yielded four factors from which the first two could clearly be linked to the five perceptual quality dimensions. The interpretation of the other two, especially factor 3, however, was more vague. Therefore, acoustic correlates of these factors should be investigated to support their interpretation.

Exploring the influence of the size of the speech corpus for unit selection voices within MaryTTS lead to the surprising discovery that the voice built from the second largest speech corpus was rated best in almost all dimensions. The cause of this unexpected result should be investigated more thoroughly. Therefore, join and target costs in each step of the unit selection process of voice A and B should be compared and deviations have to be explored.

8.3.3 Instrumental Quality Measurement

The poor results of the reference-based quality measures is most likely attributed to an inaccurate temporal alignment between reference and synthetic speech signal. Employing a dynamic time warping algorithm that is able to handle more extreme conditions than the time warping algorithm included in POLQA, could solve this problem.

The estimation accuracy of reference-free models reached correlations of up to .90. However, training algorithms on larger databases could even increase their accuracy and, moreover, also generate a more universal model.

Different obstacles were encountered during the attempt to use the RPM to improve the quality of MaryTTS unit selection voices. Future research should (i) pursue the current approach by generating a larger set of alternatives. This can be achieved by, e.g., choosing the second, third, or fourth best diphone only once in each synthesis process. By repeating the selection process until the second, third, and fourth best diphone have been chosen once at each possible point, leads to a set of possibly hundreds of different versions (depending on the length of the utterance). Then the predictor could chose from this large set of alternatives that most likely contains files of a higher quality than the set of five alternatives created in Chap. 7. And (ii) the possibilities of other synthesis approaches like HMM synthesis could be explored. The ability for simple adjustments in voice characteristics seems to be a promising approach for the generation of alternative versions.

Appendix
Statistical Analysis of Chap. 7

A.1 Results of the PC Test

Table A.1 Results of the t-test for voice B for the PC data

File	*Voice*	Test Value = 0					
		t	*df*	Sig. (2-tailed)	Mean Difference	Interval of the Lower	Upper
1	B	8.430	28	.000	1.328	1.005	1.650
2	B	3.008	28	.006	0.603	0.192	1.014
3	B	−1.976	28	.058	−0.397	−0.808	0.014
4	B	2.897	28	.007	0.466	0.136	0.795
5	B	4.353	28	.000	0.776	0.411	1.141
6	B	10.858	28	.000	1.328	1.077	1.578
7	B	4.921	28	.000	0.707	0.413	1.001
8	B	10.239	28	.000	1.431	1.145	1.717
9	B	−0.858	28	.398	−0.155	−0.526	0.215
10	B	5.008	28	.000	0.914	0.540	1.288
11	B	3.684	28	.001	0.466	0.207	0.724
12	B	1.647	28	.111	0.328	−0.080	0.735
13	B	−1.378	28	.179	−0.293	−0.729	0.143
14	B	7.405	28	.000	1.293	0.935	1.651
15	B	5.069	28	.000	0.845	0.503	1.186
16	B	3.691	28	.001	0.845	0.376	1.314
17	B	4.122	28	.000	0.638	0.321	0.955

F. Hinterleitner, *Quality of Synthetic Speech*, T-Labs Series
in Telecommunication Services, DOI 10.1007/978-981-10-3734-4

Table A.2 Results of the t-test for voice D for the PC data

File	*Voice*	Test Value = 0					
		t	*df*	Sig. (2-tailed)	Mean Difference	Interval of the Lower	Upper
18	D	0.793	28	.435	0.121	−0.191	0.433
19	D	1.361	28	.184	0.259	−0.131	0.648
20	D	5.791	28	.000	0.741	0.479	1.004
21	D	0.118	28	.907	0.017	−0.281	0.316
22	D	18.331	28	.000	1.948	1.731	2.166
23	D	1.258	28	.219	0.224	−0.141	0.589
24	D	6.076	28	.000	0.983	0.651	1.314
25	D	6.108	28	.000	0.879	0.584	1.174
26	D	−0.273	28	.787	−0.052	−0.440	0.337
27	D	4.505	28	.000	0.845	0.461	1.229
28	D	2.906	28	.007	0.431	0.127	0.735
29	D	7.839	28	.000	0.914	0.675	1.153
30	D	10.686	28	.000	1.500	1.212	1.788
31	D	8.573	28	.000	1.086	0.827	1.346
32	D	3.048	28	.005	0.534	0.175	0.894
33	D	1.056	28	.300	0.190	−0.178	0.558
34	D	9.001	28	.000	1.293	0.999	1.587
35	D	1.817	28	.080	0.397	−0.051	0.844
36	D	−0.720	28	.478	−0.121	−0.464	0.223

A.2 Results of the SD

Table A.3 Results of the t-test for the SD test of file 1

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the Lower	Upper			
OIMP	−22.724	−28.690	−16.758	−7.802	28	.000
NOV	−21.103	−29.894	−12.313	−4.918	28	.000
PQ	−34.724	−42.588	−26.860	−9.045	28	.000
FAI	−24.138	−35.854	−12.422	−4.220	28	.000
AOD	−18.483	−28.351	−8.615	−3.837	28	.001

Table A.4 Results of the t-test for the SD test of file 2

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−9.586	−16.373	−2.800	−2.893	28	.007
NOV	−4.828	−11.746	2.091	−1.429	28	.164
PQ	−7.276	−16.528	1.976	−1.611	28	.118
FAI	−17.483	−25.689	−9.277	−4.364	28	.000
AOD	−4.862	−11.988	−2.264	−1.398	28	.173

Table A.5 Results of the t-test for the SD test of file 3

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	5.034	−1.031	11.100	1.700	28	.100
NOV	4.000	−2.866	10.866	1.193	28	.243
PQ	21.862	11.142	32.582	4.178	28	.000
FAI	−6.966	−2.335	16.266	1.534	28	.136
AOD	−1.966	−7.913	3.982	−0.677	28	.504

Table A.6 Results of the t-test for the SD test of file 4

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−2.786	−8.806	3.235	−0.949	27	.351
NOV	−3.643	−10.963	3.677	−1.021	27	.316
PQ	−4.107	−13.674	5.460	−0.881	27	.386
FAI	−7.679	−16.480	1.123	−1.790	27	.085
AOD	−2.464	−9.848	4.920	−0.685	27	.499

Table A.7 Results of the t-test for the SD test of file 5

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−19.276	−25.122	−13.430	−6.754	28	.000
NOV	−12.000	−17.687	−6.313	−4.323	28	.000
PQ	−15.621	−24.634	−6.608	−3.550	28	.001
FAI	−20.000	−27.687	−12.313	−5.329	28	.000
AOD	−18.897	−26.664	−11.129	−4.984	28	.000

Table A.8 Results of the t-test for the SD test of file 6

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−22.862	−29.603	−16.121	−6.947	28	.000
NOV	−18.034	−26.637	−9.432	−4.294	28	.000
PQ	−17.414	−26.047	−8.780	−4.132	28	.000
FAI	−34.172	−43.050	−25.295	−7.885	28	.000
AOD	−17.000	−25.341	−8.659	−4.175	28	.000

Table A.9 Results of the t-test for the SD test of file 7

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−6.586	−13.003	−0.170	−2.103	28	.045
NOV	−4.655	−13.214	3.904	−1.114	28	.275
PQ	−6.966	−16.049	2.118	−1.571	28	.127
FAI	−12.207	−20.395	−4.019	−3.054	28	.005
AOD	−5.345	−11.740	1.050	−1.712	28	.098

Table A.10 Results of the t-test for the SD test of file 8

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−23.586	−29.573	−17.599	−8.070	28	.000
NOV	−24.207	−30.901	−17.512	−7.407	28	.000
PQ	−35.586	−44.302	−26.870	−8.363	28	.000
FAI	−19.966	−27.692	−12.239	−5.293	28	.000
AOD	−20.069	−29.238	−10.900	−4.483	28	.000

Table A.11 Results of the t-test for the SD test of file 9

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−4.828	−12.349	2.694	−1.315	28	.199
NOV	−2.759	−10.550	5.033	−0.725	28	.474
PQ	−2.069	−11.362	7.224	−0.456	28	.652
FAI	−6.310	−11.372	−1.249	−2.554	28	.016
AOD	−16.103	−22.991	−9.216	−4.789	28	.000

Table A.12 Results of the t-test for the SD test of file 10

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−17.862	−23.718	−12.006	−6.248	28	.000
NOV	−18.517	−25.643	−11.392	−5.323	28	.000
PQ	−21.000	−31.192	−10.808	−4.221	28	.000
FAI	−18.586	−26.610	−10.562	−4.745	28	.000
AOD	−12.448	−21.911	−2.986	−2.695	28	.012

Table A.13 Results of the t-test for the SD test of file 11

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−8.964	−14.136	−3.793	−3.556	27	.001
NOV	−10.500	−15.489	−5.511	−4.319	27	.000
PQ	−7.429	−12.236	−2.621	−3.170	27	.004
FAI	−14.071	−21.505	−6.637	−3.884	27	.001
AOD	−1.643	−8.318	5.032	−0.505	27	.618

Table A.14 Results of the t-test for the SD test of file 12

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−4.517	−10.263	1.229	−1.610	28	.119
NOV	0.483	−8.221	9.187	0.114	28	.910
PQ	12.517	2.890	22.145	2.663	28	.013
FAI	0.276	−7.331	7.883	0.074	28	.941
AOD	−11.621	−20.471	−2.770	−2.690	28	.012

Table A.15 Results of the t-test for the SD test of file 13

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−0.345	−5.225	4.535	−0.145	28	.886
NOV	6.379	0.240	12.518	2.129	28	.042
PQ	5.793	−4.898	16.484	1.110	28	.276
FAI	3.621	−3.368	10.610	1.061	28	.298
AOD	0.483	−8.241	9.206	0.113	28	.911

Table A.16 Results of the t-test for the SD test of file 14

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−22.207	−28.755	−15.658	−6.946	28	.000
NOV	−22.379	−30.066	−14.693	−5.964	28	.000
PQ	−11.690	−22.515	−0.865	−2.212	28	.035
FAI	−27.552	−36.806	−18.297	−6.099	28	.000
AOD	−32.897	−40.802	−24.992	−8.524	28	.000

Table A.17 Results of the t-test for the SD test of file 15

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−12.241	−17.640	−6.843	−4.645	28	.000
NOV	−21.483	−28.278	−14.688	−6.476	28	.000
PQ	−32.414	−40.014	−24.813	−8.736	28	.000
FAI	−10.862	−16.185	−5.539	−4.180	28	.000
AOD	−17.241	−25.874	−8.609	−4.091	28	.000

Table A.18 Results of the t-test for the SD test of file 16

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−14.276	−21.380	−7.171	−4.116	28	.000
NOV	−12.931	−19.448	−6.414	−4.065	28	.000
PQ	−26.655	−36.103	−17.207	−5.779	28	.000
FAI	−21.552	−31.091	−12.012	−4.628	28	.000
AOD	−9.552	−18.703	−0.400	−2.138	28	.041

Table A.19 Results of the t-test for the SD test of file 17

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−10.000	−14.370	−5.630	−4.695	27	.000
NOV	−10.893	−15.823	−5.963	−4.533	27	.000
PQ	−17.714	−26.296	−9.132	−4.235	27	.000
FAI	−8.571	−15.699	−1.444	−2.467	27	.020
AOD	−9.071	−16.066	−2.077	−2.661	27	.013

Table A.20 Results of the t-test for the SD test of file 18

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−1.724	−6.794	3.346	−0.697	28	.492
NOV	−1.448	−8.308	5.412	−0.432	28	.669
PQ	−5.138	−12.950	2.674	−1.347	28	.189
FAI	−2.931	−10.183	4.321	−0.828	28	.415
AOD	−3.483	−12.085	5.120	−0.829	28	.414

Table A.21 Results of the t-test for the SD test of file 19

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	2.414	−1.765	6.592	1.183	28	.247
NOV	6.310	−1.077	13.697	1.750	28	.091
PQ	−3.069	−11.637	5.499	−0.734	28	.469
FAI	−5.897	−13.444	1.651	−1.600	28	.121
AOD	1.517	−5.700	8.734	0.431	28	.670

Table A.22 Results of the t-test for the SD test of file 20

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−10.690	−16.180	−5.199	−3.988	28	.000
NOV	−13.069	−20.381	−5.757	−3.661	28	.001
PQ	−22.655	−32.164	−13.147	−4.881	28	.000
FAI	−4.655	−10.452	1.142	−1.645	28	.111
AOD	−4.897	−11.854	2.061	−1.442	28	.160

Table A.23 Results of the t-test for the SD test of file 21

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−1.759	−8.047	4.530	−0.573	28	.571
NOV	−1.655	−9.464	6.154	−0.434	28	.667
PQ	−3.448	−13.681	6.784	−0.690	28	.496
FAI	0.069	−8.475	8.613	0.017	28	.987
AOD	2.586	−5.184	10.357	0.682	28	.501

Table A.24 Results of the t-test for the SD test of file 22

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−32.655	−40.502	−24.809	−8.525	28	.000
NOV	−26.862	−36.729	−16.995	−5.576	28	.000
PQ	−31.966	−42.076	−21.855	−6.476	28	.000
FAI	−28.448	−40.422	−16.474	−4.867	28	.000
AOD	−28.138	−36.938	−19.338	−6.550	28	.000

Table A.25 Results of the t-test for the SD test of file 23

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−6.621	−12.893	−0.349	−2.162	28	.039
NOV	−6.379	−12.413	−0.346	−2.166	28	.039
PQ	−10.207	−17.937	−2.477	−2.705	28	.011
FAI	−11.000	−19.552	−2.448	−2.635	28	.014
AOD	−2.379	−9.085	4.327	−0.727	28	.473

Table A.26 Results of the t-test for the SD test of file 24

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−7.250	−13.717	−0.783	−2.300	27	.029
NOV	−4.607	−10.390	1.175	−1.635	27	.114
PQ	−7.179	−15.768	1.411	−1.715	27	.098
FAI	−11.893	−20.025	−3.760	−3.001	27	.006
AOD	−12.464	−18.271	−6.657	−4.404	27	.000

Table A.27 Results of the t-test for the SD test of file 25

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−9.897	−14.712	−5.081	−4.210	28	.000
NOV	−1.448	−6.178	3.282	−0.627	28	.536
PQ	8.793	1.128	16.458	2.350	28	.026
FAI	−14.034	−24.787	−3.282	−2.674	28	.012
AOD	−14.862	−21.286	−8.438	−4.739	28	.000

Table A.28 Results of the t-test for the SD test of file 26

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	3.655	−1.668	8.979	1.406	28	.171
NOV	3.862	−3.278	11.002	1.108	28	.277
PQ	−11.552	−19.756	−3.348	−2.884	28	.007
FAI	5.414	0.527	10.300	2.269	28	.031
AOD	13.966	6.950	20.981	4.078	28	.000

Table A.29 Results of the t-test for the SD test of file 27

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−6.759	−10.885	−2.632	−3.355	28	.002
NOV	−5.276	−10.399	−0.152	−2.109	28	.044
PQ	−10.345	−17.060	−3.630	−3.156	28	.004
FAI	−7.138	−12.098	−2.178	−2.948	28	.006
AOD	−9.310	−17.618	−1.003	−2.296	28	.029

Table A.30 Results of the t-test for the SD test of file 28

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−8.966	−14.063	−3.868	−3.603	28	.001
NOV	−7.276	−15.514	0.962	−1.809	28	.081
PQ	−3.517	−9.935	2.901	−1.123	28	.271
FAI	−9.931	−16.716	−3.146	−2.998	28	.006
AOD	−7.897	−14.899	−0.894	−2.310	28	.028

Table A.31 Results of the t-test for the SD test of file 29

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−8.517	−14.272	−2.762	−3.032	28	.005
NOV	−7.828	−13.811	−1.844	−2.680	28	.012
PQ	−8.828	−14.724	−2.931	−3.067	28	.005
FAI	−9.655	−16.610	−2.700	−2.844	28	.008
AOD	−5.517	−11.995	0.961	−1.745	28	.092

Table A.32 Results of the t-test for the SD test of file 30

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−20.414	−26.020	−14.807	−7.458	28	.000
NOV	−14.793	−21.718	−7.869	−4.376	28	.000
PQ	−21.379	−30.254	−12.504	−4.934	28	.000
FAI	−17.931	−25.820	−10.042	−4.656	28	.000
AOD	−18.621	−26.291	−10.951	−4.973	28	.000

Table A.33 Results of the t-test for the SD test of file 31

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−13.214	−19.581	−6.848	−4.259	27	.000
NOV	−2.571	−9.928	4.785	−0.717	27	.479
PQ	3.286	−7.631	14.202	0.618	27	.542
FAI	−12.214	−21.037	−3.391	−2.841	27	.008
AOD	−14.750	−21.432	−8.068	−4.529	27	.000

Table A.34 Results of the t-test for the SD test of file 32

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−4.310	−10.303	1.683	−1.473	28	.152
NOV	−7.621	−13.853	−1.388	−2.505	28	.018
PQ	−8.897	−18.738	0.945	−1.852	28	.075
FAI	−3.103	−7.279	1.073	−1.522	28	.139
AOD	−7.483	−14.256	−0.710	−2.263	28	.032

Table A.35 Results of the t-test for the SD test of file 33

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−1.207	−7.098	4.684	−0.420	28	.678
NOV	3.621	−3.895	11.136	0.987	28	.332
PQ	0.414	−7.315	8.142	0.110	28	.913
FAI	−4.034	−12.612	4.543	−0.964	28	.344
AOD	−4.207	−13.580	5.166	−0.919	28	.366

Table A.36 Results of the t-test for the SD test of file 34

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−22.448	−27.116	−17.780	−9.850	28	.000
NOV	−17.138	−24.258	−10.018	−4.931	28	.000
PQ	−9.069	−15.815	−2.323	−2.754	28	.010
FAI	−21.690	−28.933	−14.446	−6.133	28	.000
AOD	−33.724	−41.623	−25.825	−8.746	28	.000

Table A.37 Results of the t-test for the SD test of file 35

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	−8.276	−12.650	−3.902	−3.876	28	.001
NOV	−11.345	−17.894	−4.796	−3.548	28	.001
PQ	−4.552	−14.091	4.988	−0.977	28	.337
FAI	−8.034	−13.059	−3.010	−3.276	28	.003
AOD	−12.207	−19.723	−4.691	−3.327	28	.002

Table A.38 Results of the t-test for the SD test of file 36

	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Interval of the				
		Lower	Upper			
OIMP	0.379	−5.604	6.362	0.130	28	.898
NOV	1.897	−5.020	8.813	0.562	28	.579
PQ	−2.276	−11.367	6.815	−0.513	28	.612
FAI	−2.276	−10.474	5.922	−0.569	28	.574
AOD	−0.759	−8.497	6.979	−0.201	28	.842