Thesis for the degree of Doctor of Philosophy

Perceptually motivated speech recognition and mispronunciation detection

Christos Koniaris

KTH Computer Science and Communication

Centre for Speech Technology
Department of Speech, Music and Hearing
School of Computer Science and Communication
KTH - Royal Institute of Technology

Stockholm 2012

Cover: Swift, Jonathan. Gulliver's Travels. Chicago: Homewood Publishing Co., 1909, 81. "Whispering in giant's ear". Retrieved August 20, 2012 from http://etc.usf.edu/clipart/11900/11918/whispering_11918.htm

TRITA-CSC-A 2012:10
ISSN-1653-5723
ISRN-KTH/CSC/A-12/10-SE
ISBN 978-91-7501-468-5

KTH School of Computer Science
and Communication
SE-100 44 Stockholm
SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i talteknologi och kommunikation fredagen den 5 oktober 2012 klockan 10.00 i sal A2, Kungl Tekniska högskolan, Östermalmsgatan 26, Stockholm.

Tryck: E-Print AB

Abstract

This doctoral thesis is the result of a research effort performed in two fields of speech technology, i.e., *speech recognition* and *mispronunciation detection*. Although the two areas are clearly distinguishable, the proposed approaches share a common hypothesis based on psychoacoustic processing of speech signals. The conjecture implies that the human auditory periphery provides a relatively good separation of different sound classes. Hence, it is possible to use recent findings from psychoacoustic perception together with mathematical and computational tools to model the auditory sensitivities to small speech signal changes.

The performance of an automatic speech recognition system strongly depends on the representation used for the front-end. If the extracted features do not include all relevant information, the performance of the classification stage is inherently suboptimal. The work described in Papers A, B and C is motivated by the fact that humans perform better at speech recognition than machines, particularly for noisy environments. The goal is to make use of knowledge of human perception in the selection and optimization of speech features for speech recognition. These papers show that maximizing the similarity of the Euclidean geometry of the features to the geometry of the perceptual domain is a powerful tool to select or optimize features. Experiments with a practical speech recognizer confirm the validity of the principle. It is also shown an approach to improve mel frequency cepstrum coefficients (MFCCs) through offline optimization. The method has three advantages: i) it is computationally inexpensive, ii) it does not use the auditory model directly, thus avoiding its computational cost, and iii) importantly, it provides better recognition performance than traditional MFCCs for both clean and noisy conditions.

The second task concerns automatic pronunciation error detection. The research, described in Papers D, E and F, is motivated by the observation that almost all native speakers perceive, relatively easily, the acoustic characteristics of their own language when it is produced by speakers of the language. Small variations within a phoneme category, sometimes different for various phonemes, do not change significantly the perception of the language's own sounds. Several methods are introduced based on similarity measures of the Euclidean space spanned by the acoustic representations of the speech signal and the Euclidean space spanned by an auditory model output, to identify the problematic phonemes for a given speaker. The methods are tested for groups of speakers from different languages and evaluated according to a theoretical linguistic study showing that they can capture many of the problematic phonemes that speakers from each language mispronounce. Finally, a listening test on the same dataset verifies the validity of these methods.

Keywords: feature extraction, feature selection, auditory models, MFCCs, speech recognition, distortion measures, perturbation analysis, psychoacoustics, human perception, sensitivity matrix, pronunciation error detection, phoneme, second language, perceptual assessment.

Sammanfattning

Denna doktorsavhandling beskriver forskning inom två talteknologiområden: *taligenkänning* och *diagnostik detektion av uttalsfel.* Även om de två användningsområdena skiljer sig, har de föreslagna metoderna en gemensam hypotes som baseras på psykoakustisk talsignalbehandling. Hypotesen anger att människans perifera hörselsystem ger en relativt god separation av olika ljudklasser. Därmed kan den auditiva känsligheten för små talsignalförändringar modelleras med teorier om psykoakustisk perception och matematiska beräkningar.

Ett automatiskt taligenkänningssystems prestanda beror till stor del på den representation som används för talsignalen. Om representationen inte innehåller all relevant information är klassificeringsstadiet i sig suboptimal. Arbetet som beskrivs i Artiklarna A, B och C, motiveras av det faktum att människors taligenkänning är bättre än automatisk, särskilt i bullriga miljöer. Målet är att utnyttja kunskapen om mänsklig perception i valet och optimeringen av talsignalrepresentationer för taligenkänning. Dessa artiklar visar att maximering av likheten mellan den Euklidiska geometrin för talsignalens särdrag och geometrin av den perceptuella domänen är ett kraftfullt verktyg för att välja eller optimera särdragen. Experiment med ett taligenkänningssystem bekräftar principens giltighet. Dessutom redogörs för ett försök att offline optimera talsignalrepresentationen (MFCC-koefficienter). Metoden har tre fördelar: i) den kräver få beräkningssteg, ii) den använder den auditiva modellen på ett indirekt sätt och reducerar därmed antalet beräkningssteg, och, viktigast, iii) den ger bättre taligenkänningsresultat än med traditionella MFCC-koefficienter i både tysta och bullriga miljöer.

Avhandlingens andra del behandlar automatisk detektion av uttalsfel. Arbetet som beskrivs i Artiklarna D, E och F, motiveras av observationen att infödda lysnnare relativt enkelt tolkar akustiska särdag i det egna språket. Variationer mellan hur olika inhemska talare producerar ett fonem förändrar inte tolkningen på något betydande sätt. Flera metoder presenteras i avhandlingen för att identifiera de mest problematiska fonemen för en specifik talare. Metoderna baseras på att mäta likheten mellan den Euklidiska arean som spänns upp av talsgnalens akustiska representation och motsvarande area för den auditiva modellens utsignal. Metoderna testas för talare från olika språkgrupper och utvärderas i relation till en lingvistik studie av vilka uttalsfel talare från dessa språkgrupper har. Jämförelsen visar att metoderna kan identifiera många av de problematiska fonemen för talare från varje språkgrupp. Ett lyssningstest på samma data genomförs också för att verifiera metoderna.

Nyckelord: formulering och val av talsignalrepresentation, auditiv modell, MFCCs, taligenkänning, distorsionsmått, störningsanalys, psykoakustik, mänsklig perception, sensitivitetsmatris, automatisk bedömning av uttalsfel, fonem, andraspråksinlärning.

List of Papers

The thesis is based on the following papers:

[A] C. Koniaris, M. Kuropatwinski and W. B. Kleijn, “Auditory-Model Based Robust Feature Selection for Speech Recognition”, *Journal of Acoustical Society of America*, vol. 127, no. 2, pp. EL73–EL79, Feb. 2010

[B] S. Chatterjee, C. Koniaris and W. B. Kleijn, “Auditory Model Based Optimization of MFCCs Improves Automatic Speech Recognition Performance”, in *Proceedings of ISCA Interspeech*, pp. 2987–2990, Brighton, UK, Sep. 2009.

[C] C. Koniaris, S. Chatterjee and W. B. Kleijn, “Selecting Static and Dynamic Features Using an Advanced Auditory Model for Speech Recognition”, in *Proceedings of IEEE Int. Conf. on Acoust., Speech, Sig. Proc.*, pp. 4342–4345, Dallas, TX, USA, Mar. 2010.

[D] C. Koniaris and O. Engwall, “Perceptual Differentiation Modeling Explains Phoneme Mispronunciation by Non-Native Speakers”, in *Proceedings of IEEE Int. Conf. on Acoust., Speech, Sig. Proc.*, pp. 5704–5707, Prague, Czech Republic, May 2011.

[E] C. Koniaris, G. Salvi and O. Engwall, “On Mispronunciation Analysis of Individual Foreign Speakers Using Auditory Periphery Models”, *Speech Communication*, submitted.

[F] C. Koniaris, O. Engwall and G. Salvi, “Auditory and Dynamic Modeling Paradigms to Detect L2 Mispronunciations”, in *Proceedings of ISCA Interspeech*, Portland, OR, USA, Sep. 2012.

In addition to papers A-F, the following paper have also been produced in part by the author of the thesis:

[1] G. Tsontzos, V. Diakoloukas, C. Koniaris and V. Digalakis, "Estimation of General Identifiable Linear Dynamic Models with an Application in Speech Recognition", in *Proceedings of IEEE Int. Conf. on Acoust., Speech, Sig. Proc.*, vol. 4, pp. 453–456, Honolulu, HI, USA, Apr. 2007.

[2] T. Altosaar, L. ten Bosch, G. Aimetti, C. Koniaris, K. Demuynck and H. van den Heuvel, "A Speech Corpus for Modeling Language Acquisition: CAREGIVER", in *Proceedings of Inter. Conf. on Languag. Resourc. and Eval.*, pp. 1062–1068, Valletta, Malta, May 2010.

[3] C. Koniaris and O. Engwall, "Phoneme Level Non-Native Pronunciation Analysis by an Auditory Model-based Native Assessment Scheme", in *Proceedings of ISCA Interspeech*, pp. 1157–1160, Florence, Italy, Aug. 2011.

[4] C. Koniaris, O. Engwall and G. Salvi, "On the Benefit of Using Auditory Modeling for Diagnostic Evaluation of Pronunciations", in *Proceedings of Inter. Symp. Autom. Detec. Errors Pronunc. Train.*, pp. 59–64, Stockholm, Sweden, June 2012.

Acknowledgements

Approaching the “final destination” of this long and notoriously difficult, yet fascinating “journey”, I would like to take advantage of the unique opportunity and express my gratitude to people that helped me during my doctoral studies. I would like to thank my supervisor, Olov Engwall, and express my appreciation for his support from the very beginning and at the same time the freedom that he gave me in finding my own way in research. I am grateful for the many hours he patiently put into reading my work and for his constant reassurance, always being ready to offer his guidance when needed. I am also thankful to Giampiero Salvi for his valuable help having perfectly served as a co-supervisor. I express my gratitude for his guidance, patient and unceasing motivation during my studies.

In addition, I would like to thank Bastiaan Kleijn, who was my supervisor during the first half of my doctoral studies, for his many suggestions and valuable help. His experience and scientific background, has proven advantageous in my research. I particularly thank him for helping me improving my scientific maturity and for showing a path towards professionalism.

I am heartily thankful to Saikat Chatterjee for his invaluable help and consultation for every major or minor research problem that came along all these years. Unquestionably, I have been blessed to meet him during my early years as a researcher but nothing can be compared to the fact that he has become one of my best friends almost from the very beginning.

At this point, I feel I need to express my sincere gratitude to Stefan Östlund for his understanding, support and reassurance during my doctoral studies at KTH. I hope this dissertation to be a small recompense for the faith he showed to me. I also owe a debt of gratitude to Rikard Lingström and Lars Abrahamsson for having spent many hours to lighting an optimal pathway towards achieving my goal which is about to be reached, namely the PhD degree.

My uncle, George Zouridakis, was always there to help me. I am very grateful to him because he used his valuable experience to advise me with a calm and rational manner so as to improve my professional behavior and to take the correct decisions for my academic future.

A sincere ‘thank you’ to Arne Leijon and Jonas Beskow for giving me the opportunity to serve as a teaching assistant in their courses and the appreciation they have always been shown to my work. I also wish to thank Mats Blomberg who was kind enough to read and evaluate my thesis providing with useful criticism and comments.

I consider myself lucky to have met many interesting people during my doctoral studies. Therefore, as a token of my appreciation I would like to take the opportunity to give my heartfelt thanks to those people who have facilitated my learning and given emotional support along the way. Beginning from the senior members, I thank Björn Granstörm, Anders Askenfelt, David House and Rolf Carlsson for their rationality and for sharing with us their experience and valuable knowledge. I wish to thank Daniel Neiberg with which I have shared the same office, for his understanding and harmonic co-operation and Minyue Li (my previous office-mate), a gentle and friendly person who was always willing to help me when needed. I thank Jan Plasberg for sharing with me his experience in auditory modeling through the early months of chaos and confusion. I also thank Samer Al Moubayed, Gopal Ananthakrishnan, Jens Edlund, Anders Ekman, Gustav Henter, Janusz Klejsa, Ermin Kozica, Petko Petkov, Andreas Selamtzis, Svante Stadler, Sofia Strömbergsson, Dora Söderberg, Preben Wik, Guoqiang Zhang, David Zhao, and many other current and ex-colleagues for the interesting discussions and their friendship and understanding. Thank you Anita Kruckenberg, one of the most lovely and lively persons I have ever met, for being so kind and positive and for the many exciting conversations we have had.

Of course, I do not forget to thank my parents for providing me with all the necessary means which have helped me since my early years to become a mature and self-contained man. I hope that the joy of my graduation, would make them happy and proud.

Finally, I would like to thank Magda for her love and uninterrupted support, and for giving me the motivation to continue my work. She has been there for me during all the happy but also the bitter and stressful moments of my life in the last ten years, always available for cheering me up and encouraging me when I felt down. I would not have been able to achieve anything I did without her, therefore this thesis is dedicated to Magda.

Christos Koniaris
Stockholm, Sweden
September 2012

Contents

List of Figures

List of Tables

Acronyms

AIM	Auditory Image Model
AMFS	Auditory Model-based Feature Selection
AN	Auditory Nerve
ANNs	Artificial Neural Networks
ASR	Automatic Speech Recognition
BM	Basilar Membrane
CALL	Computer-Assisted Language Learning
CAPT	Computer-Assisted Pronunciation Training
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
EE	Excited-Excited cells
EER	Equal Error Rate
EM	Expectation-Maximization algorithm
ERB	Equivalent Rectangular Bandwidth
GOP	Goodness Of Pronunciation
HAMM	Hidden Articulatory Markov Model
HDMs	Hidden Dynamic Models
HLDA	Heteroscedastic Linear Discriminant Analysis
HMMs	Hidden Markov Models
HTK	Hidden Markov model Toolkit
HTS	HMM speech synthesizer
IHC	Inner Hair Cells
i.i.d.	independent and identically distributed
IIDs	Interaural Intensity Differences
IPA	International Phonetic Alphabet

ITD	Interaural Time Difference
KLT	Karhunen-Loève Transform
L1	First Language
L2	Second Language
LDA	Linear Discriminant Analysis
LDM	Linear Dynamic Model
LPC	Linear Predictive Coding
LSO	Lateral Superior Olive
MAP	Maximum A-posteriori Probability
MED	Maximum Entropy Discrimination
MFCCs	Mel Frequency Cepstrum Coefficients
MLP	Multi-Layer Perceptron
MLSA	Mel Log Spectrum Approximation
MMFCCs	Modified Mel Frequency Cepstrum Coefficients
mRMR	minimal-Redundancy-Maximal-Relevance
MSD-HMMs	Multi-Space probability Distribution HMMs
MSO	Medial Superior Olivary
nPAD	native-Perceptual Assessment Degree
OHC	Outer Hair Cells
OM	Outer and Middle ear
PCA	Principle Component Analysis
PDF	Probability Density Function
PDP	Parallel Distributed Processing
PED	Pronunciation Error Detection
POD	Proper Orthogonal Decomposition
SMs	Segment-based Models
SNR	Signal-to-Noise Ratio
SOC	Superior Olivary Complex
SPTK	Speech Signal Processing Toolkit
SVM	Support Vector Machine
VQ	Vector Quantization

"I had thought to find in knowledge
Light to guide me on my way;
Yet I still must walk in darkness
All that's known must soon decay.
Ignorance, I turn to thee!
Knowledge is but vanity."

August Strindberg, 1849-1912
To Damascus, III

"To have come this far is no small achievement:
what you have done is a glorious thing.
Even this first step
is a long way above the ordinary world.
To stand on this step
you must be in your own right
a member of the city of ideas."

Konstantinos P. Kavafis, 1863-1933
'The First Step', *Collected Poems*

Part I

Introduction

Introduction

The ways that humans can interact with computers have been developed since the early days of computer engineering. Nowadays, it it not unusual for this interaction to be done by *speech*. Many computer programs therefore include state-of-the-art developments of speech technology to perform various tasks. For these systems to be of practical use, i.e., to perform in a human-like manner, a thorough understanding of human speech perception is necessary. Additionally, a compact and relevant representation of speech input that would eliminate the influence of insignificant components such as background noise, is an important factor to enhance the system's performance. These two topics are dealt with in this Introduction.

Since auditory perception plays a principal role in the work presented in this thesis, the first part of the Introduction is devoted to discussing related terms and notions as a necessary background. Hence, in Sec. 1, the human auditory periphery system is introduced as well as some of the computational auditory models that have been developed to simulate different functions of hearing. Next, the underlying idea, from a high-level point of view, of the methods proposed in this thesis is discussed. Sec. 2, which deals with the task of speech recognition, introduces the front-end and acoustic modeling and then continues with a short presentation of several feature dimensionality reduction methods. In most cases, existing feature-selection and dimension-reduction methods require classified data. For speech recognition this suggests that dimension-reduction methods are sensitive to differences in training and testing conditions. To overcome this problem, but also to increase understanding about the representation of speech in the human periphery, a fundamentally different (auditory motivated) feature selection technique is proposed, a brief description of which is given in the end of this section. In Sec. 3, the task of automatic pronunciation error detection is discussed. At first, a quick overview of different approaches is presented, including three paradigms from real pronunciation training or language learning systems. Usually, this task is treated as a classification problem and several statistical methods are used. However, these statistical approaches may not include all the information relayed by the human auditory periphery which is vital for sound discrimination and classification.

Therefore, the methods may detect or accept deviations in the pronunciation that are not relevant to auditory processing. Another issue arises from the fact that sometimes the available systems are developed for specific language pairs, which leaves little flexibility in cases where the language backgrounds are not pre-determined. This thesis proposes a possible solution to the above restrictions by presenting a new, perceptually-motivated scheme to assess the non-native pronunciation. The basic idea of the new technique is therefore introduced in the end of Sec. 3. Finally, Sec. 4 summarizes the thesis contributions and presents a short description of the six papers of Part II, and Sec. 5 provides conclusions and discusses potential future extensions.

1 Speech perception

Speech has been, and will continue to be, the dominant manner of human social communication and information exchange. Within the broader area of speech communication, *speech perception*, at the level of the sound signal, deals with the process by which the sounds of a language are heard and interpreted into meaningful phonetic information that can later be used for upper level cognitive processing. In this work, the perceptual cues of the speech signals are studied so as to improve the performance of speech recognition and foreign-language pronunciation error detection systems. In doing so, psychoacoustic models of the human auditory periphery are employed. In the following paragraph, the human auditory system is presented to provide a background knowledge necessary to be able to understand the progress in auditory modeling research. Next, an overview of some of the functional auditory models is given and then an outline of the underlying idea and the primary assumption of this thesis, is presented.

1.1 Human hearing system

The human ear consists of several parts (Moore, 2003; Zwicker and Fastl, 1999; Huang et al., 2001): the *outer ear*, the *middle ear*, and the *inner ear*. The way these elements operate is not totally understood, but a good level of understanding has been reached to a considerable extent through previous research efforts. In the next, an insight of the human ear is provided but for more details and an extended analysis of the function of the human auditory system the reader is referred to (Moore, 2003; Zwicker and Fastl, 1999).

The first part of the human auditory system, as shown in Fig. 1, is the outer ear consisting of the pinna, the auditory or ear canal and the tympanic membrane or eardrum. The pinna is the only totally visible part of the system, and consists of what is often simply called the “ear”. This organ is commissioned to collect different sounds which will then travel via the auditory canal to the middle and inner ear. The pinna is also a 'natural radar' that can identify the origin of a sound, i.e., it contributes to the so called sound localization process. The auditory canal is a channel of about 26 mm in length and 7 mm in diameter, filled with air that leads to the tympanic membrane. The tympanic membrane is approximately $8-10$ mm in diameter and is formed of three layers of skin. The sound is filtered by the canal and then hits the eardrum that starts to vibrate. When this happens, the sound vibrations are passed into the middle ear.

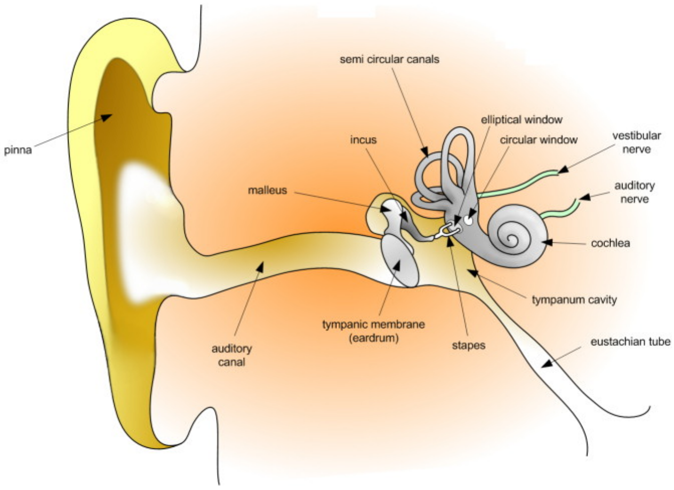


Figure 1: The anatomy of human ear.

The middle ear space, also known as the tympanic cavity lodges the ossicles, a group of three tiny bones that serve as link between the outer and the inner ear. The ossicles, called *malleus*, *incus*, and *stapes*, are the smallest bones of the body and their duty is to pass the vibrations of the tympanic membrane through the middle ear to the inner ear. The malleus, which is partially implanted in the tympanic membrane, is responsible for transferring the vibrations to the other ossicles. Inside the middle ear, there are also two very small muscles, the stapedius and the tensor tympani. Their job is to suspend and retain the ossicles within the middle ear. They also control the acoustic reflex phenomenon, namely the contraction in response to loud sound which in turn tightens the chain of ossicles to protect the sensory part of the ear from damage by loud sounds. The middle ear cavity is also connected to the back of the throat by a passage called the eustachian tube. The eustachian tube is normally closed, but opens when swallow occurs, equalizing the middle ear pressure with the external air pressure.

As a result, the tympanic membrane has equal pressure on either side to prevent hearing loss.

The inner ear consists of two parts, the cochlea and the vestibule. The cochlea is a small spiral (which looks like the shell of a snail) filled with fluid which plays a major role in hearing. Sound is transmitted as 'waves' in this fluid by vibration of the last ossicle, stapes in the 'oval window'. Inside the cochlea is an important structure known as the basilar membrane (BM) which is vibrated in various places by incoming sounds depending on their frequency range. Higher-frequency sounds vibrate the membrane near its base while lower-frequency sounds vibrate its upper part. According to place theory, humans recognize pitch based on the area of the basilar membrane that is stimulated (Moore, 2003). On the BM rests the receptor organ of hearing - the organ of Corti, which supports rows of special cells known as *hair cells*. The process of transduction (transforming mechanical vibrations into electrical signals) is performed by them. There are approximately 3 500 inner hair cells (IHC) and 11 000 outer hair cells (OHC). These hair cells connect to approximately 24 000 nerve fibers. The electrical signals produced by the hair cells travel through the auditory nerve (AN) to the brain. The AN is the nerve which arises from within the cochlea and extends to the brainstem, and carries the sound information to the cochlea nucleus, i.e., the first site of the central auditory system in which the sensory information is processed by the neural system. A sound is then considered to be perceived by the time these electrical signals reach the 'auditory cortex' of the brain where a cognitive processing is performed. Finally, the vestibule is the central part of the osseous labyrinth, and is situated in the middle of the tympanic cavity behind the cochlea and in front of the semicircular canals. It forms part of the vestibular system which contributes to the balance of the body and to the sense of spatial orientation.

The knowledge about the way in which the brain processes the extracted patterns is rather vague, but many studies have shown how individuals perceive tones and noise bands (Moore, 2003; Zwicker and Fastl, 1999). Based on that knowledge, many auditory models that simulate the functionality of the human ear, have been proposed (Moore, 2003; Zwicker and Fastl, 1999; Dau et al., 1996a; Rix et al., 1999). In the next section, a short overview of various auditory models is given.

1.2 Auditory models

Computational models of the human auditory system are widely used and a series of approaches have been proposed depending on different properties and various experimental findings. Dau (2009) presents an overview of the auditory models that are generally divided into biophysical, physiological, statistical and perceptual, depending on which properties these models focus on. It is worth noticing that most of the models are not "complete", which means that each one may simulate a specific observation or function and, consequently, can be used for certain tasks. This section concentrates on computational models of the auditory perception.

The auditory nerve has been at the epicenter of several applications. Delgutte (1990) addressed the phenomenon of physiological masking so as to show, by comparing the masking thresholds of the AN fibers, that suppression masking rises significantly for signals (known as probes) with frequency higher than the masker, while the masking is excitatory if the probe frequency is lower than the masker's. Carney (1993) presented a computational model to account for the average rate and temporal response properties of the AN fibers in cats. A more recent development was introduced in (Heinz et al., 2001b) in which non-linear properties of the cochlear amplifier were studied and a computational AN model was developed to be used in psychophysical experiments with human listeners, both normal and impaired. A stochastic method based on signal detection and computational AN models, was introduced in (Heinz et al., 2001a) to evaluate psychophysical performance limits for the task of auditory discrimination of tone frequency and level. Colburn et al. (2003) presented an analytical approach to quantify the information in AN fiber responses. The model included temporal responses as well as the non-linear phase impact of the cochlear amplifier, and was intended for the task of level discrimination.

Some approaches include information from the central auditory system. Such models use, for example, recent findings related to the superior olivary complex (SOC) which is a collection of brainstem nuclei that contribute variously in hearing. Within the SOC, a specialized nucleus called the medial superior olive (MSO) is considered to support the localization of a sound by detecting its azimuth, i.e., the angle where the sound source is located. It is also responsible for measuring the time difference of sounds' arrival between the ears which is called the interaural time difference (ITD). The interaural intensity differences (IIDs) are measured by another nucleus inside the SOC, the lateral superior olive (LSO). IIDs are also useful for the determination of the azimuth, particularly for higher frequency sounds. The so-called, binaural perceptual models are based on the concept of interaural time delay and the idea that an efficient way to estimate this delay is by introducing a coincidence network in the MSO (Jeffress, 1948). Such a network can be found in the work presented in (Colburn et al., 1990), which describes a simple way to model empirical findings from excited-excited (EE) cells (cells that are excited by signals from both ears) in the MSO. Other approaches, e.g., the models of sound localization (Blauert, 1997), focus on the interaural differences such as ITD and IIDs. A binaural signal detection model (Breebaart et al., 2001a,b,c) consisting of three parts, a peripheral preprocessing in both monaural channels, a binaural processor which produces the internal representations and a central processor to detect the sound signal was introduced as an extension of the approach presented in (Dau et al., 1996a).

Models of auditory sensations such as loudness, pitch or duration, have also received the attention of many researchers in the field (Zwicker and Fastl, 1999). Glasberg and Moore (2002) presented an example of such a model which first calculated the short-term perceived loudness (loudness that is perceived at any instant), and then computed the overall loudness impression (long-term loudness) by an averaging mechanism applied to the short-term loudness. In (Lyon and Shamma, 1996; de Cheveigné, 2005), a description of several approaches can be found to computationally model the pitch and timbre. In (Shamma and Klein, 2000), an auditory model was presented that consisted of two stages, namely the cochlear filtering and the coincidence detection, while in (Meddis and O'Mard, 1997), a pitch perception model was shown. Common for both approaches is the idea of combing spectro-temporal cues to better calculate the representation of the pitch. Further reading on this family of pitch models can be found in (Moore, 2003). In (Oxenham et al., 2004), an effort to separate temporal and place information was made by allowing temporal information of low frequency tones to be displayed in locations in the cochlea tuned for high frequencies. The experiments have shown the importance of the tonotopic place in the pitch calculation, a phenomenon that was also studied in (Shamma, 2004).

Furthermore, studies based on psychophysical experiments have revealed the importance of compression and suppression in non-simultaneous masking. In (Oxenham and Moore, 1994), an effort was made to model the additivity of non-simultaneous masking by including a compressive nonlinearity within the temporal-window model. The study was further expanded for hearing impaired listeners in (Oxenham and Moore, 1997) and it was found that cochlea damages lead to a reduced non-linearity of the basilar membrane. The latter's non-linearity was also the subject of experiments described in (Plack and Oxenham, 1998). The temporal envelope in normal-hearing individuals was investigated in (Nelson and Swain, 1996) by measuring masking at the peaks and valleys of a tone signal. In (Dau et al., 1996a,b) and (Buchholz and Mourjopoulos, 2004a,b), computational models of forward masking were presented that also included non-linear adaptation modules and signal-dependent compression of the input signal's dynamics. Both models accounted for simultaneous and non-simultaneous masking phenomena. Finally, in (Oxenham, 2001), an attempt to distinguish between neural adaptation and temporal integration as possible explanations of forward masking was performed, but the findings show similar behavior of both approaches.

A general model was discussed in (Patterson et al., 1992) that produced 'auditory images' of the sounds. The model included assumptions that are beyond the pure periphery, e.g., it included phenomena such as the phase alignment and temporal integration that occur before the formation of the initial images of the sounds. These auditory images indicated which effects of complex sounds may be explained peripherally and clarified which of them require central processing. Based on this work, a software platform of the auditory image model (AIM) was later presented in (Patterson et al., 1995). A model that also includes assumptions about the processing in more central auditory stages was presented in (Dau et al., 1997a,b), and later expanded in (Jepsen et al., 2008), in which a modulation filterbank was introduced that reflected the sensitivities to fluctuating sounds and accounted for amplitude-modulation detection and masking data, following the adaptation stage in each peripheral auditory filter.

Until now, a series of different auditory models, which simulate experimental findings or behavioral characteristics of the periphery, has been described. Next, a short presentation of the two models that were used in the papers included in Part II is given. At this point it is necessary to note that the models of human audition were used implicitly throughout this doctoral research, for the purpose of identifying the periphery's response to small signal changes and its sensitivities to distortion. Consequently, in the course of this study no development of new schemes or expansion of current ones was performed. Contrarily, the two auditory models were used as means of transformation of the speech distortion signals in the auditory perceptual field.

In (Gardner and Rao, 1995), the concept of sensitivity matrix was introduced to approximate a given distortion measure used in the problem of quantization of the linear predictive coding (LPC) parameters in speech coding systems. Later, this work was extended and generalized in (Li et al., 1999) and in (Linder et al., 1999). In (Plasberg and Kleijn, 2007), a method for deriving the sensitivity matrix for distortion measures that are relevant for audio signals was developed based on spectro-temporal auditory models. The word "sensitivity" refers to the fact that each element of this matrix represents the sensitivity of a measure of distortion to a particular small change (or error) in the input speech signal. A mathematical detailed description is shown in Part II of the thesis. It is however interesting to unveil the usage of this matrix and the valuable information that is provided. As mentioned above, the speech signal distortion is transformed into the perceptual domain in which the perceived error is measured by an auditory model-output distortion measure. A major effect of this transformation is the reorganization of the perturbation vectors' direction. More specific, the perturbation vectors that are orthogonal to the transformation matrix will lead to a perceptual domain error which will essentially become irrelevant for the auditory model output domain, therefore not perceivable. The two auditory models, i.e., the so called van de Par and Dau models, which were used in the experiments of this thesis will now be described.

The van de Par auditory model (van de Par et al., 2002) is a psychoacoustic masking model that accounts for simultaneous processing of sound signals. One channel of the model is shown in Fig. 2. The first filter (h_{om}

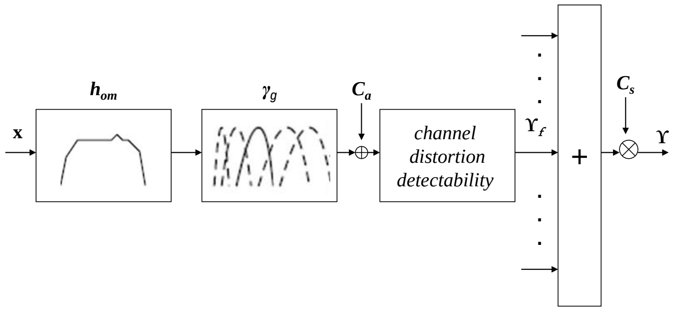


Figure 2: Block diagram of a channel of the van de Par psychoacoustic model.

in Fig. 2) which models the outer and middle ear (OM filter), is approximated by the inverse of the threshold of hearing in quiet. The output of the OM filter is then filtered by a gammatone filterbank (γ_g in Fig. 2) which models the BM in the inner ear. The center frequencies of the gammatone filterbank are spaced linearly on a equivalent rectangular bandwidth (ERB) scale. The model consists of several channels f, in each of which the ratio of the distortion $\mathbf{x} - \hat{\mathbf{x}}$ to masker $\mathbf{x}$ is estimated, where $\mathbf{x}$ denotes the magnitude spectrum of speech and $\hat{\mathbf{x}}$ its perturbation. In the end, all ratios are combined together, to account for the spectral integration property of the human auditory system. The complete model is then described by

$$\Upsilon(\mathbf{x}, \hat{\mathbf{x}}) = C_s L_e \sum_{g \in \mathcal{G}} \frac{\frac{1}{N} \sum_{f=0,\cdots,N-1} |h_{om}(f)|^2 |\gamma_g(f)|^2 |x(f) - \hat{x}(f)|^2}{\frac{1}{N} \sum_{f=0,\cdots,N-1} |h_{om}(f)|^2 |\gamma_g(f)|^2 |x(f)|^2 + C_a}, \quad (1)$$

where C_s and C_a are constants calibrated based on experimental data, L_e is the effective duration of the segment according to the temporal integration time of the human auditory system, the integer g labels the gammatone filter γ_g and $\mathcal{G}$ the set of gammatone filters considered, h_{om} is the outer and middle ear transfer function which is the inverse of the threshold in quiet and finally N is the dimension of the speech segment. In Papers A, B, D and E, the van de Par model is used to obtain the sensitivity matrix in the speech frequency domain.

The second model that was used is the so-called Dau auditory model (Dau et al., 1996a,b) which is a psychoacoustic masking model that accounts for spectro-temporal processing of sound signals. Thus, in this case the signal $\mathbf{x}$ is a time-domain vector. It consists of several stages which simulate the human auditory periphery. A channel l of the Dau model, shown in Fig. 3, includes the hair-cell model consisting of a gammatone filter, a half-way rectifier, and a low-pass filter. Next, an adaptation nonlinear

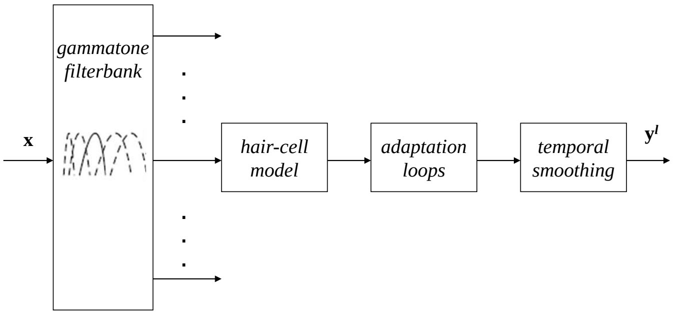


Figure 3: Block diagram of a channel of the Dau psychoacoustic model.

stage incorporates the forward masking prediction of the ear (Plasberg and Kleijn, 2007). Finally, a low-pass filter performs a temporal smoothing and the output is the so-called internal representation $\mathbf{y}^{(l)}(\mathbf{x}_j)$, where $\mathbf{x}_j$ is the j'th speech segment. The original paper (Dau et al., 1996a) included also the addition of internal noise with a level independent variance after the nonlinear transformation to simulate the loss of information that occurs in reality. However, the distortion prediction properties of the model was later performed in (Plasberg and Kleijn, 2007). In the same work a distortion measure on the internal representation was introduced as

$$\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m}) = \sum_l \parallel \mathbf{y}^{(l)}(\mathbf{x}'_j) - \mathbf{y}^{(l)}(\hat{\mathbf{x}}'_{j,m}) \parallel^2, \tag{2}$$

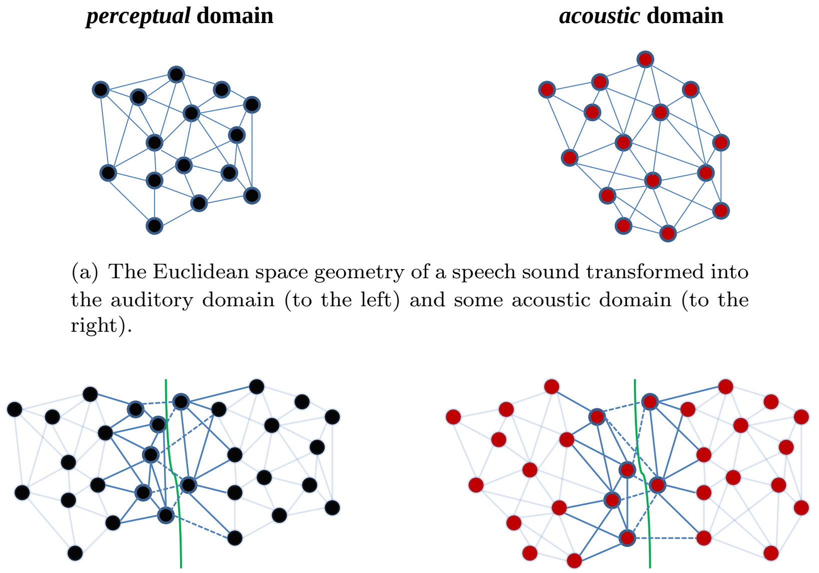
where $\mathbf{x}'_j, \hat{\mathbf{x}}'_{j,m}$ are of higher dimension than the $\mathbf{x}_j, \hat{\mathbf{x}}_{j,m}$ vectors, respectively due to the ring-out effect described in (Plasberg and Kleijn, 2007), and m is the perturbation index. The sensitivity matrix in this case is a result of a more complicated effort compared to that of the van de Par model and can be computed as the sum of the per-channel sensitivity matrices. The Dau model is used in Papers C, E and F.

The next paragraph discusses the basic concept of the thesis and provides the motivation and the conjecture of the proposed methods.

1.3 In search of the perceptual cues of speech signals

The principle of the proposed methods that constitute this thesis is based on examining the common characteristics of speech sounds when transformed into internal representations of an auditory model and into some acoustic feature set. An insight of this is described here.

The human auditory system handles differently small alterations in the speech signal. The basic conjecture of the work presented in this thesis



(a) The Euclidean space geometry of a speech sound transformed into the auditory domain (to the left) and some acoustic domain (to the right).

(b) The hypothesis indicates the human auditory periphery to provide a relatively good separation of sounds. Emphasis is given to the local Euclidean geometry dissimilarity across the speech sounds (bold solid edges for intra- and bold dash edges for inter-sound). The curved lines, describing borders between different sounds, are drawn for illustration purposes.

Figure 4: 2D drawings of a higher dimensional Euclidean space of speech sounds in the perceptual and acoustic domains, respectively. The nodes denote speech segments (speech vectors) and the edges illustrate the geometry of the created regions around these vectors in a high dimensional Euclidean space.

(that has also been verified, e.g., in (Chatterjee and Kleijn, 2011) for optimal design of speech recognition features) is the existence, for each class of sounds, of a region inside the human auditory system which is particularly sensitive to speech signal changes. Consider a speech sound that is transformed into an auditory model output domain and some acoustic feature domain. The above speech sound is composed of a finite number of speech segments that are represented by high dimensional speech vectors. Next, each speech vector is perturbed by a small amount of additive noise to create a finite number of nearby perturbation vectors. It follows that in this high dimensional Euclidean space, a number of regions that is equal to the total amount of the speech segments, are formed. Fig. 4(a) illustrates broadly, the geometry of this speech sound in a high dimensional Euclidean space for each of the considered domains. The nodes represent the high dimensional speech vectors corresponding to the segments that the speech sound consists of, and the edges symbolize the geometry of the created regions in a high dimensional Euclidean space, which for pictorial illustration are sketched on a two-dimensional plane.

The human audition has a significant role in the identification of various sounds. Moreover, it is considered that all the relevant information for sound separation is preserved in the mapping from the acoustic domain to the perceptual domain. Investigating the global geometry (large distances) of the speech sounds would be restrictive for designing any engineering scheme as it is almost impossible to find a mathematical tool apt to handle that. Also, as the sound class boundary is crucial, it might not be so important to preserve the global geometry. The focus has to be on preserving those distances that are short relative to the sound boundary curve, as illustrated in Fig. 4(b), assuming that the similarity between the local geometries in the acoustic and the auditory domains facilitates a human-like classification of the speech sounds. If the two spaces have similar geometry, then the norms, including the Euclidean distances, are preserved. Furthermore, the consideration of small distances helps in reducing complexity by using tools such as perturbation theory. Based on the aforementioned, several algorithms are proposed that are applied in two different areas: i) robust feature selection and optimization of the sound signal representations for speech recognition and ii) diagnostic evaluation of the perceptually relevant differences between native and non-native speech signals for mispronunciation detection.

The next section deals with the first area of application of the concept described above. For practical reasons, i.e., as a prerequisite knowledge necessary for the comprehension of the work described in Part II, the section begins with a general overview of a speech recognition system and a literature review. Next, a more detailed description of the components of a speech recognition system that are connected to the acoustic signal is given. Then, the problem of dimensionality reduction in the front-end is presented. Two of the most popular techniques, which are compared with the thesis feature selection approach, are presented in details, and the section ends with a general description of the proposed, perceptually-motivated, algorithm.

2 Speech recognition

Automatic speech recognition (ASR) deals with the development of techniques that transcribe human speech into written text. Fig. 5 illustrates the main blocks of such a system. These are the *front-end*, the *acoustic models*, the *language model*, the *lexicon* and the *search algorithm* (Rosti, 2004). The front-end or feature extraction part is the part of the ASR system in which the incoming speech signal is processed in a way to derive its meaningful characteristics. The acoustic models are then built from the extracted features together with the text transcriptions of the speech files

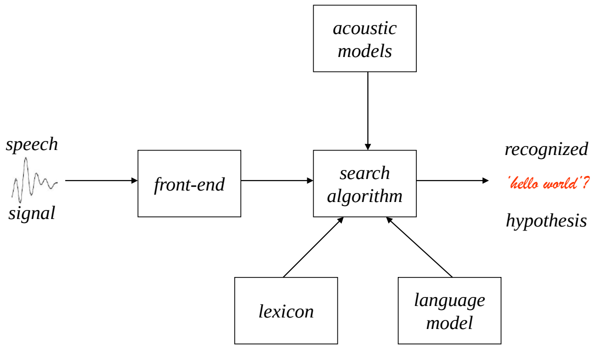


Figure 5: A real speech recognition system.

of the database. Usually, stochastic models and statistical tools are used to create the acoustic models. A language model aims at capturing the statistical properties of a language by estimating the likelihood of a following word, phone or segment in a speech sequence. The lexicon includes all the phones, words and other symbols and the search algorithm is used for finding the most probable sequence of phones or words that has been uttered by the speaker.

In recent years, the performance of ASR systems has improved drastically. One of the main reasons is the development of new acoustic modeling schemes. On the other hand it is generally accepted that an appropriate parametric representation of the acoustic data is an important issue in the design and performance of any ASR system. In other words, if the extracted speech features do not include all relevant information, the performance of the recognition stage degrades significantly.

2.1 Acoustic processing components

Focusing only on the acoustic part of the ASR system, namely the front-end and the acoustic modeling, it is discussed in the next two paragraphs the approaches that have been followed during the course of this study as well as some other known techniques.

Front-End

During the first step in the feature extraction process the speech waveform is sliced up into frames, which are transformed into spectral features as shown in Fig. 6. In this paragraph, it is briefly described the process of extracting mel-frequency cepstrum coefficients (MFCCs) (Davis and Mermelstein, 1980). These features are broadly used in the included papers in

Part II, as they are considered to be the standard acoustic representations of the speech signal for speech recognition as well as for other applications. MFCCs have two important characteristics that are desirable for any system, i.e., they follow a perceptually motivated frequency filtering rather than a linear one and use a logarithmic function to approximate the nonlinearity of the auditory system. It is noted however, that this auditory knowledge incorporated in MFCCs is not in accordance with the most recent findings in hearing research. This is why the MFCCs have been used in this thesis as representations of the acoustic properties of the speech signals with an objective either to select the most relevant coefficients or to optimize them according to more sophisticated psychoacoustic models.

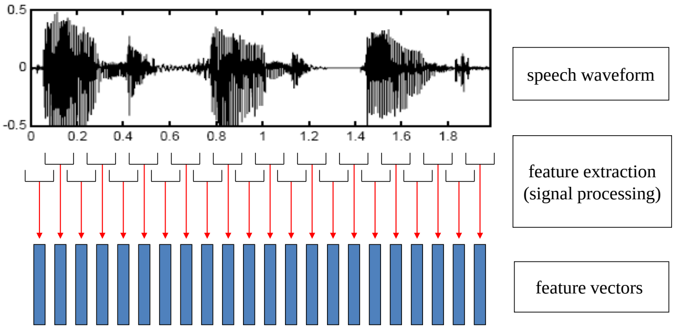


Figure 6: Extracting features from a speech signal.

Mel frequencies are based on the knowledge that the human ear resolves frequencies in a nonlinear manner. Researchers have noticed that the cochlea of the inner ear acts as a spectrum analyzer. The complex mechanism of the inner ear and auditory nerve indicates that the sound perception at different frequencies is not entirely linear (Huang et al., 2001). The response is linear at frequencies below 1 kHz and becoming logarithmic with increasing frequency (Stevens and Volkman, 1940). This behavior is modeled with a filterbank with triangular filters. The amplitude of the triangular filters, shown in Fig. 7, is computed as

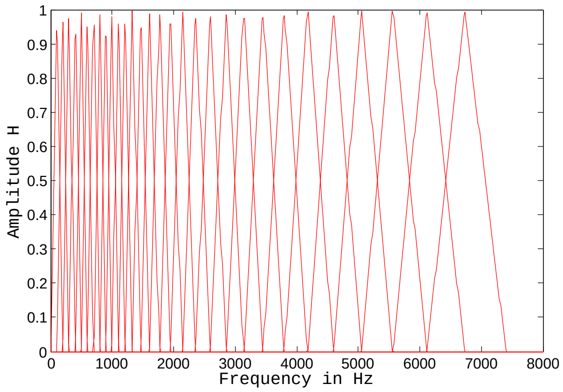


Figure 7: The mel filterbank.

$$\mathbf{H}_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{f(m-1)-k}{f(m+1)-f(m)}, & f(m) \leq k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{3}$$

which satisfies $\sum_{m=1}^{M} \mathbf{H}_m(k) = 1$ according to (Huang et al., 2001).

The speech signal is first pre-emphasized $\mathbf{x}(n) = \check{\mathbf{x}}(n) - \varrho\check{\mathbf{x}}(n-1)$, where $\check{\mathbf{x}}(n)$ is the original speech and $\varrho = 0.97$ (ETSI, 2000), and then a Hamming window (other types of windows can also be used, e.g., Blackman) is applied to the output of the pre-emphasised speech frame

$$\mathbf{x}'(n) = \left\{\alpha - \beta\cos\left\{\frac{2\pi[N-1]}{N-1}\right\}\right\}\mathbf{x}(n), n = 1...N, \tag{4}$$

where $\alpha = 0.54$, $\beta = 1 - \alpha = 0.46$ and N is the length of the window (usually 10-30 ms). A discrete Fourier transform (DFT) is applied to the windowed frame to compute the magnitude spectrum of the signal

$$\mathbf{X}(k) = \sum_{n=0}^{N-1} \mathbf{x}'(n)e^{-j2\pi kn/N}, k = 1...K, \tag{5}$$

where K is the length of the DFT. Next, the DFT power spectrum is computed which is then multiplied with the triangular mel-weighted filterbank.

The result is summed to give the logarithmic mel spectrum

$$\mathbf{s}(m) = \ln\left[\sum_{k=0}^{K-1} |\mathbf{X}(k)|^2 \mathbf{H}_m(k)\right], \tag{6}$$

where $|\mathbf{X}(k)|^2$ is the periodogram, $\mathbf{H}_m(k)$ is the m'th triangular filter, and M denotes the number of triangular bandpass filters used. In the end, the discrete cosine transform (DCT) of the logarithmic filterbank energies is considered to get the uncorrelated MFCCs (Davis and Mermelstein, 1980) as

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \mathbf{s}(m) \cos\left\{q[m + \frac{1}{2}]\frac{\pi}{M}\right\}, q = 1...Q, \tag{7}$$

where Q is the number of cepstrum coefficients, and $\mathbf{s}(m)$ represents the logarithmic mel spectrum of the m'th filter of the filterbank.

Usually, the first, $\Delta\mathbf{c}$, and the second, $\Delta\Delta\mathbf{c}$, time derivatives are added to the speech vector to better capture time dependencies (Young et al., 2002). These are calculated as

$$\Delta\mathbf{c}_t = \frac{\sum\limits_{\vartheta=1}^{\Theta} \vartheta(\mathbf{c}_{t+\vartheta} - \mathbf{c}_{t-\vartheta})}{2\sum\limits_{\vartheta=1}^{\Theta} \vartheta^2}, \tag{8}$$

and

$$\Delta\Delta\mathbf{c}_t = \frac{\sum\limits_{\vartheta=1}^{\Theta} \vartheta(\Delta\mathbf{c}_{t+\vartheta} - \Delta\mathbf{c}_{t-\vartheta})}{2\sum\limits_{\vartheta=1}^{\Theta} \vartheta^2}, \tag{9}$$

respectively. A typical configuration used is $\Theta = 3$ for a delta window, and $\Theta = 2$ for an acceleration window size.

Acoustic modeling

The feature extraction part (a typical paradigm of which is described above) is the first step in building an automatic speech recognition system. The acoustic modeling has likewise a significant role in an ASR system and naturally, is important in improving accuracy. The most popular approach in acoustic modeling is based on statistical methods. Before getting into details, it is necessary to define what an acoustic model is.

Consider a sequence of acoustic input or *observations* O, defined as $O = o_1, o_2, ..., o_T$ where o_t is the observation at time t. (The successive o_t can be considered as indicating temporally consecutive slices of the acoustic input (Jurafsky and Martin, 2000)). The goal of speech recognition is to find the corresponding word sequence $W = w_1, w_2, ..., w_T$ that has the maximum a-posteriori (MAP) probability $P(W|O)$

$$\hat{W} = \arg\max P(W|O) = \frac{P(O|W)P(W)}{P(O)}. \tag{10}$$

The above formula is known as *Bayes' theorem.* Usually, the likelihood of the observation sequence in the denominator, $P(O) : P(O) = \sum P(O|W)P(W)$, is omitted since it is independent of the word sequence. The conditional likelihood $P(O|W)$ is called the *acoustic model* and the $P(W)$ is called the *language model.*

In reality, the most difficult task is to build robust acoustic models to decode/recognize the spoken utterance. For small-vocabulary applications the task is not very complicated, and the unit that usually is modeled is a word. However, for large-vocabulary speech recognition tasks, words are not convenient to be modeled and hence sub-word units, e.g., *phones*, are considered. In all cases, the goal is to have optimal acoustic models to reflect the speech production mechanism, and to be able to model contextual effects such as co-articulation.

Hidden Markov models (HMMs) is the most popular approach for acoustic modeling. *Artificial neural networks* (ANNs) is another stochastic method that has been used in speech recognition. *Segment-based models* (SMs) have also been developed for acoustic modeling. The latest seem to overcome some of the problems met with HMMs and ANNs, though they are of higher computational complexity. This section continues with a short presentation, initially of the HMMs (the approach that is used in the speech recognition Papers A, B, and C), and then with other approaches.

HMMs as a method for acoustic modeling of speech is a flexible and successful statistical approach and hence very popular in speech recognition (Bahl et al., 1990; Jelinek, 1976; Rabiner, 1989). In HMMs, it is assumed that the sequence of observed vectors which correspond to a word or phone is generated by a Markov model (Young et al., 2002) as shown in Fig. 8. Therefore, the HMM approach is a double-embedded stochastic process with a not-directly-observable underlying stochastic process, namely the state sequence. Hence, the name 'hidden'. This hidden process is probabilistically linked with the observable stochastic process which produces the sequence of features that are seen (Huang et al., 2001).

Typically, an HMM can be defined by the following elements:

- Number of states: N ($N = 3$ in Fig. 8)

- Number of distinct observation symbols: M for discrete HMMs and ∞ for continuous HMMs ($M = 3$ in Fig. 8)

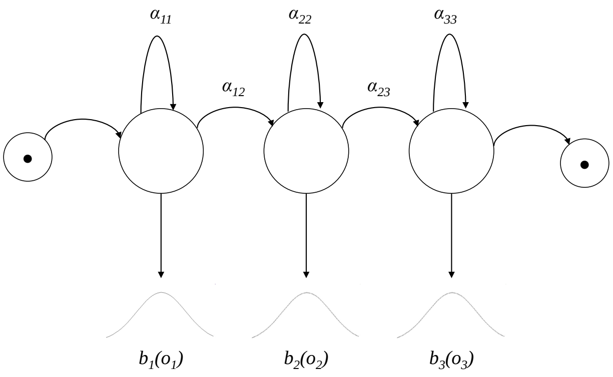


Figure 8: A hidden Markov model.

- State transition probability distribution: α_{ij}
- Output distribution of state j: $b_j(o_t)$
- Initial state probability: π_i.

To summarize, a complete specification of an HMM includes two constant parameters, N and M, that represent the total number of states and the size of observation alphabets respectively, and three sets of probability measures: the state transition matrix A, the output distribution matrix O and the initialization matrix π. For convenience, the following notation is used

$$\lambda = (A, O, \pi) \tag{11}$$

to denote the whole parameter set of an HMM (Huang et al., 2001).

In accordance with the elements of the observation matrix O, HMMs are grouped in different categories (Cole et al., 1998) according to the distribution function they follow. The HMMs are called *discrete* HMMs if the observation sequence consists of vectors of symbols in a finite alphabet of N different elements, i.e., the distributions are defined on finite spaces. If the observation is not derived from a finite set, but rather from a continuous space, limitations on the functional form of the distributions should be imposed to achieve a reasonable number of statistical parameters that need to be estimated. A common solution to this matter is the categorization of the model transitions to mixtures of known densities g of a family G that have a simple parametric form. These densities $g \in G$ are usually Gaussian or Laplacian, and can be easily characterized by two parameters, the mean vector and the covariance matrix. HMMs of this type are referred to as *continuous* HMMs. To model more complex distributions, a rather larger number of base densities has to be used in every mixture. This may require a very large training set of data to effectively estimate the parameters of the distribution. Problems arise when the available corpus is not large enough, although this can be resolved with sharing distributions among transitions of different models. Finally, in *semi-continuous* HMMs, all mixtures are expressed in terms of a common set of a base density. Different mixtures can be characterized only by different weights.

The parameters of the HMMs can be estimated by iterative learning algorithms (Rabiner, 1989) in which the likelihood of a set of training data is increased in each step. As a result of their higher complexity, the continuous HMMs need a significantly larger amount of time to compute their probability densities in comparison to the discrete HMMs. However, it is possible to speed up the computations by applying vector quantization (VQ) to initialize the Gaussian mixtures (Bocchieri, 1993).

The HMMs are based on two assumptions. The first is the Markov chain assumption in which it is assumed that the current state depends only on the previous state, given the current state (in the simplest case of a first-order Markov chain). The second is the output independence assumption in which a particular symbol that is emitted at time t depends only on the state s_t given this state, and is conditionally independent of the past observations. Dynamic information can be included in HMMs through the time-derivatives (velocity and acceleration coefficients) in the observation vector, however under the false frame-independence assumption. Although the above assumptions allow the model to become easier to use, they introduce some limitations that principally influence the accuracy of the model (Merhav and Ephraim, 1991; Digalakis, 1992). For this, other methods have been proposed for acoustic modeling, many of which are described in the next few paragraphs as they may be promising alternatives to HMMs.

Artificial neural networks (ANNs), also known as *connectionist models* or *parallel distributed processing* (PDP) were introduced by McCulloch and Pitts (1943). Due to their nature, ANNs are of great interest for tasks that require a series of constraints to be satisfied, such as ASR. Their ability to evaluate in parallel many clues and facts and their interpretation in the light of numerous interrelated constraints (Huang et al., 2001) have been appreciated by many ASR researchers. The simplest type of ANNs consists of a number of nodes or units, connected with each other by links (Russel and Norving, 1995). Each link has a probabilistic weight, and the learning procedure is performed by updating these weights. Some of the units are connected to the external environment; these are the input or output units. Each unit has a set of input links from other units, a set of output links to other units, a current activation level, and a means of computing the activation level at the next step in time, given its inputs and weights. The units depend only on their neighbors and all the computations they perform are independent of the rest units. For computational reasons, many implementations have used a synchronous control to update all the units in a fixed sequence. Other types of ANNs are described in (Hinton, 1989; Rummelhart and McClelland, 1986; Huang et al., 2001; Waibel and Lee, 1990). Finally, some hybrid HMMs/ANNs (Robinson et al., 1993; Zavaliagkos et al., 1994; Cook and Robinson, 1998; Fritsch and Finke, 1998; Morgan and Bourlard, 1995; Robinson, 1994) methods have been developed for ASR.

Segment models (SMs) have been extensively used for various applications, among them speech recognition (Digalakis, 1992; Frankel, 2003). HMMs generate a single observation that is conditionally independent from other observations given the hidden state at the current time. Hence it is difficult to model relative durations within a phone segment since it may be possible to have some parts of a segment stretched and others compressed. On the other hand, SMs generate a variable-length sequence of observations (Ostendorf et al., 1996; Rosti, 2004). A segment may be a variable-length part of the speech waveform that usually corresponds to a language unit, e.g., a word, a phone or a sub-phone (Digalakis, 1992). Segment-based models (Bocchieri and Doddington, 1986; Bush and Kopec, 1987; Ostendorf and Roukos, 1989; Digalakis et al., 1993; Kimball, 1995; Roweis and Ghahramani, 1999) have been proposed as HMMs alternatives, offering a more suitable and flexible scheme to model the dynamics of speech signals. In all cases, several modeling restrictions were applied to ensure that the model is identifiable. In (Tsontzos et al., 2007) an effort to relax these constraints was taken which allowed the choice of full noise covariances and state vectors. The use of the canonical form of the system's matrices proposed in (Ljung, 1998) ensured the system's identifiability. The parameters were estimated using a maximum likelihood, element-wise, process based on the Expectation-Maximization (EM) algorithm and the proposed system was applied in a speech recognition task using the AURORA2 (Hirsch and Pearce, 2000) speech database. Significant performance gains were found compared to HMMs, particularly in highly noisy conditions.

In recent years, a variation of segment models called *hidden dynamic models* (HDMs) (Deng and Ma, 1999; Ma and Deng, 2004; Picone et al., 1999; Richards and Bridle, 1999; Zhou et al., 2003) has been proposed. The main focus in this approach was to efficiently model the co-articulation phenomenon and improve the transitions between neighboring phones. The hidden dynamic space consisted of a single vector target per phone in which the speech trajectories were produced by a dynamic system. The observation process in HDMs was implemented by a global *multi-layer perceptron* (MLP). The model was simple and flexible, and also able to capture important aspects of the relation between the phonetic labels and the acoustic patterns. The major drawback of the method was that the inference algorithms were not tractable. Hence, a number of methods have been proposed to improve the algorithms (Lee et al., 2003; Ma and Deng, 1999; Ma and Deng, 2000; Ma and Deng, 2001; Seide et al., 2003).

Alternatively, an idea of inserting articulatory knowledge into acoustic models (Richardson et al., 2000a,b; Richardson et al., 2003) called the *hidden articulatory Markov model* (HAMM) has been applied in speech recognition. The model, based on (Erler and Freeman, 1996), is essentially an HMM in which each articulatory configuration is modeled by a separate state. The state transitions aim to naturally reflect human articulation.

The aforementioned alternative approaches for the acoustic modeling have, in some cases, accomplished better performance in ASR compared to HMMs. However, most of these methods introduce an increased computational cost and become particularly complex in larger contexts with continuous speech. In Papers A, B and C, the HMMs are preferred instead of the alternatives offered. The choice of the HMMs is based on: a) The fact that in this work the focus is mostly on the observation input and not on the acoustic model per se. In other words, the goal is to optimize the front-end of the ASR. b) The HMMs are still the standard method used in ASR and the HTK toolkit (Young et al., 2002) is a very popular environment to build a speech recognizer. c) The relatively low complexity of HMMs. d) Finally, the admittedly easier way to compare the proposed methods with other techniques when a known (and established) configuration for the recognition system is used.

2.2 Reducing feature dimensionality

In the previous section, two of the most important parts of an ASR system were described in general, i.e., the front-end and the acoustic model. This section deals with methods and techniques that have been used to lower the cardinality of the input feature vectors without loosing the maximum available information for sound class discrimination.

The effective process of the speech signal and the careful extraction of the necessary, acoustic-relative, features is essential for applications such as speech recognition. Although it seems natural to consider that a high dimensional feature vector would lead to high performance in a speech recognition system, in practice it is not always the case (Hughes, 1968; Kanal and Chandrasekaran, 1971). The phenomenon of *curse of dimensionality* (Bellman, 1957) refers to the problem caused by the exponential increase in volume associated with adding extra dimensions to a mathematical space. The performance of a speech recognition system may decrease in case the system is feeded with very large feature vectors. A series of different techniques and methods have been proposed in order to optimally reduce the dimensionality of the feature representations and improve the performance of the classification system.

In the remainder of this section, three popular methods to reduce dimensionality are described, i.e., linear discriminant analysis (LDA), heteroscedastic linear discriminant analysis (HLDA) and principal component analysis (PCA). The first two are among the techniques used for comparison when evaluating the proposed method described in this thesis (the other two are the average performance of five randomly selected MFCC feature subsets and the initial n MFCCs depending on the cardinality n that is considered). In addition, some other techniques in feature selection are discussed and finally, the proposed auditory model-based feature selection method (AMFS) is presented. The latter is presented in more details in Papers A and C.

Linear discriminant analysis

Linear discriminant analysis (Fisher, 1936; Fisher, 1938; Rao, 1965; Duda et al., 2000) has been applied in feature reduction problems for speech recognition tasks (Brown, 1987; Hunt and Lefebvre, 1989; Haeb-Umbach and Ney, 1992; Aubert et al., 1993; Siohan, 1995; Demuynck et al., 1999; Abbasian et al., 2008). In (Sharma et al., 2000) a study of combined feature sets including, among other, LDA transformations was performed. In this section, an outline of the method is given since it has been one of the major techniques that competed with the AMFS method but was not thoroughly described in the corresponding papers A and C.

The goal of LDA is to find an optimal transformation matrix ϕ^T to reduce the dimensionality of the feature space and, at the same time, to maximize the necessary information to distinguish between different classes in a classification task problem. The above can be expressed as

$$\mathbf{y} = \phi^T \mathbf{c}, \tag{12}$$

where $\mathbf{y}$ is the p-dimensional feature vector in the reduced feature domain $\mathbb{R}^p$, $\phi \in \mathbb{R}^{q \times p}$ is a transformation matrix and $\mathbf{c}$ is the q-dimensional feature vector in the original feature domain $\mathbb{R}^q$. The method requires data associated to class labels before the analysis starts. In the problem of speech recognition, it is necessary to use a transcription alignment (label) file of the recorded data in combination with feedback from the recognizer, e.g., the HMMs statistical properties in case of an HMM recognizer. To formulate mathematically the optimization procedure, the mean vector and the covariance matrix for each class can be computed as

$$\mu_j = \frac{1}{N_j} \sum_{i=1}^{N_j} c_i, \tag{13}$$

$$\mathbf{\Sigma}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} [c_i - \mu_j][c_i - \mu_j]^T, \tag{14}$$

where N_j denotes the number of training tokens in class j. Then, the mean and the covariance of all the data are computed as

$$\mu = \frac{1}{N} \sum_{i=1}^{N} c_i, \tag{15}$$

$$\mathbf{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} [c_i - \mu][c_i - \mu]^T, \tag{16}$$

where $N = \sum_{j=1}^{J} N_j$ is the total number of training tokens.

Based on the above statistics, the transformation matrix can be calculated using the following optimization criterion

$$\hat{\phi} = \arg\max_{\phi_p} \frac{|\phi_p^T \, \mathbf{\Sigma} \, \phi_p|}{|\phi_p^T \, \mathbf{S} \, \phi_p|}, \tag{17}$$

where

$$\mathbf{S} = \frac{1}{N} \sum_{j=1}^{J} N_j \mathbf{\Sigma}_j. \tag{18}$$

The maximization criterion, Eq. (17), is a measure of how well the matrix $\hat{\phi}$ maximizes the distances between classes and at the same time minimizes their size. It can be shown that $\hat{\phi}$ consists of those eigenvectors of $\mathbf{S}^{-1}\mathbf{\Sigma}$ that correspond to the p largest eigenvalues (Dillon and Goldstein, 1984; Kumar and Andreou, 1998).

In Appendix I, a short description of the implementation of the LDA method used in Papers A and C is given.

Heteroscedastic linear discriminant analysis

Heteroscedastic linear discriminant analysis (Kumar and Andreou, 1996; Kumar, 1997) is an extension of the forementioned LDA method, that has also been tested against AMFS in Papers A and C. Although the basic idea remains the same, i.e, to find the best linear discriminant, HLDA differs from LDA in the underlying assumptions. The main weakness of the LDA method is the assumption of equal covariance matrices for all classes in the parametric model. For most applications, the above assumption does not cause major problems. The class assignment problem (Kumar and Andreou, 1998) is the second shortcoming of LDA. Therefore, HLDA was developed to overcome these limitations.

In HLDA, the transformation matrix ϕ is a $q \times q$ matrix, and hence differs from the LDA, which is applied in the original feature vector as in Eq. (12). The transformation ϕ is applied to the original feature vector, however from the resulting transformed vector $\mathbf{y}$, only the first p elements are retained. This choice is based on the assumption that only the first p components of $\mathbf{y}$ may carry the classification information (Kumar and Andreou, 1998). The data are modeled as a Gaussian distribution (Kumar, 1997) and the parameters of the probability density function (PDF) are

$$\mu_j = \begin{bmatrix} \mu_j^p \\ \mu \end{bmatrix}, \tag{19}$$

and

$$\mathbf{\Sigma}_j = \begin{bmatrix} \mathbf{\Sigma}_j^p & 0 \\ 0 & \mathbf{\Sigma}^{q-p} \end{bmatrix}, \tag{20}$$

where μ_j, $\mathbf{\Sigma}_j$ are the mean and covariance for the class j, respectively. The parameters μ_j^p and $\mathbf{\Sigma}_j^p$ are different for each class while μ and $\mathbf{\Sigma}$ are common. Then, the Gaussian PDF of $\mathbf{c}_i$ is given by the following equation

$$P(\mathbf{c}_i) = \frac{|\phi|}{\sqrt{(2\pi)^q|\mathbf{\Sigma}_{g(i)}|}} \exp\left\{-\frac{1}{2}[\mathbf{y}_i - \mu_{g(i)}]^T \mathbf{\Sigma}_{g(i)}^{-1}[\mathbf{y}_i - \mu_{g(i)}]\right\}, \tag{21}$$

where $\mathbf{y}_i = \phi^T\mathbf{c}_i$, and $g(i) = j$ denotes the mapping of the observations i to classes j.

The log-likelihood function, necessary to find the best estimator for ϕ, is then

$$\log P(\mu_j, \mathbf{\Sigma}_j, \phi; \{\mathbf{c}_i\}) = N\log|\phi| - \\ -\frac{1}{2}\sum_{i=1}^{N}\left\{\log[(2\pi)^q|\mathbf{\Sigma}_{g(i)}|] + [\phi^T\mathbf{c}_i - \mu_{g(i)}]^T\mathbf{\Sigma}_{g(i)}^{-1}[\phi^T\mathbf{c}_i - \mu_{g(i)}]\right\}. \tag{22}$$

Considering the derivatives versus μ_j and $\mathbf{\Sigma}_j$, and setting them equal to zero, the following estimates arise

$$\hat{\mu}_j = \phi_p^T\mathbf{c}_j, \tag{23}$$

$$\hat{\mu} = \phi_{q-p}^T\mathbf{c}, \tag{24}$$

$$\mathbf{\Sigma}_j = \phi_p^T\mathbf{\Sigma}_j\phi_p, \tag{25}$$

and

$$\mathbf{\Sigma} = \phi_{q-p}^T\mathbf{\Sigma}\phi_{q-p}, \tag{26}$$

where $j = 1, ..., J$. Next, the above estimates can be substituted into the log-likelihood function, Eq. (22), and then it can be shown (Kumar, 1997) that the final estimate of ϕ is given by

$$\hat{\phi} = \arg\max_{\phi}\left\{-\frac{N}{2}\log|\phi_{q-p}^T\mathbf{\Sigma}\phi_{q-p}| - \sum_{j=1}^{J}\frac{N_j}{2}\log|\phi_p^T\mathbf{\Sigma}_j\phi_p| + N\log|\phi|\right\}. \tag{27}$$

Kumar and Andreou (1998) maximized the above equation using numerical methods. The $\hat{\phi}$ is initialized by the previously (from the LDA method) computed ϕ.

Principal component analysis

Principal component analysis (Pearson, 1901) is an old, non-parametric, but still interesting method to reduce data dimensionality. Even though it is not used in this thesis, PCA is widely applied in all forms of analysis (also in speech recognition (Ding and Liming, 2001; Tsai and Lee, 2003; Wanfeng et al., 2003)) due to its simplicity to extract relevant information from confusing data sets, and therefore is mentioned in this chapter. Depending on the field of application, it is also named the *discrete Karhunen-Loève transform* (KLT), the *hotelling transform* or *proper orthogonal decomposition* (POD).

The goal of PCA is to compute the most meaningful and relevant basis by transforming a set of, usually, correlated data. In doing so, the next steps are followed: firstly, the mean value is subtracted from each of the data dimensions and the covariance matrix is calculated. Next, an eigenvalue decomposition is applied and the eigenvectors and eigenvalues of the covariance matrix are calculated. The eigenvector with the highest eigenvalues is the direction with the greatest variance. The k eigenvectors with the highest eigenvalues are considered to form a matrix ψ with these eigenvectors in the columns. Finally, the feature vectors are transformed using the resulted transformation matrix ψ.

Kumar (1997) attempted to compare the discussed dimension-reduction approaches by pointing out each one's advantages in relation to that of the others. In PCA, the first principal component of a sample vector represents the direction with the largest variance over all samples. All the chosen principal components – k in total, corresponding to the k larger eigenvalues – are linear combinations of the feature vectors with the largest variance, and every newly chosen component is uncorrelated to the prior. As it is not necessarily based on vector properties related to classification, this approach includes a somewhat high risk of failure. It is not always the case that the chosen principal components involve all the required information to discriminate the classes in a pattern classification task.

Assume, for example, a classification task that consists of two Gaussian distributions with equal variance in a two-dimensional sample space, which need to be discriminated. The general form of the problem is shown in Fig. 9. The line called "PCA" is, according to the theory, in the direction of maximum variance for each of the two distributions, and in the direction of the maximum variance of the mixture of these two Gaussians, and hence in the direction of the first principal component. The line labeled "LDA" shows how the linear discriminant analysis can easily distinguish the two classes by choosing the correct direction. This is not the case with "PCA",

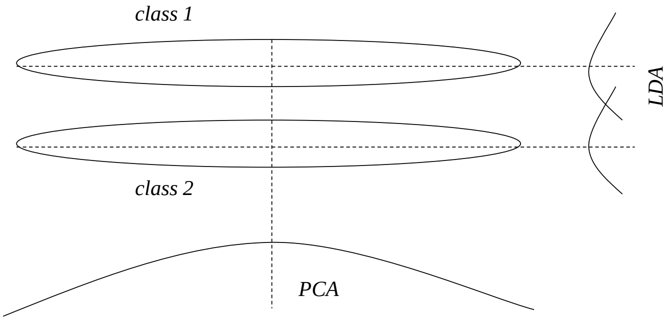


Figure 9: A two-class Gaussian classification problem where PCA fails to discriminate correctly. Adapted from (Kumar, 1997).

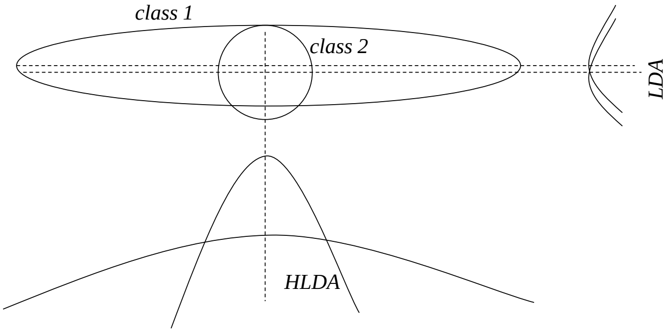


Figure 10: A two-class Gaussian classification problem where LDA fails to discriminate correctly. Adapted from (Kumar, 1997).

the projection of which, gives no discrimination result. The HLDA method would on the other hand work well, just as LDA.

Fig. 10 shows another example in which, this time, the LDA method fails. This is the case where the within class distributions are *heteroscedastic*. In this particular case, the means of the two classes are close but the variance of the one distribution is significantly larger than the other. As discussed previously, LDA considers the within-class variances. This is not sufficient for this case. A heteroscedastic model such as HLDA, can indeed obtain the best discriminant result as shown in Fig. 10. For this problem however, even PCA would result in the best discrimination of the two classes.

Other methods

To further reduce the dimensionality of feature sets, a series of other algorithms have also been proposed to select optimal subsets. An approach was to find the maximum statistical dependency between a feature subset and a class by computing the mutual information (e.g., Yang et al., 2000; Scanlon et al., 2003). This method was computationally intractable. An alternative approach proposed in (Ding and Peng, 2003) and extended in (Peng et al., 2005), combined the minimal-redundancy-maximal-relevance (mRMR) criterion with a wrapper, a comparably fast method to minimize the classification error for a particular classifier. The algorithm is particularly useful for large-scale feature selection problems where a large number of features is available, e.g., in medical tasks (Herskovits et al., 2004; Xiong et al., 2001). Crudely speaking, the mRMR approach tries to maximize the dependency. Typically, this would involve the computation of multivariate joint probability, a somehow difficult and inaccurate computation. mRMR combines both Min-Redundancy and Max-Relevance criteria to estimate multiple bivariate probabilities, a much easier way to maximize the dependency than the estimation of multivariate joint probability that would otherwise be imposed by the dependency maximization criterion per se. At each step, the approach selects those features that follow the mRMR criterion and is hence intended for features that are not independent of each other. In (Peng et al., 2005), the authors claimed that the whole process was faster than other closely-related methods due to the relatively lower computational complexity.

In (Valente and Wellekens, 2003), the maximum entropy discrimination (MED) (Jaakkola et al., 1999; Jebara and Jaakkola, 2000; Jebara, 2001) feature selection was proposed for ASR. The results were comparable to a wrapper but the algorithm was less computationally expensive. In MED, each feature was associated to a probability weight value. Then, the M out of N most important features were considered based on their probability values. This condition can be incorporated in the optimal prior formulation to help the process in finding the M most relevant features. Compared to wrapper methods, the MED feature selection is faster. Finally, since MED is a Bayesian discriminative algorithm, it usually has a high recognition rate capacity.

2.3 Auditory model-based feature selection

In all the above methods, the relation between features and target classes was investigated and different criteria were applied to reduce the classification error. In this section, the novel feature selection method for speech recognition, which is based on human perception, is presented epigrammatically (further information can be found in Papers A and C).

The auditory model-based feature selection (AMFS) is a fundamentally different approach to feature selection in which an exploitation of the knowledge implicit in the human auditory system is performed instead of optimizing the classification performance. Humans perform better at speech recognition than machines, particularly for noisy environments, suggesting that the signal representation in the human auditory periphery is both effective and robust, and thus the usage of computational models of the periphery are likely to improve the properties of the acoustic features. The motivation to study the selection and design of robust acoustic features that maximize the similarity of the Euclidean geometry of the feature set and the human auditory representation of the signal comes from the accuracy of recent methods for auditory modeling (e.g., Dau et al., 1996a; van de Par et al., 2002). The goal is to better understand the relation between human and machine-based recognition and to find a path towards better performance. The features are selected without knowledge of the meaning of speech and without the use of a specific speech recognizer, a distinctive attribute that allows the method to remain system independent.

The implementation of AMFS relies on perturbation theory. While the method does not use classified data, it is based on the following property: for two features sets to perform similarly in classification, "small" Euclidean distances must be similar in the two domains (except for a scale factor), i.e., the auditory model output and the feature domain. The similarity of "large" distances is immaterial for the classification. The results show that maximizing the similarity of the Euclidean geometry of the features to the geometry of the perceptual domain is a powerful tool to select features (Papers A and C) as well as to investigate new features (Paper B).

The focus on small distances allows complex perceptual distortion measures to be reduced to quadratic distortion measures using the so-called *sensitivity matrix*. This theme was first developed in the context of rate-distortion theory (Gardner and Rao, 1995; Linder et al., 1999; Li et al., 1999) and was later used for audio coding (Plasberg and Kleijn, 2007). In the feature domain, it is possible to have analogous distortion measures that also use the notion of sensitivity matrix. Consider the mapping $\mathbf{c}_i : \mathbb{R}^N \rightarrow \mathbb{R}^L$ from a signal segment $\mathbf{x}_j$ to a set of L features $\mathbf{c}_i(\mathbf{x}_j)$ with set index i. For a sufficiently small distortion $[\hat{\mathbf{x}}_{j,m} - \mathbf{x}_j]$, the Taylor series can be used to make a local approximation around $\mathbf{x}_j$ as

$$\mathbf{c}_i(\hat{\mathbf{x}}_{j,m}) \approx \mathbf{c}_i(\mathbf{x}_j) + \mathbf{A}[\hat{\mathbf{x}}_{j,m} - \mathbf{x}_j], \tag{28}$$

where $\mathbf{A} = \left.\frac{\partial \mathbf{c}_i(\mathbf{x}_j)}{\partial \hat{\mathbf{x}}_{j,m}}\right|_{\hat{\mathbf{x}}_{j,m}=\mathbf{x}_j}$, and $\hat{\mathbf{x}}_{j,m}$ is a perturbation of $\mathbf{x}_j$ with perturbation index m. An L^2 distance measure in the feature domain then leads to a signal domain sensitivity matrix $\mathbf{D}_\Gamma(\mathbf{x}_j) = \mathbf{A}^T\mathbf{A}$. Thus, the feature distortion measure can also be written in an analogous quadratic form as the

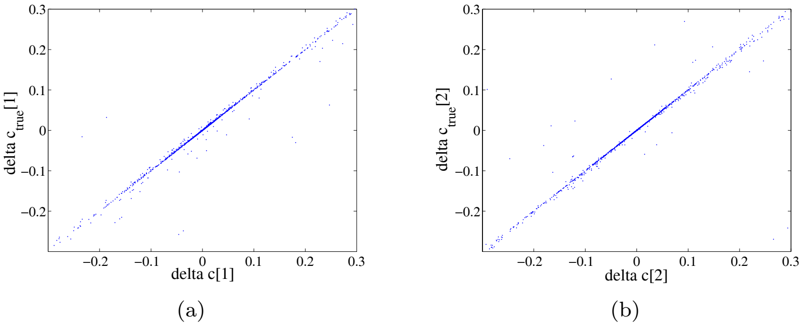


Figure 11: Scatter plots of the estimated distances (cf. Eq. (28)) between the cepstra of the original and the distorted signals, $\delta\mathbf{c}$'s, vs. the actual distances, $\delta\mathbf{c}_{true}$'s, for the first (a) and second (b) MFCC, respectively.

perceptual distortion measure (the mathematical relationships are shown in Paper A).

As already stated in Sec. 2.1, the method is applied to MFCCs[1]. The range of validity for the linearization assumption, Eq. (28), between the cepstrum and the speech is investigated with the following experiment. The speech signal is distorted with independent and identically distributed (i.i.d.) Gaussian noise at different signal-to-noise ratios (SNRs) ranging from 30 to 90 dB with a step of one. Fig. 11(a-b) shows the change in the features computed from the linearized relation Eq. (28) versus the true difference between the cepstra of the original and distorted signals, for the first (a) and second (b) MFCC, respectively. The linearity assumption is reasonable at a scale that is meaningful for sound discrimination and therefore significant for the studied task. The outliers result from regions where the power of the signal is low.

AMFS is related to other approaches that use auditory models as a front-end for ASR (e.g., Seneff, 1988; Ghitza, 1991; Jeon and Juang, 2005; Haque et al., 2007). The performance for such front-ends is, generally, particularly robust to various environmental conditions. AMFS has a significant advantage over an auditory-model-based front-end, as it avoids the computational complexity associated with pre-processing the signal with an auditory model, and also the difficulty of formatting the auditory-model output for classification.

An analytical description of the method can be found in the Part II of this thesis. Finally, Appendix II presents the derivation of the $\mathbf{A}$ matrix for

[1] The method generalizes well, hence other type of features may considered.

the case of MFCCs in both the frequency and time domains.

An application to speech synthesis

As previously mentioned, AMFS is a general approach and can be applied in ASR systems independently of their type. In the following, an example is given on how the method can be adapted even in different areas, e.g., speech synthesis, to emphasize this generalization property.

In this experiment, the HMM-based speech synthesis system[2] HTS (Tokuda et al., 2000) is used to train the HMM models. HTS does this in a unified framework by combining spectrum parameters with the fundamental frequency and duration. The spectrum part consists of the MFCCs together with the zeroth coefficient C_0 and their velocity V and acceleration A coefficients. The excitation part is composed of the logarithmic fundamental frequency $\log F_0$ that consists of a continuous value sequence for the voiced regions and discrete symbols for the unvoiced ones, and its Δ's and $\Delta\Delta$'s coefficients.

The mel-cepstral analysis technique (Fukada et al., 1992) that exploits the mel log spectrum approximation (MLSA) filter is applied to directly synthesize the speech from MFCCs. A multi-space probability distribution HMMs[3] (MSD-HMMs) (Tokuda et al., 1999) is then used to account for the continuous-discrete mixed observation sequence that is considered (Tokuda et al., 2002). A decision-tree based context clustering technique (Odell, 1995; Shinoda and Watanabe, 2000) is applied, in which context-dependent HMMs are used to model the presence of contextual factors (Tokuda et al., 2002), such as those related to phone identity and stress, which affect all the parameters of HMMs. To further reduce complexity, a clustering algorithm that has been extended for MSD-HMMs is chosen (Yoshimura, 2002). Each group of factors influence separately the spectrum, F_0 and the duration, which therefore are clustered independently.

The text from which the speech is to be synthesized is first converted to a context-based label sequence and an HMM sequence is then produced as a result of the concatenation of these context-dependent HMMs. The speech parameters, i.e., the MFCCs and the $\log F_0$, are used to maximize the output probability of the HMMs using the generation algorithm presented in (Tokuda et al., 2000). Finally, the synthesis is performed directly from the generated MFCCs and $\log F_0$ values by using the MLSA filter and 5-state context-dependent HMMs that correspond to phone units. In addition, the

[2] A speech synthesis environment built on HTK toolkit (Young et al., 2002), together with the Speech Signal Processing Toolkit (SPTK) (SPTK, 2003).

[3] MSD-HMMs are a mixture of continuous and discrete HMMs that are more flexible than regular HMMs in modeling various observation vectors, independently of their dimensionality. Each of the consisting HMMs uses a a multivariate Gaussian distribution to model the duration of the state (Yoshimura et al., 1998).

contextual factors, described in (Tokuda et al., 2002), are extracted from the data using the Festival speech synthesis system (Black et al., 2001). The training process results in trees of spectrum models, F_0 models and state duration models. The run-time core engine consists of 8 modules (Tokuda et al., 2002), the decision trees for spectrum, F_0, and duration as well as their distributions. Finally, the extracted features are converted into a context dependent label sequence and then the synthesizer generates the waveform that corresponds to this sequence.

For the experiments, 450 sentences of the TIMIT corpus (Lamel et al., 1986) sampled at 16 kHz are considered. A 25 ms Blackman window with a 5 ms shift is used. The full set consists of 75 MFCCs in total, i.e., 24 static coefficients plus 24 velocity and 24 acceleration coefficients as well as the zeroth one with its dynamics. The static AMFS algorithm (described in Paper A), is applied to the static 24 MFCCs to find the subset vectors of $18-$, $12-$ and $6-$dimension. The zeroth coefficient, its dynamics and the dynamic features of the selected MFCCs are then added. The time needed for training the HMMs, synthesizing and generating the speech is reduced as the cardinality of the MFCCs subsets is lowered. The time reduction (mainly concerning the stage in which the HMM models are built and also the context clustering phase) is between 4.3%, for the case of 18 selected MFCCs and 17.3% when only 6 MFCCs are selected.

To evaluate the quality of the synthesized speech, a perceptual experiment is performed in which the generated speech from the full set is compared to the corresponding synthesized speech from the three different subsets. For the listening experiment, a set of 20 synthesized speech files is used. The participants (five experts in speech technology and one non-expert) listen to the same sentences generated from $75-, 57-, 39-$, and $21-$dimensional MFCC vectors without knowing which utterance corresponds to each case. The synthesized speech for the full set of parameters receives, as expected, the best score from all the listeners. All subjects agree that the quality of the synthesized speech remains relatively high even for the first subset, i.e., when selecting 18 out of 24 coefficients (which essentially means a reduction of totally 18 coefficients including their dynamic features). When the dimensionality is halved, the quality of the sound is degraded, but is still at an acceptable level. Finally, the experiment shows that for the case of 6 MFCCs and their velocity and acceleration coefficients, all participants agree that the synthesized speech is of low quality.

With this paradigm, the part of the Introduction devoted to the perceptually relevant feature selection and optimization is concluded. The application to speech recognition is described in the included Papers A, B and C.

3 Pronunciation error detection

The second area of interest in this thesis, is the automatic *pronunciation error detection* (PED) in foreign language learning. The subject is relatively new with an increasing number of researchers developing original methods and novel techniques to assist second language (L2) learners. For many L2 speakers the production of the second language is problematic, in particular for target phonemes which are unrelated to sounds in the native language (Flege, 1995; Guion et al., 2000). This idiomorphism is widely related to higher-level cognition by which humans develop the ability to harmonize their hearing (and thereby, their production) system to the sounds of their native language (Werker and Tees, 1984; Kuhl, 1993). This behavior is revealed in a more emphatic manner when comparing the perceptual ability of non-native speakers with that of native ones under various noisy conditions (Cutler et al., 2008; Cooke et al., 2008; Lecumberri et al., 2010). Sometimes, L2 learners tend to interfere some speech sounds from their L1 or to ignore unfamiliar ones (Piske et al., 2001). Moreover, many L2 speakers usually adopt a reduced oral rate compared to the native speakers, which leads to unnatural expansion in the duration of their utterances. Those who on the other hand avoid reducing the speech rate, often fall into articulatory inconveniences including, e.g., unfamiliar expressional elements that cause additional difficulties to native speakers in handling speech sounds produced by L2 speakers (Munro and Derwing, 1995; Schmid and Yeni-Komshian, 1999).

Commonly, PED is treated as a problem of classification. Secs. 3.1-3.3 briefly present some of the approaches found in literature and provide the necessary preliminary background in this area. Then, Sec. 3.4 presents how the problem is dealt in this thesis.

3.1 Classification-based approaches

In (Neumeyer et al., 1996; Franco et al., 1997; Neumeyer et al., 2000) an attempt was described to directly convert a speech recognizer to a system of automatic scoring of pronunciations. A set of different pronunciation scoring algorithms were developed, namely the HMM phone log-likelihood, HMM phone log-posterior probability, segment classification, segment duration, timing, and finally, a combination of scores, were all compared to human listeners' evaluation and found that certain such scores, like the log-posterior and the normalized duration correlated well with human grades. In (Witt and Young, 2000), the goodness of pronunciation (GOP) measure was presented, a likelihood-based algorithm to calculate the likelihood ratio of a phoneme realization by an L2 speaker to its canonical pronunciation. GOP measures the quality of pronunciations by non-native speakers and gives a score to each phone of an utterance according to its closeness to that of native speakers. In (Park and Rhee, 2004), the method was knowledge-based (including acoustic-phonetic, linguistic, and expertise knowledge) accompanied with an analysis of the correlation of human listeners and machine-based rating. In (Truong et al., 2005), classifiers using LDA and decision trees were developed for three of the most mispronounced phonemes of Dutch by foreign speakers. The authors examined the acoustic properties of the considered pronunciation errors to extract acoustic features that were used to train the classifiers and were later capable of distinguishing erroneously produced phones by non-native speakers. An alternative approach was introduced in (Tepperman and Narayanan, 2008), in which articulatory information was used to improve automatic detection of typical phoneme-level errors made by non-native speakers. For this, a new version of the Hidden-articulator Markov Model (Richardson et al., 2003), adapted for pronunciation evaluation, was presented. The articulatory information concerned features that were derived by concatenating articulatory recognition results over eight streams representative of the constituents of the vocal tract and by calculating multidimensional articulatory confidence scores within these representations based on general linguistic knowledge of articulatory variants. In (Ito et al., 2005), the pronunciation error rules were grouped in a decision-tree based clustering scheme. Each cluster was bestowed with a different threshold, which led the algorithm to achieve good detection results. In (Strik et al., 2009), four different classifiers used for mispronunciation detection were examined: a GOP-based, one combining cepstral coefficients and LDA, a method based on the work described in (Weigelt et al., 1990), which is an algorithm that discriminates voiceless fricatives from voiceless plosives, and an LDA-acoustic-phonetic feature classifier. Experiments showed that the best results were obtained for the two LDA classifiers. Finally, in (Wei et al., 2009) the problem was addressed within a support vector machine (SVM) framework, with pronunciation space models to improve performance. In short, each phone was modeled with several parallel acoustic models to represent pronunciation variations of that phone at different proficiency levels which helped the system outperform the traditional posterior probability based methods.

3.2 Other approaches

In the previous section, the methods that were mostly discussed were based purely on classification. This section deals with approaches that are of a somewhat different nature. For example, one may find a series of pronunciation evaluation approaches designed for a specific first language (L1) and a certain L2. In (Kawai and Hirose, 1998), recognition results were combined with knowledge of phonetics, phonology and pedagogy (for a certain L1) to show to L2 learners which phones were mispronounced. In (Moustroufas and Digalakis, 2007), the authors developed a method that uses characteristics of the L1 of the speakers to build a system that evaluates utterances without any previous linguistic knowledge of the content.

Research on pronunciation assessment has revealed that the quality of the pronunciation ratings may be affected by several aspects of speech characteristics (Anderson-Hsieh et al., 1992). Non-native speech can deviate from native speech in e.g., fluency, syllable structure, word stress, intonation, prosody and segmental quality. In (Delmonte, 2000), a prosodic module of a computer-assisted language learning (CALL) system called SLIM, was presented to improve the L2 learners pronunciation by dealing with phonetic and prosodic problems at word and segmental level. Prosodic scoring techniques have been described in (Yamashita et al., 2005), in which a multiple regression model was employed to predict the prosodic proficiency of L2 learners using new prosodic measures that were based on F0, power and duration, of L2 and L1 speech. A fluency rating experiment was conducted in (Cucchiarini et al., 2000). Roughly speaking, a series of scoring measures was tested on 60 non-native speakers of Dutch, namely the rate of speech, the phonation/time ratio, the articulation rate, the number of silent pauses, the total duration and the mean length of pauses, the mean length of runs (i.e., the average number of phonemes occurring between unfilled pauses of no less than 0.2 s), the amount of filled pauses and finally, the number of dysfluencies. The results showed that expert ratings of fluency in read speech were reliable and that the automatic measure of speech rate was a good predictor of the human judgment. Later in (Cucchiarini et al., 2002), the authors performed two experiments to explore the relationship between objective properties of speech and perceived fluency in read and spontaneous speech.

In (Raux and Kawahara, 2002), an effort was described to incorporate the intelligibility of L2 learners into a diagnostic system on mispronunciations. The authors deduced a probabilistic algorithm to derive intelligibility from error rates and also defined a function of error priority to indicate which errors were most critical to intelligibility.

In (Xu et al., 2009), an approach was developed for pronunciation error detection that uses linguistic knowledge, obtained from non-native speakers' common mistakes, and pronunciation space constructed by using revised log-posterior probability vectors. An SVM classifier was then applied for the pronunciation error detection of L2 learners of Mandarin Chinese.

For a good overview of various PED approaches, the interested reader is encouraged to read (Eskenazi, 1996) and (Witt, 2012). Before presenting the thesis contribution in PED, some examples of different PED approaches in existing pronunciation training and language learning systems are given, as it is interesting for the future development of the proposed approach to examine how this integration becomes possible in real systems. The following section presents three such cases.

3.3 PED paradigms in real systems

Several computer-assisted pronunciation training (CAPT) systems have been developed to be used for foreign language learning and practising. One such system is the FLUENCY (Eskenazi and Hansma, 1998) which utilizes the SPHINX II speech recognition system (Ravishankar, 1996) to detect L2 mispronunciations. The FLUENCY system aims to help non-native speakers to improve their pronunciations by practising and interacting with the system. Detected mispronunciations are analyzed and feedback is sent to the learner. The system was initially built for English as a target language, but it may be adjusted for other languages. As mentioned, FLUENCY uses a speech recognition system to detect L2 mispronunciations concentrating on errors in duration (Eskenazi and Hansma, 1998) and later on phonetic errors (Eskenazi et al., 2000). For each phone or word that the user produces, FLUENCY performs an evaluation of the duration of the utterance according to previously trained duration patterns and sends relevant feedback to the learner, i.e., whether the phone was shorter or longer than the native pattern. Additionally, the system performs a phonetic error detection by comparing the non-native speech with trained native phone models. Its goal is not only to show the location of the errors but additionally to give the user the possibility to hear the target phone inside context and also instructions on how to place the articulators so as to improve performance. In (Probst et al., 2002), an effort was made to use more than one native voice in order to find the 'golden speaker' (the closest possible to one's own voice based on several features) which could be imitated by the user in order to improve his/her pronunciation skills.

A system called ISLE (Menzel et al., 2000) is founded on a large set of question-answering exercises in which the learner's response is constructed from small sets of pre-specified building blocks. It includes a speech recognizer (Morton et al., 1999) that handles the received utterance with the objective to localize potential pronunciation errors. For each phone or whole word, a confidence score is computed based upon three quantities, i.e., the acoustic likelihood of the recognized path, the output probability of the most likely state in the model set and the acoustic likelihood of the background model. The examined phone or word is then considered to be mispronounced if its score violates a defined threshold. ISLE's PED module includes both *localization*, i.e., identification of the area of an utterance that contains an error, and *diagnosis* of an error, i.e., detection of a mispronunciation on the phone-level.

Another interesting system is EduSpeak® (Franco et al., 2010) which is a toolkit that can be used for language learning and pronunciation practising which utilizes ASR technology together with PED software. The system identifies the mispronunciations and provides feedback on the overall quality of the learner's production capacity. EduSpeak® supports a phone-level PED functionality that captures the problematic phone segments. The toolkit, based on the Decipher™ large-vocabulary, speaker-independent continuous speech recognition system (Digalakis et al., 1996) consists of acoustic models adjusted for L2 speech recognition, pronunciation scoring algorithms, and other modules. The pronunciation evaluation part includes several, text-independent algorithms (Franco et al., 1997; Neumeyer et al., 2000; Neumeyer et al., 1996), i.e., without the pronunciation scoring being tailored according to specific contents. These scoring measures, described in the beginning of Sec. 3.1, are spectral, durational or prosodic and are combined in a nonlinear fashion to estimate the general pronunciation grade, simulating essentially the judgment that expert tutors would have given if a listening test would have been available. In EduSpeak®, the relationship between the machine-driven and the human assessment is established using large-scale data of non-native speech that have been scored by human experts on the sentence level (Bernstein et al., 1990; Franco et al., 2000). In (Franco et al., 2010), the authors introduced an additional word duration score to evaluate the L2 pronunciations. This new score is used either alone or in combination with the already existing ones. The PED approach that is used in EduSpeak® is built on HMM-based ASR that trains acoustic models of both native and non-native speech and then, using statistical models, compares various features of L1 and L2 utterances and creates a set of pronunciation scores for different phonetic segments. In the end, all the automatic scores are combined to account for the overall evaluation score.

So far, the discussion concentrated around the different approaches that have been followed for the automatic mispronunciation detection task. The next paragraph presents the contribution of this thesis in diagnosing problematic phonemes for this task.

3.4 Native perceptual assessment degree

The methods proposed in this thesis concentrate on the speech sound level, and describe an attempt to mathematically formulate the native speakers' ability to distinguish the non-native pronunciation with an objective to develop an automatic diagnostic evaluation scheme for pronunciation error detection. Common in all cases is the underlying idea that is based on the description given in Sec. 1.3, which compares the Euclidean space geometry of the auditory perceptual domain with the Euclidean space geometry of the acoustic domain. In this case, the goal is not to select or optimize the speech representation properties according to auditory perception as it was with the AMFS for speech recognition (cf. Sec. 2.3). On the contrary, the objective is to examine to what extent a non-native feature set has similar properties – meaningful for human perception – as the native speech signal representation. The speech sounds of a specific phoneme class that are produced by native speakers are generally without much discrepancy

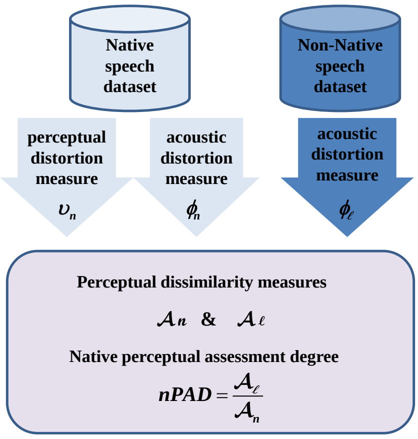

Native speech dataset
Non-Native speech dataset
perceptual distortion measure
υ_n
acoustic distortion measure
ϕ_n
acoustic distortion measure
ϕ_ℓ
Perceptual dissimilarity measures
$\mathcal{A}_n$ & $\mathcal{A}_\ell$
Native perceptual assessment degree
$nPAD = \frac{\mathcal{A}_\ell}{\mathcal{A}_n}$

Figure 12: Applying dissimilarity criteria between the native speech perceptual domain and the non-native speech acoustic domain and comparing with the native only case, the L2 mispronunciations can be detected.

as the corresponding speech sounds produced by non-native speakers in the event of occurring mispronunciations. In terms of ASR, the native speech signal can, in this case, be considered as *clean* speech and the non-native as *noisy* speech. The approach followed in Paper D, measures the dissimilarity of the local geometries between the acoustic and the auditory domain for each phoneme class, considering first the native speech and then the non-native speech. In the end, the method quantifies to what extent the non-native dissimilarity differs from that of the native, considering that the larger this difference is, the more problematic, for the L2 speakers, the examined phoneme is.

The above idea is modified in Paper E and further extended in Paper F, to directly measure the dissimilarity between the non-native acoustic domain and the native auditory model-output domain. Fig. 12 shows a block diagram of the methods' scheme. An HMM-based aligner (Sjölander, 2003) generates a phone-level transcription from the speech signal and the text file, which separates the native speech stimuli into phoneme categories. For each considered phoneme class, the native signal is transformed into the auditory domain and the non-native signal into the acoustic domain. The spatial dissimilarity

$$\mathcal{A}_\ell = \frac{1}{\mathcal{I}} \sum_{i\in\mathcal{I}} \frac{1}{\mathcal{J}_i} \sum_{j\in\mathcal{J}_i} \left[\upsilon_n(\mathbf{x}_{n_i}, \hat{\mathbf{x}}_{n_{i,j}}) - \phi_\ell(\mathbf{x}_{\ell_i}, \hat{\mathbf{x}}_{\ell_{i,j}})\right]^2, \tag{29}$$

between these two domains is measured to investigate, quantitatively, to what extent the non-native acoustic feature distortion measure $\phi_\ell(\mathbf{x}_{\ell_i}, \hat{\mathbf{x}}_{\ell_{i,j}})$ differs from the native perceptual distortion measure $\upsilon_n(\mathbf{x}_{n_i}, \hat{\mathbf{x}}_{n_{i,j}})$. The native stimuli $\mathbf{x}_n$ is then transformed into the acoustic domain to calculate the following dissimilarity measure

$$\mathcal{A}_n = \frac{1}{\mathcal{I}} \sum_{i \in \mathcal{I}} \frac{1}{\mathcal{J}_i} \sum_{j \in \mathcal{J}_i} \left[\upsilon_n(\mathbf{x}_{n_i}, \hat{\mathbf{x}}_{n_{i,j}}) - \phi_n(\mathbf{x}_{n_i}, \hat{\mathbf{x}}_{n_{i,j}}) \right]^2, \tag{30}$$

between the native perceptual and acoustic distortion domains, and the two measures $\mathcal{A}_n$ and $\mathcal{A}_\ell$ are compared to identify the most problematic phonemes. The perceptual-based method ends by considering for each phoneme class, the native-perceptual assessment degree $nPAD$ that is computed for every L1 background as

$$nPAD = \frac{\mathcal{A}_\ell}{\mathcal{A}_n}, \tag{31}$$

considering the most mispronounced phonemes to have the largest $nPAD$ values. Eq. (31) is a normalized ratio that shows the degree of the dissimilarity between the native perceptual outcome (using a spectral or a spectro-temporal auditory model) and the non-native acoustic speech signal representations (the magnitude spectrum of speech or its acoustic static and dynamic features, depending on the application, as shown in Part II) as compared to the native-only case. That is, nPAD quantifies the non-native speaker's additional pronunciation variation for each phoneme, which is meaningful for human speech perception.

nPAD is one way to detect problematic L2 phonemes aiming at profiting from recent findings in auditory research but at the same time avoiding to increase the computational complexity. Depending on the available data, nPAD could be measured using two different groups of L1 speakers to investigate the variety within the native pronunciations. Alternative methods would, e.g., transform both L1 and L2 speech signals into the auditory-model output domain to find dissimilarities or appropriately format the output for phoneme classification.

Evaluation of the method

For the experiments, a speech database was recorded which includes data from 37 (23 male and 14 female) non-native speakers from eleven different language backgrounds and also recordings from 11 (9 male and 2 female) native Swedish speakers. The stimuli consists of 23 phonetically rich single words and 55 sentences of varying complexity and length. A more detailed description of the corpus can be found in Papers D and E.

The evaluation of the automatic method can be done in various ways. One such option is by setting up a listening experiment to associate human assessment with the score of the automatic method. There are however many obstacles to follow this procedure. It is, for example, particularly difficult for a human listener to remain focused on a certain phone when hearing a whole sentence or, even impossible, to listen only to one specific phone as this would mean just a few milliseconds of speech. In addition, even if the above problem could be solved, it would still be difficult to perform a complete test with all samples of the L2 phonemes from all the examined language groups of non-native speakers[4]. Alternatively, the automatic method can be compared against a theoretical linguistic study performed for the same or similar task. Hence in the work described in this thesis, nPAD is juxtaposed against the findings of Bannert (1984). The objective for the refereed study was to be able to identify and analyze common errors made in Swedish by adult immigrants of various language backgrounds so as to improve their pronunciation performance. The study was performed through linguistic analysis and subjective observations and the informants were chosen for speaking Swedish with a foreign accent that would demonstrate their L1 backgrounds. Bannert did not use any objective measure to evaluate the mispronunciations, however in his work, which is an extensive, long-term survey in second language research, various cases of errors are examined from a linguistic perspective.

As the nPAD method is quantitative, with phonemes being ordered according to their nPAD values, it is important to set an objective threshold by which the problematic phonemes can be distinguished. This threshold is defined using the equal error rate (EER) so that the probabilities of false acceptance (mispronounced phonemes according to (Bannert, 1984) accepted by the automatic method as correctly produced) and false rejection (non-problematic phonemes according to (Bannert, 1984) judged by the computational methods to have been mispronounced) to become equal.

Considering all language groups of speakers, the EER value for the spectral nPAD version (using van de Par model and the speech power spectrum) is 41% at a threshold value of $T_1 \approx 1.0005$, and for the spectro-temporal nPAD version (using Dau model and the MFCCs) it is 31% with a threshold value at $T_2 \approx 1.0000$. The value of the EER is high for the spectral nPAD and significantly reduces when the method additionally includes temporal information of the speech signals. However, the EER remains relatively high. There are several reasons that contribute to this behavior. Bannert (1984) examines the task from a broader angle compared to nPAD, including in his analysis various linguistic and phonological factors such as coarticula-

[4] An intermediate solution has been followed in Papers E and F in which the most problematic phonemes according to the spectro-temporal automatic evaluation were tested as well as one phoneme for each group that, according to Bannert (1984), would have caused seriously mispronunciation problems but for which nPAD indicated no error. A description of the test and a further discussion can be found in Papers E and F and in Appendix IV.

tion, epenthesis, prosody, which contribute to a different methodology. His study considers L2 speakers who have a strong accent, representative for their language background, and who speak Swedish based purely on their own capacity. The subjects used in the work described in this thesis did on the other hand repeat each utterance directly after a virtual language tutor. In addition, it is important to note at this point that the nPAD is used in this thesis with the objective to, first, introduce a novel idea of estimating L2 mispronunciations based only on models of auditory perception and perturbation analysis of the speech signal and, second, perform a diagnostic evaluation of the pronunciation of L2 speakers showing an ordered list of problematic phonemes that can later be used for mispronunciation analysis. On this account, the comparison of the proposed method with (Bannert, 1984) is done with a scope to investigate if the new scheme is suited for diagnosing most of the mispronunciations which are generally considered in the literature. Furthermore, the evaluation of the nPAD with teachers of Swedish as a foreign language resulted in an agreement between the human and the automatic assessment of 86.5% for the spectro-temporal nPAD and 65.4% for the spectral nPAD method as described further in Papers E and F. The aforementioned evaluation is based on a relatively small number of phonemes and can hence not be used to perform a complete comparison but only to verify some important findings of nPAD. It is noted that the choice of the tested phonemes where mostly based on the spectro-temporal method as this version achieves better experimental results.

Another way of comparing the quantitative and the theoretical findings is to consider the same number of mispronounced phonemes per language as Bannert found, and check for potential mismatches between the two lists. The latter method is selected for the experiment described in the next section because it is more suitable for comparisons between the examined measures.

The benefit of auditory knowledge

Intuitively arises the question whether the auditory input is fruitful for the system to perform an automatic pronunciation evaluation. In order to investigate this query, a measure $\mathcal{B}_\ell$ is considered (Koniaris et al., 2012) that only evaluates the Euclidean geometric similarities as

$$\mathcal{B}_\ell = \frac{1}{\mathcal{I}} \sum_{i \in \mathcal{I}} \frac{1}{\mathcal{J}_i} \sum_{j \in \mathcal{J}_i} \left[\phi_n(\mathbf{x}_{n_i}, \hat{\mathbf{x}}_{n_{i,j}}) - \phi_\ell(\mathbf{x}_{\ell_i}, \hat{\mathbf{x}}_{\ell_{i,j}}) \right]^2 , \tag{32}$$

between the native and the non-native power spectrum distortion. With this definition, $\mathcal{B}_n$ is zero.

Tables 1 and 2 list the vowels and the consonants, respectively, found to be difficult for the different groups of non-native speakers according to the

Table 1: Problematic vowels per language background. The quantitative results show phonemes in decreasing order, starting from the one with the highest spectral $nPAD$ or $\mathcal{B}_\ell$. Phonemes that differ from the linguistic evaluation are listed in parentheses. The theoretical results are not ordered. Seriously mispronounced vowels are bold marked, and vowels that are not captured by the automatic evaluation are underlined.

non-native speakers L1 background		automatic evaluation (quantitative results)	linguistic study (Bannert, 1984) (theoretical results)
English (US)	$nPAD$	æː, ɛ, yː, uː, ʊ, œː, ɛː, ø, ɵ, øː, (iː), ɑː, (ə), eː, e, ɔ, a, ʉː	**a**, **eː**, **e**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **ɛː**, **ɛ**, **æː**, **øː**, **ø**, **œː**, ɑː, ʊ, ɔ
	$\mathcal{B}_\ell$	ø, ɛ, ɑː, æː, a, œː, ɛː, e, øː, eː, ʏ, (i), ɔ, (iː), (ə), ʉː, yː, ɵ	**a**, **eː**, **e**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **ɛː**, **ɛ**, **æː**, **øː**, **ø**, **œː**, ɑː, ʊ, ɔ
German	$nPAD$	æː, (ɛ), yː, uː, (ʊ), ɛː, (ø), œː, øː, iː, ə, ɑː	**ɑː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, ə, ʏ, ɛː, æː, øː, œː
	$\mathcal{B}_\ell$	(ø), æː, ɑː, (ɛ), œː, (a), øː, ɛː, (e), ɵ, ʏ, (i)	**ɑː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, ə, ʏ, ɛː, æː, øː, œː
French	$nPAD$	æː, ɛ, yː, uː, œː, (ʊ), ɛː, (ø), ɵ, øː, iː, ə, ɑː, eː, e, ɔ, (a)	**ɑː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **oː**, **æː**, eː, e, ə, ʏ, ɔ, ɛː, ɛ, øː, œː
	$\mathcal{B}_\ell$	(ø), ɛ, ɑː, æː, (a), œː, ɛː, e, øː, eː, ʏ, (i), ɔ, iː, ə, ʉː, yː	**ɑː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **oː**, **æː**, eː, e, ə, ʏ, ɔ, ɛː, ɛ, øː, œː
Polish	$nPAD$	æː, (ɛ), yː, uː, œː, (ʊ), ɛː, ø, ɵ, øː, iː, ɑː, ə, eː, (e), (ɔ)	**ɑː**, **eː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **ɛː**, **æː**, **øː**, **ø**, a, ə, œː
	$\mathcal{B}_\ell$	ø, (ɛ), ɑː, æː, a, œː, ɛː, (e), øː, eː, ʏ, (i), (ɔ), iː, ə, ʉː	**ɑː**, **eː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **ɛː**, **æː**, **øː**, **ø**, a, ə, œː
Russian	$nPAD$	ɛ, yː, ɛː, ø, øː, iː, ɑː, ɵ, eː, e, ʉː, a, (ɔ), ʏ, (ə), (i), œː	**ɑː**, **eː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **ɛː**, **æː**, **øː**, **ø**, a, e, oː, ɛ, œː
	$\mathcal{B}_\ell$	ø, œː, ɛː, øː, ʏ, ɑː, a, e, ɛ, æː, eː, ʉː, uː, yː, oː, iː, ɵ	**ɑː**, **eː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **ɛː**, **æː**, **øː**, **ø**, a, e, oː, ɛ, œː
Greek	$nPAD$	æː, (ɛ), yː, uː, œː, (ʊ), ɛː, ø, ɵ, øː, iː, ɑː, ə, eː, e, (ɔ)	**ɑː**, **eː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **ɛː**, **øː**, **ø**, e, ə, æː, œː
	$\mathcal{B}_\ell$	ø, (ɛ), ɑː, æː, (a), œː, ɛː, e, øː, eː, ʏ, (i), (ɔ), iː, ə, ʉː	**ɑː**, **eː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **ɛː**, **øː**, **ø**, e, ə, æː, œː
Spanish	$nPAD$	uː, æː, ɛ, ø, iː, ɵ, eː, e, øː, (a), ʉː, ɑː, (i), ə, ʏ, (ɔ), oː	**ɑː**, **eː**, **e**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **ɛː**, **ɛ**, **øː**, **ø**, **œː**, ə, æː
	$\mathcal{B}_\ell$	uː, æː, yː, ø, (a), œː, e, (ʊ), ɛ, oː, ɛː, eː, (i), iː, ʏ, ʉː, øː	**ɑː**, **eː**, **e**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **ɛː**, **ɛ**, **øː**, **ø**, **œː**, ə, æː
Turkish	$nPAD$	æː, (ɛ), yː, (ɛː), (ø), uː, ɵ, ʊ, øː, iː, ɑː, eː, (ə)	**eː**, **e**, **iː**, **uː**, **ʊ**, **ʉː**, **ɵ**, **yː**, **oː**, **æː**, **øː**, ɑː, œː
	$\mathcal{B}_\ell$	(ø), (ɛ), æː, (a), ɑː, œː, (ɛː), øː, e, eː, (ʏ), (i), (ɔ)	**eː**, **e**, **iː**, **uː**, **ʊ**, **ʉː**, **ɵ**, **yː**, **oː**, **æː**, **øː**, ɑː, œː
Arabic	$nPAD$	æː, ɛ, yː, uː, œː, (ʊ), ɛː, ø, ɵ, øː, iː, ɑː, ə, eː, e, (ɔ), a, ʉː	**ɑː**, **eː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **ɛː**, **æː**, **øː**, **ø**, a, e, ə, ɛ, œː
	$\mathcal{B}_\ell$	ø, ɛ, ɑː, æː, a, œː, ɛː, e, øː, eː, ʏ, (i), (ɔ), iː, ə, ʉː, yː, oː	**ɑː**, **eː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **ɛː**, **æː**, **øː**, **ø**, a, e, ə, ɛ, œː
Chinese	$nPAD$	ɵ, æː, ɛ, yː, uː, ɛː, ø, øː, iː, ɑː, eː, e, ɔ, (ə), a, ʉː, (i), oː	**a**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **øː**, ɑː, eː, e, iː, ɔ, ɛː, ɛ, æː, ø, œː
	$\mathcal{B}_\ell$	ɵ, ʏ, ø, ɑː, æː, ɛ, œː, a, e, ɛː, øː, (i), eː, ɔ, iː, ʉː, oː, yː	**a**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **øː**, ɑː, eː, e, iː, ɔ, ɛː, ɛ, æː, ø, œː
Persian	$nPAD$	ɵ, æː, yː, (ɛ), ø, øː, ʏ, (i), ʉː, eː, a, (e), (ɑː), oː, (ɔ)	**a**, **eː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **ɛː**, **æː**, **øː**, **ø**, **œː**, ə
	$\mathcal{B}_\ell$	ɵ, æː, yː, oː, ə, a, (ʊ), iː, (e), (ɔ), ø, (ɑː), (ɛ), eː, uː	**a**, **eː**, **iː**, **uː**, **ʉː**, **ɵ**, **yː**, **ʏ**, **oː**, **ɛː**, **æː**, **øː**, **ø**, **œː**, ə

Table 2: Problematic consonants per language background. The quantitative results show phonemes in decreasing order, starting from the one with the highest spectral $nPAD$ or $\mathcal{B}_\ell$. Phonemes that differ from the linguistic evaluation are listed in parentheses. The theoretical results are not ordered. Seriously mispronounced consonants are bold marked, and consonants that are not captured by the automatic evaluation are underlined.

non-native speakers L1 background		automatic evaluation (quantitative results)	linguistic study (Bannert, 1984) (theoretical results)
English	$nPAD$	ɧ, ŋ, (v), m, n, (b), r, (d), l, k, ʂ, t	**l**, **ŋ**, **r**, **s̲**, **ɧ**, **ɕ̲**, k, m, n, ʂ, t, ʈ̲
(US)	$\mathcal{B}_\ell$	ʂ, ɕ, (d), s, (b), (j), ɧ, (g), (h), r, t, ŋ	**l̲**, **ŋ**, **r**, **s**, **ɧ**, **ɕ**, k̲, m̲, n̲, ʂ, t, ʈ̲
German	$nPAD$	ɧ, ŋ, v, n, (m), b, r, d, (l), k, ʂ, t, p, (h), f, ɕ, s	**b**, **d**, **g̲**, **ŋ**, **r**, **s**, **ɧ**, **ʂ**, **ʈ̲**, **v**, ɕ, f, j̲, k, n, p, t
	$\mathcal{B}_\ell$	ɕ, ʂ, s, t, k, g, (l), (h), ɧ, r, d, b, f, j, v, n, ŋ	**b**, **d**, **g**, **ŋ**, **r**, **s**, **ɧ**, **ʂ**, **ʈ̲**, **v**, ɕ, f, j, k, n, p̲, t
French	$nPAD$	ŋ, ɧ, (v), m, n, b, r, (l), d, ʂ, k, t, p, h, ɕ, g	**b**, **d**, **g**, **k**, **ŋ**, **p**, **r**, **t**, **ɕ**, ɧ, h, m, n, s̲, ʂ, ʈ̲
	$\mathcal{B}_\ell$	ɕ, ʂ, s, k, t, (l), g, h, (j), r, b, d, ɧ, (f), p, (v)	**b**, **d**, **g**, **k**, **ŋ̲**, **p**, **r**, **t**, **ɕ**, ɧ, h, m̲, n̲, s, ʂ, ʈ̲
Polish	$nPAD$	ɧ, ŋ, v, (m), n, b, (r), d, (l), k, ʂ, t, p, s, (f)	**b**, **d**, **g̲**, **h̲**, **k**, **ŋ**, **p**, **s**, **ɧ**, **ʂ**, **t**, **v**, ɕ̲, n, ʈ̲
	$\mathcal{B}_\ell$	ɕ, ʂ, s, t, k, g, (l), ɧ, h, (j), (r), b, d, (f), v	**b**, **d**, **g**, **h**, **k**, **ŋ̲**, **p̲**, **s**, **ɧ**, **ʂ**, **t**, **v**, ɕ, n̲, ʈ̲
Russian	$nPAD$	v, ŋ, (m), (n), (r), d, (l), h, b, k, t, g, (s), f, j	**b**, **d**, **g**, **k**, **ŋ**, **p̲**, **ɧ̲**, **ʂ̲**, **t**, **v**, ɕ̲, f, h, j, ʈ̲
	$\mathcal{B}_\ell$	ɕ, ɧ, ʂ, g, (s), k, b, ŋ, h, f, t, (l), (r), (m), j	**b**, **d̲**, **g**, **k**, **ŋ**, **p̲**, **ɧ**, **ʂ**, **t**, **v̲**, ɕ, f, h, j, ʈ̲
Greek	$nPAD$	ɧ, ŋ, (v), m, n, b, (r), d, l, ʂ, k, t, p, ɕ, (f), s	**b**, **d**, **g̲**, **h̲**, **k**, **n**, **ŋ**, **p**, **s**, **ɧ**, **ʂ**, **t**, **ɕ**, l, m, ʈ̲
	$\mathcal{B}_\ell$	ɕ, ʂ, s, t, k, g, l, ɧ, (j), (r), b, d, h, (f), (v), n	**b**, **d**, **g**, **h**, **k**, **n**, **ŋ̲**, **p̲**, **s**, **ɧ**, **ʂ**, **t**, **ɕ**, l, m̲, ʈ̲
Spanish	$nPAD$	ŋ, v, (r), n, (l), b, t, ʂ, (f), k, g, d, j, p, ɕ, s, h	**b**, **d**, **g**, **h**, **k**, **ŋ**, **p**, **s**, **ɧ̲**, **ʂ**, **t**, **ɕ**, **v**, j, m̲, n, ʈ̲
	$\mathcal{B}_\ell$	ɧ, ɕ, s, h, ʂ, (l), b, ʈ, (r), (f), k, d, p, t, j, g, ŋ	**b**, **d**, **g**, **h**, **k**, **ŋ**, **p**, **s**, **ɧ**, **ʂ**, **t**, **ɕ**, **v̲**, j, m̲, n̲, ʈ
Turkish	$nPAD$	ŋ, v, (m), n, b, r, l, d, k, (ʂ), t, p, f, h, g, ʈ, j, s	**b**, **d**, **g**, **k**, **n**, **ŋ**, **p**, **ɧ̲**, **t**, **ʈ**, **ɕ̲**, **v**, f, h, j, l, r, s
	$\mathcal{B}_\ell$	ɕ, (ʂ), t, ɧ, k, s, l, g, h, r, j, b, d, f, v, n, p, ŋ	**b**, **d**, **g**, **k**, **n**, **ŋ**, **p**, **ɧ**, **t**, **ʈ̲**, **ɕ**, **v**, f, h, j, l, r, s
Arabic	$nPAD$	ɧ, ŋ, v, (m), (n), (b), r, d, (l), k, ʂ, t, p	**f̲**, **k**, **ŋ**, **p**, **r**, **s̲**, **ɧ**, **ʂ**, **t**, **ɕ̲**, **v**, d, ʈ̲
	$\mathcal{B}_\ell$	ɕ, ʂ, s, t, k, (g), (l), ɧ, (h), (j), r, (b), d	**f̲**, **k**, **ŋ̲**, **p̲**, **r**, **s**, **ɧ**, **ʂ**, **t**, **ɕ**, **v̲**, d, ʈ̲
Chinese	$nPAD$	ɧ, ŋ, v, m, n, b, r, l, d, k, t, f, g, ʈ, p, j, (h), (s)	**j**, **k**, **l**, **n**, **ŋ**, **p**, **r**, **ɧ**, **ʂ̲**, **t**, **ɕ̲**, **v**, b, d, f, g, m, ʈ
	$\mathcal{B}_\ell$	ɕ, ʂ, (s), d, b, t, k, ɧ, g, (h), j, l, r, ʈ, f, v, n, p	**j**, **k**, **l**, **n**, **ŋ̲**, **p**, **r**, **ɧ**, **ʂ**, **t**, **ɕ**, **v**, b, d, f, g, m̲, ʈ
Persian	$nPAD$	b, d, (ɧ), v, (f), g, (h), t, s, (j), ɕ, p, ʈ, k, l, r	**k**, **n̲**, **ŋ̲**, **p**, **s**, **ʂ̲**, **t**, **ʈ**, **ɕ**, **v**, b, d, g, l, m̲, r
	$\mathcal{B}_\ell$	ɕ, ʂ, b, (ɧ), s, g, (h), k, t, r, d, (j), (f), l, v, ʈ	**k**, **n̲**, **ŋ̲**, **p̲**, **s**, **ʂ**, **t**, **ʈ**, **ɕ**, **v**, b, d, g, l, m̲, r

two considered evaluation methods described by Eq. (31) and Eq. (32) (to the left) and the previous linguistic observations (Bannert, 1984) (to the right). For each L2 speaker group, the first line to the left shows, in order, the most deviating phonemes according to the spectral nPAD method (described in Paper E). Correspondingly, the second line to the left shows the evaluation of the spectral dissimilarity measure $\mathcal{B}_\ell$. The experimental results are ordered, beginning with the phonemes that are the most problematic for each L2 speaker group. The corresponding theoretical list to the right is not ordered, since no quantitative measure was used. For the scope of this evaluation, the mispronounced phonemes found in (Bannert, 1984) have been divided into two categories. The first is the *seriously problematic* phonemes (marked in bold in the list to the right), i.e., that are totally mispronounced under all circumstances, while the second category consists of the *problematic* phonemes that are mispronounced in special cases or lying in the edge of being characterized as native-like according to Bannert (1984). From the tables, it can be seen that in most cases the results of the perceptual-based method are in better agreement with previous linguistic observations (Bannert, 1984). Divergences from the theoretical findings are reported in parentheses (false rejections) and underlying (false accepts). It can be seen that the majority of the parentheses are located on the right side, meaning that most of the false rejects are for lower nPAD values. In general, this also holds for $\mathcal{B}_\ell$, but it is worth mentioning that for three language groups the first vowels are misdetections according to Bannert (1984). Comparing the algorithms, it can be seen that, generally speaking, for as many as eight language groups, i.e., *English US, German, French, Polish, Greek, Turkish, Arabic,* and *Persian* the nPAD method gets quantitatively better results. For three language groups, namely *Russian, Spanish,* and *Chinese*, the results of the perceptual-based method become worse in comparison to the spectral dissimilarity measure. The perceptual-based method not only performs better in terms of a lower number of false rejections compared to the linguistic study, but also in terms of detection of seriously mispronounced phonemes. In short, the nPAD method has one mismatch less for the German and Arabic speakers, two less for French, Greek and Turkish speakers and three less for the English speakers. In addition, concentrating mainly on the seriously problematic phonemes, nPAD captures one more seriously problematic phoneme for German, Polish, Greek, Arabic and Persian speakers, two more for French and three for Turkish speakers. The spectral dissimilarity measure has one mismatch less for Chinese speakers and four for Russian speakers. Finally, it captures one more seriously problematic phoneme for English, Spanish and Chinese speakers and two more for Russian speakers. To summarize, the nPAD method has in total 69 false rejects (34 for vowels and 35 for consonants) and the spectral measure 75 (37 for vowels and 38 for consonants) out of 350 mispronunciations listed in (Bannert, 1984), i.e., 19.7% for the nPAD method and 21.4% for the spectral $\mathcal{B}_\ell$. Finally, the nPAD method has 49 missed seriously problematic phonemes (27 for vowels and 22 for consonants) and the spectral measure 54 (34 for vowels and 20 for consonants) out of 245 in total, i.e., 20.0% for the nPAD method and 22.0% for the spectral $\mathcal{B}_\ell$. This type of evaluation gives the opportunity to look into some quality aspects of the findings, e.g., the number of seriously mispronounced phonemes that have not been captured by the automatic evaluation.

The quantitative difference in the total numbers of mismatches for the two methods is arguably small. However, the qualitative differences are more important. Tables 1 and 2 reveal some important weaknesses of the measure $\mathcal{B}_\ell$ in identifying major mispronunciations for some of the most problematic phonemes. In most cases for example, it does not recognize the Swedish long rounded vowel /uː/ as problematic, except for the Spanish group. The linguistic findings have shown that all foreign groups mispronounce the Swedish /uː/ because they produce it either as a short vowel or with inadequate lip rounding. The nPAD measure on the other hand is not only able to capture this vowel, but additionally to rank it high on the problematic vowels list. In most cases, except for Chinese and Persian speakers, the spectral distortion measure $\mathcal{B}_\ell$ further fails in detecting the vowel /ɵ/, which is one of the seriously problematic Swedish vowels for non-native speakers, often confused by /ʊ/ (Bannert, 1984). Additionally, the measure $\mathcal{B}_\ell$ misses the vowel /yː/, which is produced with protruded instead of compressed lips, and according to Bannert (1984), is mainly confused with the short unrounded /i/. On the contrary, the nPAD measure succeeds to detect both the aforementioned vowels and indeed to classify them among the most problematic vowels.

There are several consonants that appear to be problematic for many foreign speakers, four of which are in particular difficult for almost all of the language groups. The velar nasal /ŋ/ is one of the consonants that most speakers are inclined to mispronounce. In (Bannert, 1984) it is noted that it is often replaced by /ŋg/. Table 2 show that the nPAD method can better detect this error than the spectral distortion measure. Additionally, the Swedish consonant /v/, which is very often mispronounced by non-native speakers either as /f/, /b/ or /w/, is also detected by the nPAD measure. The Swedish fricatives /ɧ/ and /ç/ are probably the most difficult Swedish consonants for non-native speakers due to their uniqueness and their large variety depending on the neighboring sounds. The nPAD approach is more capable in capturing the problems related to the /ɧ/, while the spectral distortion measure is more sensible in detecting errors only related to /ç/.

This comparatively better capacity of the nPAD measure to detect problematic phonemes has presumably its roots in the information that the auditory model provides through the sensitivity matrix. It seems that the small distances in the power spectrum between the native and the non-native speech signals become clearer in the auditory domain where only the perceptually relevant elements of the two spectrums are considered. This enriches the nPAD method's potential to identify the meaningful details that reveal the pronunciation divergences between native and non-native speakers. In addition, it is worth noting that when computing the total value of $\mathcal{B}_\ell$ for all of the problematic consonants for each language, the part that corresponds to the value of the first one is very high. In other words, the method has limited capability to detect the problematic consonants in general and can mainly focus on the detection of the most mispronounced consonant.

Even though the performance of nPAD is better than that of the spectral evaluation, it appears to have some disagreements with the theoretical study concerning the problematic phonemes. This can be explained by the nature and the context of the data since the recordings were made with the subjects repeating, in two sessions, text after a natively speaking virtual language tutor. Hence, it is likely that the speakers have avoided some otherwise occurring mispronunciations, that usually accompanying spontaneous speech.

A more analytic and detailed description of the work introduced in this chapter can be found in the Part II of the thesis.

4 Summary of contributions

This thesis makes the following major contributions:

- Two novel methods to select conventional acoustic features for speech recognition based on the knowledge of human perception (Papers A and C).
- An optimization and design of improved MFCCs using a spectral psychoacoustic auditory model for speech recognition (Paper B).
- A method to automatically and quantitatively measure the perceptually relevant differences between native and non-native speakers in distinguishing the target phonemes (Paper D).
- Three general, novel methods for automatic diagnostic assessment of the pronunciation of individual non-native speakers based on models of the human auditory periphery (Papers E and F).

This work is described in more detail in six original research papers that are included in Part II. The initial concept in Papers A, B and C comes from Bastiaan Kleijn who had the overall supervision and helped with the writing of the papers. In Paper A, the author did the theoretical derivations of the **A** matrix (see Appendix II), the implementation of the method and conducted all the experiments. Marcin Kuropatwinski helped with the van de Par model and the algorithm. The author together with Bastiaan Kleijn wrote Paper A. In Paper B, the author did the word recognition experiments, provided the **A** matrix and the van de Par model, and helped with the writing of the paper. The main contributor of Paper B is Saikat Chatterjee who implemented the method, did the phone recognition experiments, and wrote the major part of it. In Paper C the author did the implementation of the algorithm and the experiments and wrote the major part of the paper. Saikat Chatterjee helped with the algorithm. The main concept in Papers D, E and F comes from the author who is also the principal contributor. In Paper D, the author performed the implementation of the algorithm, did the experiments and wrote the major part of it. Olov Engwall supervised the research effort and helped with the writing of the paper. In Paper E, the author proposed and implemented the algorithms, derived the $\mathbf{W}_\ell$ matrix that linearizes the non-native and native speech signals, performed the experiments and wrote the major part of the paper. Giampiero Salvi and Olov Engwall provided valuable feedback concerning the structure of the paper. In addition, Giampiero Salvi verified the mathematical validity of the proposed methods and Olov Engwall helped with the writing of the paper. Finally, in Paper F the author proposed and implemented the algorithm, performed the listening test and wrote the major part of the paper. Olov Engwall provided consultation on the setup of the listening test and helped with the paper's writing. Giampiero Salvi commented on the manuscript. A short summary of each paper is presented below.

Paper A: Auditory-Model Based Robust Feature Selection for Speech Recognition

We show that robust feature selection for speech recognition can be based on a model of the human auditory system. Our approach is fundamentally different from the established selection methods: instead of optimizing classification performance, we exploit knowledge implicit in the human auditory system to select good features. The method finds the acoustic feature set that maximizes the similarity of the Euclidian geometry of the feature domain and the perceptual domain, as represented by an auditory model. As only small distances are critical for correct sound discrimination, we use a perturbation analysis for the selection process. Using a static auditory model and static features, experiments with a practical speech recognizer confirm that the human auditory system can be used for feature selection. The results are robust and generalize to unseen environmental conditions.

Paper B: Auditory Model Based Optimization of MFCCs Improves Automatic Speech Recognition Performance

We use a spectral auditory model and perturbation analysis to develop a new framework to optimize a set of features for speech recognition. The proposed framework tries to reflect the way human perception performs recognition. The optimization of the features is done offline based on the assumption that the local geometries of the feature domain and the perceptual auditory domain should be similar. In our effort to modify and optimize the static mel frequency cepstrum coefficients (MFCCs), no feedback from the speech recognition system was used. The results show improvement in speech recognition accuracy under all environmental conditions, clean and noisy.

Paper C: Selecting Static and Dynamic Features Using an Advanced Auditory Model for Speech Recognition

We extend our previous work in feature selection for speech recognition exploiting a sophisticated quantitative model of the human auditory periphery. Motivated by the success of the method proposed in Paper A, we expand the system in two ways: we use a spectro-temporal auditory model to include the effect of time-domain masking, and consider the first and second order time derivatives in the feature selection algorithm. The new selected subsets consist of features able to capture their time dependencies in a more efficient way. In parallel, the method remains still independent of the automatic speech recognizer. The experimental results show a significantly better performance of the extended selection algorithm compared to discriminant analysis.

Paper D: Perceptual Differentiation Modeling Explains Phoneme Mispronunciation by Non-Native Speakers

Influenced by the underlying idea of comparing the geometry between auditory and acoustic signal domains and motivated by human sound perception, we investigate the similarity of the Euclidean space spanned by the power spectrum of a native speech signal and the Euclidean space spanned by the auditory model output, for a certain phoneme category. We then repeat the procedure, this time considering only non-native speech signals from second language learners. Comparing the two similarity measurements, we find problematic phonemes for a given set of speakers. The method, which is general, totally automatic and quantitative, is applied to different groups of non-native speakers of various language backgrounds and compared to theoretical linguistic findings. The results are promising as they are in accordance with theory.

Paper E: On Mispronunciation Analysis of Individual Foreign Speakers Using Auditory Periphery Models

We introduce two general, automatic, diagnostic, pronunciation evaluation methods of non-native speakers based on models of the human auditory system. Both approaches are based on one of the major difficulties in second language learning, namely the discrimination between the acoustic diversity within an L2 phoneme category and between different categories. We model the native perception by measuring the geometric shape similarity between the native auditory and acoustic representation domains. Then for each phoneme, we compare the native perception with the dissimilarities found between the Euclidean geometry of the native perceptual domain and the non-native acoustic domain. Our approaches are tested with different non-native speaker groups from various language backgrounds. The experimental results are in agreement with linguistic findings, particularly when the spectro-temporal cues of the speech signal are considered, instead of simply its spectral characteristics.

Paper F: Auditory and Dynamic Modeling Paradigms to Detect L2 Mispronunciations

We expand the work presented in Paper E on automatic pronunciation error detection by using the state dynamic representations of a trained linear dynamic model with native speech stimuli. In addition, we perform a pronunciation analysis by considering the task of mispronunciation detection as a classification problem. A linear dynamic model is employed to model the native speech phoneme classes and then data from non-native speech are classified according to the trained models. Finally, we evaluate the proposed methods with a listening test on the same speech material and compare the results with the automatic methods. The spectro-temporal approach is found to have the best agreement with the human evaluation, particularly with that of teachers of Swedish as a foreign language, and with previous exhaustive linguistic studies.

5 Conclusions and future work

The goal of the first half of this thesis was to investigate the use of auditory modeling in the front-end of an ASR system. The proposed methods incorporated a combination of knowledge from models of the human auditory periphery, speech signal processing, perturbation analysis techniques and acoustic modeling. It is concluded that the selection or optimization of speech features based on human perception results in robust features that generalize well to various environmental conditions. Furthermore, the successful experimental results can be considered as the "proof" of the underlying assumption that the output of the auditory system is useful for increasing the accuracy of the modern speech recognition engines.

The second half of this thesis was directed on investigating the problem of automatic pronunciation error detection from a fundamentally different perspective than conventional methods. Exploiting the underlying idea utilized in the speech recognition part, the proposed methods performed a diagnostic evaluation of the most problematic speech sounds for various groups of non-native speakers. The conclusion is that the utilization of auditory models and the use of perturbation theory can lead to results that are generally close to linguistic studies. A listening test with native listeners, experts and non-experts, on the same data as the automatic methods strengthened the claim that the method is both efficient and valid.

The research presented in this thesis is somewhat unconventional and the two areas of study, the speech recognition and the mispronunciation detection, are treated differently compared to the common approaches. This suggests that there is room for further development in both fields. Hence for the feature selection for ASR applications, the method can be modified to account for other type of acoustic features or even for a combination of different sets of features. In this case, the method can be used as it currently is or it could be converted to fully benefit from the speech recognizer by, e.g., combining AMFS with HLDA to first select the optimal feature subset (out of a large feature vector) and then apply the HLDA to receive feedback from the recognizer and find a set of transformed features to maximize the performance of the system. Moreover, the information from the sophisticated (and enriched with recent findings in hearing technology) auditory models could also be applied to other type of features to develop new robust, optimized features as it was the case with the MMFCCs.

For the automatic assessment of the non-native pronunciation, an interesting extension could be to optimally integrate such a module into a computer-assisted pronunciation training program to perform an online mispronunciation detection. Before that, a further evaluation of the proposed methods would be necessary. For this, the methods have to be tested on larger datasets so as to obtain an adequate number of utterances from all the groups of speakers and for all the target phonemes. Furthermore, a listening test with more native experts should be considered, to evaluate all the target phonemes instead of only judging the most mispronounced according to the automatic assessment output. The results of this test would then be examined in relation to all the proposed methods and also to state-of-the-art methods for mispronunciation detection, in order to find the most optimal method that would be integrated into a CAPT program. Depending on the application, the methods can easily be adjusted for similar tasks, e.g., for comparison among regional accent within a target language or to additionally account for prosodic features such as syllable length or loudness, and hence to be used for a broader pronunciation and language learning purposes.

Appendix I LDA implementation

In this appendix, the implementation of the LDA is presented (it also concerns the HLDA method since, as it is stated in the related paragraph in Sec. 2.2, the LDA matrix was used to initialize the HLDA algorithm). Before applying the LDA method, the feature extraction and the speech recognition tasks should be performed. In this case, the generated features were the MFCCs (Davis and Mermelstein, 1980). Using the HTK toolkit (Young et al., 2002), the digits were modeled as whole word HMMs with 16 states (HTK's notation is 18 states including the beginning and end states) and three Gaussian mixture components per state with full covariance matrices. An initial model with global data means and covariances, identical for each digit was used, and then 16 iterations were necessary to build the final model. Two recognition tasks were considered. In the first, the training was performed on the clean train set of 8440 sentences and the testing on the 4004 clean data of the so-called AURORA2 Test set A. In the second, the training was performed on the multi-conditioning noisy train set consisting of 6752 files and the testing on the 24024 noisy data of the AURORA2 Test set A.

Statistics computation

Using again the HTK toolkit, a master label file was created by reading through the MFCCs and the HMMs that were trained during the recognition stage. A short sample of the master label file is

```
"MAE_12A.lab"
0          1000000    sil_s2    sil
1000000    1900000    sil_s3
1900000    2000000    sil_s4
2000000    2100000    one_s2    one
2100000    2200000    one_s3
..............................
```

where the `.lab` is the file's name and the numbers represent the start and end times in 100 ns units. Next, new label files for each word-state were created followed by start and end points of each occurrence of this class, containing all of its different realizations in the database. For example, the file for the word-state `eight_s2` (referring to the word *eight* at HMM state 2) that includes the filename, followed by start and end point of each occurrence of the word-state is as follows

```
"MAJ_1978213A.lab"
11300000    11800000
```

```
"MAJ_4487A.lab"
6200000    6300000
"MAW_2568Z23A.lab"
13100000    13800000
..............................
```

Then, the word-class label files accompanied by the MFCCs were read serially, and the class and the overall data statistics $\mu_j, \mathbf{\Sigma}_j, \mu, \mathbf{\Sigma}$ were computed, respectively. The procedure started by reading a label file (e.g., the $eight\_s2$ as mentioned above) and by opening the MFCC file that was named first (in the aforementioned label file). In each iteration, one frame is read for each sample vector according to the time indices specified in the label file. A context size $C = 5$, defined in (Kumar, 1997; Kumar and Andreou, 1998) as the number of feature vectors before and after the current feature vector that are used to incorporate dynamic information, was considered. When the reading of all frames had finished, the next MFCC file was considered and the procedure continued with all the MFCCs that included tokens of the considered word-class. The number of tokens in each class as well as the total number of tokens counted. Thereafter, the next word-class label file considered and the same procedure was repeated. Both the means of each class and the whole database were calculated after reading through all the data once. To compute the covariance matrices $\mathbf{\Sigma}_j$ and $\mathbf{\Sigma}$, a second run through the whole corpus was necessary, because the mean vectors, indispensable for the computation, were not available during the first run.

Transformation computation

At the end, as the statistics to compute the optimization criterion Eq. (17) were finally known, i.e., both the within-class and total scatter matrices, the LDA transform was computed by accumulating the eigenvectors of $\mathbf{S}^{-1}\mathbf{\Sigma}$ in a matrix that corresponds to the p largest eigenvalues. The output is the transformation matrix ϕ^T.

New LDA representations

The new - reduced in size - representations of the original MFCC features were extracted in the last stage of the process. The procedure was similar as in the first stage, when the label files were read one after the other, but the difference was that, this time, no computation was performed. The tokens were just read in, multiplied by the ϕ^T, and written to a new feature file with the same name. To ensure that the files were stored in a “HTK-friendly” format, the function `writehtk.m` from the VOICEBOX toolbox (VOICEBOX, 1999) was used. The new transformed features were then used as input to HTK and new HMM models were trained. Then, the recognizer used the transformed test data to complete the word recognition task.

Discussion

The performance of the LDA features (Papers A and C) although reasonable in clean conditions, was not very promising when noisy conditions were considered. Apart from the straightforward reason of the presence of the noise per se, a possible explanation of this behavior is the computation of a global LDA transformation which, for the *multi-to-multi* case, is trying to compensate noises of subway, babble, car, and exhibition in several SNR values of $20, 15, 10, 5, 0$ and -5 dB. Naturally, this transformation considers all the different noisy aspects of noise type and noise level and leads in a general transformation ϕ^T. On the other hand, if someone would try to have a separate transformation for each individual case, a single ϕ^T should be computed for each one of the 4 noise types and for each of the 6 noise levels leading to a total number of 24 different transformation matrices for each experiment i.e., for every reduced feature subspace. Note also that this approach does not guarantee a better performance of the analysis. On the other hand, for the case of *clean-to-clean* no such phenomenon occured since all the data were clean, and hence the transformation was computed based on a homoeomorphous data set.

Appendix II Derivation of the A matrix

In this appendix, the derivation of the **A** matrix is shown both in spectral and time domains. The linearized relation between a small distortion in the speech frame $\delta\mathbf{x} = \hat{\mathbf{x}} - \mathbf{x}$ and the corresponded distortion in MFCCs $\delta\mathbf{c} = \hat{\mathbf{c}} - \mathbf{c}$ is described by Eq. (28) or to write it in a simpler way, $\delta\mathbf{c} \approx \mathbf{A}\,\delta\mathbf{x}$. The formula for the computation of MFCCs can be written in different ways by simply following the steps of computing MFCCs, beginning from the latest. Hence,

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \mathbf{s}(m) \cos\left\{ q[m + \frac{1}{2}] \frac{\pi}{M} \right\}, q = 1...Q, \tag{33}$$

where Q is the number of cepstrum coefficients, and $\mathbf{s}(m)$ represents the logarithmic mel spectrum of the m'th filter of the filterbank or

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \ln \mathbf{z}(m) \cos\left\{ q[m + \frac{1}{2}] \frac{\pi}{M} \right\}, \tag{34}$$

where $\mathbf{z}(m)$ is the product of the power spectrum with the triangular mel weighted filters or

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \ln\left\{ \sum_{k=0}^{K-1} \mathbf{Y}(k)\mathbf{H}_m(k) \right\} \cos\left\{ q[m + \frac{1}{2}] \frac{\pi}{M} \right\}, \tag{35}$$

where $\mathbf{Y}(k)$ is the periodogram, $\mathbf{H}_m(k)$ is the m'th triangular mel-filter or

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \ln\left\{ \sum_{k=0}^{K-1} |\mathbf{X}(k)|^2 \mathbf{H}_m(k) \right\} \cos\left\{ q[m + \frac{1}{2}] \frac{\pi}{M} \right\}, \tag{36}$$

in which $\mathbf{X}(k)$ denotes the DFT of the signal or finally as

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \ln\left\{ \sum_{k=0}^{K-1} \left| \sum_{n=0}^{N-1} \mathbf{x}'(n) e^{-\frac{j2\pi kn}{N}} \right|^2 \mathbf{H}_m(k) \right\} \cos\left\{ q[m + \frac{1}{2}] \frac{\pi}{M} \right\}, \tag{37}$$

where $\mathbf{x}'(n)$ is the windowed speech frame and $\mathbf{x}(n)$ the pre-emphasized speech block. From the above, **A** can be calculated as the product of the following derivatives

$$\mathbf{A}(q, n) = \frac{\partial \mathbf{c}(q)}{\partial \mathbf{s}(m)} \frac{\partial \mathbf{s}(m)}{\partial \mathbf{z}(m)} \frac{\partial \mathbf{z}(m)}{\partial \mathbf{Y}(k)} \frac{\partial \mathbf{Y}(k)}{\partial \mathbf{x}'(n)} \frac{\partial \mathbf{x}'(n)}{\partial \mathbf{x}(n)}. \tag{38}$$

In Paper A, the **A** matrix is shown in frequency domain. This covers the first three derivatives in Eq. (38). In this paragraph, it is further found the corresponding formula for in time domain. For the fourth factor, it can be shown that the periodogram $\mathbf{Y}(k)$ is given by

$$\mathbf{Y}(k) = \sum_{n=0}^{N-1} \mathbf{x}'^2(n) + 2\sum_{n=0}^{N-2}\sum_{m=n+1}^{N-1} \mathbf{x}'(n)\mathbf{x}'(m)\cos\left\{\frac{2\pi k}{N}[n-m]\right\}. \tag{39}$$

Then, its derivative $\frac{\partial \mathbf{Y}(k)}{\partial \mathbf{x}'(n)}$, i.e., the derivative of the periodogram with respect to the windowed signal is

$$\frac{\partial \mathbf{Y}(k)}{\partial \mathbf{x}'(n)} = 2\mathbf{x}'(n) + 2\sum_{\substack{h=0,\\ h\neq n}}^{N-1} \mathbf{x}'(h)\cos\left\{\frac{2\pi k}{N}[n-h]\right\}. \tag{40}$$

One can see that

$$\begin{aligned}
\frac{\partial \mathbf{Y}(k)}{\partial \mathbf{x}'(n)} = 2\mathbf{x}'(n) + 2\sum_{\substack{h=0,\\ h\neq n}}^{N-1} \mathbf{x}'(h)\cos\left\{\frac{2\pi k}{N}[n-h]\right\} &= \\
2\sum_{h=0}^{N-1} \mathbf{x}'(h)\cos\left\{\frac{2\pi k}{N}[n-h]\right\} &= \\
2\Re\left\{\sum_{h=0}^{N-1} \mathbf{x}'(h)\left\{\cos\left\{\frac{2\pi k}{N}[n-h]\right\} + j\sin\left\{\frac{2\pi k}{N}[n-h]\right\}\right\}\right\} &= \\
2\Re\left\{\sum_{h=0}^{N-1} \mathbf{x}'(h)e^{j\frac{2\pi k}{N}[n-h]}\right\} &= \\
2\Re\left\{\left\{\sum_{h=0}^{N-1} \mathbf{x}'(h)e^{j\frac{2\pi k}{N}h}\right\}e^{-j\frac{2\pi k}{N}n}\right\} &= \\
2\Re\left\{\mathbf{X}^*(k)e^{-j\frac{2\pi k}{N}n}\right\}&, \qquad (41)
\end{aligned}$$

where $\mathbf{X}^*(k)$ is the conjugate of the DFT of the signal. Finally, the formula of matrix $\mathbf{A}$ in time domain is given by

$$\mathbf{A}_{qn} = \sum_{m=0}^{M-1}\cos\left\{q[m+\frac{1}{2}]\frac{\pi}{M}\right\}\frac{1}{\mathbf{z}(m)}\mathbf{H}_m(n)2\Re\left\{\mathbf{X}^*(k)e^{-j\frac{2\pi k}{N}n}\right\}\mathbf{w}(n), \tag{42}$$

where $\mathbf{w}(n)$ is the hamming window.

Appendix III Linear dynamic model implementation

In Paper F, the linear dynamic model (LDM) - that was used for mispronunciation detection and analysis - was outlined. In the first case of mispronunciation detection, the model was used so as to include additional knowledge from classified data, aiming at enriching the nPAD system's acoustic information. In the second case of analysis, the LDM was used as an ordinary classifier aiming not to compare with nPAD, but to test the ability of the model in analyzing the detected mispronunciation. In the following, the implementation of the model is described.

Training phoneme models

For each phoneme class, the mel cepstra from native speech $\mathbf{c}_n$ were extracted to be used as the observation vector for the LDM described by the following pair of equations

$$\mathbf{x}_{n_{k+1}} = \mathbf{F}\mathbf{x}_{n_k} + \mathbf{w}_k \tag{43}$$

$$\mathbf{c}_{n_k} = \mathbf{H}\mathbf{x}_{n_k} + \mathbf{u}_k, \tag{44}$$

where $\mathbf{x}_{n_k}$, $\mathbf{c}_{n_k}$ are the state and the observation vectors at the time frame k, respectively, and $\mathbf{w}_k, \mathbf{u}_k$ are uncorrelated, zero-mean Gaussian vectors with covariances $E\{\mathbf{w}_k\mathbf{w}_l^T\} = \mathbf{P}\delta_{kl}$ and $E\{\mathbf{u}_k\mathbf{u}_l^T\} = \mathbf{R}\delta_{kl}$. The initial state $\mathbf{x}_{n_0}$ was Gaussian with known mean $\mu_{\mathbf{x}_{n_0}}$ and covariance $\Sigma_{\mathbf{x}_{n_0}}$. Eq. (43) describes the state dynamics, and Eq. (44) gives an observation prediction based on the state estimation. The size of the state-space (state vectors) was considered equal to the size of the observation vectors. The initial state-transition matrices, and the covariance of the initial state $\mathbf{x}_{n_0}$ were directly estimated from the observation vectors. The noise covariance matrices of the system were randomly initialized. The parameters of the LDM $\theta = \{\mathbf{F}, \mathbf{H}, \mathbf{P}, \mathbf{R}\}$ were calculated using an Expectation-Maximization (EM) based algorithm introduced in (Digalakis, 1992; Digalakis et al., 1993), according to which their estimated values can be computed as

$$\hat{\mathbf{F}} = \Gamma_4\Gamma_3^{-1} \tag{45}$$

$$\hat{\mathbf{H}} = \Gamma_6\Gamma_1^{-1} \tag{46}$$

$$\hat{\mathbf{P}} = \Gamma_2 - \Gamma_4\Gamma_3^{-1}\Gamma_4^T = \Gamma_2 - \hat{\mathbf{F}}\Gamma_4^T \tag{47}$$

$$\hat{\mathbf{R}} = \Gamma_5 - \Gamma_6\Gamma_1^{-1}\Gamma_6^T = \Gamma_5 - \hat{\mathbf{H}}\Gamma_6^T \tag{48}$$

where

$$\begin{aligned}\Gamma_1 &= \frac{1}{N+1}\sum_{k=0}^{N}\mathbf{x}_{n_k}\mathbf{x}_{n_k}^T && (49)\\ \Gamma_2 &= \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_{n_k}\mathbf{x}_{n_k}^T && (50)\\ \Gamma_3 &= \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_{n_{k-1}}\mathbf{x}_{n_{k-1}}^T && (51)\end{aligned}$$

$$\begin{aligned}\Gamma_4 &= \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_{n_k}\mathbf{x}_{n_{k-1}}^T && (52)\\ \Gamma_5 &= \frac{1}{N+1}\sum_{k=0}^{N}\mathbf{c}_{n_k}\mathbf{c}_{n_k}^T && (53)\\ \Gamma_6 &= \frac{1}{N+1}\sum_{k=0}^{N}\mathbf{c}_{n_k}\mathbf{x}_{n_k}^T. && (54)\end{aligned}$$

The above sufficient statistics required the following quantities at iteration p:

$$\begin{aligned}E_{\theta^{(p)}}\{\mathbf{c}_{n_k}\mathbf{x}_{n_k}^T|\mathbf{Y}\} &= \mathbf{c}_{n_k}\hat{\mathbf{x}}_{n_{k|N}} && (55)\\ E_{\theta^{(p)}}\{\mathbf{c}_{n_k}\mathbf{c}_{n_k}^T|\mathbf{Y}\} &= \mathbf{c}_{n_k}\mathbf{c}_{n_k}^T && (56)\\ E_{\theta^{(p)}}\{\mathbf{x}_{n_k}\mathbf{x}_{n_{k-1}}^T|\mathbf{Y}\} &= \mathbf{\Sigma}_{k,k-1|N} + \hat{\mathbf{x}}_{n_{k|N}}\hat{\mathbf{x}}_{n_{k-1|N}}^T && (57)\\ E_{\theta^{(p)}}\{\mathbf{x}_{n_k}\mathbf{x}_{n_k}^T|\mathbf{Y}\} &= \mathbf{\Sigma}_{k|N} + \hat{\mathbf{x}}_{n_{k|N}}\hat{\mathbf{x}}_{n_{k|N}}^T. && (58)\end{aligned}$$

At each EM iteration (5 in total), the sufficient statistics described previously were computed using the fixed interval smoothing form of the Kalman filter (RTS smoother) (Rauch et al., 1965) (it consisted of a backward pass that followed the standard Kalman filter forward recursions (Kalman, 1960)) and the old estimates of the model parameters. Then during the maximization step, the new estimates were obtained from these statistics and the Eqs. (45)-(48). Moreover, the cross covariances were computed in both the forward and backward pass (Digalakis, 1992; Digalakis et al., 1993). Since the estimates of the parameters of the state linear equation are mutually dependent, it was considered an additional iterative estimation process based on the same sufficient statistics obtained in the expectation step. This iterative process was terminated when a predefined small threshold (equal to 0.001) of the distance $d(\theta,\hat{\theta}) = \frac{\|\theta-\hat{\theta}\|}{\|\theta\|}$ between two successive estimates θ and $\hat{\theta}$ was reached. All the necessary recursions of the forward - backward from the Kalman smoother are shown in the next page.

Forward Recursions

$$\hat{\mathbf{x}}_{n_{k|k}} = \hat{\mathbf{x}}_{n_{k|k-1}} + \mathbf{K}_k \mathbf{e}_k \tag{59}$$

$$\hat{\mathbf{x}}_{n_{k+1|k}} = \mathbf{F}\hat{\mathbf{x}}_{n_{k|k}} \tag{60}$$

$$\mathbf{e}_k = \mathbf{c}_{n_k} - \mathbf{H}\hat{\mathbf{x}}_{n_{k|k-1}} \tag{61}$$

$$\mathbf{K}_k = \boldsymbol{\Sigma}_{k|k-1}\mathbf{H}^T\boldsymbol{\Sigma}_{\mathbf{e}_k}^{-1} \tag{62}$$

$$\boldsymbol{\Sigma}_{\mathbf{e}_k} = \mathbf{H}\boldsymbol{\Sigma}_{k|k-1}\mathbf{H}^T + \mathbf{R} \tag{63}$$

$$\boldsymbol{\Sigma}_{k|k} = \boldsymbol{\Sigma}_{k|k-1} - \mathbf{K}_k\boldsymbol{\Sigma}_{e_k}\mathbf{K}_k^T \tag{64}$$

$$\boldsymbol{\Sigma}_{k,k-1|k} = (\mathbf{I} - \mathbf{K}_k\mathbf{H})\mathbf{F}\boldsymbol{\Sigma}_{k-1|k-1} \tag{65}$$

$$\boldsymbol{\Sigma}_{k+1|k} = \mathbf{F}\boldsymbol{\Sigma}_{k|k}\mathbf{F}^T + \mathbf{P} \tag{66}$$

Backward Recursions

$$\hat{\mathbf{x}}_{n_{k-1|N}} = \hat{\mathbf{x}}_{n_{k-1|k-1}} + \mathbf{A}_k[\hat{\mathbf{x}}_{n_{k|N}} - \hat{\mathbf{x}}_{n_{k|k-1}}] \tag{67}$$

$$\boldsymbol{\Sigma}_{k-1|N} = \boldsymbol{\Sigma}_{k-1|k-1} + \mathbf{A}_k[\boldsymbol{\Sigma}_{k|N} - \boldsymbol{\Sigma}_{k|k-1}]\mathbf{A}_k^T \tag{68}$$

$$\mathbf{A}_k = \boldsymbol{\Sigma}_{k-1|k-1}\mathbf{F}^T\boldsymbol{\Sigma}_{k|k-1}^{-1} \tag{69}$$

$$\boldsymbol{\Sigma}_{k,k-1|N} = \boldsymbol{\Sigma}_{k,k-1|k} + [\boldsymbol{\Sigma}_{k|N} - \boldsymbol{\Sigma}_{k|k}]\boldsymbol{\Sigma}_{k|k}^{-1}\boldsymbol{\Sigma}_{k,k-1|k} \tag{70}$$

The estimated model parameters were then used for two purposes. Firstly, to compute the state vectors that served as the acoustic input for the hybrid nPAD method, and secondly, to proceed to the classification of non-native speech and through this, to the mispronunciation analysis presented in Paper F. The classification stage is described in the next paragraph.

Classification of the non-native observations

It was then used the trained phoneme-models to classify the non-native speech features[5]. Hence, the classification process began by considering the non-native signal representations $\mathbf{c}_\ell$ of a specific phoneme class and the parameters of each phoneme category θ that were previously estimated during the training stage using the native speech signal. For example, consider the model parameters for the Swedish phoneme /n/. The system was initialized by the model parameters $\theta_{/n/}$ and the MFCCs from a specific language group ℓ were first considered. The data were classified according to the

[5] In the case of LDM pronunciation analysis (Paper F), two different acoustic features were used, namely the MMFCCs and the MFCCs. As the extraction of both features are based on the same process, what is here referring to MFCCs was also applied for the MMFCCs.

value of the log-likelihood

$$L(\mathbf{C}_\ell, \theta) = -\sum_{k=0}^{\mathcal{I}} \left\{ \log |\mathbf{\Sigma}_{\mathbf{e}_k}(\theta)| + \mathbf{e}_k^T(\theta) \mathbf{\Sigma}_{\mathbf{e}_k}^{-1}(\theta) \mathbf{e}_k(\theta) \right\}, \qquad (71)$$

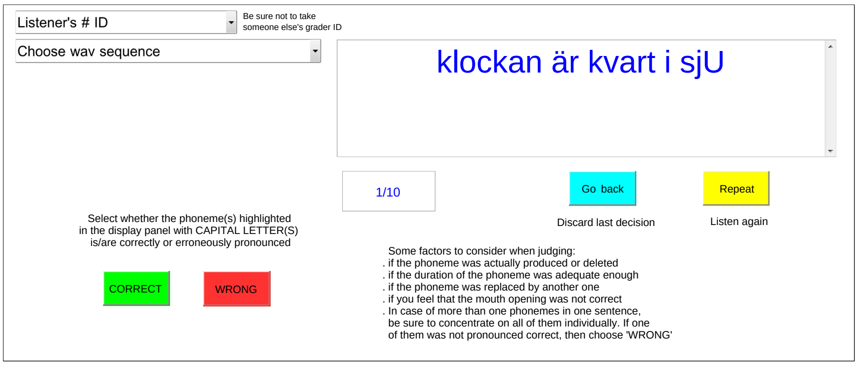
where $\mathbf{e}_k(\theta)$ is the prediction error and $\mathbf{\Sigma}_{\mathbf{e}_k}(\theta)$ its covariance, obtained from the linear quadratic estimation.

Appendix IV The listening test setup

This appendix section presents the setup of the listening test, described in Paper F, so as to evaluate the proposed automatic methods for the scope of pronunciation error detection. It should be noted that performing such a listening test was not a straightforward assignment, since the proposed methods were evaluated for each L2 phoneme on the segment level, while any extra non-speech signal information was discarded. Therefore, the simulation of such a process by human evaluation involvement was difficult. Ideally, the test should let the listeners hear the phonemes, which each file consists of, individually. However, this would likely lead the experiment in failure as it is humanly impossible to judge such a short, often unclear, sound. Hence, it was decided to let the subjects listen to entire utterances and request them to pay attention to a specific phoneme at each time. Another problem, which needed to be solved, was the total amount of the speech material that each native listener would have to listen to, since the total amount of data (even though not large for computer programs) exceeded the physical limitation of a listening test with humans. It was hence decided to consider the phonemes which were found to be the most problematic according to the spectro-temporal nPAD Ξ_ℓ and the hybrid nPAD Ξ_{h_ℓ} methods for each language group of L2 speakers as well as one phoneme for each group that, according to Bannert (1984), would have caused seriously mispronunciation problems but for which both nPAD methods or at least one of them, indicated no error.

Eight native listeners were chosen to participate in the test. Three of them were teachers of Swedish as a foreign language, qualified with great experience from students from different language backgrounds and hence highly sensitive in identifying pronunciation errors. Those were the 'experts' group. The remaining five listeners were well-educated native speakers from different areas of Sweden which, for the purpose of the experiment, were considered as the 'naive' listeners group[6]. All listeners were seated in front of a screen displaying the graphical user interface of a computer program, and were asked to evaluate (accept or reject) if a specific target phoneme was pronounced natively within various contexts (words or sentences). Fig. 13 shows an example of the graphical environment used. Each listener was associated with a numerical identity before starting to listen through all of the considered material that was divided into 52 categories depending on the target phoneme and the L2 speaker group. The content of all speech files was displayed on the screen with the target phoneme shown in uppercase letter independently of its position in the sentence. In case there were more than one instance of the considered phoneme, the listener was asked to focus on each one of them and press the correct button only in case all the

[6] Simply because they are not teachers of Swedish as a foreign language.



Be sure not to take
someone else's grader ID

Select whether the phoneme(s) highlighted
in the display panel with CAPITAL LETTER(S)
is/are correctly or erroneously pronounced

Some factors to consider when judging:
. if the phoneme was actually produced or deleted
. if the duration of the phoneme was adequate enough
. if the phoneme was replaced by another one
. if you feel that the mouth opening was not correct
. In case of more than one phonemes in one sentence,
be sure to concentrate on all of them individually. If one
of them was not pronounced correct, then choose 'WRONG'

Figure 13: The listening's test graphical interface. Here an example of the phrase "klockan är kvart i sjU" (the time is a quarter to seven). The target phoneme is the short U (/ɵ/ in the IPA notation) in the word 'sjU', displayed in uppercase.

target phonemes were pronounced natively. It is important to note that the listener was encouraged to concentrate only on the target phoneme and try to reduce the influence of other components, e.g., fluency of the L2 speaker in Swedish. Explicit instructions were also written in the graphical interface to consider factors such as the duration of the phoneme or if the phoneme was deleted by the speaker or even replaced by another one. To avoid prejudge and reduce biased decisions, the information regarding the language background of the L2 speakers was not available to the listeners.

The duration of the experiment varied significantly, from 35 minutes up to almost 120 minutes, though the majority completed the test in about 50 minutes. For many listeners, it was necessary to repeatedly listen to most of the phrases, as it was difficult for them to judge any potential error immediately from the first hearing. At this point, it is interesting to concentrate on the participants' feedback concerning the experiment. It was widely admitted that the task was difficult since humans usually combine a variety of factors to judge the pronunciation of L2 speakers. The concentration on a specific phoneme, especially on consonants for which the duration is usually shorter than that of the vowels, was not an easy assignment, especially for the 'naive' group. For the 'experts', the daily experience with foreign students of Swedish has helped them to develop the ability to identify relatively easily pronunciation difficulties and weaknesses. This is probably the basic reason that explains both the agreement within the 'expert' and with the automatic methods, which by nature, are based on small geometric distortion differences in order to assess the utterances. On the other hand, the 'naive' group appeared to be more divided and sometimes incapable to capture all the errors. The participants declared that they mainly judged based on their own way of pronouncing Swedish and factors, such as dialect origin and subjectiveness, were at the top of their evaluation criteria in lieu of the standard, dialect-free, official Swedish pronunciation.

Appendix V International phonetic alphabet for Swedish

Tables 3-4 present the International Phonetic Alphabet (IPA) for Swedish vowels and consonants, respectively, that has been used in the work presented in this thesis. The phoneme symbols are shown to the left accompanied by an example word in Swedish, and its translation in English to the right.

Table 3: List of IPA symbols for Swedish vowels.

IPA symbol	Example
/ɑː/	gl**a**s *(glass)*
/a/	k**a**n *(can)*
/eː/	park**e**ra *(park)*
/e/	d**e**mokrati *(democracy)*
/ɛː/	v**ä**g *(road)*
/ɛ/	t**ä**ndsticksask *(matchbox)*
/æː/	h**ä**r *(here)*
/æ/	n**ä**rmast *(nearest)*
/iː/	v**i**la *(rest)*
/i/	f**i**nna *(find)*
/uː/	g**o**d *(good)*
/ʊ/	t**o**mat *(tomato)*
/ʉː/	f**u**l *(ugly)*
/ɵ/	**u**pp *(up)*
/yː/	b**y** *(village)*
/ʏ/	k**y**kling *(chicken)*
/oː/	fr**å**n *(from)*
/ɔ/	l**å**ng *(long)*
/øː/	k**ö**pa *(buy)*
/ø/	mj**ö**lk *(milk)*
/œː/	f**ö**r *(for)*
/œ/	f**ö**rstå*(understand)*
/ə/	ring**e**r *(call)*

Table 4: List of IPA symbols for Swedish consonants.

IPA symbol	Example
/b/	**b**etala *(pay)*
/d/	**d**yr *(expensive)*
/ɖ/	bo**rd**e *(should)*
/ɡ/	**g**las *(glass)*
/k/	å**k**a *(go; drive)*
/ŋ/	ri**ng**a *(ring)*
/p/	**p**otatis *(potato)*
/h/	**h**a *(have)*
/l/	vi**l**a *(rest)*
/s/	**s**tuga *(cottage)*
/ɧ/	**sj**u *(seven)*
/ʂ/	tandbo**rs**te *(toothbrush)*
/ɕ/	**tj**ock *(thick)*
/ɭ/	hä**rl**ig *(lovely)*
/ʈ/	sna**rt** *(soon)*
/v/	**v**inna *(win)*
/n/	**n**är *(when)*
/m/	**m**iddag *(dinner; noon)*
/ɳ/	gä**rn**a *(willingly)*
/t/	**t**a *(take)*
/j/	gre**j** *(thing)*
/f/	**f**laska *(bottle)*
/r/	va**r**a *(be)*

Bibliography

Abbasian, H., Nasersharif, B., Akbari, A., Rahmani, M., and Moin, M. S. (2008). Optimized linear discriminant analysis for extracting robust speech features. In *Int. Symp. on Comm., Control, Sig. Proc. (ISCCSP), St. Julians, Malta*, pages 819–824.

Anderson-Hsieh, J., Johnson, R., and Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, 42(4):529–555.

Aubert, X., Haeb-Umbach, R., and Ney, H. (1993). Continuous mixture densities and linear discriminant analysis for improved context-dependent acoustic models. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Minneapolis, MN, USA*, volume 2, pages 648–651.

Bahl, L. R., Jelinek, F., and Mercer, R. L. (1990). *A maximum likelihood approach to continuous speech recognition. In Readings in speech recognition.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Bannert, R. (1984). Problems in learning Swedish pronunciation and in understanding foreign accent. *Folia Linguistica*, 18(1-2):193–222.

Bellman, R. E. (1957). *Dynamic Programming.* Princeton University Press, Princeton, NJ, USA, Republished Dover 2003, ISBN: 0-486-42809-5.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., and Weintraub, M. (1990). Automatic evaluation and training in english pronunciation. In *ISCA Int. Conf. on Spoken Lang. Proc., Kobe, Japan*, pages 1185–1188.

Black, A. W., Taylor, P., and Caley, R. (2001). The Festival Speech Sythesis System. `http://www.festvox.org/festival/`.

Blauert, J. (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, Cambridge, MA, USA, ISBN: 0-262-02413-6.

Bocchieri, E. (1993). Vector quantization for the efficient computation of continuous density likelihoods. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Minneapolis, MN, USA*, volume 2, pages 692–694.

Bocchieri, E. and Doddington, G. (1986). Frame-specific statistical features for speaker-independent speech recognition. *IEEE Trans. Acoust. Speech and Sig. Proc.*, 34(4):755–764.

Breebaart, J., van de Par, S., and Kohlrausch, A. (2001a). Binaural processing model based on contralateral inhibition. I. model structure. *J. Acoust. Soc. Am.*, 110(2):1074–1088.

Breebaart, J., van de Par, S., and Kohlrausch, A. (2001b). Binaural processing model based on contralateral inhibition. II. dependence on spectral parameters. *J. Acoust. Soc. Am.*, 110(2):1089–1104.

Breebaart, J., van de Par, S., and Kohlrausch, A. (2001c). Binaural processing model based on contralateral inhibition. III. dependence on temporal parameters. *J. Acoust. Soc. Am.*, 110(2):1105–1117.

Brown, P. F. (1987). *The Acoustic-Modelling Problem in Automatic Speech Recognition.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA.

Buchholz, J. M. and Mourjopoulos, J. (2004a). A computational auditory masking model based on signal dependent compression. I. model description and performance analysis. *Acustica - Acta Acust.*, 90(5):873–886.

Buchholz, J. M. and Mourjopoulos, J. (2004b). A computational auditory masking model based on signal dependent compression. II. model simulations and analytical approximations. *Acustica - Acta Acust.*, 90(5):887–900.

Bush, M. A. and Kopec, G. E. (1987). Network-based connected digit recognition. *IEEE Trans. Acoust. Speech and Sig. Proc.*, 35(10):1401–1413.

Carney, L. H. (1993). A model for the responses of low-frequency auditory-nerve fibers in cat. *J. Acoust. Soc. Am.*, 93(1):401–417.

Chatterjee, S. and Kleijn, W. B. (2011). Auditory model-based design and optimization of feature vectors for automatic speech recognition. *IEEE Tr. Audio, Speech, Lang. Proc.*, 19(6):1813–1825.

Colburn, H. S., Carney, L. H., and Heinz, M. G. (2003). Quantifying the information in auditory-nerve responses for level discrimination. *J. Assoc. Otolaryngol.*, 4(3):294–311.

Colburn, H. S., Han, Y. A., and Cullota, C. P. (1990). Coincidence model of MSO responses. *Hear. Res.*, 49(1-3):335–346.

Cole, R., Mariani, J., Uszkoreit, H., Varile, G. B., Zaenen, A., and Zampolli, A. (1998). *Survey of the state of the Art in Human Language Technology (Studies in Natural Language Processing)*. Cambridge University Press.

Cook, G. and Robinson, T. (1998). Transcribing broadcast news with the 1997 Abbot system. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Seattle, WA, USA*, volume 2, pages 917–920.

Cooke, M., Lecumberri, M. L. G., and Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *J. Acoust. Soc. Am.*, 123(1):414–427.

Cucchiarini, C., Strik, H., and Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *J. Acoust. Soc. Am.*, 107(2):989–999.

Cucchiarini, C., Strik, H., and Boves, L. (2002). Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech. *J. Acoust. Soc. Am.*, 111(6):2862–2873.

Cutler, A., Lecumberri, M. L. G., and Cooke, M. (2008). Consonant identification in noise by native and non-native listeners: Effects of local context. *J. Acoust. Soc. Am.*, 124(2):1264–1268.

Dau, T. (2009). Auditory processing models. In Havelock, D., Kuwano, S., and Vorländer, M., editors, *Handbook of Signal Processing in Acoustics*, pages 175–196. Springer New York, NY, USA.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997a). Modeling auditory processing of amplitude modulation. I. modulation detection and masking with narrowband carriers. *J. Acoust. Soc. Am.*, 102(5):2892–2905.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997b). Modeling auditory processing of amplitude modulation. II. spectral and temporal integration in modulation detection. *J. Acoust. Soc. Am.*, 102(5):2906–2919.

Dau, T., Püschel, D., and Kohlrausch, A. (1996a). A quantitative model of the effective signal processing in the auditory system. I. model structure. *J. Acoust. Soc. Am.*, 99(6):3615–3622.

Dau, T., Püschel, D., and Kohlrausch, A. (1996b). A quantitative model of the effective signal processing in the auditory system. II. simulations and measurements. *J. Acoust. Soc. Am.*, 99(6):3623–3631.

Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Sig. Proc.*, 28(4):357–366.

de Cheveigné, A. (2005). Pitch perception models. In Plack, C., Oxenham, A. J., Popper, A. N., and Fay, R. R., editors, *Pitch: Neural Coding and Perception*, pages 169–233. Springer New York, NY, USA.

Delgutte, B. (1990). Physiological mechanisms of psychophysical masking: observations from auditory-nerve fibers. *J. Acoust. Soc. Am.*, 87(2):791–809.

Delmonte, R. (2000). SLIM prosodic automatic tools for self-learning instruction. *Speech Communication*, 30(2-3):145–166.

Demuynck, K., Duchateau, J., and Compernolle, D. V. (1999). Optimal feature sub-space selection based on discriminant analysis. In *Europ. Conf. Speech Comm. and Tech. (EUROSPEECH), Budapest, Hungary*, pages 1311–1314.

Deng, L. and Ma, J. (1999). A statistical coarticulatory model for the hidden vocal-tract-resonance dynamics. In *Europ. Conf. Speech Comm. and Tech. (EUROSPEECH), Budapest, Hungary*, pages 1499–1502.

Digalakis, V. (1992). *Segment-based stochastic models of spectral dynamics for continuous speech recognition*. PhD thesis, Boston University, Boston, MA, USA.

Digalakis, V., Monaco, P., and Murveit, H. (1996). Genones: Generalised mixture tying in continuous hidden markov model-based speech recognizers. *IEEE Trans. Speech and Audio Proc.*, 4(4):281–289.

Digalakis, V., Rohlicek, J. R., and Ostendorf, M. (1993). ML estimation of a Stochastic Linear System with the EM Algorithm and its application to Speech Recognition. *IEEE Trans. Speech and Audio Proc.*, 1(4):431–442.

Dillon, W. R. and Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications*. John Wiley and Sons, New York, NY, USA.

Ding, C. and Peng, H. C. (2003). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.*, pages 523–529.

Ding, P. and Liming, Z. (2001). Speaker recognition using principal component analysis. In *Int. Conf. on Neural Inf. Proc. (ICONIP), Shanghai, China*.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classifcation.* Wiley-Interscience; 2nd edition, New York, NY, USA.

Erler, K. and Freeman, G. H. (1996). An HMM-based speech recognizer using overlapping articulatory features. *J. Acoust. Soc. Am.*, 100:2500–2513.

Eskenazi, M. (1996). An overview of spoken language technology for education. *Speech Communication*, 51:832–844.

Eskenazi, M. and Hansma, S. (1998). The fluency pronunciation trainer. In *Speech Tech. Lang. Learn. Marholmen, Sweden*, pages 77–80.

Eskenazi, M., Ke, M., Albornoz, J., and Probst, K. (2000). The fluency pronunciation trainer: Update and user issues. In *Proc. InSTIL, Dundee, UK*, pages 73–76.

ETSI (2000). Speech processing transmission and quality aspects; distributed speech recognition; front-end feature extraction algorithm; compression algorithms.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics*, 8:376–386.

Flege, J. E. (1995). *Second-language speech learning: theory, findings, and problems.* Strange, W. (Ed.), Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research. Timonium, MD: York Press Inc.

Franco, H., Bratt, H., Rossier, R., Gadde, V. R., Shriberg, E., Abrash, V., and Precoda, K. (2010). Eduspeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27(3):401–418.

Franco, H., Neumeyer, L., Digalakis, V., and Ronen, O. (2000). Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication*, 30(2-3):121–130.

Franco, H., Neumeyer, L., Kim, Y., and Ronen, O. (1997). Automatic pronunciation scoring for language instruction. In *IEEE Int. Conf. Acoust., Speech, Sig. Proc., Munich, Germany*, pages 1471–1474.

Frankel, J. (2003). *Linear dynamic models for automatic speech recognition.* PhD thesis, The Centre for Speech Technology Research, Edinburgh University, Edinburgh, UK.

Fritsch, J. and Finke, M. (1998). ACID/HNN: Clustering hierarchies of neural networks for context-dependent connectionist acoustic modeling. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Seattle, WA, USA*, volume 1, pages 505–508.

Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., San Francisco, CA, USA*, volume 1, pages 137–140.

Gardner, W. R. and Rao, B. D. (1995). Theoretical analysis of the high-rate vector quantization of LPC parameters. *IEEE Trans. Speech, Audio Proc.*, 3(5):367–381.

Ghitza, O. (1991). Auditory nerve representation as a basis for speech processing. In *Advances in Speech Signal Proc.*, pages 453–485. Marcel Dekker.

Glasberg, B. R. and Moore, B. C. J. (2002). A model of loudness applicable to time-varying sounds. *J. Audio Eng. Soc.*, 50(5):331–342.

Guion, S. G., Flege, J. E., Ahahane-Yamada, R., and Pruitt, J. C. (2000). An investigation of current models of second language speech perception: the case of japanese adults' perception of english consonants. *J. Acoust. Soc. Am.*, 107(5):2711–2724.

Haeb-Umbach, R. and Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., San Francisco, CA, USA*, volume 1, pages 13–16.

Haque, S., Togneri, R., and Zaknich, A. (2007). A temporal auditory model with adaptation for automatic speech recognition. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Honolulu, HI, USA*, volume 4, pages 1141–1144.

Heinz, M. G., Colburn, H. S., and Carney, L. H. (2001a). Evaluating auditory performance limits: I. one-parameter discrimination using a computational model for the auditory nerve. *Neural Comput.*, 13(10):2273–2316.

Heinz, M. G., Zhang, X., Bruce, I. C., and Carney, L. H. (2001b). Auditory-nerve model for predicting performance limits of normal and impaired listeners. *Acoust. Research Lett. Online*, 2(3):91–96.

Herskovits, E., Peng, H. C., and Davatzikos, C. (2004). A Bayesian morphometry algorithm. *IEEE Trans. in Medical Imaging*, 23(6):723–737.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, 40:185–234.

Hirsch, H. G. and Pearce, D. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions. In *ISCA Tutorial and Research Workshop ASR2000*, pages 29–32, Paris, France.

Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, Upper Saddle River, NJ, USA.

Hughes, G. F. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Trans. on Information Theory*, 14(1):55–63.

Hunt, M. J. and Lefebvre, C. (1989). A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Glasgow, UK*, volume 1, pages 262 – 265.

Ito, A., Lim, Y. L., Suzuki, M., and Makino, S. (2005). Pronunciation error detection method based on error rule clustering using a decision tree. In *ISCA Interspeech, Lisbon, Portugal*, pages 173–176.

Jaakkola, T., Meila, M., and Jebara, T. (1999). Maximum Entropy Discrimination. In *Advances in Neural Information Processing Systems 12*, pages 470–476. MIT Press.

Jebara, T. (2001). *Discriminative, generative and imitative learning*. PhD thesis, Media Laboratory MIT, Cambridge, MA, USA.

Jebara, T. and Jaakkola, T. (2000). Feature selection and Dualities in Maximum Entropy Discrimination. *Conf. in Uncertainity in Artificial Intell., Stanford, CA, USA*, pages 291–300.

Jeffress, L. A. (1948). A place theory of sound localization. *J. Compar. Physiol. Psychol.*, 41(1):35–39.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proc. of the IEEE*, 64(4):532–556.

Jeon, W. and Juang, B. H. (2005). A study of auditory modeling and processing for speech signals. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Philadelphia, PA, USA*, volume 1, pages 929 – 932.

Jepsen, M. L., Ewert, S. D., and Dau, T. (2008). A computational model of human auditory signal processing and perception. *J. Acoust. Soc. Am.*, 124(1):422–438.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing*. Prentice Hall, Englewood Cliffs, NJ, USA.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Tr. ASME, J. Basic Eng.*, 82:35–45.

Kanal, L. and Chandrasekaran, B. (1971). On dimensionality and sample size in statistical pattern classification. *Pattern Recognition*, 3:225–234.

Kawai, G. and Hirose, K. (1998). A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training. In *Inter. Conf. Spoken Lang. Proc., Sydney, Australia*, pages 1823–1826.

Kimball, O. A. (1995). *Segment Modeling Alternatives for Continuous Speech Recognition*. PhD thesis, Boston University, Boston, MA, USA.

Koniaris, C., Engwall, O., and Salvi, G. (2012). On the benefit of using auditory modeling for diagnostic evaluation of pronunciations. In *Inter. Symp. Autom. Detec. Errors Pronunc. Train. (IS ADEPT), Stockholm, Sweden*, pages 59–64.

Kuhl, P. K. (1993). Early linguistic experience and phonetic perception: implications for theories of developmental speech perception. *J. Phonetics*, 21:125–139.

Kumar, N. (1997). *Investigation of Silicon Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*. PhD thesis, Johns Hopkins University, Baltimore, MD, USA.

Kumar, N. and Andreou, A. G. (1996). A generalization of linear discriminant analysis in maximum likelihood framework. In *Joint Meeting of Amer. Statist. Assoc., Chicago, IL, USA*.

Kumar, N. and Andreou, A. G. (1998). Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication*, 26(4):283–297.

Lamel, L. F., Kassel, R. H., and Seneff, S. (1986). Speech database development: Design and analysis of the acoustic-phonetic corpus. In *DARPA Speech Recognition Workshop*, pages 100–109.

Lecumberri, M. L. G., Cooke, M., and Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, 52(11-12):864–886.

Lee, L. J., Attias, H., and Deng, L. (2003). Variational inference and learning for segmental switching state space models of hidden speech dynamics. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Hong Kong, Hong Kong*, volume 1, pages I–872 – I–875.

Li, J., Chaddha, N., and Gray, R. M. (1999). Asymptotic performance of vector quantizers with a perceptual distortion measure. *IEEE Trans. Inform. Theory*, 45(4):1082–1091.

Linder, T., Zamir, R., and Zeger, K. (1999). High-resolution source coding for non-difference distortion measures: multidimensional companding. *IEEE Trans. Inform. Theory*, 45(2):548–561.

Ljung, L. (1998). *System Identification: Theory for the User (2nd Edition)*. Prentice Hall PTR, Englewood Cliffs, NJ, USA.

Lyon, R. and Shamma, S. (1996). Auditory representations of timbre and pitch. In Hawkins, H. L., McMullen, T. A., Popper, A. N., and Fay, R. R., editors, *Auditory Computation*, pages 221–270. Springer New York, NY, USA.

Ma, J. and Deng, L. (1999). Optimization of dynamic regimes in a statistical hidden dynamic model for conversational speech recognition. In *Europ. Conf. Speech Comm. and Tech. (EUROSPEECH), Budapest, Hungary*, volume 3, pages 1339–1342.

Ma, J. and Deng, L. (2000). A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamic model of speech. *Computer Speech and Language*, 14(2):101–114.

Ma, J. and Deng, L. (2001). Efficient decoding strategy for conversational speech recognition using state-space models for vocal-tract-resonance dynamics. In *Europ. Conf. Speech Comm. and Tech. (EUROSPEECH), Aalborg, Denmark*, pages 603–606.

Ma, J. and Deng, L. (2004). A mixed-level switching dynamic system for continuous speech recognition. *Computer Speech and Language*, 18(1):49–65.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

Meddis, W. S. and O'Mard, L. (1997). A unitary model of pitch perception. *J. Acoust. Soc. Am.*, 102:1811–1820.

Menzel, W., Herron, D., Bonaventura, P., and Morton, R. (2000). Automatic detection and correction of non-native english pronunciations. In *Work. Intergr. Speech Tech. Lang. Learn. Ass. Inter., Dundee, UK*, pages 49–56.

Merhav, N. and Ephraim, Y. (1991). Hidden Markov modeling using the most likely state sequence. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Toronto, ON, Canada*, pages 469–472.

Moore, B. C. (2003). *An Introduction to the Psychology of Hearing*. London WC1X 8RR, UK: Academic Press.

Morgan, N. and Bourlard, H. (1995). Continuous speech recognition: An introduction to hybrid HMM/Connectionist approach. *IEEE Sig. Proc. Magazine*, 2:25–42.

Morton, P., Whitehouse, D., and Ollason, D. (1999). *The* IHAPI *book*. Entropic Cambridge Research Laboratory, Ltd., Cambridge, UK.

Moustroufas, N. and Digalakis, V. (2007). Automatic pronunciation evaluation of foreign speakers using unknown text. *Computer Speech & Language*, 21(1):219–230.

Munro, M. J. and Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1):73–97.

Nelson, D. A. and Swain, A. C. (1996). Temporal resolution within the upper accessory excitation of a masker. *Acustica - Acta Acust.*, 82(2):328–334.

Neumeyer, L., Franco, H., Digalakis, V., and Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication*, 30:83–93.

Neumeyer, L., Franco, H., Weintraub, M., and Price, P. (1996). Automatic text-independent pronunciation scoring of foreign language student speech. In *Int. Conf. Spoken Lang. Proc., Philadelphia, PA , USA*, pages 1457–1460.

Odell, J. J. (1995). *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, Cambridge University, Cambridge, UK.

Ostendorf, M., Digalakis, V., and Kimball, O. A. (1996). From HMMs to Segment Models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. on Speech and Audio Proc.*, 4(5):360–378.

Ostendorf, M. and Roukos, S. (1989). A stochastic segment model for phoneme-based continuous speech recognition. *IEEE Trans. Acous. Speech and Sig. Proc.*, 37(12):1857–1869.

Oxenham, A. J. (2001). Forward masking: adaptation or integration? *J. Acoust. Soc. Am.*, 109(2):732–741.

Oxenham, A. J., Bernstein, J. G., and Penagos, H. (2004). Correct tonotopic representation is necessary for complex pitch perception. In *Proc. Natl. Acad. Sci.*, volume 101, pages 1421–1425.

Oxenham, A. J. and Moore, B. C. J. (1994). Modeling the additivity of nonsimultaneous masking. *Hear. Res.*, 80(1):105–118.

Oxenham, A. J. and Moore, B. C. J. (1997). Modeling the effects of peripheral nonlinearity in listeners with normal and impaired hearing. In Jesteadt, W., editor, *Modeling Sensorineural Hearing Loss*, pages 273–288. Erlbaum, Hillsdale, NJ, USA.

Park, J. G. and Rhee, S. C. (2004). Development of the knowledge-based spoken english evaluation system and its application. In *ISCA Interspeech, Jeju Island, South Korea*, pages 1681–1684.

Patterson, R. D., Allerhand, M., and Giguère, C. (1995). Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform. *J. Acoust. Soc. Am.*, 98(4):1890–1894.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (1992). Complex sounds and auditory images. In Cazals, Y., Demany, L., and Horner, K., editors, *Auditory Physiology and Perception, Proc. of 9th Int. Sympos. on Hear.*, pages 429–446. Pergamon, Oxford, UK.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Patt. Analys., Mach. Intellig.*, 27(8):1226–1238.

Picone, J., Pike, S., Regan, R., Kamm, T., Bridle, J., Deng, L., Ma, Z., Richards, H., and Schuster, M. (1999). Initial evaluation of hidden dynamic models on conversational speech. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Phoenix, AZ, USA*, volume 1, pages 109–112.

Piske, T., Flege, J., and MacKay, I. (2001). Factors affecting degree of foreign accent in an l2: a review. *J. Phonetics*, 29(2):191–215.

Plack, C. J. and Oxenham, A. J. (1998). Basilar-membrane nonlinearity and the growth of forward masking. *J. Acoust. Soc. Am.*, 103(3):1598–1608.

Plasberg, J. H. and Kleijn, W. B. (2007). The sensitivity matrix: Using advanced auditory models in speech and audio processing. *IEEE Trans. Audio, Speech, Lang. Proc.*, 15(1):310–319.

Probst, K., Ke, Y., and Eskenazi, M. (2002). Enhancing foreign language tutors - in search of the golden speaker. *Speech Communication*, 37(3-4):161–173.

Rabiner, L. A. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rao, C. R. (1965). *Linear Statistical Inference and Its Applications.* John Wiley and Sons, New York, USA, 2nd edition 2001.

Rauch, H. E., Tung, F., and Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3:1445–1450.

Raux, A. and Kawahara, T. (2002). Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning. In *Int. Conf. Spoken Lang. Proc., Denver, CO, USA*, pages 737–740.

Ravishankar, M. (1996). *Efficient algorithms for speech recognition.* PhD thesis, Carnegie Mellon University, Pittsburg, PA, USA.

Richards, H. B. and Bridle, J. S. (1999). The HDM: A segmental hidden dynamic model of coarticulation. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Phoenix, AZ, USA*, volume 1, pages 357–360.

Richardson, M., Bilmes, J., and Diorio, C. (2000a). Hidden-articulator Markov models for speech recognition. In *ISCA Tutorial and Research Workshop ASR2000*, pages 133–139.

Richardson, M., Bilmes, J., and Diorio, C. (2000b). Hidden-articulator Markov models: performance improvements and robustness to noise. In *Inter. Conf. on Spoken Lang. Proc., Beijing, China*, volume 3, pages 131–134.

Richardson, M., Bilmes, J., and Diorio, C. (2003). Hidden-articulator Markov models for speech recognition. *Speech Communication*, 41:511–529.

Rix, A., Bourret, A., and Hollier, M. (1999). Models of human perception. *BT Technology Journal*, 17:24–34.

Robinson, A. J. (1994). An application of recurrent nets to phone probability estimation. *IEEE Trans. on Neural Networks*, 5:298–305.

Robinson, T., Almeida, L., Boite, J. M., Bourlard, H., Fallside, F., Hochberg, M., Kershaw, D., Kohn, P., Konig, Y., Morgan, N., Neto, J. P., Renals, S., Saerens, M., and Wooters, C. (1993). A neural network based, speaker independent, large vocabulary, continuous speech recognition system: The WERNICKE Project. In *Proc. of the European Conf. on Speech Comm. and Tech., Berlin, Germany*, pages 1941–1944.

Rosti, A.-V. I. (2004). *Linear Gaussian Models for Speech Recognition.* PhD thesis, University of Cambridge, Wolfson College, UK.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of the linear gaussian models. *Neural Computation*, 11(2).

Rummelhart, D. E. and McClelland, J. L. (1986). *Parallel Distributed Processing - Explorations in the Microstructure of Cognition, Volume* I*: Foundations.* Cambridge, MA, MIT Press, USA.

Russel, S. and Norving, P. (1995). *Artificial Intelligence: A modern Approach.* Prentice Hall PTR, Englewood Cliffs, NJ, USA.

Scanlon, P., Ellis, D. P. W., and Reilly, R. (2003). Using mutual information to design class specific phone recognizers. In *Proc. of the European Conf. on Speech Comm. and Tech., Geneva, Switzerland*, pages 857–860.

Schmid, P. M. and Yeni-Komshian, G. H. (1999). The effects of speaker accent and target predictability on perception of mispronunciations. *J. Speech, Lang., Hear. Res.*, 42:56–64.

Seide, F., Zhou, J. L., and Deng, L. (2003). Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM - MAP decoding and evaluation. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Hong Kong, Hong Kong*, volume 1, pages I–748 – I–751.

Seneff, S. (1988). A joint synchrony/mean rate model of auditory speech processing. *J. Phonetics*, 16:55–76.

Shamma, S. and Klein, D. (2000). The case of the missing pitch templates: how harmonic templates emerge in the early auditory system. *J. Acoust. Soc. Am.*, 107:2631–2644.

Shamma, S. A. (2004). Topographic organization is essential for pitch perception. In *Proc. Natl. Acad. Sci.*, volume 101, pages 1114–1115.

Sharma, S., Ellis, D., Kajarekar, S., Jain, P., and Hermansky, H. (2000). Feature extraction using non-linear transformation for robust speech recognition on the AURORA database. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Istanbul, Turkey*, volume 2, pages 1117–1120.

Shinoda, K. and Watanabe, T. (2000). DL-based context-dependent subword modeling for speech recognition. *J. Acoust. Soc. Jpn.(E)*, 21(2):79–86.

Siohan, O. (1995). On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Detroit, MI, USA*, volume 1, pages 125–128.

Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. In *Fonetik*, pages 93–96.

SPTK (2003). SPEECH SIGNAL PROCESSING TOOLKIT. `http://sp-tk.sourceforge.net`.

Stevens, S. S. and Volkman, J. (1940). The relation of the pitch to frequency: a revised scale. *Amer. J. Psychology*, 53:329–353.

Strik, H., Truong, K., de Wet, F., and Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication*, 51(10):845–852.

Tepperman, J. and Narayanan, S. (2008). Using articulatory representations to detect segmental errors in nonnative pronunciation. *IEEE Tr. Audio, Speech, Lang. Proc.*, 16(1):8–22.

Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (1999). Hidden markov models based on multi-space probability distribution for pitch pattern modeling. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Phoenix, AZ, USA*, volume 1, pages 229–232.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Istanbul, Turkey*, volume 3, pages 1315–1318.

Tokuda, K., Zen, H., and Black, A. W. (2002). An HMM-based speech synthesis system applied to english. In *IEEE Workshop on Speech Synthesis - WSS*, pages 227– 230.

Truong, K. P., Neri, A., de Wet, F., Cucchiarini, C., and Strik, H. (2005). Automatic detection of frequent pronunciation errors made by L2-learners. In *ISCA Interspeech, Lisbon, Portugal*, pages 1345–1348.

Tsai, S. N. and Lee, L. S. (2003). Improved robust features for speech recognition by integrating time-frequency principal components (TFPC) and histogram equalization (HEQ). In *IEEE Works. on Autom. Speech Recogn. and Underst., St. Thomas, U.S. Virgin Islands*, pages 297–302.

Tsontzos, G., Diakoloukas, V., Koniaris, C., and Digalakis, V. (2007). Estimation of general identifiable linear dynamic models with an application in speech recognition. In *IEEE Int. Conf. on Acoust., Speech and Sig. Proc., Honolulu, HI, USA*, volume 4, pages IV–453–IV–456.

Valente, F. and Wellekens, C. (2003). Maximum entropy discrimination (MED) feature subset selection for speech recognition. In *IEEE Works. on Autom. Speech Recogn. and Underst., St. Thomas, U.S. Virgin Islands*, pages 327–332.

van de Par, S., Kohlrausch, A., Charestan, G., and Heusdens, R. (2002). A new psychoacoustical masking model for audio coding applications. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Orlando, FL, USA*, volume 2, pages 1805–1808.

VOICEBOX (1999). Speech processing toolbox for matlab. `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

Waibel, A. H. and Lee, K. F. (1990). *Readings in Speech Recognition*. Morgan Kaufman Publishers, San Mateo, CA, USA.

Wanfeng, Z., Yingchun, Y., Zhaohui, W., and Lifeng, S. (2003). Experimental evaluation of a new speaker identification framework using PCA. In *IEEE Int. Conf. on Systems, Man and Cybern., Washington, DC, USA*, volume 5, pages 4147–4152.

Wei, S., Hu, G., Hu, Y., and Wang, R.-H. (2009). A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication*, 51(10):896–905.

Weigelt, L. F., Sadoff, S. J., and Miller, J. D. (1990). Plosive/fricative distinction: the voiceless case. *J. Acoust. Soc. Am.*, 87:2729–2737.

Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49–63.

Witt, S. M. (2012). Automatic error detection in pronunciation training: Where we are and where we need to go. In *Inter. Symp. Autom. Detec. Errors Pronunc. Train. (IS ADEPT), Stockholm, Sweden*, pages 1–8.

Witt, S. M. and Young, S. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30:95–108.

Xiong, M., Fang, Z., and Zhao, J. (2001). Biomarker identification by feature wrappers. *Genome Research*, 11:1878–1887.

Xu, S., Jiang, J., Chen, Z., and Xu, B. (2009). Automatic pronunciation error detection based on linguistic knowledge and pronunciation space. In *IEEE Int. Conf. Acoust. Speech Sig. Proc. (ICASSP), Taipei, Taiwan*, pages 4841–4844.

Yamashita, Y., Kato, K., and Nozawa, K. (2005). Automatic scoring for prosodic proficiency of english sentences spoken by japanese based on utterance comparison. *IECE Trans. Inform. Systems*, E88-D:496–501.

Yang, H. H., Vuuren, S. V., Sharma, S., and Hermansky, H. (2000). Relevance of time-frequency features for phonetic and speaker channel classification. *Speech Communication*, 31:35–50.

Yoshimura, T. (2002). *Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for* HMM*-based Text-To-Speech systems*. PhD thesis, Nagoya Institute of Technology, Nagoya, Japan.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1998). Duration modeling in HMM-based speech synthesis system. In *Inter. Conf. Spoken Lang. Proc. (ICSLP), Sydney, Australia*, volume 2, pages 29–32.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2002). *The HTK Book (for HTK Version 3.2)*. Cambridge University, Engineering Department, Cambridge, UK.

Zavaliagkos, G., Zhao, Y., Schwartz, R., and Makhoul, J. (1994). A Hybrid Segmental Neural Net/Hidden Markov Model system for continuous speech recognition. *IEEE Trans. on Speech and Audio Proc.*, 2:151–160.

Zhou, J. L., Seide, F., and Deng, L. (2003). Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM - model and training. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Hong Kong, Hong Kong*, volume 1, pages 744–747.

Zwicker, E. and Fastl, H. (1999). *Psychoacoustics, Facts and Models.* Springer, Heidelberg, Germany.

Part II

Included papers