Thesis for the degree of Licentiate

A Study on Selecting and Optimizing Perceptually Relevant Features for Automatic Speech Recognition

Christos Koniaris

KTH Electrical Engineering

Sound and Image Processing Laboratory
School of Electrical Engineering
KTH (Royal Institute of Technology)

Stockholm 2009

Koniaris, Christos
A Study on Selecting and Optimizing Perceptually Relevant Features for Automatic Speech Recognition

ISBN 978-91-7415-478-8
TRITA-EE 2009:049
ISSN 1653-5146

Sound and Image Processing Laboratory
School of Electrical Engineering
KTH (Royal Institute of Technology)
SE-100 44 Stockholm, Sweden
Telephone + 46 (0)8-790 7790

Abstract

The performance of an automatic speech recognition (ASR) 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. This work is motivated by the fact that humans perform better at speech recognition than machines, particularly for noisy environments. The goal of this thesis is to make use of knowledge of human perception in the selection and optimization of speech features for speech recognition.

Papers A and C show that robust feature selection for speech recognition can be based on models of the human auditory system. These papers show that maximizing the similarity of the Euclidian geometry of the features to the geometry of the perceptual domain is a powerful tool to select features. Whereas conventional methods optimize classification performance, the new feature selection method exploits knowledge implicit in the human auditory system, inheriting its robustness to varying environmental conditions. The proposed algorithm show how the feature set can be learned from perception only by establishing a measure of goodness for a given feature based on a perturbation analysis and distortion criteria derived from psycho-acoustic models. Experiments with a practical speech recognizer confirm the validity of the principle.

In Paper B the perceptually relevant objective criterion is used to define new features. Again the motivation has its origin at the human peripheral auditory system which plays a major role to the input speech signal until it reaches the central auditory system of the brain where the recognition occurs. While many feature extraction techniques incorporate knowledge of the auditory system, the procedures are usually designed for a specific task, and they lack of the most recently gained knowledge on human hearing. Paper B shows an approach to improve mel frequency cepstrum coefficients (MFCCs) through off-line optimization. The method has three advantages: i) it is computational 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.

Keywords: feature extraction, feature selection, auditory models, MFCCs, speech recognition, distortion measures, perturbation analysis, psychoacoustics, human perception, sensitivity matrix.

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," submitted.

[B] S. Chatterjee, C. Koniaris, and W. B. Kleijn, "Auditory Model Based Optimization of MFCCs Improves Automatic Speech Recognition Performance," in *Proceedings of Interspeech*, 2009.

[C] C. Koniaris, S. Chatterjee, and W. B. Kleijn, "Selecting Static and Dynamic Features Using an Advanced Auditory Model for Speech Recognition," submitted.

Acknowledgements

I would like to thank Bastiaan Kleijn, my supervisor, for his many suggestions and valuable help and support during this research. His experience and scientific background, has proven advantageous in my research.

I am also grateful to my papers' co-authors Marcin Kuropatwinski, Saikat Chatterjee and of course Bastiaan Kleijn. We had long interesting discussions and a lot of scientific deliberation during the course of this study. I learned many new valuable things and have been able to develop my research skills and experience.

I consider myself lucky to share office with Minyue Li, a gentle and friendly person who was always willing to help me when needed. During my studies, I had the pleasure of meeting Jan Plasberg who was kind enough to share with me his experience in auditory modeling through the early months of chaos and confusion. At this point, I feel I need to express many thanks to all my current and past colleges for their help, understanding, and support but also for their friendship and all the interesting discussions. These include Obada Alhaj Moussa, Anders Ekman, Gustav Henter, Janusz Klejsa, Ermin Kozica, Arne Leijon, Petko Petkov, Thippur Sreenivas, Svante Stadler, Dora Söderberg, Guoqiang Zhang, David Zhao and many other members of Sound and Image Processing Laboratory.

Of course, I do not forget to thank my parents for providing me with all the necessary means, spiritual and material, during the previous years that have brought me at the point where I can claim to be a 'normal person'!

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. His intellectuality has been a paradigm to be copied.

Finally, I would like to thank Magda for her love and uninterrupted support, and for giving me the motivation to continue my work. To her, I dedicate this thesis.

Christos Koniaris
Stockholm, September 2009

Contents

List of Figures

List of Tables

Acronyms

ANNs	Artificial Neural Networks
AMFS	Auditory Model-based Feature Selection
ASR	Automatic Speech Recognition
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
ERB	Equivalent Rectangular Bandwidth
EM	Expectation-Maximization algorithm
IHC	Inner Hair Cells
KLT	Karhunen-Loève Transform
LDA	Linear Discriminant Analysis
MAP	Maximum A-Posteriori Probability
MED	Maximum Entropy Discrimination
MFCCs	Mel Frequency Cepstrum Coefficients
mRMR	minimal-Redundancy-Maximal-Relevance
MMFCCs	Modified Mel Frequency Cepstrum Coefficients
MLP	Multi-Layer Perceptron
HLDA	Heteroscedastic Linear Discriminant Analysis
HAMM	Hidden Articulatory Markov Model
HDMs	Hidden Dynamic Models
HMMs	Hidden Markov Models
HTK	Hidden Markov Model Toolkit

OHC	Outer Hair Cells
OM	Outer and Middle Ear
PDP	Parallel Distributed Processing
PCA	Principle Component Analysis
PDF	Probability Density Function
POD	Proper Orthogonal Decomposition
SMs	Segment-based Models
SNR	Signal-to-Noise Ratio
VQ	Vector Quantization

“Λίγο ακόμα ϑα ιδούμϵ
τις αμυγδαλιές ν’ ανϑίζουν.
Λίγο ακόμα ϑα ιδούμϵ
τα μάρμαρα να λάμπουν,
να λάμπουν στον ήλιο
κι η ϑάλασσα να κυματίζϵι.
Λίγο ακόμα, να σηκωϑούμϵ
λίγο ψηλότϵρα.”

Γεώργιος Σεφέρης, ‘Λίγο ακόμα’
Νομπέλ Λογοτεχνίας, 1963

“Just a little more and we shall see
the almond trees in blossom.
The marbles shining in the sun,
the sea, the curling waves.
Just a little more, let us rise
just a little higher.”

Georgios Seferis, ‘Just a little more’
Nobel Prize in Literature, 1963

Part I

Introduction

Introduction

The way humans interact with computers has been developed since the early days of computer engineering. In nowadays, it is not unusual this interaction to be done by *speech*. Different systems have been designed to perform this task. Due to its inherent difficulty, a thorough understanding of human perception is needed. Additionally, a compact and relevant representation of speech input is an important factor to enhance the system's performance. This chapter deals with the above. In Sec. 1 the human auditory system is introduced as well as two different auditory models. Next, the front-end and the acoustic models are discussed. Sec. 2 deals with features dimensionality reduction methods. In the end, a short description of the proposed, auditory motivated, feature selection technique is given. Sec. 3 presents the thesis contributions and a short description of the three papers of Part II. Finally, Sec. 4 provides conclusions.

1 Perception and speech recognition

Speech communication has been, and will continue to be, the dominant manner of human social communication and information exchange. This is reflected in the way humans prefer to interact with computers and other technological artifacts. Within the broader area of *speech communication*, i.e., the science of communication between humans and computers, *speech recognition* deals with the development of new techniques that transcribe human speech into written text.

In recent years, the performance of automatic speech recognition (ASR) systems has improved dramatically. 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.

In the next section, the human auditory system is presented as a background knowledge necessary to be able to understand the progress in the auditory modeling community.

1.1 Human hearing system

The human ear consists of several parts [38, 62, 98]: the *outer ear*, the *middle ear*, and the *inner ear*. The way these elements operate is not totally understood, although a series of studies have reached a good level of comprehension to a considerable extent. In the next, we provide an insight of the human ear but for more details and an extended analysis of the function of the human auditory system the reader is referred to [62, 98].

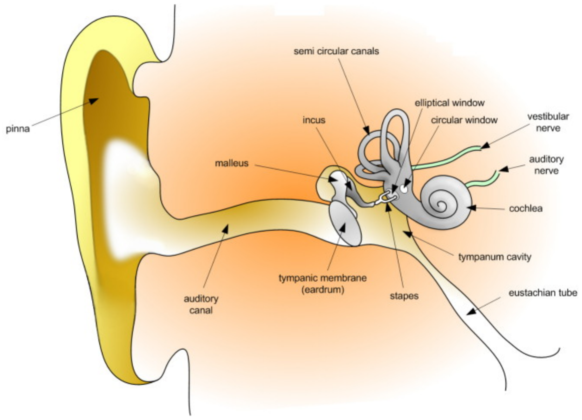

Figure 1: The anatomy of human ear.

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 humans simply call 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., performs 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 which is filtered by the canal, hits the eardrum and the latest starts to vibrate. When this happens, the sound vibrations are passed into an area known as the middle ear.

The middle ear space, also known as tympanic cavity is connected to the back of the throat by the eustachian tube. This space 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.

As mentioned above, 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 we swallow, equalizing the middle ear pressure with the external air pressure. As a result, the tympanic membrane has equal pressure on either side and this helps it to work properly. In special occasions when the outside pressure changes abruptly and e.g., when travelling or flying, this mechanical pressure equalization does not work automatically and people need to swallow from time to time to equalize the pressure across their eardrums. Finally, when a person suffers from a cold, the eustachian tube can become clogged with mucus. In such case, air and fluid are trapped inside the ear, and can cause a temporarily impaired hearing or even a painful ear infection.

The inner ear has two parts, the cochlea and the vestibule. The cochlea is a small spiral (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 on which 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 to the brain. 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 way in which the brain processes the extracted patterns is rather vague. Many studies though have shown how individuals perceive tones and noise bands [62, 98]. Based on that knowledge, many auditory models that simulate the functionality of the human ear, have been proposed [2, 13, 62, 98]. In the next section, a short introduction on two of them is given, namely the van de Par [91] and the Dau [13] auditory models.

1.2 Auditory models

In [31] the concept of the 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 [54] and in [53]. In [69], a method for deriving the sensitivity matrix for distortion measures that are relevant for audio signals was developed based on spectro-temporal auditory models.

Let $\mathbf{x}_j \in \mathbb{R}^N$ be a N-dimensional speech signal vector characterizing a segment with time index $j \in \mathbb{Z}$ and let $\hat{\mathbf{x}}_{j,m}$ be a perturbation of $\mathbf{x}_j$ with perturbation index m. Furthermore, let $\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})$ be a distortion measure between $\mathbf{x}_j$ and $\hat{\mathbf{x}}_{j,m}$. For small distortions, we perform a Taylor series expansion of Υ

$$\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m}) = \Upsilon(\mathbf{x}_j, \mathbf{x}_j) + \left.\frac{\partial \Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})}{\partial \hat{\mathbf{x}}_{j,m}}\right|_{\hat{\mathbf{x}}_{j,m}=\mathbf{x}_j} [\hat{\mathbf{x}}_{j,m} - \mathbf{x}_j] +$$

$$\frac{1}{2}[\hat{\mathbf{x}}_{j,m} - \mathbf{x}_j]^T \left.\frac{\partial^2 \Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})}{\partial \hat{x}_\kappa \partial \hat{x}_\mu}\right|_{\hat{\mathbf{x}}_{j,m}=\mathbf{x}_j} [\hat{\mathbf{x}}_{j,m} - \mathbf{x}_j]^T + \mathbf{O}[\| \hat{\mathbf{x}}_{j,m} - \mathbf{x}_j \|^3]. (1)$$

In the above expansion we know that $\Upsilon(\mathbf{x}_j, \mathbf{x}_j) = 0$, and because $\hat{\mathbf{x}}_{j,m}$ is a unique minimum of $\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})$, the term $\left.\frac{\partial \Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})}{\partial \hat{\mathbf{x}}_{j,m}}\right|_{\hat{\mathbf{x}}_{j,m}=\mathbf{x}_j}$ vanishes. Moreover, all the terms that are of order three and above $\mathbf{O}[\| \hat{\mathbf{x}}_{j,m} - \mathbf{x}_j \|^3]$, are approximated to zero. Hence, the distortion measure is approximated [31] as

$$\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m}) \approx [\hat{\mathbf{x}}_{j,m} - \mathbf{x}_j]^T \mathbf{D}_\Upsilon(\mathbf{x}_j)[\hat{\mathbf{x}}_{j,m} - \mathbf{x}_j]. \tag{2}$$

The matrix $\mathbf{D}_{\Upsilon,\kappa\mu}(\mathbf{x}_j) = \left.\frac{\partial^2 \Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})}{\partial \hat{x}_\kappa \partial \hat{x}_\mu}\right|_{\hat{\mathbf{x}}_{j,m}=\mathbf{x}_j}$ is called *sensitivity matrix.* The word "sensitivity" refers to the fact that each element of this matrix represents the sensitivity of the distortion $\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})$ to a particular $[\hat{\mathbf{x}}_{j,m} - \mathbf{x}_j]$.

In the next two paragraphs, two different auditory models are presented that are used to extract the sensitivity matrix.

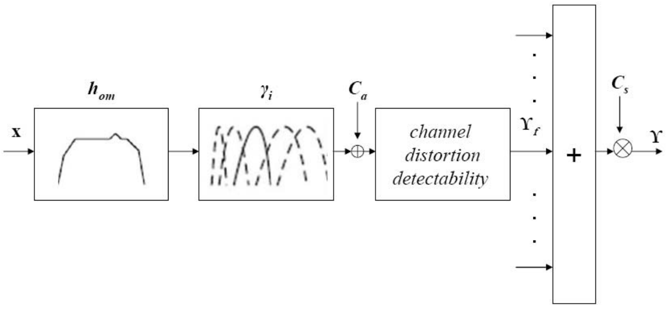

Figure 2: Block diagram of a channel of the van de Par psycho-acoustic model.

van de Par model

The van de Par [91] auditory model is a psycho-acoustic masking model that accounts for simultaneous processing of sound signals. One channel of the model is shown in Fig. 2. The first filter 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 which models the basilar membrane 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. 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

$$\Upsilon(\mathbf{x},\hat{\mathbf{x}}) = C_s L_e \sum_{g\in\mathcal{G}} \frac{\frac{1}{N}\sum_{f=0,\cdots,N-1} |h_{om}(f)|^2|\gamma_i(f)|^2|x(f)-\hat{x}(f)|^2}{\frac{1}{N}\sum_{f=0,\cdots,N-1} |h_{om}(f)|^2|\gamma_i(f)|^2|x(f)|^2 + C_a}, \quad (3)$$

where C_s and C_a are constants calibrated based on measurement 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 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 γ_i is the i'th gammatone filter.

In Papers A and B, the van de Par model is used to obtain the sensitivity matrix in the speech frequency domain. It is a diagonal matrix with the diagonal element for row and column f given by

$$\mathbf{D}_{\Upsilon,ff}(\mathbf{x}) \approx 2C_sL_e \sum_i \frac{\frac{1}{N}\sum_f |h_{om}(f)|^2|\gamma_i(f)|^2}{\frac{1}{N}\sum_f |h_{om}(f)|^2|\gamma_i(f)|^2|x(f)|^2 + C_a}. \tag{4}$$

Dau model

The Dau [13, 14] auditory model is a psycho-acoustic 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 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 stage incorporates the forward masking prediction of the ear [69]. Finally, a low-pass filter performs a temporal smoothing and the output is the so-called internal representation $\mathbf{a}^{(l)}(\mathbf{x}_j)$, where $\mathbf{x}_j$ is the j'th speech segment. The original paper [13] did not study the distortion prediction properties of the model, an investigation that was later performed in [69]. In the same work a distortion measure on the internal representation was introduced as

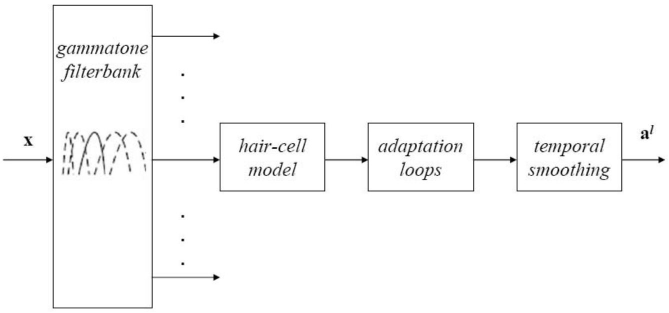

Figure 3: Block diagram of a channel of the Dau psycho-acoustic model.

$$\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m}) = \sum_l \|\mathbf{a}^{(l)}(\mathbf{x}'_j) - \mathbf{a}^{(l)}(\hat{\mathbf{x}}'_{j,m})\|^2, \tag{5}$$

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 [69]. The sensitivity matrix in this case is a result of a complicated and sophisticated effort. Crudely speaking, the sensitivity matrix can be computed as the sum of per-channel sensitivity matrices $\mathbf{D}_{\Upsilon}^{(l)}(\mathbf{x}_j)$

$$\mathbf{D}_{\Upsilon}(\mathbf{x}_j) = \sum_{l} \mathbf{D}_{\Upsilon}^{(l)}(\mathbf{x}_j), \tag{6}$$

where

$$\mathbf{D}_{\Upsilon}^{(l)}(\mathbf{x}_j) = 2 \left[\prod_{k} \mathbf{J}_k^{(l)} \right]^H \mathbf{J}_k^{(l)}, \tag{7}$$

and $\mathbf{J}_k^{(l)}$ is the Jacobian for stage k in channel l.

At this point, the discussion has mainly been focused on the area of auditory modeling. The next paragraph introduces the area of speech recognition. It starts with the feature extraction process, an important part of an ASR system associated to auditory knowledge.

1.3 Front-End

During the first step in the feature extraction process the speech waveform is sliced up into frames, which are transformed to spectral features as is shown in Fig. 4. In this paragraph, we briefly describe the process of extracting mel-frequency cepstrum coefficients (MFCCs).

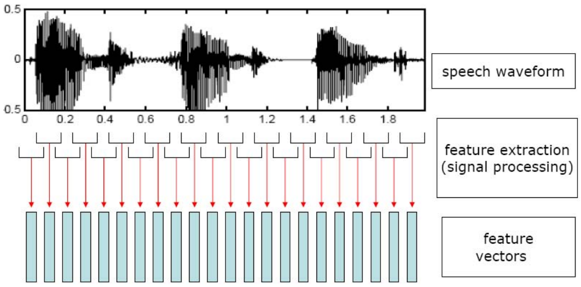

Figure 4: Extracting features from speech signal.

Mel frequencies are based on the knowledge of the human auditory system. 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 [38]. The response is linear at frequencies below 1 kHz and becoming logarithmic with increasing frequency [86]. This behavior is with a filter bank with triangular filters. The amplitudes of the triangular filters, shown in Fig. 5, are computed as

$$\mathbf{H}_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{f(m-1)-k}{f(m+1)-f(m)}, & f(m) \leq k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{8}$$

which satisfies $\sum_{m=1}^{M} \mathbf{H}_m(k) = 1$ according to [38].

The speech signal is first pre-emphasized $\mathbf{x}(n) = \breve{\mathbf{x}}(n) - \alpha\breve{\mathbf{x}}(n-1)$, where $\breve{\mathbf{x}}(n)$ is the original speech and $\alpha = 0.97$ [89], 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

$$\mathbf{x}'(n) = \left\{0.54 - 0.46\cos\left\{\frac{2\pi[N-1]}{N-1}\right\}\right\}\mathbf{x}(n), n = 1...N, \tag{9}$$

where 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

$$\mathbf{X}(k) = \sum_{n=0}^{N-1} \mathbf{x}'(n)e^{-j2\pi kn/N}, k = 1...K, \tag{10}$$

where K is the length of the DFT. Next, the DFT power spectrum is computed which then is multiplied with the triangular mel-weighted filterbank. The result is summed to give the logarithmic mel spectrum

$$\mathbf{s}(m) = \ln\left[\sum_{k=0}^{K-1} |\mathbf{X}(k)|^2 \mathbf{H}_m(k)\right], \tag{11}$$

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 [15] as

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \mathbf{s}(m)\cos\left\{q[m - \frac{1}{2}]\frac{\pi}{M}\right\}, q = 1...Q, \tag{12}$$

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.

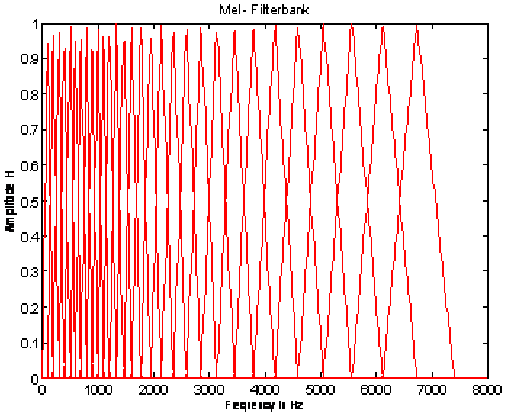

Figure 5: The mel filterbank.

Usually, the first and the second time derivatives are added to the speech vector, $\Delta\mathbf{c}$ and $\Delta\Delta\mathbf{c}$ to better capture time dependencies [26]. These are calculated as

$$\Delta\mathbf{c}_t = \frac{\sum\limits_{\theta=1}^{\Theta}\theta(\mathbf{c}_{t+\theta} - \mathbf{c}_{t-\theta})}{2\sum\limits_{\theta=1}^{\Theta}\theta^2}, \tag{13}$$

and

$$\Delta\Delta\mathbf{c}_t = \frac{\sum\limits_{\theta=1}^{\Theta}\theta(\Delta\mathbf{c}_{t+\theta} - \Delta\mathbf{c}_{t-\theta})}{2\sum\limits_{\theta=1}^{\Theta}\theta^2}, \tag{14}$$

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

1.4 Acoustic modeling

The feature extraction part (a typical paradigm of which described above) is the first step in building an automatic speech recognition system. Fig. 6 shows all the main blocks of such a system. These are the *front-end*, the *acoustic models*, the *language model*, the *lexicon* and the *search algorithm* [77]. The acoustic modeling has a significant role in an ASR system and naturally, is important in improving accuracy. The most popular approach

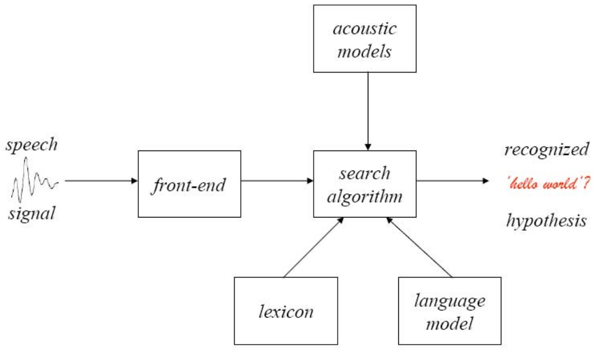

Figure 6: A real speech recognition system.

in acoustic modeling is based on statistical methods. Before getting into details, let us first give the definition of an acoustic model.

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. (We can consider the successive o_t indicating temporally consecutive slices of the acoustic input [45].) 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 probability (MAP) $P(W|O)$

$$\hat{W} = \arg\max P(W|O) = \frac{P(O|W)P(W)}{P(O)}. \tag{15}$$

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 the sub-word units, called *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) are the most popular approach to 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. These models seem to overcome some of the problems we meet in HMMs and ANNs, though they are of higher computational complexity. In the following, we begin by presenting the HMMs (the approach that used in all Papers A, B, and C) and then continue with other approaches.

Hidden Markov models

HMMs method is a flexible and successful statistical approach and hence very popular for acoustic modeling in speech recognition [5, 43, 70]. In HMMs, it is assumed that the sequence of observed vectors which correspond to each word or phone is generated by a Markov model [26] as shown in Fig. 7. Hence, the HMM approach is a double-embedded stochastic process with an not-directly-observable underlying stochastic process, namely the state sequence. Hence, the name 'hidden' has been adopted due to this fact. This hidden process is probabilistically linked with the observable stochastic process which produces the sequence of features we see [38].

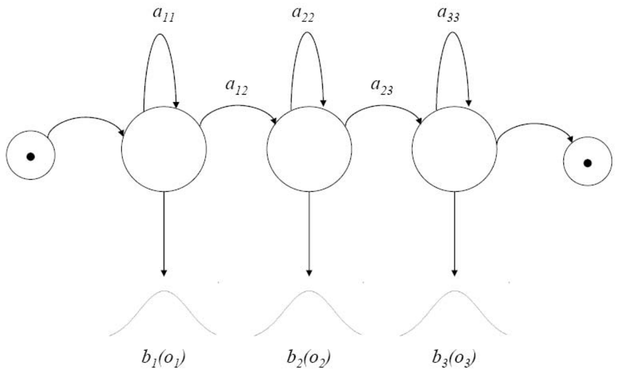

Figure 7: A hidden Markov model.

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

- Number of states: N
- Number of distinct observation symbols: M for discrete HMMs and ∞ for continuous HMMs
- State transition probability distribution: $\alpha_{i,j}$
- Output distribution of state j: $b_j(o_t)$
- Initial state probability: π_i.

To summarize, a complete specification of a 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 A, O, and π, the state transition matrix, the output distribution matrix and the initialization matrix, respectively. For convenience, we use the following notation

$$\lambda = (A, O, \pi) \tag{16}$$

to denote the whole parameter set of a HMM [38].

Types of HMMs

In accordance to the elements of the observation matrix O, HMMs are grouped in different categories [11] according to the distribution function that 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 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. This can be resolved though by 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 [70] 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 [8].

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 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. Although the above assumptions allow the model to become easier to use, they introduce some limitations that principally reflect on the accuracy of the model [18,61]. For this, other methods have been proposed to be applied in acoustic modeling.

Other approaches

Although HMMs predominate in most speech recognition systems, they still have many modeling inadequacies as a result of the assumptions that are accompanying HMMs to simplify the speech recognition problem [88]. Dynamic information can be included in HMMs through the time-derivatives (delta and acceleration coefficients) in the observation vector, though under the false frame-independence assumption.

Artificial neural networks (ANNs), also known as *connectionist models* or *parallel distributed processing* were introduced in 1943 by McCulloch and Pitts [60]. 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 [38] 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 [80]. 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 [36,38,79, 92]. Finally, some hybrid HMMs/ANNs [12,25,30,63,76,96] methods have been developed for ASR.

Segment models (SMs) have been extensively used for various applications, among them in speech recognition [18,29]. HMMs generate a single observation that is conditionally independent from the other. 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 [64,77]. A segment may be a variable-length part of the speech waveform [18], that usually corresponds to a language unit, e.g., a word, a phone or a sub-phone. Segment-based models [7,10,19,47,65,78] have been proposed as HMMs alternatives, offering a more suitable and flexible scheme to model the dynamics of speech signal. In all cases, several modeling restrictions were applied to ensure that the model is identifiable. In [48] an effort was made to relax these constraints, and allow to choose full noise covariances and state vectors that have arbitrary increased dimension compared to the size of the observation vector. The use of the canonical form of the system's matrices proposed in [55] ensures the system's identifiability. Furthermore, an investigation of the use of an extra control input in the state equation was performed. The parameters estimation performed with novel maximum likelihood, element-wise, parameter estimation processes based on the Expectation-Maximization (EM) algorithm. In [88], the proposed system applied in speech recognition task. The classification experiments on the AURORA2 [37] speech database show performance gains compared to HMMs, particularly on highly noisy conditions.

In recent years, a variation of segment models called *hidden dynamic models* (HDMs) [17, 59, 68, 72, 97] have been proposed. The main focus in this approach is to efficiently model the co-articulation phenomenon and improve the transitions between neighboring phones. The hidden dynamic space consists of a single vector target per phone in which the trajectories of the speech are produced by a dynamic system. The observation process in HDMs is implemented by a global *multi-layer perceptron* (MLP). The model is simple and flexible, able to capture important aspects of the relation between the phonetic labels and the acoustic patterns. The major drawback of the method is that the inference algorithms are not tractable. A number of approximate methods have been proposed [52, 56–58, 82] to improve the algorithms.

Another approach, from the family of segment-based models, was the idea of inserting articulatory knowledge into acoustic models [73–75] called the *hidden articulatory Markov model* (HAMM). The model, based on the [24], is essentially a HMM in which each articulatory configuration is modeled by a separate state. The state transitions aim to naturally reflect human articulation.

2 Reducing features dimensionality

In the previous section we described two of the most important parts of an ASR system, namely the front-end and the acoustic model. In this section, we study methods and techniques to lower the cardinality of the feature vectors while keeping the maximum available information for discriminating different sounds.

The initial process and the careful extraction of the necessary, acoustic relative, features is essential. 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 [39, 46]. In [6] the phenomenon of *curse of dimensionality* is described. It 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 we feed the system 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, we discuss the method of linear discriminant analysis (LDA) in Sec. 2.1 and the heteroscedastic linear discriminant analysis (HLDA) in Sec. 2.2. In Sec. 2.3 we give a short description of the principal component analysis (PCA) and in Sec. 2.4 we discuss other techniques in feature selection. Finally, Sec. 2.5 introduces the proposed auditory model-based feature selection method (AMFS). The latest is presented in more details in Papers A and C.

2.1 Linear discriminant analysis

LDA [23,27,28,71] has been applied in feature reduction problems for speech recognition tasks [3,4,9,16,33,85]. In [84] a study of combined feature sets including among other LDA transformations, was performed. The goal of LDA is to find an optimal transformation matrix ϕ^T to reduce the dimensionality of the feature space and in the same time, to maximize the necessary information to distinguish between different classes in a classification task problem. The above can be expressed as

$$\mathbf{y} = \phi^T \mathbf{c}, \tag{17}$$

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 a HMM recognizer. To formulate mathematically the optimization procedure, the mean vector and the covariance matrix for each class can be computed as

$$\mu_j = \frac{1}{N_j} \sum_{i=1}^{N_j} c_i, \tag{18}$$

$$\mathbf{\Sigma}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} [c_i - \mu_j][c_i - \mu_j]^T, \tag{19}$$

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

$$\mu = \frac{1}{N} \sum_{i=1}^{N} c_i, \tag{20}$$

$$\mathbf{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} [c_i - \mu][c_i - \mu]^T, \tag{21}$$

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

$$\hat{\phi} = \arg\max_{\phi_p} \frac{|\phi_p^T \, \mathbf{\Sigma} \, \phi_p|}{|\phi_p^T \, \mathbf{S} \, \phi_p|}, \tag{22}$$

where

$$\mathbf{S} = \frac{1}{N} \sum_{j=1}^{J} N_j \mathbf{\Sigma}_j. \tag{23}$$

The maximization criterion (22) 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 [20, 51].

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

2.2 Heteroscedastic linear discriminant analysis

Heteroscedastic linear discriminant analysis (HLDA) [49, 50] is an extension of the forementioned LDA method. Although the basic idea is 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 is the second shortcoming [51] of LDA. Hence, HLDA was developed to overcome these limitations.

In HLDA, the transformation matrix ϕ is a $q \times q$ matrix, and thus differs from the LDA, that again is applied in the original feature vector as

$$\mathbf{y} = \phi^T \mathbf{c}, \tag{24}$$

with $\mathbf{y} \in \mathbb{R}^p$ where p is the dimension of the feature vector in the reduced feature domain and $\mathbf{c} \in \mathbb{R}^q$ where q refers to the dimension of the feature vector in the original feature domain. 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. The latest is based on the assumption that only the first p components of $\mathbf{y}$ may carry the classification information [51]. The data are modeled as a Gaussian distribution [49] and the parameters of the probability density function (PDF) are

$$\mu_j = \begin{bmatrix} \mu_j^p \\ \mu \end{bmatrix}, \tag{25}$$

and

$$\mathbf{\Sigma}_j = \begin{bmatrix} \mathbf{\Sigma}_j^p & 0 \\ 0 & \mathbf{\Sigma}^{q-p} \end{bmatrix}, \tag{26}$$

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

$$P(\mathbf{c}_i) = \frac{|\phi|}{\sqrt{(2\pi)^q |\mathbf{\Sigma}_{g(i)}|}} \exp\left\{ -\frac{1}{2}[\mathbf{y}_i - \mu_{g(i)}]^T \mathbf{\Sigma}_{g(i)}^{-1} [\mathbf{y}_i - \mu_{g(i)}] \right\}, \tag{27}$$

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

$$\log P(\mu_j, \mathbf{\Sigma}_j, \phi; \{\mathbf{c}_i\}) = N \log|\phi| - \\ -\frac{1}{2}\sum_{i=1}^{N} \left\{ \log[(2\pi)^q |\mathbf{\Sigma}_{g(i)}|] + [\phi^T \mathbf{c}_i - \mu_{g(i)}]^T \mathbf{\Sigma}_{g(i)}^{-1} [\phi^T \mathbf{c}_i - \mu_{g(i)}] \right\}. \tag{28}$$

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

$$\hat{\mu}_j = \phi_p^T \mathbf{c}_j, \tag{29}$$

$$\hat{\mu} = \phi_{q-p}^T \mathbf{c}, \tag{30}$$

$$\mathbf{\Sigma}_j = \phi_p^T \mathbf{\Sigma}_j \phi_p, \tag{31}$$

and

$$\mathbf{\Sigma} = \phi_{q-p}^T \mathbf{\Sigma} \phi_{q-p}, \tag{32}$$

where $j = 1, ..., J$. Next, the above estimates can be substituted into the log-likelihood function (28), and then it can be shown [49] that the final estimate of ϕ is given by

$$\hat{\phi} = \arg\max_{\phi} \left\{ -\frac{N}{2} \log|\phi_{q-p}^T \mathbf{\Sigma} \phi_{q-p}| - \sum_{j=1}^{J} \frac{N_j}{2} \log|\phi_p^T \mathbf{\Sigma}_j \phi_p| + N \log|\phi| \right\}. \tag{33}$$

In [51], the maximization of the above equation was performed using numerical methods. The $\hat{\phi}$ is initialized by the ϕ computed previously by the LDA method.

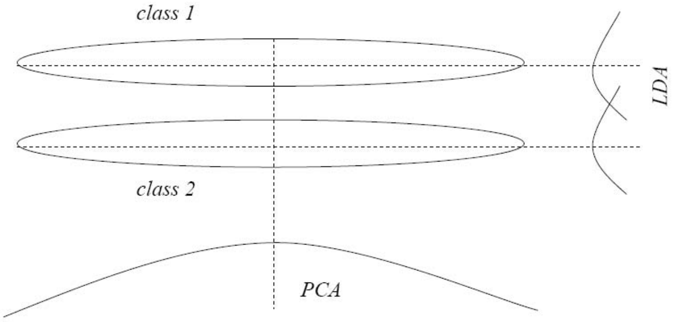

Figure 8: A two-class gaussian classification problem where PCA fails to discriminate correctly. Adapted from [49].

2.3 Principle component analysis

Principal component analysis (PCA) [66] is an old, non-parametric, but still interesting method to reduce data dimensionality. PCA is widely used in all forms of analysis (also in speech recognition [22,87,93]) due to its simplicity to extract relevant information from confusing data sets. 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 to transform a set of, usually, noisy data by keeping only the clean components of it and disclosing the hidden structure. 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.

Comparison of LDA, HLDA and PCA

In [49] an attempt is made to compare the discussed approaches in feature vector dimension reduction by pointing out their advantages. In PCA, the

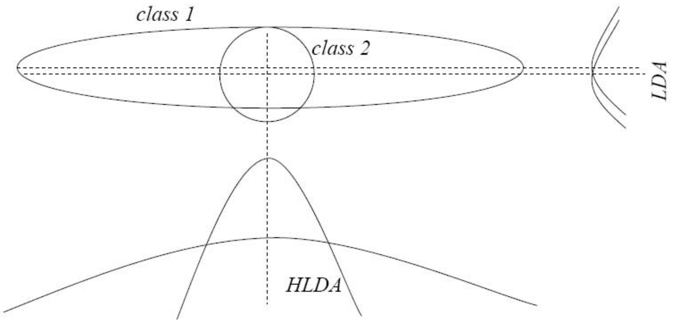

Figure 9: A two-class gaussian classification problem where LDA fails to discriminate correctly. Adapted from [49].

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 in the mean time every newly chosen component is uncorrelated to the prior. As it is not based on vector properties that are necessarily related to classification, this approach includes a somehow high risk of failure. It is not always the case that the chosen principal components involve the necessary information to discriminate essentially the classes in a pattern classification task.

Let assume that the classification task consists of two Gaussian distributions with equal variance in a two-dimensional sample space, that need to be discriminated. The general form of the problem is shown in Fig. 8. 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 choosing the correct direction. This is not the case with "PCA" in which the projection on it gives no discrimination result. The HLDA method would however work well.

Fig. 9 shows another example in which the LDA method fails this time. 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 in Sec. 2.1, 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. 9. For this problem though, even PCA would result in the best discrimination of the two classes.

2.4 Other methods

To further reduce the dimensionality of features sets, algorithms have been proposed to select optimal subsets. An approach is to find the maximum statistical dependency between a features subset and a class by computing the mutual information, e.g., [81, 95]. This method is computationally intractable. An alternative approach proposed in [21] and extended in [67], combines 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 especially useful for large-scale feature selection problems where a large number of features are available, e.g., in medical tasks [35, 94]. 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 (densities) - which is much easier - resulting in a better way to maximize the dependency. At each step, the approach selects those features that follow the mRMR criterion and hence is intended for features that are not independent of each other. In [67], the authors claim that the whole process is faster than other closely-related methods due to the lower computational complexity.

In [90], the maximum entropy discrimination (MED) [40–42] feature selection proposed for ASR. The results are comparable to a wrapper but the algorithm is less computationally expensive. In MED, each feature is associated to a probability weight value. Then, the M out of N most important features are 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, MED feature selection is faster. Finally, since MED is a Bayesian discriminative algorithm appears to have a good recognition rate.

2.5 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 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. The motivation to study the selection and design of robust acoustic features that maximize the similarity of the Euclidian geometry of the feature set and the human auditory representation of the signal comes from the accuracy of recent methods for auditory modeling [13, 91]. 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.

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" Euclidian distance must be similar in the two domains (except for a scaling). The similarity of "large" distances is immaterial for the classification. The results show that maximizing the similarity of the Euclidian 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).

Let consider Eq. (2) to be the perceptual-domain distortion measure. We can define a similar distance measure $\Gamma_i : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^+$ for the feature domain of feature set i. Let $\mathbf{c}_i : \mathbb{R}^N \to \mathbb{R}^L$ be the mapping from a signal segment $\mathbf{x}_j$ to a set of L features $\mathbf{c}_i(\mathbf{x}_j)$ with set index i. Then, the feature-domain distortion measure is

$$\Gamma_i(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m}) = \|\mathbf{c}_i(\mathbf{x}_j) - \mathbf{c}_i(\hat{\mathbf{x}}_{j,m})\|^2. \tag{34}$$

We define the similarity measure $G(i)$ in the perceptual-domain distortion and the feature-domain distortion as

$$G(i) = \sum_{j \in \mathcal{J}, m \in \mathcal{M}_j} \left[\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m}) - \lambda \Gamma_i(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})\right]^2, \tag{35}$$

where

$$\lambda = \frac{\sum_{j \in \mathcal{J}, m \in \mathcal{M}_j} \Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m}) \Gamma_i(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})}{\sum_{j \in J, m \in M_j} \Gamma_i(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})^2} \tag{36}$$

is an optimal scaling of the acoustic feature criterion. Given a finite sequence of frames $j \in \mathcal{J}$ and a finite set of acoustic perturbations $m \in \mathcal{M}_j$, our objective is to find the particular set of features i that minimizes Eq. (35).

The focus on small distances allows complex perceptual distortion measures to be reduced to quadratic distortion measures using the so-called *sensitivity matrix* (see Eq. (2)). This theme was first developed in the context of rate-distortion theory [31, 53, 54] and was later used for audio coding [69]. In the feature domain, it is possible to have analogous distortion measures that also use the sensitivity matrix. Let consider the mapping $\mathbf{c}_i$ to the feature domain. If the mapping $\mathbf{c}_i$ is analytic, the Taylor series can be used to make a local approximation around $\mathbf{x}_j$:

$$\mathbf{c}_i(\hat{\mathbf{x}}_{j,m}) \approx \mathbf{c}_i(\mathbf{x}_j) + \mathbf{A}[\hat{\mathbf{x}}_{j,m} - \mathbf{x}_j], \tag{37}$$

where $\mathbf{A} = \left.\frac{\partial \mathbf{c}_i(\mathbf{x}_j)}{\partial \hat{\mathbf{x}}_{j,m}}\right|_{\hat{\mathbf{x}}_{j,m}=\mathbf{x}_j}$. An L^2 distance measure in the feature domain then leads to a signal domain sensitivity matrix

$$\mathbf{D}_\Gamma(\mathbf{x}_j) = \mathbf{A}^T\mathbf{A}. \tag{38}$$

Thus, we can write the distortion $\Gamma_i(\mathbf{x}_j, \hat{\mathbf{x}}_{j,m})$ in the form of Eq. (2). The sensitivity matrix based expressions facilitate a fast evalution of Eq. (35). Appendix II presents the derivation of the **A** matrix for the MFCCs in both the frequency and time domains.

AMFS is related to other approaches that use auditory models as a front-end for ASR, e.g., [32, 34, 44, 83]. 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.

3 Summary of contributions

This thesis makes two major contributions:

- A novel method 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 psycho-acoustic auditory model for speech recognition (Paper B).

This work is described in more detail in three research papers that are included in the thesis. The main concept in all papers comes from Bastiaan Kleijn who had the overall supervision. Bastiaan Kleijn helped with the writing of the papers. In Paper A, the author did the theoretical derivations of **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. Finally, in Paper C the author did the implementations and the experiments and wrote the major part of the paper. Saikat Chatterjee helped with the algorithm. 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 the human perception performs. The optimization of the features is done off-line 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.

4 Conclusions

The goal of this thesis was to investigate the use of auditory modeling in the front-end of an ASR system. The proposed methods incorporate a combination of knowledge of the human periphery, speech signal processing, perturbation analysis techniques and acoustic modeling. The study of selecting features for speech recognition was explored using two sophisticated auditory models of different nature, i.e., a spectral only and a spectro-temporal psycho-acoustic model. Depending on the model used, we performed a selection of acoustic features considering static features only (Paper A) and a combination of static and dynamic features (Paper C). We conclude that the selection of speech features based on human perception results in robust features that generalize well to various environmental conditions. Furthermore, the proposed perceptual-distance preserving measure was also used to optimize the commonly used MFCCs in speech recognition (Paper B). The experimental results indicated a success of this optimization and the new features called *modified* MFCCs (MMFCCs) can be considered as the "proof" of our underlying assumption that the output of the auditory system is useful in increasing the accuracy of the modern speech recognition engines.

Appendix I LDA implementation

Before applying the LDA method, the features extraction and the speech recognition tasks should be performed. In our case, the generated features were the MFCCs [15]. Using the HTK toolkit [26], 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_s*2 (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
.............................
```

Thence, 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 named first in it. 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 [49, 51] 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. The mean of each class and of the whole database was 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 found 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 (22) 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 to the first part when the label files were read one after the other, but the difference was that no computation was performed in this phase of the method. 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 [1] 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. 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

$$\delta\mathbf{c} = \mathbf{A}\ \delta\mathbf{x}. \tag{39}$$

The steps of computing MFCCs starting from the end are:

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \mathbf{s}(m) \cos\left\{ q[m - \frac{1}{2}]\frac{\pi}{M} \right\}, q = 1...Q, \tag{40}$$

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

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \ln \mathbf{z}(m) \cos\left\{ q[m - \frac{1}{2}]\frac{\pi}{M} \right\}, \tag{41}$$

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

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \ln\left\{ \sum_{k=0}^{K-1} \mathbf{Y}(k)\mathbf{H}_m(k) \right\} \cos\left\{ q[m - \frac{1}{2}]\frac{\pi}{M} \right\}, \tag{42}$$

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

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \ln\left\{ \sum_{k=0}^{K-1} |\mathbf{X}(k)|^2 \mathbf{H}_m(k) \right\} \cos\left\{ q[m - \frac{1}{2}]\frac{\pi}{M} \right\}, \tag{43}$$

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

$$\mathbf{c}(q) = \sum_{m=0}^{M-1} \ln\left\{ \sum_{k=0}^{K-1} \left| \sum_{n=0}^{N-1} \mathbf{x}'(n) e^{-\frac{j2\pi kn}{N}} \right|^2 \mathbf{H}_m(k) \right\} \cos\left\{ q[m - \frac{1}{2}]\frac{\pi}{M} \right\}, \tag{44}$$

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

$$\mathbf{A}(q,n) = \frac{\partial\mathbf{c}(q)}{\partial\mathbf{s}(m)} \frac{\partial\mathbf{s}(m)}{\partial\mathbf{z}(m)} \frac{\partial\mathbf{z}(m)}{\partial\mathbf{Y}(k)} \frac{\partial\mathbf{Y}(k)}{\partial\mathbf{x}'(n)} \frac{\partial\mathbf{x}'(n)}{\partial\mathbf{x}(n)}. \tag{45}$$

In Paper A, the **A** matrix is shown in frequency domain. This covers the first three derivatives in Eq. (45). For the fourth factor, it can be shown that the periodogram $\mathbf{Y}(k)$ is given by

$$\mathbf{Y}(k) = \sum_{n=0}^{N-1} \mathbf{x}'^2(n) + 2\sum_{n=0}^{N-2} \sum_{m=n+1}^{N-1} \mathbf{x}'(n)\mathbf{x}'(m) \cos\left\{ \frac{2\pi k}{N}[n-m] \right\}. \tag{46}$$

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

$$\frac{\partial \mathbf{Y}(k)}{\partial \mathbf{x}'(n)} = 2\mathbf{x}'(n) + 2\sum_{\substack{h=0,\\ h\neq n}}^{N-1} \mathbf{x}'(h)\cos\left\{\frac{2\pi k}{N}[n-h]\right\}. \tag{47}$$

One can see that

$$\begin{aligned}\frac{\partial \mathbf{Y}(k)}{\partial \mathbf{x}'(n)} = 2\mathbf{x}'(n) + 2\sum_{\substack{h=0,\\ h\neq n}}^{N-1} \mathbf{x}'(h)\cos\left\{\frac{2\pi k}{N}[n-h]\right\} &= \\ 2\sum_{h=0}^{N-1} \mathbf{x}'(h)\cos\left\{\frac{2\pi k}{N}[n-h]\right\} &= \\ 2\Re\left\{\sum_{h=0}^{N-1} \mathbf{x}'(h)\left\{\cos\left\{\frac{2\pi k}{N}[n-h]\right\} + j\sin\left\{\frac{2\pi k}{N}[n-h]\right\}\right\}\right\} &= \\ 2\Re\left\{\sum_{h=0}^{N-1} \mathbf{x}'(h)e^{j\frac{2\pi k}{N}[n-h]}\right\} &= \\ 2\Re\left\{\left\{\sum_{h=0}^{N-1} \mathbf{x}'(h)e^{j\frac{2\pi k}{N}h}\right\}e^{-j\frac{2\pi k}{N}n}\right\} &= \\ 2\Re\left\{\mathbf{X}^*(k)e^{-j\frac{2\pi k}{N}n}\right\}&,\end{aligned} \tag{48}$$

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

$$\mathbf{A}_{qn} = \sum_{m=0}^{M-1} \cos\left\{q[m-\frac{1}{2}]\frac{\pi}{M}\right\}\frac{1}{\mathbf{z}(m)}\mathbf{H}_m(n)2\Re\left\{\mathbf{X}^*(k)e^{-j\frac{2\pi k}{N}n}\right\}\mathbf{w}(n), \tag{49}$$

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

References

[1] VOICEBOX. `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

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

[3] H. Abbasian, B. Nasersharif, A. Akbari, M. Rahmani, and M. S. Moin. Optimized linear discriminant analysis for extracting robust speech features. *ISCCSP 2008, Malta*, March 2008.

[4] X. Aubert, R. Haeb-Umbach, and H. Ney. Continuous mixture densities and linear discriminant analysis for improved context-dependent acoustic models. *IEEE Proc. of Internat. Conf. on Acoustics, Speech, and Signal Processing, Minneapolis, Minnesota, USA*, 2:648–651, 1993.

[5] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *Readings in speech recognition*, pages 308–319, 1990.

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

[7] E. Bocchieri and G. Doddington. Frame-specific statistical features for speaker-independent speech recognition. *IEEE Trans. Acoust. Speech and Signal Processing*, 34(4):755–764, August 1986.

[8] E. L. Bocchieri. Vector quantization for the efficient computation of continuous density likelihoods. *IEEE Proc. of Internat. Conf. on Acoustics, Speech, and Signal Processing, Minneapolis, Minnesota, USA*, pages 692–694, April 1993.

[9] P. F. Brown. *The* A*coustic-*M*odelling Problem in Automatic Speech Recognition.* PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A, 1987.

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

[11] R. Cole, J. Mariani, H. Uszkoreit, G. B. Varile, A. Zaenen, and A. Zampolli. *Survey of the state of the Art in Human Language Technology (Studies in Natural Language Processing).* Cambridge University Press, March 1998. ISBN 0521592771.

[12] G. Cook and A. Robinson. Transcribing broadcast news with the 1997 Abbot system. *Int. Conf. on Acoustics, Speech and Signal Processing, Seattle, WA, USA*, pages 917–920, 1998.

[13] T. Dau, D. Püschel, and A. Kohlrausch. A quantitative model of the effective signal processing in the auditory system. I. Model structure. *Acoustical Society of America*, 99(6):3615–3622, June 1996.

[14] T. Dau, D. Püschel, and A. Kohlrausch. A quantitative model of the effective signal processing in the auditory system. II. Simulations and measurements. *Acoustical Society of America*, 99(6):3623–3631, June 1996.

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

[16] K. Demuynck, J. Duchateau, and D. Van Compernolle. Optimal feature sub-space selection based on discriminant analysis. *EUROSPEECH'99, Budapest, Hungary*, pages 1311–1314, 1999.

[17] L. Deng and J. Ma. A statistical coarticulatory model for the hidden vocal-tract-resonance dynamics. *In Proceedings Eurospeech*, 4:1499–1502, 1999.

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

[19] V. Digalakis, J. R. Rohlicek, and M. Ostendorf. ML estimation of a Stochastic Linear System with the EM Algorithm and its application to Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 1(4), October 1993.

[20] W. R. Dillon and M. Goldstein. *Multivariate Analysis.* John Wiley and Sons, New York, NY, USA, 1984.

[21] C. Ding and H. C. Peng. Minimum redundancy feature selection from microarray gene expression data. *Proc. Second IEEE Computational Systems Bioinformatics Conf.*, pages 523–528, Aug 2003.

[22] P. Ding and Z. Liming. Speaker recognition using principal component analysis. *Proceedings of ICONIP 2001, 8th International Conference on Neural Information Processing, Shanghai*, 2001.

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

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

[25] A. J. Robinson et al. A neural network based, speaker independent, large vocabulary. *Proc. of the European Conf. on Speech Communication and Technology, Berlin, Germany*, pages 1941–1944, 1999.

[26] S. Young et al. *The HTK Book (for HTK Version 3.2).* Cambridge University, Engineering Department, UK, December 2002.

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

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

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

[30] J. Fritsch and M. Finke. ACID/HNN: Clustering hierarchies of neural networks for context-dependent connectionist acoustic modeling. *Int. Conf. on Acoustics, Speech and Signal Processing, Seattle, WA, USA*, pages 505–508, 1998.

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

[32] O. Ghitza. Auditory nerve representation as a basis for speech processing. In *Advances in Speech Signal Process.*, pages 453–485. Marcel Dekker, 1991.

[33] R. Haeb-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. *IEEE Int. Conf. on Acoust., Speech, Sig. Proc.*, 1:13–16, 1992.

[34] S. Haque, R. Togneri, and A. Zaknich. A temporal auditory model with adaptation for automatic speech recognition. volume 4, pages 1141–1144, 2007.

[35] E. Herskovits, H. C. Peng, and C. Davatzikos. A Bayesian morphometry algorithm. *IEEE Transactions in Medical Imaging*, 24(6):723–737, 2004.

[36] G. E. Hinton. Connectionist learning procedures. *Artificial Intelligence*, 40:185–234, 1989.

[37] H. G. Hirsch and D. Pearce. The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions. In *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millenium*, Paris, France, 2000.

[38] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* Prentice Hall, New Jersey, USA, 2001. ISBN 0-13-022616-5.

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

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

[41] T. Jebara. *Discriminaive, generative and imitative leaming.* PhD thesis, Media Laboratory MIT, Cambridge, MA, U.S.A, December 2001.

[42] T. Jebara and T. Jaakkola. Feature selection and Dualities in Maximum Entropy Discrimination. *UAI 2000*, July 2000.

[43] F. Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556, April 1976.

[44] W. Jeon; B.-H. Juang. A study of auditory modeling and processing for speech signals. volume 1, pages 929 – 932, 2005.

[45] D. Jurafsky and J. H. Martin. *Speech and Language Processing.* Prentice Hall, New Jersey, USA, 2000. ISBN 0-13-095069-6.

[46] L. Kanal and B. Chandrasekaran. On dimensionality and sample size in statistical pattern classification. *Pattern Recognition*, pages 225–234, 1971.

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

[48] C. Koniaris. Estimation of general identifiable state-space models. Master's thesis, Techical University of Crete (TUC), Chania, Greece, September 2006.

[49] N. Kumar. *Investigation of Silicon Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition.* PhD thesis, Johns Hopkins University, USA, 1997.

[50] N. Kumar and A. G. Andreou. A generalization of linear discriminant analysis in maximum likelihood framework. *Proceedings of Joint Meeting of American Statistical Association, Chicago, IL, USA*, August 1996.

[51] N. Kumar and A. G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Commun.*, 26(4):283–297, 1998. ISSN 0167-6393.

[52] L.J. Lee, H. Attias, and L. Deng. Variational inference and learning for segmental switching state space models of hidden speech dynamics. *In Proceedings ICASSP*, 1:920–923, 2003.

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

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

[55] L. Ljung. *System Identification: Theory for the User (2nd Edition)*. Prentice Hall PTR, New Jersey, USA, December 1998. ISBN 0136566952.

[56] J. Ma and L. Deng. Optimization of dynamic regimes in a statistical hidden dynamic model for conversational speech recognition. *In Proceedings Eurospeech, Budapest, Hungary*, 3:1339–1342, 1999.

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

[58] J. Ma and L. Deng. Efficient decoding strategy for conversational speech recognition using state-space models for vocal-tract-resonance dynamics. *In Proceedings Eurospeech, Aalborg, Denmark*, pages 603–606, 2001.

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

[60] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(2):225–238, June 1986.

[61] N. Merhav and Y. Ephraim. Hidden Markov modeling using the most likely state sequence. *IEEE International Conf. Acoust. Speech Signal Processing, Toronto, Canada*, pages 469–472, may 1991.

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

[63] N. Morgan and H. Bourlard. Continuous speech recognition: An introduction to hybrid HMM/Connectionist approach. *IEEE Signal Processing Magazine*, pages 25–42, 1995.

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

[65] M. Ostendorf and S. Roukos. A stochastic segment model for phoneme-based continuous speech recognition. *IEEE Trans. Acoustic Speech and Signal Processing*, 37(12):1857–1869, December 1989.

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

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

[68] J. Picone, S. Pike, R. Regan, T. Kamm, J. Bridle, L. Deng, Z. Ma, H. Richards, and M. Schuster. Initial evaluation of hidden dynamic models on conversational speech. *In Proceedings ICASSP, Phoenix, Arizona, USA*, 1:109–112, 1999.

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

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

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

[72] H. B. Richards and J. S. Bridle. The HDM: A segmental hidden dynamic model of coarticulation. *In Proceedings ICASSP*, 1:357–360, 1999.

[73] M. Richardson, J. Bilmes, and C. Diorio. Hidden-articulator Markov models for speech recognition. *ASR*, pages 133–139, 2000.

[74] M. Richardson, J. Bilmes, and C. Diorio. Hidden-articulator Markov models: performance improvements and robustness to noise. *ICSLP, Beijing, China*, 3:131–134, 2000.

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

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

[77] A.-V. I. Rosti. *Linear Gaussian Models for Speech Recognition.* PhD thesis, University of Cambridge, Wolfson College, UK, May 2004.

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

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

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

[81] P. Scanlon, D. P. W. Ellis, and R. Reilly. Using mutual information to design class specific phone recognizers. *Proceedings European Conference Speech Technology, Geneva, Switzerland*, pages 857–860, 2003.

[82] F. Seide, J. L. Zhou, and L. Deng. Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM - MAP decoding and evaluation. *In Proceedings ICASSP, Hong Kong, Hong Kong*, 1:748–751, 2003.

[83] S. Seneff. A joint synchrony/mean rate model of auditory speech processing. *J. Phonet.*, 16:55–76, 1988.

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

[85] O. Siohan. On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition. *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Detroit, Michigan, USA*, 1:125–128, 1995.

[86] S. S. Stevens and J. Volkman. The relation of the pitch to frequency. *Journal of Phychology*, 53:329, 1940.

[87] S. N. Tsai and L. S. Lee. Improved robust features for speech recognition by integrating time-frequency principal components (TFPC) and histogram equalization (HEQ). *IEEE Workshop on Automatic Speech Recognition and Understanding, Virgin Islands, USA*, pages 297–302, 2003.

[88] G. Tsontzos, V. Diakoloukas, C. Koniaris, and V. Digalakis. Estimation of general identifiable linear dynamic models with an application in speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, Honolulu, Hawaii, USA*, 4:IV–453–IV–456, April 2007. ISSN 1520-6149.

[89] ETSI ES 201 108 v1.1.2. Speech processing transmission and quality aspects; distributed speech recognition; front-end feature extraction algorithm; compression algorithms. 2000-2004.

[90] F. Valente and C. Wellekens. Maximum entropy discrimination (MED) feature subset selection for speech recognition. *IEEE Works. on ASRU, Virgin Islands, USA*, pages 327–332, December 2003.

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

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

[93] Z. Wanfeng, Y. Yingchun, W. Zhaohui, and S. Lifeng. Experimental evaluation of a new speaker identification framework using PCA. *IEEE International Conference on Systems, Man and Cybernetics*, 5:4147–4152, 2003.

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

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

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

[97] J. L. Zhou, F. Seide, and L. Deng. Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM - model and training. *In Proceedings ICASSP, Hong Kong, Hong Kong*, 1:744–747, 2003.

[98] E. Zwicker and H. Fastl. *Psychoacoustics, Facts and Models.* Springer, Heidelberg, Germany, 1999.