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

Christos Koniaris

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

Stockholm 2009
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:


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 ix
List of Tables xi

I Introduction 1
1 Perception and speech recognition .......................... 1
  1.1 Human hearing system .................................. 2
  1.2 Auditory models ........................................ 4
  1.3 Front-End ............................................... 7
  1.4 Acoustic modeling ...................................... 9
2 Reducing features dimensionality ............................ 14
  2.1 Linear discriminant analysis .............................. 15
  2.2 Heteroscedastic linear discriminant analysis .......... 16
  2.3 Principle component analysis ............................ 18
  2.4 Other methods .......................................... 20
  2.5 Auditory model-based feature selection ................. 20
3 Summary of contributions ................................... 22
4 Conclusions .................................................. 24
Appendix I LDA implementation ................................. 25
Appendix II Derivation of the A matrix ......................... 28
References ..................................................... 30

II Included papers 39

A Auditory-model based robust feature selection for speech recognition A1
  1 Introduction .............................................. A1
  2 Maximizing similarity of feature and perceptual domain .. A2
    2.1 A distance preservation measure ....................... A3
    2.2 Perturbation analysis ................................ A4
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The anatomy of human ear.</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Block diagram of a channel of the van de Par psycho-acoustic model.</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>Block diagram of a channel of the Dau psycho-acoustic model.</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>Extracting features from speech signal.</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>The mel filterbank.</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>A real speech recognition system.</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>A hidden Markov model.</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>A two-class gaussian classification problem where PCA fails to discriminate correctly. Adapted from [49].</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>A two-class gaussian classification problem where LDA fails to discriminate correctly. Adapted from [49].</td>
<td>19</td>
</tr>
<tr>
<td>10</td>
<td>Scatter plots of the estimated $\delta c$'s vs. $\delta c_{true}$'s for the first and second MFCC, respectively.</td>
<td>A6</td>
</tr>
<tr>
<td>11</td>
<td>The $J$'th superframe consists of three overlapped subframes $j-1, j, j+1$.</td>
<td>C6</td>
</tr>
</tbody>
</table>
List of Tables

1. AURORA2 clean training results. The \textit{amfs} selected coefficients with indices \(1,2,3,4,9,10,11,12\) for \(n=8\) and \(1,2,3,12\) for \(n=4\), respectively. A7

2. AURORA2 multi-conditioning noisy training results. The \textit{amfs} selected coefficients with indices \(1,2,3,4,5,6,11,12\) for \(n=8\) and \(1,2,3,4\) for \(n=4\), respectively. A8

1. Phone recognition accuracy (in \%) of static 12-dimensional MFCC and MMFCC features using TIMIT. B8

2. Robust word recognition accuracy (in \%) of 39-dimensional MFCC and MMFCC features using Aurora 2. B8

3. Robust phone recognition accuracy (in \%) of 39-dimensional MFCC and MMFCC features at 10 dB SNR. B9

1. AURORA2 clean training results. C8

2. AURORA2 multi-conditioning noisy training results. C9
## Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNs</td>
<td>Artificial Neural Networks</td>
</tr>
<tr>
<td>AMFS</td>
<td>Auditory Model-based Feature Selection</td>
</tr>
<tr>
<td>ASR</td>
<td>Automatic Speech Recognition</td>
</tr>
<tr>
<td>DCT</td>
<td>Discrete Cosine Transform</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>ERB</td>
<td>Equivalent Rectangular Bandwidth</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation-Maximization algorithm</td>
</tr>
<tr>
<td>IHC</td>
<td>Inner Hair Cells</td>
</tr>
<tr>
<td>KLT</td>
<td>Karhunen-Loève Transform</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum A-Posteriori Probability</td>
</tr>
<tr>
<td>MED</td>
<td>Maximum Entropy Discrimination</td>
</tr>
<tr>
<td>MFCCs</td>
<td>Mel Frequency Cepstrum Coefficients</td>
</tr>
<tr>
<td>mRMR</td>
<td>minimal-Redundancy-Maximal-Relevance</td>
</tr>
<tr>
<td>MMFCCs</td>
<td>Modified Mel Frequency Cepstrum Coefficients</td>
</tr>
<tr>
<td>MLP</td>
<td>Multi-Layer Perceptron</td>
</tr>
<tr>
<td>HLDA</td>
<td>Heteroscedastic Linear Discriminant Analysis</td>
</tr>
<tr>
<td>HAMM</td>
<td>Hidden Articulatory Markov Model</td>
</tr>
<tr>
<td>HDMs</td>
<td>Hidden Dynamic Models</td>
</tr>
<tr>
<td>HMMs</td>
<td>Hidden Markov Models</td>
</tr>
<tr>
<td>HTK</td>
<td>Hidden Markov Model Toolkit</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>OHC</td>
<td>Outer Hair Cells</td>
</tr>
<tr>
<td>OM</td>
<td>Outer and Middle Ear</td>
</tr>
<tr>
<td>PDP</td>
<td>Parallel Distributed Processing</td>
</tr>
<tr>
<td>PCA</td>
<td>Principle Component Analysis</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>POD</td>
<td>Proper Orthogonal Decomposition</td>
</tr>
<tr>
<td>SMs</td>
<td>Segment-based Models</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>VQ</td>
<td>Vector Quantization</td>
</tr>
</tbody>
</table>
“Λίγο ακόμα θα ιδούμε
tις αμυγδαλίες ν’ ανθίζουν.
Λίγο ακόμα θα ιδούμε
tα μάρμαρα να λάμπουν,
να λάμπουν στον ήλιο
και τη 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.](image)

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
Introduction

4 Introduction

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

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

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

\[
\Upsilon(x_j, \hat{x}_{j,m}) \approx [\hat{x}_{j,m} - x_j]^T D_\Upsilon(x_j)[\hat{x}_{j,m} - x_j]. \tag{2}
\]

The matrix \( D_\Upsilon(x_j) = \frac{\partial^2 \Upsilon(x_j, \hat{x}_{j,m})}{\partial \hat{x}_\kappa \partial \hat{x}_\mu} \bigg|_{\hat{x}_{j,m} = x_j} \) is called sensitivity matrix. The word “sensitivity” refers to the fact that each element of this matrix represents the sensitivity of the distortion \( \Upsilon(x_j, \hat{x}_{j,m}) \) to a particular \( [\hat{x}_{j,m} - x_j] \).

In the next two paragraphs, two different auditory models are presented that are used to extract the sensitivity matrix.
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 an equivalent rectangular bandwidth (ERB) scale. The model consists of several channels $f$, in each of which the ratio of the distortion $x - \hat{x}$ to masker $x$ is estimated, where $x$ 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

$$\Upsilon(x, \hat{x}) = C_s L_e \sum_{g \in G} \sum_{f=0, \ldots, N-1} \frac{1}{N} |h_{om}(f)|^2 |\gamma_i(f)|^2 |x(f) - \hat{x}(f)|^2,$$

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 $\gamma_i$ 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

\[
D_{T,f,f}(x) \approx 2C_sL_e \sum_i \frac{1}{N} \frac{\sum_f |h_{om}(f)|^2 |\gamma_i(f)|^2}{\sum_f \sum_f |\delta_{om}(f)|^2 |\gamma_i(f)|^2 |x(f)|^2 + C_a}.
\] (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 \( x \) 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 \( a(l)(x_j) \), where \( x_j \) 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

\[
\Upsilon(x_j, x_{j,m}) = \sum_l \| a^{(l)}(x'_j) - a^{(l)}(x'_{j,m}) \|^2,
\] (5)

where \( x'_j, x'_{j,m} \) are of higher dimension than the \( x_j, x_{j,m} \) 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 $D^{(l)}_Y(x_j)$

$$D_Y(x_j) = \sum_l D^{(l)}_Y(x_j), \tag{6}$$

where

$$D^{(l)}_Y(x_j) = 2 \left[ \prod_k J^{(l)}_k \right]^H J^{(l)}_k, \tag{7}$$

and $J^{(l)}_k$ 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.](image-url)

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

\[
H_m(k) = \begin{cases} 
0, & k < f(m - 1) \\
\frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m - 1) \leq k \leq f(m) \\
\frac{f(m-1)-k}{f(m+1)-f(m)}, & f(m) \leq k \leq f(m + 1) \\
0, & k > f(m + 1) 
\end{cases}
\]  

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

The speech signal is first pre-emphasized \( x(n) = \hat{x}(n) - \alpha \hat{x}(n - 1) \), where \( \hat{x}(n) \) 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

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

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

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

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

\[
s(m) = \ln \left[ \sum_{k=0}^{K-1} |X(k)|^2 H_m(k) \right], \]  

where \( |X(k)|^2 \) is the periodogram, \( H_m(k) \) 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

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

where \( Q \) is the number of cepstrum coefficients, and \( s(m) \) represents the logarithmic mel spectrum of the \( m \)’th filter of the filterbank.
Usually, the first and the second time derivatives are added to the speech vector, $\Delta c$ and $\Delta \Delta c$ to better capture time dependencies [26]. These are calculated as

$$\Delta c_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2},$$  \hspace{1cm} (13)$$

and

$$\Delta \Delta c_t = \frac{\sum_{\theta=1}^{\Theta} \theta (\Delta c_{t+\theta} - \Delta c_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2},$$  \hspace{1cm} (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
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) $P(W|O)$

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

The above formula is known as Bayes’ theorem. Usually, the likelihood of the observation sequence in the denominator, $P(O) = \sum P(O|W)P(W)$, 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].

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

- Number of states: \( N \)
- Number of distinct observation symbols: \( M \) for discrete HMMs and \( \infty \) for continuous HMMs
- State transition probability distribution: \( a_{i,j} \)
- Output distribution of state \( j \): \( b_j(o) \)
- Initial state probability: \( \pi_i \)

![Figure 7: A hidden Markov model.](image-url)
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 $\pi$, the state transition matrix, the output distribution matrix and the initialization matrix, respectively. For convenience, we use the following notation

$$\lambda = (A, O, \pi)$$

(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 restric-
tions 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 $\phi^T$ 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

$$y = \phi^T c,$$

where $y$ 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 $c$ 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

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

$$\Sigma_j = \frac{1}{N_j} \sum_{i=1}^{N_j} (c_i - \mu_j)(c_i - \mu_j)^T,$$

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

\[ \mu = \frac{1}{N} \sum_{i=1}^{N} c_i, \]  

(20)

\[ \Sigma = \frac{1}{N} \sum_{i=1}^{N} (c_i - \mu)(c_i - \mu)^T, \]  

(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

\[ \hat{\phi} = \arg \max_{\phi} \frac{\phi^T \Sigma \phi_p}{\phi^T S \phi_p}, \]  

(22)

where

\[ S = \frac{1}{N} \sum_{j=1}^{J} N_j \Sigma_j. \]  

(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 \( S^{-1} \Sigma \) 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 \( \phi \) is a \( q \times q \) matrix, and thus differs from the LDA, that again is applied in the original feature vector as

\[ y = \phi^T c, \]  

(24)

with \( y \in \mathbb{R}^p \) where \( p \) is the dimension of the feature vector in the reduced feature domain and \( c \in \mathbb{R}^q \) where \( q \) refers to the dimension of the feature vector in the original feature domain. The transformation \( \phi \) is applied to
the original feature vector, however from the resulting transformed vector \( y \), only the first \( p \) elements are retained. The latest is based on the assumption that only the first \( p \) components of \( y \) 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

\[
\mu_j = \begin{bmatrix} \mu^p_j \\ \mu \end{bmatrix},
\]

and

\[
\Sigma_j = \begin{bmatrix} \Sigma^p_j & 0 \\ 0 & \Sigma^{q-p} \end{bmatrix},
\]

where \( \mu_j, \Sigma_j \) are the mean and covariance for the class \( j \), respectively.

The parameters \( \mu^p_j \) and \( \Sigma^p_j \) are different for each class while \( \mu \) and \( \Sigma \) are common. Then, the Gaussian PDF of \( c_i \) is given by the following equation

\[
P(c_i) = |\varphi| \sqrt{(2\pi)^q |\Sigma_g(i)|} \exp \left\{ -\frac{1}{2} (y_i - \mu_g(i))^T \Sigma_g^{-1}(i) (y_i - \mu_g(i)) \right\},
\]

where \( y_i = \varphi^T c_i \), 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 \( \varphi \), is then

\[
\log P(\mu_j, \Sigma_j, \varphi; \{c_i\}) = N \log |\varphi| - \frac{1}{2} \sum_{i=1}^{N} \left\{ \log[(2\pi)^q |\Sigma_g(i)|] + [\varphi^T c_i - \mu_g(i)]^T \Sigma_g^{-1}(i) [\varphi^T c_i - \mu_g(i)] \right\}.
\]

Considering the derivatives versus \( \mu_j \) and \( \Sigma_j \), and setting them equal to zero, the following estimates arise

\[
\hat{\mu}_j = \varphi^T c_j,
\]

\[
\hat{\Sigma}_j = \varphi^T \Sigma_j \varphi,
\]

and

\[
\hat{\Sigma} = \varphi^T \Sigma \varphi,
\]

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 \( \varphi \) is given by

\[
\hat{\varphi} = \arg \max_{\varphi} \left\{ -\frac{N}{2} \log |\varphi^T \Sigma \varphi| - \sum_{j=1}^{J} \frac{N_j}{2} \log |\varphi^T \Sigma_j \varphi| + N \log |\varphi| \right\}.
\]
In [51], the maximization of the above equation was performed using numerical methods. The \( \hat{\phi} \) is initialized by the \( \hat{\phi} \) computed previously by the LDA method.

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 \( \psi \) with these eigenvectors in the columns. Finally, the feature vectors are transformed using the resulted transformation matrix \( \psi \).

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
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 combina-
tions 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 distri-
butions 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 maxi-
mum 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 feature 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 effec-
tive 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 $\Gamma_i$:

\[
\Gamma_i(x_j, \hat{x}_{j,m}) = \left\| c_i(x_j) - c_i(\hat{x}_{j,m}) \right\|^2.
\]  

(34)

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

\[
G(i) = \sum_{j \in J, m \in M_j} \left[ \Upsilon(x_j, \hat{x}_{j,m}) - \lambda \Gamma_i(x_j, \hat{x}_{j,m}) \right]^2,
\]  

(35)

where

\[
\lambda = \frac{\sum_{j \in J, m \in M_j} \Upsilon(x_j, \hat{x}_{j,m}) \Gamma_i(x_j, \hat{x}_{j,m})}{\sum_{j \in J, m \in M_j} \Gamma_i(x_j, \hat{x}_{j,m})^2}
\]  

(36)

is an optimal scaling of the acoustic feature criterion. Given a finite sequence of frames $j \in J$ and a finite set of acoustic perturbations $m \in M_j$, 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 $c_i$ to the feature domain. If the mapping $c_i$ is analytic, the Taylor series
can be used to make a local approximation around $x_j$:

$$c_i(\hat{x}_{j,m}) \approx c_i(x_j) + A[\hat{x}_{j,m} - x_j],$$

(37)

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

$$D \Gamma(x_j) = A^T A.$$ 

(38)

Thus, we can write the distortion $\Gamma_i(x_j, \hat{x}_{j,m})$ in the form of Eq. (2). The sensitivity matrix based expressions facilitate a fast evaluation 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., \cite{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 Euclidean 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 signifi-
cantly 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 combi-
nation of knowledge of the human periphery, speech signal processing, per-
turbation analysis techniques and acoustic modeling. The study of selecting
features for speech recognition was explored using two sophisticated audi-
tory models of different nature, i.e., a spectral only and a spectro-temporal
psycho-acoustic model. Depending on the model used, we performed a se-
lection 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. Further-
more, 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 sys-
tem 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_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"
Thence, the word-class label files accompanied by the MFCCs were read serially, and the class and the overall data statistics $\mu_j$, $\Sigma_j$, $\mu$, $\Sigma$ 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 $\Sigma_j$ and $\Sigma$, 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 $S^{-1}\Sigma$ in a matrix that corresponds to the $p$ largest eigenvalues. The output is the transformation matrix $\phi^T$.

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 $\phi^T$, 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 $\phi^T$. On the other hand, if someone would try to have a separate transformation for each individual case, a single $\phi^T$ 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 occurred since all the data were clean, and hence the transformation was computed based on a homomorphous 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 $\delta x = \hat{x} - x$ and the corresponded distortion in MFCCs $\delta c = \hat{c} - c$ is

$$\delta c = A \delta x.$$  \hfill (39)

The steps of computing MFCCs starting from the end are:

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

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

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

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

$$c(q) = \sum_{m=0}^{M-1} \ln \left\{ \sum_{k=0}^{K-1} |X(k)|^2 H_m(k) \right\} \cos \left\{ q[m - \frac{1}{2}] \frac{\pi}{M} \right\},$$  \hfill (42)

in which $X(k)$ denotes the DFT of the signal or finally as

$$c(q) = \sum_{m=0}^{M-1} \ln \left\{ \sum_{k=0}^{K-1} x'(n) e^{-j2\pi nk/N} H_m(k) \right\} \cos \left\{ q[m - \frac{1}{2}] \frac{\pi}{M} \right\},$$  \hfill (43)

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

$$A(q,n) = \frac{\partial c(q)}{\partial s(m)} \frac{\partial s(m)}{\partial z(m)} \frac{\partial z(m)}{\partial Y(k)} \frac{\partial Y(k)}{\partial x'(n)} \frac{\partial x'(n)}{\partial x(n)}.$$  \hfill (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 $Y(k)$ is given by

$$Y(k) = \sum_{n=0}^{N-1} x'^2(n) + 2 \sum_{n=0}^{N-2} \sum_{m=n+1}^{N-1} x'(n)x'(m) \cos \left\{ \frac{2\pi k}{N} |n-m| \right\}.$$  \hfill (46)
Then its derivative \( \frac{\partial Y(k)}{\partial x'(n)} \), i.e., the derivative of the periodogram with respect to the windowed signal is

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

(47)

One can see that

\[
\frac{\partial Y(k)}{\partial x'(n)} = 2x'(n) + 2 \sum_{h=0, h \neq n}^{N-1} x'(h) \cos \left\{ \frac{2\pi k}{N} [n - h] \right\} = 2 \sum_{h=0}^{N-1} x'(h) \cos \left\{ \frac{2\pi k}{N} [n - h] \right\} = 2 \sum_{h=0}^{N-1} x'(h) e^{j \frac{2\pi k}{N} [n - h]} = 2 \sum_{h=0}^{N-1} x'(h) e^{j \frac{2\pi k}{N} [n - h]} = 2 \sum_{h=0}^{N-1} x'(h) e^{j \frac{2\pi k}{N} h} e^{-j \frac{2\pi k}{N} n} = 2 \Re \left\{ X^*(k) e^{-j \frac{2\pi k}{N} n} \right\}.
\]

(48)

where \( X^*(k) \) is the conjugate of the DFT of the signal.

Finally, the formula of matrix \( A \) in time domain is given by

\[
A_{qn} = \sum_{m=0}^{M-1} \cos \left\{ q[m - \frac{1}{2}] \frac{\pi}{M} \right\} \frac{1}{z(m)} H_m(n) 2 \Re \left\{ X^* (k) e^{-j \frac{2\pi k}{N} n} \right\} w(n),
\]

(49)

where \( w(n) \) is the hamming window.
References


Part II

Included papers
Auditory-model based robust feature selection for speech recognition

Christos Koniaris, Marcin Kuropatwinski, W. Bastiaan Kleijn

submitted
The layout has been revised for thesis consistency
Auditory-model based robust feature selection for speech recognition

Christos Koniaris, Marcin Kuropatwinski, W. Bastiaan Kleijn

Abstract

It is shown that robust dimension-reduction of a feature set for speech recognition can be based on a model of the human auditory system. Whereas conventional methods optimize classification performance, the new method exploits knowledge implicit in the auditory periphery, inheriting its robustness. Features are selected to maximize the similarity of the Euclidian geometry of the feature domain and the perceptual domain. Recognition experiments confirm the effectiveness of the approach, showing that features can be selected without labeled training data. For noisy data the method outperforms commonly used dimension-reduction methods that rely on labeling; for clean data the performance of all systems is similar.

1 Introduction

The extraction of acoustic features is an essential component of automatic speech recognition (ASR). It enables the classification of speech signals at a reasonable computational complexity based on training with speech databases of a practical size. However, the data processing inequality implies that the extraction of acoustic features from a speech signal at best preserves information relevant for phone discrimination. Thus, careful selection of the acoustic features is essential.

The existing approach for selecting features from a larger set of candidate features is based on direct optimization of classification performance, using labeled training databases. Many algorithms have been developed to select features for classification [5, 12, 13, 15]. In ASR it is common to use dimension-reduction procedures [5, 9, 17], a more general paradigm where input features are combined into a new set of lower cardinality. In general, existing feature-selection and dimension-reduction methods require classi-
A2 Auditory-model based robust feature selection for speech recognition

fied training data. For ASR this means that dimension-reduction methods are sensitive to differences in training and testing conditions.

In this paper, a fundamentally different principle is proposed for feature selection for ASR: to exploit the knowledge implicit in the human auditory system. Importantly, this means the new method does not require labeled training data. Humans perform better at speech recognition than machines, particularly for noisy environments. Recently, accurate models of the periphery have become available [1, 18]. This motivates the selection of a subset of acoustic features from a larger set by maximizing the similarity of the Euclidian geometry of the selected feature set and the human auditory representation of the signal.

The implementation of our approach relies on perturbation theory. For two features sets to perform similarly in classification, “small” Euclidian distances must be similar in the two domains (except for a scaling). The similarity of “large” distances is immaterial for the classification. The implementation is based on the so-called sensitivity matrix, which was first developed in the context of rate-distortion theory [3, 10, 11] and has been used for audio coding [14].

The present work is related to the many studies on the usage of auditory models as a front-end for ASR, e.g., [4, 6, 8, 16]. The performance for such front-ends is generally robust to variations in environmental conditions. Importantly, the new approach removes the computational complexity associated with pre-processing the signal with an auditory model. It also avoids the difficulty of formatting the auditory-model output for classification.

A side outcome of our work is that it provides a measure of relative importance of a set of features. In this first study, the most commonly used set of static features, the mel-frequency cepstral coefficients (MFCC) [2], are used. The results confirm that the human auditory model is a good guide for the selection of robust acoustic features. They also show that the initial set of MFCCs corresponds to perceptually important information.

This paper is organized as follows. Sec. 2 discusses a similarity measure for the perceptual and feature domains. Sec. 3 applies the method to ASR. Sec. 4 confirms with experiments that the selected features are effective and robust and Sec. 5, provides conclusions.

2 Maximizing similarity of feature and perceptual domain

Our objective is to select, from a larger set of features, a subset of features that provides a separation of sound classes that is close to that obtained by state-of-the-art auditory models. Ideally, this implies an isometry between the perceptual domain and the selected acoustic feature domain. The mapping from the perceptual domain to the acoustic-feature domain would then
be distance preserving. To obtain the best approximation to this ideal scenario, we define a new, objective criterion in this section. Thus, we avoid the ad-hoc nature of many auditory-system inspired features.

The motivation for the objective is that human recognition performance indicates that the human auditory periphery provides a relatively good separation of sound classes. We postulate that little information relevant for sound classification is lost in the mapping from the acoustic domain to the perceptual domain. It is, however, not clear if the representation is redundant.

2.1 A distance preservation measure

It is not possible to design acoustic features that are a distance-preserving mapping from the perceptual domain. For accurate classification, the preservation of the data geometry near the class boundaries is most critical. The preservation of distances that are short relative to classification boundary curvature is important, whereas the preservation of “long” distances is not important.

Distance measures must be defined in both the perceptual and feature domains. Let \( x_j \in \mathbb{R}^N \) denote the \( N \)-dimensional speech signal vector characterizing a segment with time index \( j \in \mathbb{Z} \) and let \( \hat{x}_{j,m} \) be a perturbation of \( x_j \) with perturbation index \( m \). A perceptual-domain distortion is defined as a surjective mapping of two signals: \( \Upsilon : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+ \), where \( \mathbb{R}^+ \) are the nonnegative reals. Perceptual distortion measures are commonly based on the \( L^2 \) norm of the difference between the perceptual-domain signals \( y(x_j) \) and \( y(\hat{x}_{j,m}) \), where \( y : \mathbb{R}^N \rightarrow \mathbb{R}^K \) is a mapping to the \( (K\text{-dimensional}) \) perceptual domain, \( \Upsilon(x_j, \hat{x}_{j,m}) = \|y(x_j) - y(\hat{x}_{j,m})\|^2 \). This measure is the desired distance measure in the perceptual domain.

A similar distortion measure \( \Gamma_i : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+ \) can be defined for the feature domain of feature set \( i \). Let \( c_i : \mathbb{R}^N \rightarrow \mathbb{R}^L \) be the mapping from a signal segment \( x_j \) to a set of \( L \) features \( c_i(x_{j}) \) with set index \( i \). An \( L^2 \) norm based measure is then: \( \Gamma_i(x_j, \hat{x}_{j,m}) = \|c_i(x_j) - c_i(\hat{x}_{j,m})\|^2 \).

Given a finite sequence of frames \( j \in J \) and a finite set of acoustic perturbations \( m \in M_j \), the distance-preservation objective leads to the objective to find the particular set of features \( i \) that minimizes a measure of dissimilarity in the perceptual-domain distortion and the feature-domain distortion,

\[
G(i) = \sum_{j \in J, m \in M_j} \left[ \Upsilon(x_j, \hat{x}_{j,m}) - \lambda \Gamma_i(x_j, \hat{x}_{j,m}) \right]^2, \tag{1}
\]

where \( \lambda = \frac{\sum_{j \in J, m \in M_j} \Upsilon(x_j, \hat{x}_{j,m}) \Gamma_i(x_j, \hat{x}_{j,m})}{\sum_{j \in J, m \in M_j} \Gamma_i(x_j, \hat{x}_{j,m})^2} \) is an optimal scaling of the acoustic feature criterion. Eq. (1) can be interpreted as a measure of proximity to isometry.
2.2 Perturbation analysis

While it is possible to evaluate Eq. (1) directly even for complex distortion measures, this can be computationally expensive. For short distances, the perceptual distortion measure $T(x_j, \hat{x}_{j,m})$ and the feature-based distortion measure $\Gamma(x_j, \hat{x}_{j,m})$ can be approximated with simpler quadratic measures, reducing the computational complexity. The approach is based on the sensitivity matrix framework \[3,10,11,14\].

The perturbation analysis for the perceptual domain and the feature domain is identical, and we only describe the first case. For notational brevity we omit the subscripts indicating frame number and perturbation where no ambiguity exists. First, let us consider $\Upsilon(x, \hat{x})$ to be known. We assume that $\Upsilon(x, x) = 0$ and that this forms a minimum. We furthermore assume that $\Upsilon(x, \hat{x})$ is analytic in $\hat{x}$. Then, for sufficiently small perturbations $\hat{x} - x$, we can make the approximation

$$\Upsilon(x, \hat{x}) \approx [\hat{x} - x]^T D\Upsilon(x)[\hat{x} - x],$$

(2)

where $D\Upsilon_{ij}(x) = \frac{\partial^2 \Upsilon(x, \hat{x})}{\partial \hat{x}_i \partial \hat{x}_j} \bigg|_{\hat{x}=x}$ is the sensitivity matrix.

It is common that the mapping from $x$ to the perceptual or feature domain is given, rather than the distortion criterion. Consider the mapping $c$ to the feature domain. If the mapping $c$ is analytic, the Taylor series can be used to make a local approximation around $x$:

$$c(\hat{x}) \approx c(x) + A[\hat{x} - x],$$

(3)

where $A = \frac{\partial c(x)}{\partial x} \bigg|_{x=x}$. An $L^2$ distance measure in the feature domain then leads to a signal domain sensitivity matrix

$$D\Gamma(x) = A^T A.$$

(4)

Thus, we can write the distortion $\Gamma(x, \hat{x})$ in the form of Eq. (2). The sensitivity matrix based expressions facilitate a fast evaluation of Eq. (1).

3 Application to speech recognition

In our algorithm the perceptual domain is the domain of the output vectors of the auditory model used. This section illustrates the application of the method to a specific auditory model and specific type of acoustic features.

3.1 van de Par auditory model

The van de Par [18] auditory model is a static psycho-acoustic masking model. As it uses the magnitude spectrum as input, the vector $x_j$ characterizing speech segment $j$ is now the (square-root) periodogram. The model
consists of channels $f$, in each of which the ratio of the distortion $x - \hat{x}$ to masker $x$ is estimated, where $x$ 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

$$\Upsilon(x, \hat{x}) = C_s L_e \sum_{g \in \mathcal{G}} \frac{1}{N} \sum_{f=a_0, \ldots, a_{N-1}} |h_{om}(f)|^2 |\gamma_i(f)|^2 |x(f) - \hat{x}(f)|^2,$$

where $C_s$ and $C_a$ are constants calibrated using 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 $\gamma_i$ is the $i$'th gammatone filter.

Combining Eq. (2) and Eq. (5), the sensitivity matrix $D_\Upsilon(x)$ can be obtained. It is a diagonal matrix with the diagonal element for row and column $f$ given by

$$D_{\Upsilon,ff}(x) \approx 2C_s L_e \sum_{r \in \mathcal{R}} \frac{1}{N} \sum_{f} |h_{om}(f)|^2 |\gamma_i(f)|^2 |x(f) + C_a. (6)$$

### 3.2 Local linearization of the MFCCs

In our experiments, the mel-frequency cepstral coefficients (MFCCs) [2] were used since they are the most commonly used acoustic features. The MFCCs are defined as

$$c(q) = \sum_{m=0}^{M-1} \ln \left\{ \sum_{n=0}^{N-1} x(n)H_m(n) \right\} \cos \left\{ q[m - \frac{1}{2} \frac{\pi}{M}] \right\}, q = 1, \ldots, Q, (7)$$

where $x(n)$ is the periodogram, $H_m(n)$ is the $m$’th triangular mel-filter, $m$ is the filter index, $M$ is the number of triangular bandpass filters used, and $Q$ is the number of cepstrum coefficients.

Sec. 2.2 introduced the matrix $A$ that characterizes the local relation between the features and the signal $x$. For the MFCCs, the matrix $A$ is

$$A_{qn} = \sum_{m=0}^{M-1} \cos \left\{ q[m - \frac{1}{2} \frac{\pi}{M}] \right\} \frac{H_m(n)}{\sum_{n=0}^{N-1} x(n)H_m(n)}.$$


4 Experimental results

This section examines the plausibility of the linearity assumption used in the perturbation method and verifies the robust performance of the selected feature sets. All experiments were performed on MFCCs.

The MFCCs were extracted from the AURORA2 [7] database, sampled at 8 kHz, using a Hamming window of 25 ms with an overlap of 12.5 ms. The DFT dimensionality was 256 and the number of filters used was 23. A set of 12 conventional MFCCs was extracted.

4.1 Range of linearity

The range of validity for the linearization assumption between the cepstrum and the speech was examined first. The speech was distorted with i.i.d. Gaussian noise at different SNRs ranging from 30 to 90 dB.

Fig. 1 shows the change in the features computed from the linearized relation Eq. (3) versus the true difference between the cepstra of the original and distorted signals, for the first and second MFCC, respectively. The linearity assumption is reasonable at a scale that is meaningful for sound discrimination.

4.2 Speech recognition experiments

We performed recognition experiments on features derived from the standard set of 12 MFCCs. We compared five types of feature sets for identical dimensionalities \( n < 12 \). The first set of features results from the auditory-model based feature selection (amfs) method introduced in this paper. The second set of features was obtained using standard (homoscedastic) linear discriminant analysis (lda) [5]. The third set of features was obtained using standard heteroscedastic linear discriminant analysis (hllda) [9]. The average performance of five randomly selected MFCC feature subsets is displayed.
as 5-rsfs. The fifth and final set is simply the set of the first $n$ MFCCs, denoted as initial.

Table 1: AURORA2 clean training results. The amfs selected coefficients with indices 1,2,3,4,9,10,11,12 for $n=8$ and 1,2,3,12 for $n=4$, respectively.

<table>
<thead>
<tr>
<th>feature set</th>
<th>Data Test Set A</th>
<th>clean 1</th>
<th>clean 2</th>
<th>clean 3</th>
<th>clean 4</th>
<th>avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>12+E</td>
<td>full,static</td>
<td>98.1</td>
<td>97.6</td>
<td>97.7</td>
<td>98.1</td>
<td>97.9</td>
</tr>
<tr>
<td>12+E+V+A</td>
<td>full,dynamic</td>
<td>99.9</td>
<td>99.1</td>
<td>98.8</td>
<td>99.2</td>
<td>99.0</td>
</tr>
<tr>
<td>8+E</td>
<td>amfs</td>
<td>97.8</td>
<td>97.6</td>
<td>97.1</td>
<td>97.8</td>
<td>97.6</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>98.3</td>
<td>97.8</td>
<td>98.4</td>
<td>98.3</td>
<td>98.2</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
<td>98.7</td>
<td>98.1</td>
<td>98.5</td>
<td>99.0</td>
<td>98.6</td>
</tr>
<tr>
<td></td>
<td>5-rsfs</td>
<td>95.8</td>
<td>95.7</td>
<td>95.5</td>
<td>96.0</td>
<td>95.8</td>
</tr>
<tr>
<td></td>
<td>initial</td>
<td>97.5</td>
<td>97.3</td>
<td>97.6</td>
<td>97.9</td>
<td>97.6</td>
</tr>
<tr>
<td>8+E+V+A</td>
<td>amfs</td>
<td>99.1</td>
<td>98.9</td>
<td>98.8</td>
<td>99.2</td>
<td>99.0</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>98.4</td>
<td>98.1</td>
<td>98.2</td>
<td>98.6</td>
<td>98.3</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
<td>98.6</td>
<td>98.5</td>
<td>98.5</td>
<td>99.1</td>
<td>98.7</td>
</tr>
<tr>
<td></td>
<td>initial</td>
<td>98.9</td>
<td>98.9</td>
<td>98.8</td>
<td>99.1</td>
<td>98.9</td>
</tr>
<tr>
<td>4+E</td>
<td>amfs</td>
<td>96.9</td>
<td>96.5</td>
<td>96.6</td>
<td>97.3</td>
<td>96.8</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>96.3</td>
<td>95.5</td>
<td>95.7</td>
<td>96.7</td>
<td>96.1</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
<td>95.9</td>
<td>95.0</td>
<td>96.0</td>
<td>96.5</td>
<td>95.9</td>
</tr>
<tr>
<td></td>
<td>5-rsfs</td>
<td>85.5</td>
<td>85.3</td>
<td>85.6</td>
<td>85.6</td>
<td>85.5</td>
</tr>
<tr>
<td></td>
<td>initial</td>
<td>97.0</td>
<td>96.5</td>
<td>96.7</td>
<td>97.3</td>
<td>96.9</td>
</tr>
<tr>
<td>4+E+V+A</td>
<td>amfs</td>
<td>98.6</td>
<td>98.9</td>
<td>98.7</td>
<td>99.0</td>
<td>98.8</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>98.3</td>
<td>97.9</td>
<td>98.2</td>
<td>98.2</td>
<td>98.2</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
<td>98.7</td>
<td>98.3</td>
<td>98.3</td>
<td>98.9</td>
<td>98.6</td>
</tr>
<tr>
<td></td>
<td>initial</td>
<td>98.6</td>
<td>98.8</td>
<td>98.6</td>
<td>99.2</td>
<td>98.8</td>
</tr>
</tbody>
</table>

Note that lda and hlda have two advantages over amfs: i) they are dimension-reduction methods, rather than subset-selection methods and ii) they require classified data as training input. The amfs method has as advantage that it can rely on knowledge inherent in the auditory periphery.

To build the recognizer we used the HTK [19] toolkit. 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. To minimize modeling artefacts, the results are for full covariance matrices, but the use of diagonal covariance matrices gives essentially the same results. An initial model with global data means and covariances, identical for each digit, was used and 16 iterations were used to build the final model.

Table 1 shows the recognition accuracy for training and testing on clean
Table 2: AURORA2 multi-conditioning noisy training results. The \textit{amfs} selected coefficients with indices 1,2,3,4,5,6,11,12 for \textit{n}=8 and 1,2,3,4 for \textit{n}=4, respectively.

<table>
<thead>
<tr>
<th>feature set</th>
<th>Data Test Set A</th>
<th>20 dB</th>
<th>15 dB</th>
<th>10 dB</th>
<th>5 dB</th>
<th>0 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>12+E</td>
<td>full,static</td>
<td>94.8</td>
<td>93.1</td>
<td>88.4</td>
<td>75.0</td>
<td>47.4</td>
</tr>
<tr>
<td>12+E+V+A</td>
<td>full,dynamic</td>
<td>97.4</td>
<td>96.5</td>
<td>94.0</td>
<td>86.0</td>
<td>59.6</td>
</tr>
<tr>
<td>8+E</td>
<td>\textit{amfs}</td>
<td>93.6</td>
<td>91.4</td>
<td>85.7</td>
<td>65.1</td>
<td>34.9</td>
</tr>
<tr>
<td></td>
<td>\textit{lda}</td>
<td>84.0</td>
<td>78.1</td>
<td>63.2</td>
<td>40.8</td>
<td>17.8</td>
</tr>
<tr>
<td></td>
<td>\textit{hlda}</td>
<td>85.0</td>
<td>80.9</td>
<td>71.7</td>
<td>55.7</td>
<td>33.2</td>
</tr>
<tr>
<td></td>
<td>5-rsfs</td>
<td>88.0</td>
<td>84.6</td>
<td>75.6</td>
<td>53.0</td>
<td>24.6</td>
</tr>
<tr>
<td></td>
<td>\textit{initial}</td>
<td>93.1</td>
<td>91.0</td>
<td>86.0</td>
<td>69.2</td>
<td>38.1</td>
</tr>
<tr>
<td>8+E+V+A</td>
<td>\textit{amfs}</td>
<td>97.1</td>
<td>96.0</td>
<td>93.4</td>
<td>84.1</td>
<td>56.8</td>
</tr>
<tr>
<td></td>
<td>\textit{lda}</td>
<td>94.4</td>
<td>92.8</td>
<td>87.8</td>
<td>76.5</td>
<td>52.0</td>
</tr>
<tr>
<td></td>
<td>\textit{hlda}</td>
<td>94.8</td>
<td>93.4</td>
<td>89.4</td>
<td>79.3</td>
<td>58.4</td>
</tr>
<tr>
<td></td>
<td>\textit{initial}</td>
<td>97.5</td>
<td>96.1</td>
<td>92.9</td>
<td>83.1</td>
<td>53.2</td>
</tr>
<tr>
<td>4+E</td>
<td>\textit{amfs}</td>
<td>91.1</td>
<td>87.4</td>
<td>76.9</td>
<td>50.9</td>
<td>20.9</td>
</tr>
<tr>
<td></td>
<td>\textit{lda}</td>
<td>40.5</td>
<td>32.5</td>
<td>19.6</td>
<td>9.9</td>
<td>7.7</td>
</tr>
<tr>
<td></td>
<td>\textit{hlda}</td>
<td>44.0</td>
<td>38.8</td>
<td>28.6</td>
<td>16.6</td>
<td>8.4</td>
</tr>
<tr>
<td></td>
<td>5-rsfs</td>
<td>66.3</td>
<td>59.3</td>
<td>43.6</td>
<td>24.0</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>\textit{initial}</td>
<td>91.1</td>
<td>87.4</td>
<td>76.9</td>
<td>50.9</td>
<td>20.9</td>
</tr>
<tr>
<td>4+E+V+A</td>
<td>\textit{amfs}</td>
<td>96.0</td>
<td>94.6</td>
<td>90.9</td>
<td>79.4</td>
<td>50.1</td>
</tr>
<tr>
<td></td>
<td>\textit{lda}</td>
<td>93.6</td>
<td>92.1</td>
<td>86.3</td>
<td>73.9</td>
<td>48.9</td>
</tr>
<tr>
<td></td>
<td>\textit{hlda}</td>
<td>94.2</td>
<td>92.8</td>
<td>88.9</td>
<td>78.8</td>
<td>58.5</td>
</tr>
<tr>
<td></td>
<td>\textit{initial}</td>
<td>96.0</td>
<td>94.6</td>
<td>90.9</td>
<td>79.4</td>
<td>50.1</td>
</tr>
</tbody>
</table>

data for dimensionality \textit{n} = 8 and for \textit{n} = 4. The caption of the table provides the MFCC subsets selected by \textit{amfs}. For the experiments we added the energy feature “+E”. We also performed the experiments with feature sets that were augmented with their velocity (“+V”) and acceleration (“+A”). For clarity we note that the subset-selection and dimension-reduction operations were always performed on the static features. Training was performed on the clean training set of 8440 sentences and the testing on the 4004 clean data of test set A.

For clean data the \textit{amfs} selected feature set performs similarly to the \textit{lda} and \textit{hlda} feature sets and to the \textit{initial} set of MFCCs. All these features sets perform significantly better than the average of randomly selected feature sets 5-rsfs. It is interesting to note that multiple distinct MFCC subsets perform well. Consistent with the recognition results shown in Table 1, the score of the measure of proximity to isometry, given by Eq. (1), is similar for the set \textit{initial} and for the \textit{amfs} selected features.

Table 2 shows the recognition accuracy for noisy data. Again the table
caption provides the MFCC subsets selected by amfs. The training was performed on the multi-conditioned noisy training set consisting of 6752 files and the testing on the 24024 noisy data of test set A. The results shown in Table 2 are averaged over subway, babble, car, and exhibition additive noise for several SNR values. The 5-rsfs configuration is the same as for the clean case (the same MFCCs subsets were considered).

For the noisy data, the performance of the amfs subsets are in all cases better than lda, hlda and 5-rsfs. The performance of initial is similar to that of amfs, although it does not consistently use the same subset. The results indicate that the new amfs method is more robust to environmental noise than lda and hlda. This increased robustness was confirmed in other experiments where we trained and tested on different environmental conditions. This result is not unexpected as amfs is based on an auditory periphery that is robust over a large range of environmental conditions. In contrast, lda and hlda must rely on the training data only.

As for the clean results shown in Table 1, the results of Table 2 indicate that low quefrency MFCCs represent well the important component of the perceived information.

5 Conclusions

We conclude that the selection of speech features based on human perception results in robust features that perform well for speech recognition over a range of environmental conditions. Our results suggest that the method results in features that are more robust to noise than either homoscedastic or heteroscedastic discriminant analysis. This implies that effective dimension reduction of feature sets for speech recognition is possible without knowledge of the meaning of the signal (without the availability of classified data). Our results indicate that the human auditory periphery has a parsimonious output representation, as significant redundancy would have made the measure of proximity to isometry, Eq. (1), ineffective for classification.

References


Auditory Model Based Optimization of MFCCs Improves Automatic Speech Recognition Performance

Saikat Chatterjee, Christos Koniaris and W. Bastiaan Kleijn

Refereed article published in Proceedings of Interspeech 2009
Auditory Model Based Optimization of MFCCs Improves Automatic Speech Recognition Performance

Saikat Chatterjee, Christos Koniaris and W. Bastiaan Kleijn

Abstract

Using a spectral auditory model along with perturbation based analysis, we develop a new framework to optimize a set of features such that it emulates the behavior of the human auditory system. The optimization is carried out in an off-line manner based on the conjecture that the local geometries of the feature domain and the perceptual auditory domain should be similar. Using this principle, we modify and optimize the static mel frequency cepstral coefficients (MFCCs) without considering any feedback from the speech recognition system. We show that improved recognition performance is obtained for any environmental condition, clean as well as noisy.

1 Introduction

An automatic speech recognition (ASR) system comprises two main tasks: feature extraction and pattern recognition. The feature extraction stage is designed to transform the incoming speech signal into a representation that serves as the input to a later pattern recognition stage. Feature extraction is a dimensionality reduction problem where the output representation should preserve the important aspects of the input speech signal relevant for speech recognition in any environmental condition, clean as well as noisy.

Different feature sets have been proposed in the literature, but the solutions remain ad hoc. We propose to define the features based on a perceptually relevant objective criterion. The human peripheral auditory system enhances the input speech signal for further processing by the central auditory system of the brain. Pre-processing of the input speech signal by the human auditory periphery forms a useful basis for designing an efficient
Auditory Model Based Optimization of MFCCs Improves Automatic Speech Recognition Performance

feature set. Commonly used features use knowledge of the auditory system in an ad hoc manner. For example, several feature extraction methods perform auditory frequency filtering on a perceptually motivated frequency scale than a linear scale. Another example is the use of a logarithmic function to approximate the non-linear dynamic compression in the auditory system, which allows us to cover the large dynamic range between hearing threshold and uncomfortable loudness level. Using these two auditory motivated signal processing techniques, MFCCs were designed a few decades ago [1]. They are still universally used due to their computational simplicity as well as good performance. Importantly, the MFCCs do not use up-to-date quantitative knowledge of the auditory system.

Several attempts have been made to use quantitative auditory models in a practical ASR system processing chain [2]- [7]. In these techniques, the input speech signal is first processed through a readily available auditory model and then the output signal of the auditory model is formatted to use as an input to the pattern recognition stage of the ASR system. The direct use of an auditory model was shown to provide better speech recognition performance, but at the expense of higher computational complexity. In recent years, the research in quantitative modeling of the complex peripheral auditory system has reached a high level of sophistication [8]- [13], and it is appealing to use a sophisticated auditory model for designing efficient features. The feature set should not incur the higher computational complexity associated with a full auditory model.

In this paper, instead of the direct on-line use, we investigate the use of an auditory model to design improved MFCCs through off-line optimization. The optimized MFCCs are referred to as modified MFCCs (MMFCCs). The off-line approach helps to retain the computational simplicity of MMFCCs. Also, it avoids the difficulty of formatting the output of the auditory model for recognition. Comparing to traditional MFCCs, the MMFCCs have a similar structure as well as computational simplicity.

In our approach, the feature set is optimized in such a way that it emulates the behavior of the human auditory system. The implementation of our method relies on perturbation theory and does not consider any feedback from the ASR system. We conjecture that human-like classification of speech sounds is facilitated by similarity between the local geometries of two domains, the feature domain and the perceptual domain. For improved classification, the preservation of the data geometry near the class boundaries is most critical. This means that ‘small’ Euclidean distances must be similar in the two different domains, except for an overall scaling. The focus on small distances allows a complex perceptual distance to be reduced to a quadratic distance measure using a sensitivity matrix based analysis. The sensitivity matrix based analysis was first developed in the context of source coding [14]. In [15], the sensitivity matrix was used to simplify an auditory distance measure for audio coding. Here, we extend
the sensitivity matrix paradigm to optimize a feature set. Using HTK, the optimized MMFCCs are shown to provide better recognition performance than traditional MFCCs for both clean and noisy acoustic conditions.

2 Maximizing Similarity between Spaces

Improvement in sound classification requires a feature representation that provides a good separation of sound classes in the feature space. Noting the high human recognition performance, it can be expected that the output of a sophisticated auditory model provides good separation of sound classes. Therefore, we optimize a feature set to better describe the intersound distances of a state-of-the-art auditory model. We conjecture that if the Euclidean distance between two acoustic features approximates the corresponding perceptual distortion for two different speech sounds, then the use of that acoustic feature generally leads to better classification in an ASR system. Ideally, this implies an isometry between the perceptual and feature domains. The mapping from the perceptual to feature domain would then be distance preserving.

2.1 Distance Preserving Measure

In practice, it is not be possible to design a feature set that leads to an accurate distance-preserving mapping from perceptual domain to feature domain. However, it is not required to preserve all the distances. For good classification, the preservation of the data geometry near the class boundaries is most critical. More generally, the preservation of small distances (reflecting the local geometry) near the classification boundary is important, whereas the preservation of large distances (reflecting the global geometry) is not required. In principle, to achieve better sound classification, we then simply desire to have the same small distances for the auditory domain and for the feature domain.

A feature set is a function of an input speech signal segment (or frame) and some adjustable design parameters. For example, to design MMFCCs, these design parameters can be the frequency warping parameter to change the shape of a filter bank (such as heights, widths, center frequency of filters), a parameter to change the shape of a compressing function (like logarithmic function), etc. The objective is to obtain a feature set with optimum parameters for which any small perturbation of the input speech signal segment leads to a Euclidean distance in the feature domain that best approximates the perceptual distortion indicated by the auditory model. Naturally this criterion has to hold for all speech segments. To measure the similarity of the auditory model distortion and the feature domain distance, a suitable objective measure needs to be designed that will provide a
means of ensemble averaging over all speech segments and all perturbations. By optimizing the parameters, a higher similarity, through evaluating the objective measure, leads to a better feature set.

We now define an objective measure that relates between the perceptual and feature domains. Let us denote the signal vector for the $j$’th speech frame as $x_j \in \mathbb{R}^N$, where $j \in J \subset \mathbb{Z}$, and the perceptual domain representation of $x_j$ as $y : \mathbb{R}^N \rightarrow \mathbb{R}^K$. We also denote the design parameters of a feature set by a vector $p \in \mathbb{R}^S$. Then, we can denote the $Q$-dimensional feature derived from $x_j$ using $p$ as $c : \mathbb{R}^N \times \mathbb{R}^S \rightarrow \mathbb{R}^Q$. The perceptual domain distortion is defined through a mapping as $\Upsilon : \mathbb{R}^K \times \mathbb{R}^K \rightarrow \mathbb{R}^+$, where $\mathbb{R}^+$ is the set of non-negative reals. For the $j$’th speech frame, let us denote the $l$’th perturbed signal as $\hat{x}_{j,l}$. Often the perceptual distortion measure is based on the $L^2$ norm of the difference between the perceptual domain signal $y(x_j)$ and its distorted version $y(\hat{x}_{j,l})$. In that case, $\Upsilon(x_j, \hat{x}_{j,l}) = \|y(x_j) - y(\hat{x}_{j,l})\|^2$. Using the $L^2$ norm, we can define a distance measure for the feature $c(x_j, p)$ as $\Gamma(x_j, \hat{x}_{j,l}, p) = \|c(x_j, p) - c(\hat{x}_{j,l}, p)\|^2$. Now, considering the finite sequence of speech frames $j \in J$ and a finite set of acoustic perturbations $l \in L_j$, the objective is to minimize a measure of dissimilarity between perceptual domain distortion and feature domain distortion with respect to the parameter set $p$. To satisfy this objective, a suitable norm based measure can be defined as

$$O = \sum_{j \in J} \sum_{l \in L_j} [\Upsilon(x_j, \hat{x}_{j,l}) - \lambda \Gamma(x_j, \hat{x}_{j,l}, p)]^2,$$  

(1)

where

$$\lambda = \frac{\sum_{j \in J} \sum_{l \in L_j} \Upsilon(x_j, \hat{x}_{j,l}) \Gamma(x_j, \hat{x}_{j,l}, p)}{\sum_{j \in J} \sum_{l \in L_j} \Gamma(x_j, \hat{x}_{j,l}, p)^2}.$$  

(2)

Here $\lambda$ is the necessary scaling to eliminate the effect of a scale mismatch between perceptual domain and feature domain. So, the objective is to minimize the norm based distance $O$ with respect to the parameter vector $p$.

### 2.2 Perturbation Analysis

While it is possible to minimize the objective measure of eq. (1) even for complex distortion measures, this can be computationally expensive. Since we are interested in small distances, we can approximate the perceptual and feature domain distortion measure using simpler quadratic measures, leading to a significant reduction in computational complexity and an increase in mathematical tractability. This approach is based on the sensitivity matrix framework [14], [15].
Let us omit the subscripts for notational brevity where no ambiguity exists. We assume that $Υ(x, \hat{x})$ is analytic and $Υ(x, x) = 0$. Then, for a sufficiently small perturbation $\hat{x} - x$, we can write

$$Υ(x, \hat{x}) \approx \frac{1}{2} [\hat{x} - x]^T D_Υ(x) [\hat{x} - x],$$

(3)

where $D_Υ(x)$ is the sensitivity matrix whose elements are $D_Υ_{ij}(x) = \frac{\partial^2 Υ(x, \hat{x})}{\partial x_i \partial x_j} \bigg| _{\hat{x} = x}$. In certain cases, such as the spectral auditory model of section 2.3, $Υ(x, \hat{x})$ and $D_Υ(x)$ are known.

Next, we consider a simplification of the distortion in the feature domain i.e., $Γ(x, \hat{x}, p)$. If the mapping $c(x, p)$ is analytic in $x$, then we can use the Taylor series expansion to make a local approximation around $x$ as

$$c(\hat{x}, p) = c(x, p) + A(p) [\hat{x} - x],$$

(4)

where $A(p)$ is a $Q \times N$-dimensional matrix as $A(p) = \frac{∂c(x, p)}{∂x} \bigg| _{\hat{x} = x}$. We can then write the distortion in the feature domain as

$$Γ(x, \hat{x}, p) = \|c(x, p) - c(\hat{x}, p)\|^2 = [\hat{x} - x]^T A(p)^T A(p) [\hat{x} - x].$$

(5)

### 2.3 A Spectral Auditory Model

In this paper, we optimize the MMFCCs to minimize the norm based measure of eq. (1). The MMFCCs are designed using the power spectrum of the input speech signal. Therefore, for optimization, we use the spectral auditory model developed by van de Par, et al. [13] which is referred to as the van de Par auditory model (VAM). The VAM is a psycho-acoustic masking model that accounts for simultaneous processing of sound signals with different frequencies. To use the VAM, we consider the input signal $x$ as the power spectrum of a speech frame. The VAM consists of several frequency channels, in each of which the ratio of distortion power to masker power is calculated. Then, the ratios of all the frequency channels are combined together to account for the spectral integration property of the human auditory system. Let $H$ be a diagonal $N$-dimensional matrix whose diagonal is formed by the frequency response of the outer and middle ear filter. In the same fashion, a diagonal $G_i$ is defined, so that the frequency response of the $i$’th channel Gamma-tone auditory filter forms its diagonal. For the VAM, the diagonal sensitivity matrix is

$$D_Υ(x) \approx \frac{2 C_s L_e}{N} \sum_i \frac{[G_i, H]^T[G_i, H]}{[G_i, H]^T[G_i, H]} + C_a,$$

(6)

where $C_s$ and $C_a$ are constants calibrated based on measurement data, and $L_e$ is a constant to account for the influence of temporal integration time in the human auditory system on frame duration.
It is important to mention that each speech frame is independently analyzed in the VAM. Therefore, the use of VAM is appropriate for optimizing a static feature. Note that due to the inability to model the auditory response across speech frames, the use of the VAM is inappropriate for optimizing the temporal dynamic features, such as velocity and acceleration. However, it is possible to compute the dynamic features from any static feature using standard regression method.

3 Modified MFCCs

We first generalize the definition of the MFCCs to render a set of features with adjustable parameters $p$. We refer to this new set of features as modified MFCCs (MMFCCs). Let the $N$-dimensional vector $x = [x_0, x_1, \cdots, x_n, \cdots, x_{N-1}]^T$ be the power spectrum of a Hamming windowed speech frame. Then the steps of evaluating the MMFCCs are as follows:

1. Calculation of the energy in each channel:
\[
    z_m = x^T w_m(\alpha) = \sum_{n=0}^{N-1} x_n \times w_{m,n}(\alpha), 0 \leq m \leq M - 1, \tag{7}
\]

where $w_m(\alpha)$ is the $N$-dimensional vector denoting the triangular filter of the $m$'th channel and satisfies $\sum_{n=0}^{N-1} w_{m,n}(\alpha) = 1$. $M$ is the total number of channels with a typical value of $M = 26$. The shape of a triangular filter depends on the extent of frequency warping. The warped frequency scale \[16\] is given as
\[
    f_{\text{warp}} = 2595 \times \log_{10} (1 + (f/\alpha)), \tag{8}
\]

where $\alpha$ is the warping factor and $f$ is the frequency in Hz. An increase in $\alpha$ leads to a decrease in the extent of warping. For the MMFCCs, $\alpha$ is a parameter to optimize to achieve better recognition performance. In the case of MFCCs, the triangular filters are designed using the mel frequency scale where $\alpha = 700 \ [16]$.

2. Compression of the dynamic range of the energy in each channel:
\[
    s_m = \log_{10} \left[ \sum_{r=1}^{R} b_r (z_m)^r \right], 0 \leq m \leq M - 1, \tag{9}
\]

where $\sum_{r=1}^{R} b_r = 1$ and $b_r \geq 0$. For the MMFCCs, we optimize the polynomial coefficients $\{b_r\}_{r=1}^R$. In the case of MFCCs, $R = 1$ and $b_1 = 1 \ [1]$. We note that eq. (9) implies that our results are scale dependent and require proper normalization.
3. De-correlation using the DCT to evaluate $Q$-dimensional MMFCC feature vector:

$$c_q = \sum_{m=0}^{M-1} s_m \times \cos \left[ q (m + 0.5) \frac{\pi}{M} \right], 1 \leq q \leq Q. \quad (10)$$

A typical value of feature vector dimension is $Q = 12$.

### 3.1 Optimization of the MMFCCs

The parameters that we optimize to obtain the MMFCCs are $p = [\alpha, \{b_r\}_{r=1}^R]$. To optimize the parameters, we need to minimize the objective measure $O$ of eq. (1). This objective measure is a function of the sensitivity matrix based perceptual domain distortion of eq. (3) and the feature domain distortion of eq. (5). To evaluate the perceptual domain distortion, we need a closed form sensitivity matrix $D_Y(x)$ which is given by the VAM as shown in eq. (6). We also need a closed form $A(p)$ for evaluating the feature domain distortion. For an MMFCC feature, the elements of the matrix $A(p)$ are

$$A_{qn} = \frac{\partial c_q}{\partial x_n} = \frac{\partial c_q}{\partial s_m} \frac{\partial s_m}{\partial z_m} \frac{\partial z_m}{\partial z_n} \frac{\partial z_n}{\partial s_n} = \sum_{m=0}^{M-1} \cos \left[ q (m + 0.5) \frac{\pi}{M} \right] \times \frac{\sum_{r=1}^R r b_r (z_m)^{r-1}}{\ln 10} \times \frac{\sum_{r=1}^R b_r (z_m)^{r-1}}{w_m_n(\alpha)}. \quad (11)$$

It is interesting to jointly optimize all the parameters through a closed-form/iterative solution, such as using gradient descent search technique. This requires a closed form gradient expression $\frac{dO}{dp}$, which is not easy to evaluate due to the intricate relationship existing between the measure of $O$ and the parameter vector $p = [\alpha, \{b_r\}_{r=1}^R]$. Therefore, we use a simple increment-based linear search technique and optimize the parameters one by one. We first optimize $\{b_r\}_{r=1}^R$ and then $\alpha$. For both the cases of wideband (sampling frequency 16 kHz) and narrow-band (sampling frequency 8 kHz) speech, we use a 32 ms Hamming windowed speech frame with 10 ms frame shift. To evaluate the MMFCCs, we use $M = 26$ and $Q = 12$. The power spectrum of each frame is computed using a standard DFT based periodogram technique and the power spectrum is perturbed with i.i.d Gaussian noise at different SNRs ranging from 120 to 130 dB. Using an increment-based linear search, we evaluate the minimum value of the measure of eq. (1) and find that a polynomial order of $R = 2$ is sufficient; the values of the polynomial coefficients are $b_1 = 0.1$ and $b_2 = 0.9$. Next we search for the optimum $\alpha$. For wide-band speech and narrow-band speech,
Auditory Model Based Optimization of MFCCs Improves Automatic Speech Recognition Performance

Table 1: Phone recognition accuracy (in %) of static 12-dimensional MFCC and MMFCC features using TIMIT

<table>
<thead>
<tr>
<th>Feature</th>
<th>Number of Gaussian mixtures/state</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>MFCC</td>
<td>43.19</td>
</tr>
<tr>
<td>MMFCC</td>
<td>45.13</td>
</tr>
</tbody>
</table>

we find the optimum values are $\alpha = 900$ and $\alpha = 1100$, respectively. We note that standard MFCCs use $b_1 = 1$ and $\alpha = 700$ irrespective of the sampling frequency of input speech, choice of the window length and shift, and the feature dimension ($Q$) and number of channels ($M$) [1], [16].

4 Recognition Results

Table 2: Robust word recognition accuracy (in %) of 39-dimensional MFCC and MMFCC features using Aurora 2

<table>
<thead>
<tr>
<th>Feature</th>
<th>Test Set a</th>
<th>Test Set b</th>
<th>Test Set c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>set 1</td>
<td>set 2</td>
<td>set 3</td>
</tr>
<tr>
<td>MFCC</td>
<td>95.46</td>
<td>96.67</td>
<td>96.12</td>
</tr>
<tr>
<td>MMFCC</td>
<td>96.49</td>
<td>97.49</td>
<td>97.08</td>
</tr>
<tr>
<td></td>
<td>SNR = 20 dB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC</td>
<td>85.05</td>
<td>86.49</td>
<td>83.39</td>
</tr>
<tr>
<td>MMFCC</td>
<td>87.07</td>
<td>88.51</td>
<td>87.00</td>
</tr>
<tr>
<td></td>
<td>SNR = 10 dB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Using the HTK toolkit, we performed phone and word recognition experiments to compare between MFCC and MMFCC features. The static 12-dimensional MFCC feature set was extracted using the same setup as that used to extract the 12-dimensional MMFCC feature set. Using the standard approach, 39-dimensional feature vectors were evaluated. To the static features, we appended the log energy of a speech frame and the velocity and acceleration of the features.

We first compared the performance of 12-dimensional static features through a clean speech phoneme recognition experiment. In this case, we used the TIMIT database where the speech is sampled at 16 kHz. HTK training and testing were performed using the training set and the test set of TIMIT respectively. The TIMIT transcriptions are based on 61 phones. Following convention, the 61 phones were folded onto 39 phones as described...
Table 3: Robust phone recognition accuracy (in %) of 39-dimensional MFCC and MMFCC features at 10 dB SNR

<table>
<thead>
<tr>
<th>Feature</th>
<th>Clean</th>
<th>White</th>
<th>Pink</th>
<th>Babble</th>
<th>Volvo</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>68.11</td>
<td>37.03</td>
<td>40.51</td>
<td>46.25</td>
<td>59.71</td>
</tr>
<tr>
<td>MMFCC</td>
<td>68.34</td>
<td>43.65</td>
<td>46.67</td>
<td>48.94</td>
<td>61.91</td>
</tr>
</tbody>
</table>

in [17]. To train the HMMs, we used three states per phoneme and the performance is shown in Table 1 for a varying number of Gaussian mixtures per state. We used Gaussian mixtures with diagonal covariance matrices. From Table 1, it can be noted that MMFCCs outperform MFCCs for any number of mixtures. In case of the 39-dimensional feature vectors, the performance improvement of using MMFCCs over MFCCs was always positive, but small for clean speech phone recognition.

Next we considered the 39-dimensional feature vectors for robust word and phone recognition experiments where clean speech training and noisy speech testing were performed. For the robust recognition, we used cepstrum mean and variance normalization (CMVN) on the feature sets [18]. For the robust word recognition experiment, we used the Aurora 2 database where the speech is sampled at 8 kHz and the sub-datasets of test set are corrupted by different noise types at varying SNRs. The standard configuration of the HTK setup was used where HMMs were trained using 16 states per word and three Gaussian mixtures per state (diagonal covariance matrices). The robust word recognition performance for 39-dimensional MFCCs and MMFCCs are shown in Table 2 at the testing conditions of 20 dB and 10 dB SNRs. We note that MMFCCs perform better than MFCCs for all the sub-datasets corrupted with different noises. In the case of clean speech word recognition, the improvement of 39-dimensional MMFCCs over MFCCs was small like in the case of clean speech phone recognition.

Finally, we consider a robust phone recognition experiment where the clean test speech database of TIMIT was corrupted with additive noise. We used the following noise types from the NoiseX-92 database: white, pink, babble and car (volvo) noise. The test speech database was corrupted by adding each noise at 10 dB SNR. The HMMs consisted of three states per phoneme and 20 Gaussian mixtures per state. The robust phone recognition performance for MFCCs and MMFCCs are shown in Table 3 and we note that MMFCCs perform better than MFCCs for all noise types.
5 Conclusions

Our development of MMFCCs shows that the use of a sophisticated auditory model can lead to a simple feature set that provides improved speech recognition performance for any environmental condition. The success of our perceptual-distance preserving measure in optimizing features suggests that the auditory system provides as output a signal representation that is ‘efficient’ for speech recognition. As we developed the static MMFCCs using a static spectral auditory model, further investigation should consider the optimization of dynamic features using a spectro-temporal auditory model, such as that presented in [11].

6 Acknowledgments

This work is supported by EU FP6 ACORNS project. The authors wish to thank V. Surayanarayana. K of Indian Institute of Science, Bangalore, India.

References


Paper C

Selecting static and dynamic features using an advanced auditory model for speech recognition

Christos Koniaris, Saikat Chatterjee and W. Bastiaan Kleijn

submitted
The layout has been revised for thesis consistency
Selecting static and dynamic features using an advanced auditory model for speech recognition

Christos Koniaris, Saikat Chatterjee and W. Bastiaan Kleijn

Abstract

We describe a method to select features for speech recognition that is based on a quantitative model of the human auditory periphery. The method maximizes the similarity of the geometry of the space spanned by the subset of features and the geometry of the space spanned by the auditory model output. The selection method uses a spectro-temporal auditory model that captures both frequency- and time-domain masking. The selection method is blind to the meaning of speech and does not require annotated speech data. We apply the method to the selection of a subset of features from a conventional set consisting of mel cepstra and their first-order and second-order time derivatives. Although our method uses only knowledge of the human auditory periphery, the experimental results show that it performs significantly better than selection algorithms based on linear and heteroscedastic discriminant analysis that require training with annotated speech data.

1 Introduction

In speech recognition, the goal of acoustic-feature selection is to reduce input dimensionality while retaining most of the information relevant for accurate classification. Although a large feature cardinality results in principle in a better recognition accuracy, in practice it is not possible to train or use such an automatic speech recognition (ASR) system. This phenomenon is called the curse of dimensionality [1].

The selection of appropriate features is not a trivial task. Generally a large pool of possible features is used as a starting point. For a good selection algorithm, the performance of the ASR system increases when the cardinality of the feature set is reduced from that initial pool, e.g.,
Selecting static and dynamic features using an advanced auditory model for speech recognition

This increase in performance results from improved generalization as the number of system parameters is reduced. Many approaches to feature selection have been developed, e.g., [7, 10, 14]. Generally these methods require annotated speech data and feedback on classification performance from an ASR system. This means they are optimized for a particular ASR system. The methods require ASR systems to be trained for the full set first. They do not use knowledge implicit in the human auditory system, which has become available with the development of sophisticated models of the auditory periphery, e.g., [2, 15].

In our previous work [9], we presented a new acoustic-feature-selection method that was based on human perception only, called auditory model-based feature selection (AMFS) for speech recognition. Inspired by the observation that the best recognition performance can be achieved by a normal-hearing individual, we use knowledge of the human auditory periphery to select robust feature subsets. The selection of acoustic features is performed by finding the subset that maximizes the similarity of the Euclidian geometry of the feature set and the human auditory representation of the signal. In [9], we considered only static features - for our speech recognition application we applied it to mel-frequency cepstrum coefficients (MFCCs) [4]. Accordingly, a static psycho-acoustic masking model [15] was considered. The results showed a significant performance increase compared to the linear discriminant analysis (LDA) [7] method and compared to an average of randomly selected feature subsets in various noisy conditions. The performance increase was sustained when first and second-order derivatives of the selected subsets were included as input for the ASR system.

In this paper, we further develop the method to enable the selection of a subset of features from a set of both static and dynamic features. To make this possible, we consider a more sophisticated auditory model [2] that we refer to as the Dau model. This model accounts for the spectro-temporal processing of sound signals by the human auditory periphery. Let the complete feature set characterize certain (locally) audible components of the speech signal. Then, for a given subset cardinality, the auditory model is used to select the feature subset that best captures the most audible of these signal components.

To validate our approach, we used the MFCCs and their first and second time derivatives. Our results verify that the human auditory model forms an effective basis for selecting robust acoustic features from this set of static and dynamic features.

This paper is organized as follows. Sec. 2 discusses a similarity measure for the perceptual and feature domains. Sec. 3 presents the auditory model used and describes the selection algorithm. Sec. 4 shows experimental results and Sec. 5, provides conclusions.
2 Maximizing the perceptual relevance of features

The high recognition accuracy of the normal-hearing listeners indicates that the auditory periphery system is efficient in sound-class discrimination. Implicitly, this suggests that the necessary information for sound classification reaches the ‘auditory cortex’ of the brain where the cognitive processing is performed. As ASR systems require a perceptually relevant set of acoustic features to achieve a high recognition rate, it is natural to find a set that is similar to that used by humans.

Our research is aimed at finding robust feature subsets based on quantitative models of the auditory periphery. The best-case scenario would be that the selected features include all the information relayed by the human auditory periphery. This would lead to the ideal situation in which the perceptual and the selected acoustic-feature domains are isometric. In practice this is not completely possible. In this section we define a measure that has proven to be effective in maximizing similarity between the perceptual domain and the acoustic-feature domain.

2.1 A measure of dissimilarity

It is our objective to obtain a feature set for which the geometry of the data (characterized by the distances between sounds) is similar to that of the perceptual domain (the output of the auditory periphery). This is particularly the case for “short” distances, as they determine the performance of classifiers. Let us define an \( L^2 \) norm based distance measure (also known as distortion measure) in the perceptual domain as a mapping of two signals: \( \Upsilon : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+ \), where \( \mathbb{R}^+ \) are the non-negative reals. Let us also consider the perceptual-domain signals \( y(x_j) \) and \( y(\hat{x}_{j,m}) \), where \( y : \mathbb{R}^N \rightarrow \mathbb{R}^K \) is a mapping to the \((K\text{-dimensional})\) perceptual domain. Then

\[
\Upsilon(x_j, \hat{x}_{j,m}) = \|y(x_j) - y(\hat{x}_{j,m})\|_2, \tag{1}
\]

where \( x_j \in \mathbb{R}^N \) denotes the \( N\text{-dimensional} \) speech signal vector characterizing a segment with time index \( j \in \mathbb{Z} \) and \( \hat{x}_{j,m} \) is a perturbation of \( x_j \) with perturbation index \( m \).

Similarly, a distance measure for the feature domain can be defined as \( \Gamma_i : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+ \) for the feature set \( i \). Let \( c_i : \mathbb{R}^N \rightarrow \mathbb{R}^L \) be the mapping from a signal segment \( x_j \) to a set of \( L \) features \( c_i(x_j) \) with set index \( i \). An \( L^2 \) norm based measure is then

\[
\Gamma_i(x_j, \hat{x}_{j,m}) = \|c_i(x_j) - c_i(\hat{x}_{j,m})\|_2. \tag{2}
\]

Our geometry objective is equivalent to find the particular set of fea-
Selecting static and dynamic features using an advanced auditory model for speech recognition

We find the set \( i \) that minimizes a measure of dissimilarity of the perceptual-domain distances and the feature-domain distances. We find the set \( i \) that minimizes

\[
G(i) = \sum_{j \in J, m \in M_j} \left[ \Upsilon(x_j, \hat{x}_{j,m}) - \lambda \Gamma_i(x_j, \hat{x}_{j,m}) \right]^2,
\]

where \( \lambda = \frac{\sum_{j \in J, m \in M_j} \Upsilon(x_j, \hat{x}_{j,m}) \Gamma_i(x_j, \hat{x}_{j,m})}{\sum_{j \in J, m \in M_j} \Gamma_i(x_j, \hat{x}_{j,m})^2} \) is an optimal scaling of the acoustic feature criterion, and \( j \in J, m \in M_j \) represent a finite frame sequence and a finite set of acoustic perturbations, respectively.

2.2 Perturbation analysis

To reduce computational effort, we use the powerful tools of perturbation analysis and the sensitivity matrix [6, 11–13]. We compute eq. (3) by approximating the perceptual distortion measure \( \Upsilon(x_j, \hat{x}_{j,m}) \) by a simpler quadratic measure. Note that perturbation analysis is natural as we consider only small distances in eq. (3).

Let consider \( \Upsilon(x_j, \hat{x}_{j,m}) \) to be known. We assume that \( \Upsilon(x_j, x_j) = 0 \) and that this forms a minimum. We furthermore assume that \( \Upsilon(x_j, \hat{x}_{j,m}) \) is analytic in \( \hat{x}_{j,m} \). Then, for sufficiently small perturbations \( \hat{x}_{j,m} - x_j \), we can make the approximation

\[
\Upsilon(x_j, \hat{x}_{j,m}) \approx [\hat{x}_{j,m} - x_j]^T D_{\Upsilon}(x_j) [\hat{x}_{j,m} - x_j],
\]

where \( D_{\Upsilon,\kappa,\mu}(x_j) = \left. \frac{\partial^2 \Upsilon(x_j, \hat{x}_{j,m})}{\partial \kappa \partial \mu} \right|_{\hat{x}_{j,m} = x_j} \) is the sensitivity matrix.

3 Application to speech recognition

This section illustrates the application of the method to a specific auditory model and specific type of acoustic features and presents the implementation of the feature selection algorithm.

3.1 Dau auditory model

The Dau [2, 3] auditory model is a psycho-acoustic masking model that accounts for spectro-temporal processing of sound signals. Thus, the speech signal \( x \) is a time-domain vector. It consists of several stages which simulate the human auditory periphery. A channel \( l \) of the Dau model 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 [13]. Finally, a low-pass filter performs a temporal smoothing and the output is the so-called internal representation.
The original paper [2] did not study the distortion prediction properties of the model, an investigation that was later performed in [13]. In the same paper, a distortion measure on the internal representation was introduced as

$$\Upsilon(x_j, \hat{x}_{j,m}) = \sum_l \| a^{(l)}(x'_j) - a^{(l)}(\hat{x}'_{j,m}) \|_2^2,$$

where $x'_j, \hat{x}'_{j,m}$ are of higher dimension than the $x_j, \hat{x}_{j,m}$ vectors, respectively due to the ring-out effect described in [13]. The derivation of the sensitivity matrix in this case is non-trivial and tedious. The final sensitivity matrix can be computed as the sum of per-channel sensitivity matrices

$$D^{(l)} \Upsilon(x_j) = \sum_l D^{(l)} \Upsilon(x_j),$$

where

$$D^{(l)} \Upsilon(x_j) = 2 \left[ \prod_k J_k^{(l)} \right]^H J_k^{(l)},$$

and $J_k^{(l)}$ is the Jacobian for processing stage $k$ in channel $l$.

### 3.2 AMFS algorithm for a spectro-temporal model

We can now discuss the implementation of a AMFS algorithm that accounts for both spectral and temporal masking. To handle the dynamic aspect of the auditory measure we use a superframe $J$ that consists of three overlapped subframes $j-1, j, j+1$ as shown in the Fig. 1. The length of the superframe was set at 40 ms with an overlap of 10 ms. The length of each subframe was 20 ms with an overlap of 10 ms for the AURORA 2 database of 8 kHz. The superframe is used to compute the Dau sensitivity matrix and the subframes are used to compute the static MFCCs and their first- and second-order derivatives.

The feature selection operation was divided into two stages. In the first stage all necessary quantities were computed, and in the second stage the feature set was selected.

In the first stage the Dau sensitivity matrix $D \Upsilon(x_J)$ was computed for the $J$'th superframe, $x_J$. A set of 100 vectors $\hat{x}_{J,m}$ was computed by adding 100 dB SNR noise to $x_J$. Next, the 12 mel cepstrum static coefficients were calculated for the central $j$'th subframe. Their first-order ($\Delta$) and second-order ($\Delta\Delta$) derivatives were computed using the $j-1$'th and the $j+1$'th subframes. The full feature vector $c_J$ for the superframe $J$ was built as

$$c_J = [c_j \Delta c_j \Delta\Delta c_j]^T.$$
Finally, we computed the 100 distorted features vectors $\hat{c}_{J,m}$'s corresponding to the 100 distorted speech vectors that were previously calculated. The first step of the algorithm ended by performing a cepstral mean and variance normalization (CMVN) in the feature vectors.

The second stage began by reading again the speech data. For each superframe, we considered the quantities that were computed (and saved) in the previous stage. We used them to compute the distortion measure in the perceptual domain according to eq. (4). Similarly, we computed the distortion in the feature domain for all the possible feature subsets $i$ according to eq. (2). Finally, to find the particular set of features $i$ we computed the measure of dissimilarity as described in eq. (3).

### 3.3 Greedy feature selection

Due to the high cardinality of the full feature vector, the search through all combinations of possibly optimal subsets is practically impossible. For example when a subset of 32-out-of-36 features is requested, $58,905$ different combinations have to be considered. To cope with this problem we used a greedy algorithm. The second stage of the algorithm was run iteratively, reducing the dimensionality by one dimension at each iteration. A feature that was not included in the selected subset was not considered in the next round. We did not consider forward-backward algorithms.

### 4 Experiments

To test and compare the performance of the new feature-selection system, we performed speech recognition experiments using the test set A of the AU-
RORA2 [8] database, sampled at 8 kHz. We used an MFCC representation. The full set of MFCCs was extracted by using a Hamming window of 20 ms with an overlap of 10 ms. The DFT dimensionality was 256 and we considered 23 mel filters. A set of 12 conventional MFCCs was extracted plus the energy coefficient and their corresponding first and second order time derivatives (resulting in a 39 feature vector). The selection of a subset from the full feature vector was performed with three different feature-selection methods. The first method is the new auditory-model based feature selection (indicated by \textit{amfs} in the tables) described in section 3. The other two methods are the well-known linear-discriminant analysis (\textit{lda}) [7] and heteroscedastic linear discriminant analysis (\textit{hlda}) [10] methods.

For the discriminant analysis methods we used the HTK [5] toolkit to build label files for each word-state (‘word’ refers to the digit to be recognized while ‘state’ refers to the state of the hidden Markov model (HMM)). Hence, we used information from the classified data for both \textit{lda} and \textit{hlda} methods.

To build the recognizer we also used the HTK [5] toolkit. 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. An initial model with global data means and covariances, identical for each digit, was used and 16 iterations were used to build the final model.

4.1 Results

Table 1 shows the recognition accuracy for training on clean data for various feature subsets starting from 33-out-of-36 and ending at 12-out-of-36 extended with the energy coefficient indicated as “+E” and its dynamics “+ΔE” and “+ΔΔE” for the velocity and acceleration coefficients, respectively. As reference the corresponding 39-dimensional feature set is provided. The training was performed on the clean training set of 8 440 sentences and the testing on the 20 020 data of test set A, consisting of 4 004 data from each of the ‘clean’, ‘20’, ‘15’, ‘10’ and ‘5’ dB SNR groups of data. The performance of the \textit{amfs} remains relatively stable as the number of coefficients is reduced. The \textit{amfs} features generally perform better than both the \textit{lda} and \textit{hlda} feature, particularly under noisy testing conditions.

Table 2 shows the recognition accuracy for noisy only data. The training was performed on the multi-conditioned noisy training set consisting of 6 752 files and the testing on 20 020 noisy data of test set A. Note that the results for the noisy test conditions shown in tables 1 and 2 are averaged over subway, babble, car, and exhibition additive noise for several SNR values. The performance of the \textit{amfs} subsets are in general better than the \textit{lda} and \textit{hlda}. 
Table 1: AURORA2 clean training results.

<table>
<thead>
<tr>
<th>feature set</th>
<th>Data Test Set A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>clean</td>
</tr>
<tr>
<td>36+E+ΔE+ΔΔE (full 39-dim)</td>
<td>99.1</td>
</tr>
<tr>
<td>33+E+ΔE+ΔΔE</td>
<td>amfs</td>
</tr>
<tr>
<td></td>
<td>lda</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
</tr>
<tr>
<td>30+E+ΔE+ΔΔE</td>
<td>amfs</td>
</tr>
<tr>
<td></td>
<td>lda</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
</tr>
<tr>
<td>27+E+ΔE+ΔΔE</td>
<td>amfs</td>
</tr>
<tr>
<td></td>
<td>lda</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
</tr>
<tr>
<td>24+E+ΔE+ΔΔE</td>
<td>amfs</td>
</tr>
<tr>
<td></td>
<td>lda</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
</tr>
<tr>
<td>21+E+ΔE+ΔΔE</td>
<td>amfs</td>
</tr>
<tr>
<td></td>
<td>lda</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
</tr>
<tr>
<td>18+E+ΔE+ΔΔE</td>
<td>amfs</td>
</tr>
<tr>
<td></td>
<td>lda</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
</tr>
<tr>
<td>15+E+ΔE+ΔΔE</td>
<td>amfs</td>
</tr>
<tr>
<td></td>
<td>lda</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
</tr>
</tbody>
</table>

4.2 Discussion

An interesting aspect of the results is the observation of how many features are necessary to achieve at least 99% of the original recognition performance of the 39 dimensional feature vector. Studying table 1, we see that for the clean-clean case amfs needs 18 features, for clean-20 dB case 24 features, for clean-15 dB 27, for clean-10 dB 36 and for the case of clean-5 dB it cannot reach this goal. For the noisy training experiments, table 2 shows that the goal of 99% can be reached with 24 features for the cases noisy-20 dB and noisy-15 dB and 30 for noisy-10 dB. In the case of noisy-5 dB amfs needs 33 features while for the case of noisy-0 dB, it has a lower accuracy. On the other hand, both lda and hlda cannot reach the 99% goal in all cases but for the clean-clean case, where 15 dimensional feature vector suffices
Table 2: AURORA2 multi-conditioning noisy training results.

<table>
<thead>
<tr>
<th>feature set</th>
<th>Data Test Set A</th>
<th>30 dB</th>
<th>15 dB</th>
<th>10 dB</th>
<th>5 dB</th>
<th>0 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>97.7</td>
<td>96.9</td>
<td>94.8</td>
<td>87.6</td>
<td>71.4</td>
</tr>
<tr>
<td>36+E+ΔE+ΔΔE (full 39-dim)</td>
<td>amfs</td>
<td>97.6</td>
<td>96.7</td>
<td>94.5</td>
<td>86.9</td>
<td>68.9</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>94.7</td>
<td>93.1</td>
<td>88.5</td>
<td>77.7</td>
<td>53.0</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
<td>95.4</td>
<td>94.0</td>
<td>90.3</td>
<td>80.4</td>
<td>59.5</td>
</tr>
<tr>
<td>33+E+ΔE+ΔΔE</td>
<td>amfs</td>
<td>97.6</td>
<td>96.6</td>
<td>94.4</td>
<td>86.9</td>
<td>68.3</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>94.6</td>
<td>93.1</td>
<td>88.4</td>
<td>77.3</td>
<td>52.8</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
<td>95.3</td>
<td>93.9</td>
<td>90.0</td>
<td>80.1</td>
<td>59.3</td>
</tr>
<tr>
<td>30+E+ΔE+ΔΔE</td>
<td>amfs</td>
<td>97.4</td>
<td>96.4</td>
<td>94.0</td>
<td>85.8</td>
<td>66.9</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>94.3</td>
<td>92.7</td>
<td>87.7</td>
<td>76.9</td>
<td>52.2</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
<td>95.0</td>
<td>93.5</td>
<td>89.9</td>
<td>79.6</td>
<td>59.0</td>
</tr>
<tr>
<td>27+E+ΔE+ΔΔE</td>
<td>amfs</td>
<td>97.1</td>
<td>96.0</td>
<td>93.2</td>
<td>84.2</td>
<td>64.2</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>94.4</td>
<td>92.8</td>
<td>87.8</td>
<td>76.5</td>
<td>52.0</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
<td>94.8</td>
<td>93.4</td>
<td>89.4</td>
<td>79.3</td>
<td>58.4</td>
</tr>
<tr>
<td>24+E+ΔE+ΔΔE</td>
<td>amfs</td>
<td>97.1</td>
<td>95.9</td>
<td>93.2</td>
<td>84.0</td>
<td>64.7</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>94.2</td>
<td>92.7</td>
<td>87.5</td>
<td>75.7</td>
<td>51.3</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
<td>94.9</td>
<td>93.3</td>
<td>89.3</td>
<td>79.4</td>
<td>59.3</td>
</tr>
<tr>
<td>21+E+ΔE+ΔΔE</td>
<td>amfs</td>
<td>96.5</td>
<td>95.0</td>
<td>90.5</td>
<td>78.9</td>
<td>56.4</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>94.1</td>
<td>92.3</td>
<td>87.3</td>
<td>75.4</td>
<td>50.7</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
<td>94.5</td>
<td>92.9</td>
<td>88.8</td>
<td>78.6</td>
<td>58.2</td>
</tr>
<tr>
<td>18+E+ΔE+ΔΔE</td>
<td>amfs</td>
<td>96.5</td>
<td>94.8</td>
<td>90.3</td>
<td>78.6</td>
<td>55.6</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>94.1</td>
<td>92.4</td>
<td>87.0</td>
<td>75.4</td>
<td>50.1</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
<td>94.5</td>
<td>92.7</td>
<td>89.1</td>
<td>78.7</td>
<td>58.5</td>
</tr>
<tr>
<td>15+E+ΔE+ΔΔE</td>
<td>amfs</td>
<td>96.5</td>
<td>94.6</td>
<td>89.6</td>
<td>77.3</td>
<td>53.6</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>93.6</td>
<td>92.1</td>
<td>86.3</td>
<td>73.9</td>
<td>48.9</td>
</tr>
<tr>
<td></td>
<td>hlda</td>
<td>94.2</td>
<td>92.8</td>
<td>88.9</td>
<td>78.8</td>
<td>58.5</td>
</tr>
</tbody>
</table>

(although overall hlda is significant better than lda). The obvious inference is that the recognition process can become faster when amfs is used without dropping more than 1% of the maximum performance since the speed of a HMM ASR system depends on the feature vector size.

5 Conclusions

In this paper, we presented an advanced method to select robust features subsets for speech recognition based on a spectro-temporal advanced auditory model. Under the assumption of small distortion errors we used perturbation analysis and the sensitivity matrix to build the feature selection algorithm. We included the first and second order time derivatives of the
features and showed how we succeeded to associate them to the perception information obtained from the sensitivity matrix. The experimental results verified that the human auditory periphery is a powerful tool in selecting relevant features for robust speech recognition.

References


