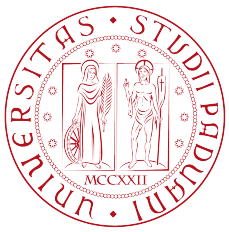

Giuliano Bernardi

IMPLICATIONS OF MODULATION FILTERBANK PROCESSING FOR AUTOMATIC SPEECH RECOGNITION

Master's Thesis, July 2011

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

DTU Electrical Engineering
Department of Electrical Engineering

THIS REPORT WAS PREPARED BY:
Giuliano Bernardi

SUPERVISORS:
Alfredo Ruggeri
Torsten Dau
Morten Løve Jepsen

ADDRESS:
Department of Electrical Engineering
Centre for Applied Hearing Research (CAHR)
Technical University of Denmark
Ørsteds plads building 352
DK-2800 Kgs. Lyngby
Denmark

http://www.dtu.dk/centre/cahr/
Tel: (+45) 45 25 39 32
E-mail: cahrinfo@elektro.dtu.dk

PROJECT PERIOD:	January 17, 2011 – July 16, 2011
CATEGORY:	1 (public)
EDITION:	First
COMMENTS:	This report is part of the requirements to achieve the Master of Science in Engineering (M.Sc.Eng.) at the Technical University of Denmark. This report represents 30 ECTS points.

PREFACE

This report documents the Master Thesis project carried out at the Center for Applied Hearing Research (CAHR) at the Technical University of Denmark (DTU) as a final work for the Master of Science in Engineering (M.Sc.Eng.). The project was carried out from January to July 2011 for a total workload of 30 ECTS.

ABSTRACT

An auditory-signal-processing-based feature extraction technique is presented as front-end for an Automatic Speech Recognition (ASR) system. The front-end feature extraction is performed using two auditory perception models, described in Dau *et al.* (1996a, 1997a), implemented to simulate results from different psychoacoustical tests. The main focus of the thesis is put on the stages of the models dealing with temporal modulations. This is done because evidence of the crucial role played by temporal modulations in speech perception and understanding were confirmed in different studies (e. g. Drullman *et al.*, 1994a,b; Drullman, 1995) and the investigation of such relevance in an ASR framework could allow to achieve better understanding of the complex processes involved in the speech analysis performed by the human auditory system.

The accuracy results on both clean and noisy utterances from a speaker-independent, digits-based speech corpus were evaluated for a control case, given by Mel-Frequency Cepstral Coefficient (MFCC) features, and for several cases corresponding to modifications applied to the auditory models. The results with the auditory-based features, encoded using the Dau *et al.* (1996a) model, showed better performance than the ones with MFCC features deriving from an additional noise robustness, confirming the findings in Tchorz and Kollmeier (1999). No improvement were apparently achieved using the features extracted with the Dau *et al.* (1997a) model, introducing a filterbank in the modulation domain, compared to the results obtained with the Dau *et al.* (1996a) model. However, it was argued that this behavior is likely to be caused by technical limitation of the framework employed to perform the ASR experiments.

Finally, an attempt to replicate the results from an ASR study (Kanedera *et al.*, 1999) validating the perceptual findings on the importance of different modulation frequency bands was performed. Some of the results were confirmed, whilst others were refuted, most likely because of the difference in the auditory signal-processing between the two studies.

ACKNOWLEDGMENTS

I would like to thank all the people that supported and helped me throughout the past six months during the development of this master project. A first acknowledgment goes to both my supervisors, Torsten Dau and Morten Løve Jepsen, for the help and the many valuable advice they gave me. Moreover, I would like to express my appreciation about the period I spent working is such a nice, but at the same time very stimulating, working environment as the Center for Applied Hearing Research in DTU is. Additional acknowledgments go to Guy J. Brown (University of Sheffield) and Hans-Günter Hirsch (Niederrhein University of Applied Sciences) — for the answers they provided to my questions about some issues with the HTK and about Automatic Speech Recognition (ASR) — as well as to Roberto Togneri (University of Western Australia) that allowed me to modify and use his HTK scripts.

I would also like to thank my family and my friends back in Italy, that have always been close during the past two years I spent in Denmark. Last but (definitely) not least, a special thanks goes to all the people I spent the last two years (and especially the last six months of my thesis) of my master with, especially the group from Engineering Acoustics 2009/2010 and all the other friend I have made in DTU. Thank you all for this great period of my life!

— Giuliano Bernardi

CONTENTS

LIST OF FIGURES

ACRONYMS

ANN Artificial Neural Network

ASR Automatic Speech Recognition

BM Basilar Membrane

BPE Band Pass Experiment

CMS Cepstral Mean Subtraction

CTK RESPITE CASA Toolkit

DauOCF Dau *et al.* (1997a) filterbank Original Center Frequencies

DauNCF Dau *et al.* (1997a) filterbank New Center Frequencies

DCT Discrete Cosine Transform

DFT Discrete Fourier Transform

DTW Dynamic Time Warping

ERB Equivalent Rectangular Bandwidth

FFT Fast Fourier Transform

FQNCF Fixed-Q filterbank New Center Frequencies

FT Fourier Transform

GT Gammatone

HMM Hidden Markov Model

HSR Human Speech Recognition

HTK Hidden Markov Models Toolkit

IHC Inner-Hair Cell

IIR Infinite Impulse Response

IR Internal Representation

J-RASTA Jah RelAtive SpecTrAl

KLT Karhunen-Loève Transform

LPC Linear Predictive Coding

M1 *Method 1*

M2 *Method 2*

MFB Modulation FilterBank

MFCC Mel-Frequency Cepstral Coefficient

MLP Modulation Low-Pass

MTF Modulation Transfer Function

PCA Principal Components Analysis

PLP Perceptual Linear Predictive

RASTA RelAtive SpecTrAl

RMS Root Mean Square

SNR Signal to Noise Ratio

TMTF Temporal Modulation Transfer Function

WAcc Word Recognition Accuracy

WER Word Error Rate

1 INTRODUCTION

Automatic Speech Recognition (ASR) refers to the process of converting spoken speech into text. From the first approaches to the problem over than seventy years ago, many improvements have been introduced, especially in the last twenty years thanks to the application of advanced statistical modeling techniques. Moreover, hardware systems upgrades together with the implementation of faster and more efficient algorithms fostered the diffusion of ASR systems in different areas of interests, as well as the possibility of having nearly real-time continuous-speech recognizers, which are nevertheless employing very large dictionaries with hundreds of thousands words.

Both changes in the features encoding processes and in the statistical modeling are narrowing down the performance gap, usually described by an accuracy measure, between humans and machines. In Lippmann (1997), an order-of-magnitude difference was reported between Human Speech Recognition (HSR) and ASR in several real life recognition conditions. After more than ten years, besides the mentioned improvements, there are still rather big differences between humans and machines recognition of speech in some critical conditions. The same level of noise robustness observed in HSR experiments is far from being achieved with the current methods and models employed in ASR and this could be due to both problems in the feature extraction procedures developed so far as well as to partially unsuited modeling paradigms.

In fact, ASR performance breaks down already at conditions and at Signal to Noise Ratios (SNRs) which only slightly affect human listeners (Lippmann, 1997; Cui and Alwan, 2005; Zhao and Morgan, 2008; Zhao *et al.*, 2009; Palomäki *et al.*, 2004). Thus, the idea of modeling speech processing in a way closer to the actual processing performed by the human auditory pathway seems to be relevant. Such approaches, namely auditory signal-processing based feature extraction techniques, have been already investigated in several studies (e. g. Brown *et al.*, 2010; Holmberg *et al.*, 2007; Tchorz and Kollmeier, 1999) and have (sometimes) shown improvements compared to the classic feature extraction techniques, such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC) or Perceptual Linear Predictive (PLP) analysis (Davis and Mermelstein, 1980; Markel and Gray, 1976 and Hermansky, 1990 respectively), especially in the case of speech embedded in noise.

The main focus of the current work is to test a new set of auditory-based features, and use the results obtained in such a case in compar-

ison to the results of a standard method (referred to as the baseline and chosen to be MFCC features). This should allow to systematically investigate the importance of different processing stages of the auditory system in the process of decoding speech. Specifically, the processing of temporal modulations (i. e. the changes of the speech envelope with time) of speech is investigated with greater detail, due to the strong importance of these speech attributes observed in several perceptual tasks (Drullman *et al.*, 1994a,b; Drullman, 1995; Houtgast and Steeneken, 1985). The investigation performed is the current work is more oriented toward hearing research. Thus, the new feature encoding strategies employing auditory models will be analyzed and their result will be interpreted to obtain further information about the importance of the mentioned stages in robust speech perception, more than just merely aiming to optimizing already existing techniques to achieve better results.

The first part of the thesis describes the tools practically exploited to perform the ASR experiments, starting with Chapter 2, that provides a description of the ASR systems used in the current work, splitting the discussion in the two traditional subsystems embodied in what is commonly referred to as a speech recognizer: front- and back-end (Rabiner, 1989). The front-end used to obtain the reference features, i. e. the MFCCs which have been used to compute the baseline results, is described and compared with an other well known method called RelAtive SpecTrAl (RASTA), Hermansky and Morgan (1994), and with the different auditory-based feature extraction techniques. The back-end section describes, in a rather simplified way, how the core of the recognition system works: the statistical concept of Hidden Markov Model (HMM) is provided and its usage in ASR explained.

In Chapter 3 of the current work, the auditory models employed to accomplish the feature extraction are presented and described. Firstly, the model based on the Dau *et al.* (1996a) study is presented. The function of each stage is briefly analyzed and complemented with figures illustrating the way the signals are processed. Subsequently, the model based on the Dau *et al.* (1997a) study is introduced. In both the cases, particular attention is drawn to the stage operating in the modulation domain, comprising the diverse versions of filterbanks.

Chapter 4 introduces the concept of auditory-based features and its usage in ASR. The methods employed to extract the feature vectors from the Internal Representations (IRs) computed via auditory models are described and the different problems encountered in this process (together with the proposed ways to solve them) illustrated. Furthermore, a brief introduction is given of the speech material adopted for the recognition experiments.

The second part of the thesis introduces and discusses the results of the current work. Chapter 5 reports the results of several simulations performed in the current study. It is divided in two parts discussing

the results of the standard ASR experiments carried out in the first part of the project, providing the accuracy scores as a function of the SNR, and the results of a different kind of experiment, inspired by the work of Kanedera *et al.* (1999), providing the accuracies as a function of lower and upper cutoff frequency of a set of band-pass filters described in the modulation domain.

In Chapter 6, the results collected from the different simulations are discussed and interpreted in order to provide a meaningful answer to the problems arisen in the previous sections. Some of the limitation encountered in the different parts of the employed framework are discussed and, based on these, some different approaches as well as new ideas for the continuation of the current work are proposed.

Finally, a summary of the work is provided in Chapter 7.

2 AUTOMATIC SPEECH RECOGNITION

In order to perform the recognition task, a recognizer is required. In ASR the word *recognizer* usually denotes the whole system, i. e. the whole sequence of stages that are gone through in the process of speech recognition from recording of the speech signal to the output of the recognized message. The two main parts that can be defined within a recognizers are the front-end and the back-end. Concisely, one can refer to the front-end as the part of the system that receives the audio signal, analyzes and converts it to a suitable format to be further processed, while the back-end is the actual recognizer mapping words or phonemes' sequences to the signal processed in the first part and testing the modeled responses.

In the current work, a freely available recognizer[1] has been employed, called Hidden Markov Models Toolkit (HTK). The program offers several tools for manipulating HMMs, the statistical models by which the actual recognition process is performed. HTK is mainly used for ASR, but it can be adapted to other problems where HMMs are employed, such as speech synthesis, written digits or letters recognition and DNA sequencing. A detailed description of the usage of HMMs for speech recognition is given, e. g. in Gales and Young (2007). A manual explaining how the HTK works and is structured can be downloaded at the HTK's website (Young *et al.*, 2006).

2.1 FRONT-ENDS

As previously mentioned, front-end is the word used to describe the preparatory methods employed by the recognizer to obtain a signal representation suitable to be further analyzed by the subsequent stages in ASR. The conversion transforms the audio signal into an alternative representation, consisting of a collection of *features*. The extraction of features, or sets of them composing the so called feature vectors, is a process required for two main reasons:

A. identifying properties of the speech signal somehow (partially) hidden in the time domain representation, i. e. enhance aspects contributing to the phonetic classification of speech;

B. reduce the data size, by leaving out those information which are not phonetically or perceptually relevant.

The first point states that, although the message carried by the audio signal is, of course, embedded within the signal itself, several other in-

1 http://htk.eng.cam.ac.uk/

formation are not directly related to the message to be extracted, thus contributing to introduce variability in the informational-distortion-free message. Without performing any transformation on the signal's samples, the classification process of the different segments extracted from the message is unfeasible with the methods currently used in ASR, mainly because the time domain representation of audio signals suffers from the aforementioned variability. Therefore, as often required in classification problems, one has to map the original data to a different dataset which guarantees a robust codification of the properties to be described. The robustness of the representation, in the case of ASR tasks, has to be required with respect to a whole set of different parameters responsible (in different ways) of the high non-stationarity of the speech signals. Amongst others, one can list speaker-dependent variabilities given by accent, age, gender etc... and prosody-dependent variabilities, i. e. rhythm, stress and intonation (Huang *et al.*, 2001).

The second point is related to the computational effort needed to sustain an ASR system. At the present day, it is not unusual to work with audio signals sampled at several kHz; for this reason, the amount of data with such high sampling frequencies is a critical issue, even considering the high computational power available. If the system has to be used for real time recognition, data reduction could be a necessity.

From the early years of ASR to the present day, several methods of feature extraction have been developed. Some of these methods have found a wide use in ASR and have been used for the past thirty years (e. g. MFCCs). These procedures will be referred to as *classical* methods. There are some similarities between several of these methods; most notably is the fact that they employ short-term speech spectral representations. This is mostly due to the fact that short-term speech representation approaches were successfully used in speech coding and compression before to be used in ASR and, considering the good results obtained in the mentioned fields, they were thought to offer a good mean to approach the problem of ASR (Hermansky, 1998).

Another important aspect relative to the processes of features encoding is given by the insertion of dynamic coefficients (i.e. changes of the features with time), which will be discussed in greater detail in one of the following section.

2.1.1 *Spectral- and temporal-based feature extraction techniques*

As previously pointed out, some of the methods introduced in the early years of ASR, were originally developed for different purposes and subsequently found an important application in the field of ASR. In speech coding and compression procedures a different kind of information is exploited that in ASR has to be rejected to offer a more

robust representation of the noise-free signal, like speaker-dependent cues and environmental information (Hermansky, 1998). Moreover, some aspects of the *classical* methods were developed to work with ASR systems different to those representing the main trend nowadays (Sakoe and Chiba, 1978; Paliwal, 1999). Amongst others, two widely used *classical* approaches are Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coding (LPC).

In the classic approaches to ASR, the preprocessing of the data fed to the pattern matching systems was mostly realized taking into consideration spectral characteristics of the speech signals. Indeed, some properties of speech can be recognized in the frequency domain in an easier way compared to the time domain, e. g. speech voicing or vowel's formants (Ladefoged, 2005). Therefore, using the spectral representation of speech to extract information about it seems to be a sensible choice. Such methods rely on the assumption that speech can be broken down into short frames (which lengths are on the order of a few tens of milliseconds) that are considered stationary and independent from each others. Such assumptions lead to tractable and efficiently implementable systems, but it is fairly straightforward to understand that such hypothesis is not fulfilled in many real life cases, as they neglect some crucial aspects of speech that are defined in longer-term temporal intervals (around a few hundredths of milliseconds). See, e. g., Hermansky (1997); Hon and Wang (2000).

Based on this consideration, methods accounting for the temporal aspects of speech have been developed since the Eighties. Dynamic cepstral coefficients, introduced in Furui (1986), represent one of the first attempts used in ASR to include temporal information within the feature vectors. These coefficients return measures of the changes in the speech spectrum with time, representing a derivative-like operation applied on the static (i. e. cepstral) coefficients. The first order coefficients are usually called *velocities* or *deltas* whereas the second order ones are defined *accelerations* or *delta-deltas*. The coefficients' estimation is often performed employing a regression technique (Furui, 1986); this approach is also implemented by the recognizer adopted in this work and it will be subsequently described. Dynamic coefficients are usually employed in ASR where they are used to build augmented MFCC feature vectors. Appending these coefficients to the static feature vectors has proved to increase the recognizer performance in many studies (e. g. Furui, 1986; Rabiner *et al.*, 1988) whereas they were found to provide worse results when used in isolation, as noted e. g. in Furui (1986).

Other strategies, lead by the pioneeristic work in Hermansky and Morgan (1994) back in the Nineties, started to employ *solely* temporal-based features of the speech signals, in order to provide robust recognition methods in real life noise conditions, which are likely to bring severe performance degradation with the *classical* methods. The RASTA

method, introduced in Hermansky and Morgan (1994), is one of the aforementioned techniques and it showed improvements in some conditions together with some degradation in others. One of the advantages introduced by this technique, can be understood by carrying out a simple analysis using concepts of homomorphic signal processing, briefly introduced in Appendix A.1.

2.1.2 *Mel-Frequency Cepstral Coefficients*

Amongst the number of feature extraction techniques that can be listed, Mel-Frequency Cepstral Coefficients (MFCCs) will be described in the following. This is done because MFCCs were selected in this work to represent the baseline used for comparison. The choice was made based on the fact that in several other studies, MFCCs were employed as a baseline to test new features encoding strategies, both auditory-modeling-based (e. g. Tchorz and Kollmeier, 1999; Holmberg *et al.*, 2006, 2007; Brown *et al.*, 2010; Jürgens and Brand, 2009) and more purely signal-processing-oriented approaches (e. g. Batlle *et al.*, 1998; Paliwal, 1999; Palomäki *et al.*, 2006).

MFCCs can be referred to as a *classical* encoding approach because it has been used ever since its introduction in the Eighties in the work of Davis and Mermelstein (1980). The name Mel-Frequency Cepstral Coefficients suggests the two key operations performed in this method. Both the concepts of *Mel scale* and *cepstrum* are exploited. The mel-frequency scale is a (nonlinear) perceptual scale of pitches, introduced in Stevens *et al.* (1937). Since perception is taken into consideration, it means that even though the MFCC method does not attempt to strictly model the auditory system processing, some meaningful perceptual measures are implemented. One of the proposed conversion formulæ between frequencies in Hertz (denoted by f) and frequencies in Mel (denoted by mel) is given by, Young *et al.* (2006):

$$mel = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{2.1}$$

An approximation of the formula can be done considering an almost linear spacing below 1 kHz and an almost logarithmic spacing above 1 kHz. The filterbank employed in the MFCC method exploits the mel-frequencies distribution, by an equally spaced set of central frequencies. An example of the mel-filterbank is shown in the fourth panel from the top of Fig. 2.1. In the MFCC case, the logarithm is taken on the different power spectra obtained filtering the power spectra of the time frames with the mel-filterbank.

The filterbank represents a very rough example (using triangular overlapping windows) of the auditory filterbank and provides the mapping of the frames' powers onto the mel-frequency scale, somehow mimicking the frequency selectivity of the auditory system. The

subsequent logarithm function provides the compression of the filterbank's outputs and it was mainly introduced in combination with the Discrete Cosine Transform (DCT) to provide a signal-processing concept very similar to the cepstrum. This was applied since it was found to be very useful for other speech processing purposes (Kolossa, 2007; Paliwal, 1999). The DCT of the log filterbank amplitudes m_j (of the single time frame) is computed by, Young *et al.* (2006):

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=0}^{N} m_j \cos\left[\frac{\pi i}{N}(j - 0.5)\right]. \tag{2.2}$$

Only a small number of coefficients c_i is usually retained (10 to 14, e. g. Davis and Mermelstein, 1980; Tchorz and Kollmeier, 1999; Brown *et al.*, 2010; Holmberg *et al.*, 2007). Further details about the DCT are provided in Appendix A.2.

A summary of the signal processing steps, whose illustration is provided in Fig. 2.1, necessary to evaluate MFCCs is the following:

1. segmentation of the signal in a sequence of overlapping frames (usually 25 ms long, 40% overlap);
2. Fourier Transform (FT) of each frame and mapping of its power spectrum onto a mel-frequency scale;
3. cepstrum of the frequency warped power spectrum (logarithm of it followed by DCT).

The procedure of MFCC feature encoding was performed internally the HTK, via the command HCopy. 14 MFCCs were retained as well as 14 deltas and delta-deltas, for a total number of 42 coefficients per feature vector. The dynamic coefficients, mentioned in the previous section, are evaluated via the formula, Young *et al.* (2006):

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta\,(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{2.3}$$

where d_t is the delta coefficient at time t computed using the $2\Theta + 1$ frames between $c_{t-\Theta}$ and $c_{t+\Theta}$. No energy terms (defined e. g. in Young *et al.*, 2006) were included as features in the current study, as they were not in some of the works used as references for the parametrical tuning of HTK (Brown *et al.*, 2010; Holmberg *et al.*, 2007). For the same reason, on the other hand, the 0^{th} cepstral coefficients were included, even though in some works they are referred to as inaccurate, Picone (1993). Figure 2.2 illustrates the MFCC representation of a speech signal corresponding to the utterance of the digit sequence "8601162".

Regarding the decorrelation properties of the DCT, see Appendix A.3, in Fig. 2.3 it is shown the correlation matrix of the MFCC features shown in Fig. 2.2.

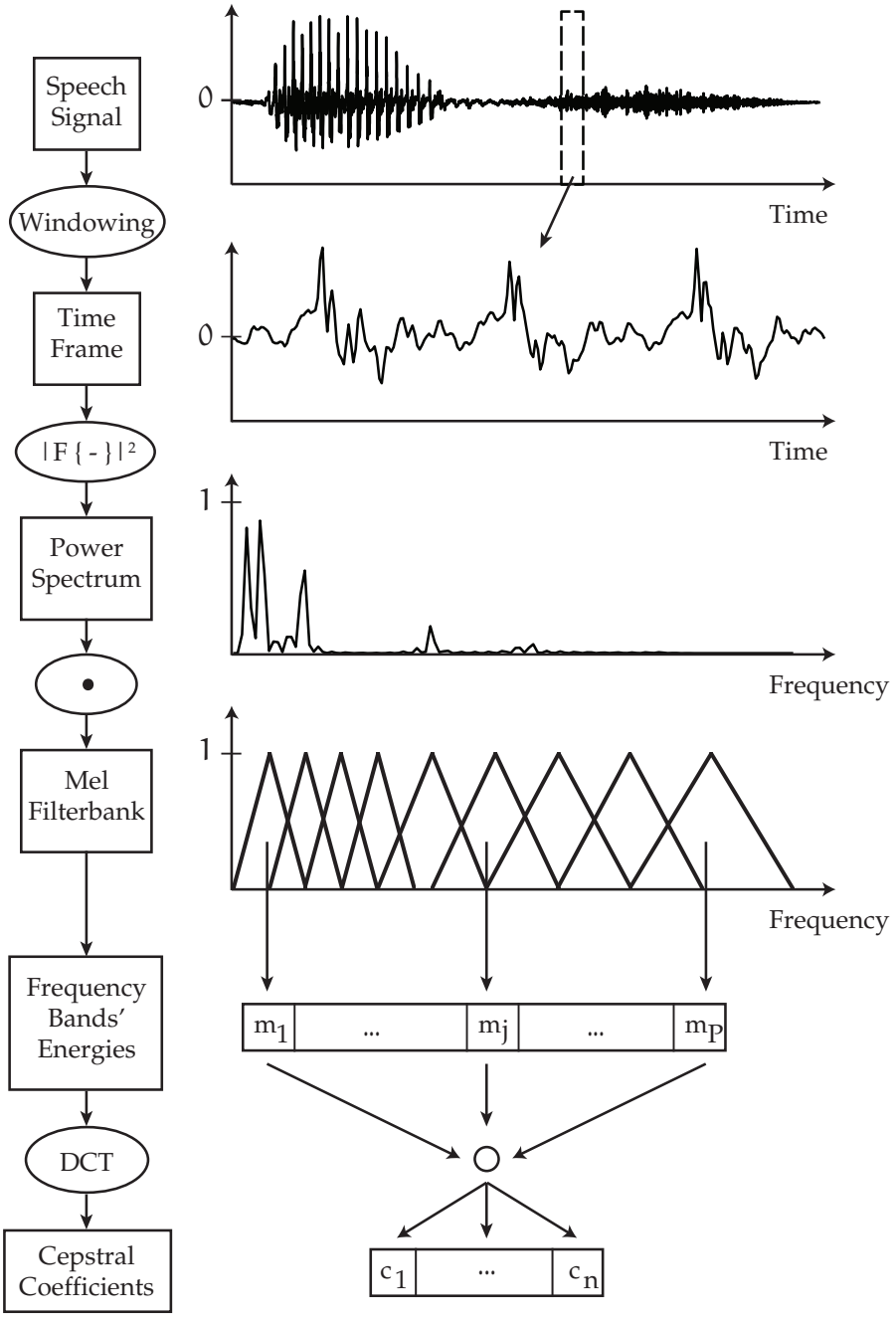

Figure 2.1: Illustration of the signal processing steps necessary to evaluate MFCC features. A detailed explanation can be found in the text.

2.1.3 *RASTA method*

Another rather popular method in the ASR field is the so called RelAtive SpecTrAl (RASTA), introduced in Hermansky and Morgan (1994). Besides the wide popularity gained by this method, its importance — regarding this project — consists in the fact that the operations performed by the RASTA algorithm are similar to the ones performed by the current auditory model. RASTA was introduced as an evolution

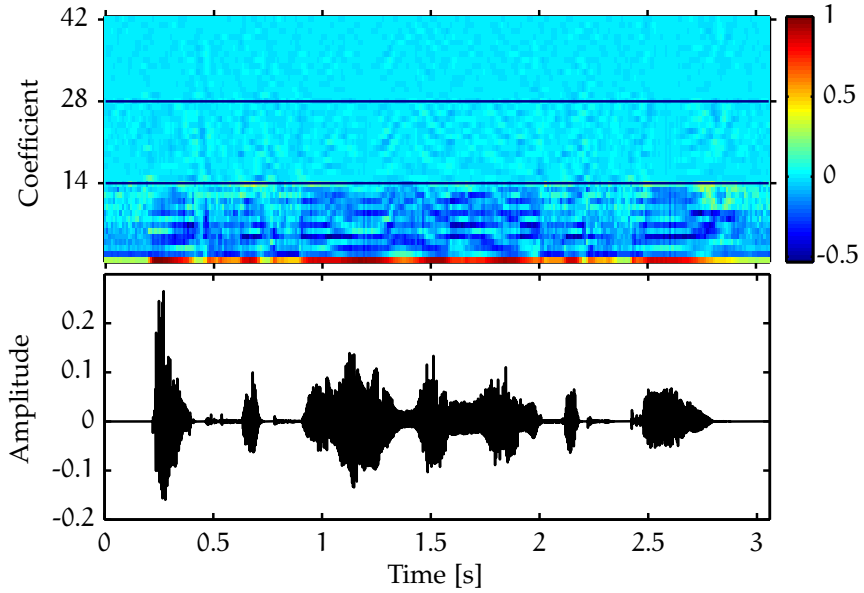

Figure 2.2: Example of MFCC feature extraction (top) on an utterance of the digit sequence "8601162" (bottom). The coefficients' sequence is given by: 14 MFCCs (c_0 to c_{13}), 14 deltas and 14 delta-deltas for a total of 42 entries.

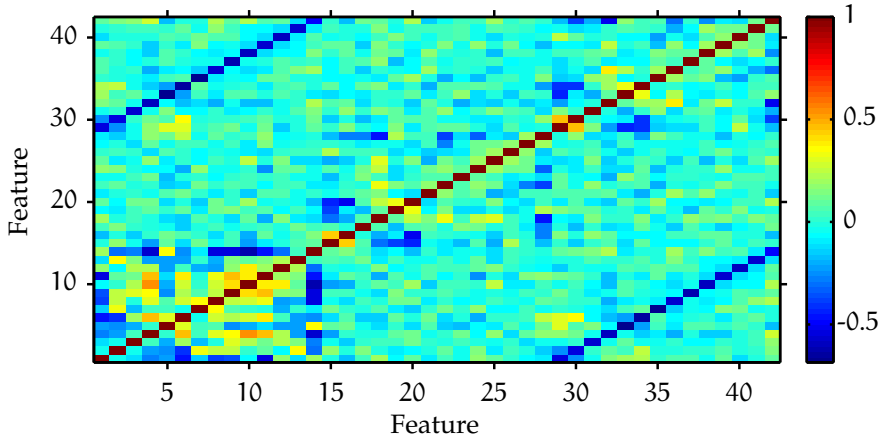

Figure 2.3: Correlation matrix of the MFCC features representation in Fig. 2.2. The high energy concentration in the diagonal and the lower energy concentration in the off-diagonal area describe the high degree of uncorrelation between the features.

of the *classical spectral oriented* methods for ASR, since it is one of the precursors of the *temporal oriented* methods. It was developed on the base of important characteristics that can be observed in real-life (i. e. somehow corrupted) speech samples. Firstly, by noting that temporal properties of disturbances affecting speech varies differently from the temporal properties of the actual speech signals (Hermansky and Morgan, 1994). In second place, the evidence that the modulation frequencies around 4 Hz were found to be perceptually more important

than lower or higher frequency in the modulation frequency domain (see e. g. Drullman *et al.*, 1994a,b, even though in Hermansky and Morgan, 1994 they refer to previous studies). Based on these general ideas, the filter developed to be used within the RASTA method had transfer function:

$$H(z) = 0.1z^4 \cdot \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98\,z^{-1}} \tag{2.4}$$

which can be expressed by the difference equation:

$$\begin{aligned} y(\omega, k) = {} & 0.2\,x(\omega, k) + 0.1\,x(\omega, k-1) - \\ & 0.1\,x(\omega, k-3) - 0.2\,x(\omega, k-4) + \\ & 0.98\,y(\omega, k-1). \end{aligned} \tag{2.5}$$

Figure 2.4 illustrates the frequency response of the filter defined in Eq. (2.4), showing the bandpass behavior of the frequency response (Hermansky and Morgan, 1994).

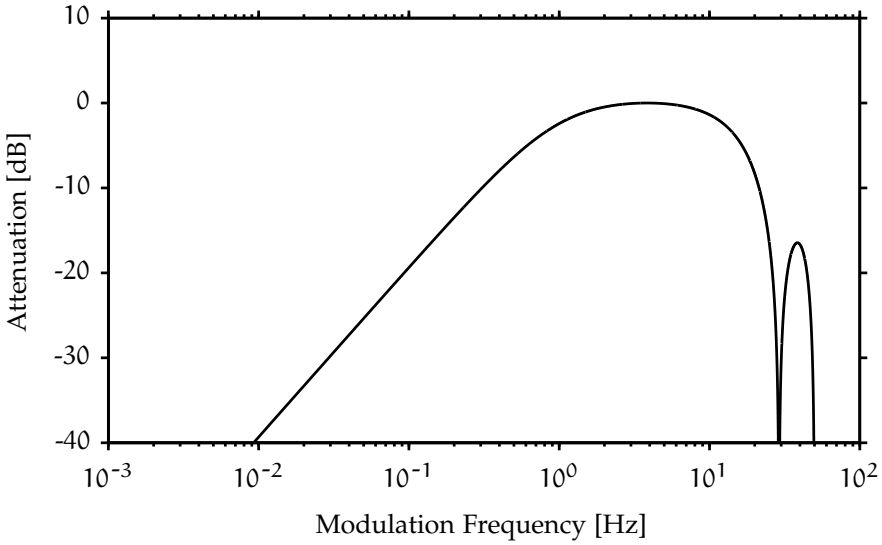

Figure 2.4: Frequency response of the RASTA filter, showing a band-pass characteristics and a approximately flat response for the frequencies in the range [2, 10] Hz. Redrawn from (Hermansky and Morgan, 1994).

The steps of the RASTA algorithm can be summarized as follows:

1. computation of the critical-band power spectrum;
2. transformation via compressive static nonlinearity;
3. filtering of the temporal trajectories, using the filter in Eq. (2.4);
4. transformation via expansive static nonlinearity (inverse of the compressive);

5. loudness equalization and Stevens' power law simulation (by raising the signal to the power 0.33);

6. computation of an all-pole model of the resulting spectrum

For the reasons described in Appendix A.1, a widely used function employed in the compressive static nonlinearity step is the logarithm. Although RASTA processing of speech turned out to be very efficient in presence of convolutional disturbances, its robustness drops for other kind of noises. Some modifications have been introduced to improve RASTA performance. In particular, in order to deal with additive disturbances a slightly modified version of the method has been proposed, the Jah RelAtive SpecTrAl (J-RASTA), presented in Morgan and Hermansky (1992); Hermansky and Morgan (1994). The modification proposed consists in the introduction of the parameter J in the log-transformation which multiplies the power spectra in the different frames.

$$y = \log(1 + Jx). \tag{2.6}$$

The value of J depends on some estimates of the SNR, Morgan and Hermansky (1992). By taking the Taylor expansion of Eq. (2.6), it can be seen that for $J \gg 1$ the function has a quasi-logarithmical behavior; if $J \ll 1$ the function can be approximated by a linear operator.

In Chapter 6, an explanation of the reasons which have guaranteed the success of this method for almost the past 20 years are discussed and used as a mean of comparison with the auditory-based feature extraction methods.

2.1.4 *Auditory-signal-processing-based feature extraction*

Conveniently adapted auditory models have been employed in several studies (e. g. Tchorz and Kollmeier, 1999; Brown *et al.*, 2010; Holmberg *et al.*, 2006, 2007) to process the audio signal in the correspondent feature extraction procedures. The auditory models employed in the different experiments will be discussed in greater detail in Section 4.1 due to the relevance of the auditory-model-based approach for the current project.

2.2 BACK-END

In ASR, the back-end is the stage that models the encoded speech signal and realizes its conversion in a sequence of predefined symbols (e.g. phonemes, syllables or words) via some kind of deterministic or statistical model. From the early developments of ASR until the beginning of the Nineties, there was a strong disagreement regarding the proper acoustic models to be used. Several different approaches have been proposed through the years such as Dynamic Time Warping

(DTW), Artificial Neural Networks (ANNs) and Hidden Markov Models (HMMs).

Currently, HMMs-based modeling has become one of the most popular techniques employed in ASR (Morgan *et al.*, 2004). Furthermore, the toolkit employed in the current work, the HTK (Young *et al.*, 2006), exploits the concept of HMMs and their application in ASR. Therefore, a brief introduction to this statistical tool, as well as some words regarding HMMs modeling in ASR will be shortly discussed in the following.

2.2.1 *Hidden Markov Models*

A complete description of the theory behind HMMs is not the purpose of this section. However, introducing the topic can be helpful to better understand why the choice of using HMM for ASR is sensible; moreover, it will be pointed out one of the characteristics that somehow limited the application of HTK for the goal of the current project, namely the constraint on the covariance matrixes.

An HMM can be generally defined as a mathematical model that can predict the probability with which a given sequence of values was generated by a state system. Regarding the ASR problem, speech units (such as phones, phonemes, syllable or words) can be associated to sets of parameters describing them. These parameters are embedded within a statistical model built from multiple utterances of each unit. A probabilistic modeling framework allows to obtain a much more generalizable parametrical representation than using directly the speech units (or derived sets of their features), Morgan *et al.* (2004).

The HMM framework applied to ASR relies on a simple (yet approximated) assumption: the possibility of interpreting speech, a highly nonstationary process per definition, as a sequence of piecewise stationary processes whose characteristics can be modeled on the base of short-term observations. Thus, speech units are characterized by statistical models of collections of stationary speech segments (Morgan *et al.*, 2004; Bourlard *et al.*, 1996). A summary of other assumption that have to be taken into account when adopting a statistical HMM framework is provided in Morgan *et al.* (2004).

In order to understand the idea behind HMMs, the concept of Markov model has to be introduced. A Markov model[2] is a stochastic model describing the behavior of a system composed by a set of states, undergoing state-to-state probabilistic transitions at discrete times, Rabiner (1989). Unlike other state-based stochastic models, a Markov model assumes the Markov property specifying that the state q_t

2 For speech recognition application the interest is focused on discrete Markov models or Markov chains.

occupied at a given time instant t only depends on the value of the previous state and not on the whole transition history. Thus:

$$P\left[q_t = S_j | q_{t-1} = S_i, \ldots, q_0 = S_k\right] = P\left[q_t = S_j | q_{t-1} = S_i\right]. \tag{2.7}$$

The state transition probabilities given by the right hand side of Eq. (2.7) are usually denoted as a_{ij}, Rabiner (1989).

A Markov model is too restrictive for the purposes of ASR (Rabiner, 1989) and it is required to be generalized to an HMM. In such a case, the states of the process are hidden, i. e. not observable anymore, and they are only accessible via observation variables $\mathbf{o}_t$ stochastically related to them. In ASR, the observations are usually coarse representations of short-term power spectra, meaning that the HMMs combines the model of an observable process accounting for spectral variability together with an underlying Markov process accounting for temporal variability. Among the different ways that have been employed to characterize the distributions of observation variables given a state, continuous Gaussian Mixture models will be considered, as they are adopted by the HTK (Young *et al.*, 2006). E. g. the probability $b_j(\mathbf{o}_t)$ of obtaining the observation $\mathbf{o}_t$ from the state j is:

$$b_j(\mathbf{o}_t) = P\left[\mathbf{o}_t | q_t = S_j\right] \tag{2.8}$$

$$= \sum_{k=1}^{M} c_{jk} \mathcal{N}\left[\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}\right] \tag{2.9}$$

where c_{jk} is the k^{th} component of the mixture and

$$\mathcal{N}\left[\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}\right] = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}_{jk}|}} e^{(\mathbf{o}_t - \mu_{jk})^T \boldsymbol{\Sigma}^{-1} (\mathbf{o}_t - \mu_{jk})}. \tag{2.10}$$

The parameters $\boldsymbol{\mu}_{jk}$ and $\boldsymbol{\Sigma}_{jk}$ are, respectively, mean and covariance of the multivariate distribution. Often, $\boldsymbol{\Sigma}_{jk}$ is constrained to be diagonal to reduce the number of parameters and properly train the HMMs using a smaller amount of training data, Gales and Young (2007). Furthermore, a reduction of the computational load and time is achieved. Whether diagonal covariance matrixes are used, the observations must be uncorrelated between each other, otherwise the estimated $\boldsymbol{\Sigma}_{jk}$ will only represent a poor approximation of the covariance matrix describing the real probability distribution (Gales and Young, 2007; Young *et al.*, 2006). As it will later be seen, this represents one of the limitations of the usage of HMMs for ASR (especially with auditory-oriented features). The quantities $\mathbf{A} = \{a_{ij}\}$, $\mathbf{B} = \{b_j(\mathbf{o}_t)\}$ and the initial state distribution $\boldsymbol{\pi} = \{\pi_i\}$[3] represent the model parameters of an HMM.

By now restricting the possible modeling scenarios to an isolated word recognition experiment, as it is in the current study, and given

3 Describing the probability of each state to be occupied at the initial time instant.

T observations $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T$ composing the word w_k, the whole recognition problem can be boiled down to the computation of the most likely word given the set of observations $\mathbf{O}$, Gales and Young (2007); Young *et al.* (2006):

$$\hat{w}_k = \arg\max_k \{P(w_k|\mathbf{O})\} \tag{2.11}$$

where the probability to be maximized can be expressed as:

$$P(w_k|\mathbf{O}) = \frac{P(\mathbf{O}|w_k)\,P(w_k)}{P(\mathbf{O})}. \tag{2.12}$$

Thus, given a set of priors $P(w_k)$ and provided an acoustics model $M_k = (\mathbf{A}_k, \mathbf{B}_k, \boldsymbol{\pi}_k)$ describing the word w_k, i. e. $P(\mathbf{O}|M_k) = P(\mathbf{O}|w_k)$, the maximization of $P(\mathbf{O}|M_k)$ returns the most likely[4] word $\hat{w}_k$. Figure 2.5 illustrates the concept of ASR by means of an example: the audio signal from an utterance of the word "yes" (bottom) is converted to its features representation (middle) and each observation is associated to the most likely phoneme (top).

The ways to actually perform the estimation of model parameters and probabilities in Eq. (2.12) or the maximization in Eq. (2.11) are not discussed here but it can just be mentioned that sophisticated dynamic programming techniques as well as advanced statistical tools are exploited in the task. Detailed explanations are offered in literature (e. g. Rabiner, 1989; Young *et al.*, 2006; Gales and Young, 2007).

After a set of HMMs is trained to the provided speech material and the models have been tested, a measure of the recognition accuracy is necessary to describe the goodness of the modeling. In the HTK, given a number N of total units to recognize and after the numbers of substitution errors (S), deletion errors (D) and insertion errors (I) are calculated (after dynamic string alignment), they are combined to obtain the percentage accuracy defined as, Young *et al.* (2006):

$$WAcc = \frac{N - D - S - I}{N} \times 100\%. \tag{2.13}$$

This measure will be employed in all the recognition results shown in the current study, as it has been used for comparing performance in several other studies (e. g. Brown *et al.*, 2010; Holmberg *et al.*, 2006; Tchorz and Kollmeier, 1999)[5].

4 In a maximum likelihood sense, for instance.

5 In literature, a related performance measure, called Word Error Rate (WER), is also employed. WER is defined as the complement of Word Recognition Accuracy (WAcc), i. e. $WER = 1 - WAcc$.

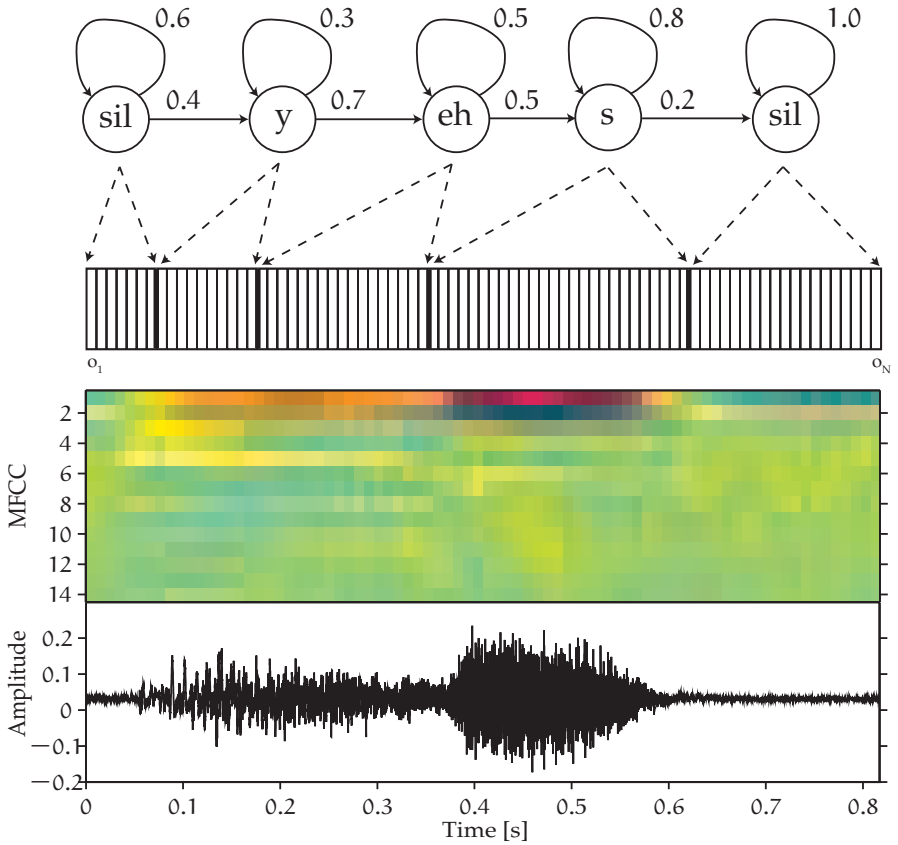

Figure 2.5: Schematic example of the main steps undergone during an ASR process. The signal (bottom) represents an utterance of the word "yes". It is converted to an MFCC features representation (bottom) and each observation is then associated with the most likely phoneme (top). A grammatical constraint, represented by the forced positioning of a silence at both the beginning and at the end of the word (denoted by `sil`), is also illustrated. The probabilities shown represent how likely is the transition from a state to the subsequent (or to remain in the same state), i.e. the transition probabilities a_{ij}.

3 AUDITORY MODELLING

The auditory model employed in the current study is a slightly modified version of the auditory model developed by Dau *et al.* (1996a) to simulate the results of different psychoacoustical tasks, such as spectral and forward masking as well as temporal integration. This modified version of the Dau *et al.* (1996a) model will be referred to as Modulation Low-Pass (MLP) throughout the current work[1]. It includes all the stages of the Dau *et al.* (1996a) model up to the optimal detector, not considered here since the detection process to be performed in ASR differs from the one needed in psychoacoustical tests and it is carried out by the statistical back-end. A subsequent version of the model that includes a modulation filterbank instead of a single low-pass modulation filter and is capable of simulating modulation-detection and modulation-masking experiments, is described in Dau *et al.* (1997a). This more recent version (again, the optimal detector is left out) is employed in some of the tested conditions and will be referred to as Modulation FilterBank (MFB).

3.1 MODULATION LOW-PASS

The processing stages of the first two models are here briefly described, with a visual description designed to guide the reader through the stages given in Fig. 3.1.

3.1.1 *Gammatone filterbank*

The first stage of the model accounts for the frequency separation of sounds performed within the cochlea from the basilar membrane. Thus, no outer- and middle-ear transfer function are considered. The frequency-place paradigm is a well known phenomenon of audition, see Békésy (1949), stating that the Basilar Membrane (BM) acts as a bank of continuous filters, each tuned to different frequencies within the range of audible frequencies spanned in a non-linear way.

Unlike the original model presented in Dau *et al.* (1996a), the current model implements a Gammatone (GT) filterbank in the form of the one found in Dau *et al.* (1997a). GT filter shapes were proven to give better fits to physiological data and a more efficient computation, Patterson *et al.* (1988), even though the model is purely phenomelogical unlike

1 Not to be confused with the acronym often employed to refer to the Multi-Layer Perceptron architecture of an ANNs. This is pointed out since ANN has been also used in many ASR studies and the same acronym could generate misunderstanding.

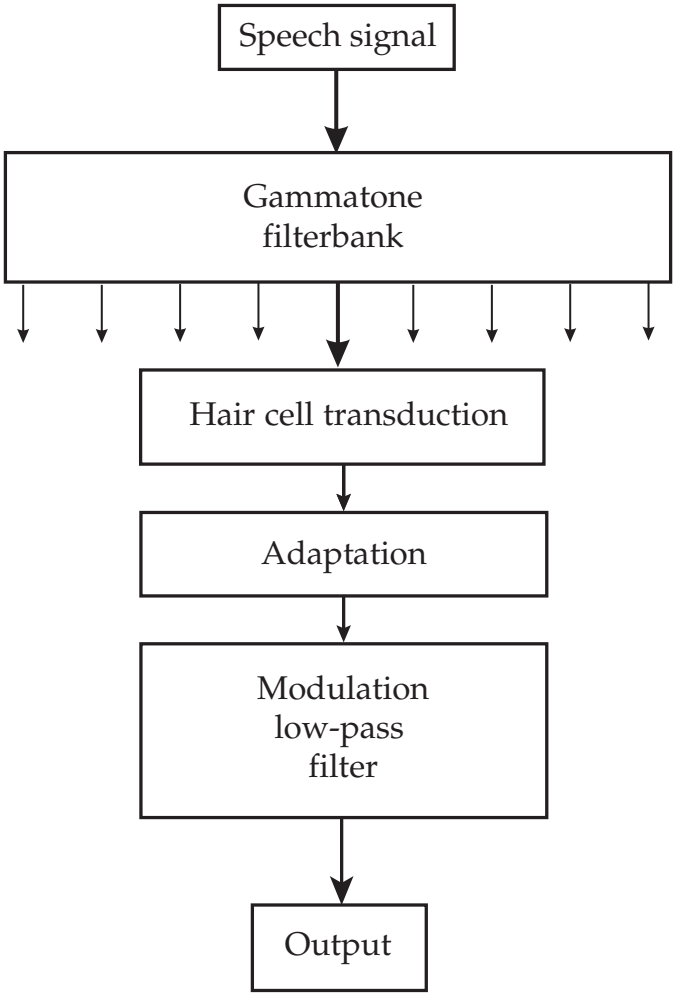

Figure 3.1: Block diagram of the MLP model.

the transmission-line cochlea models. The impulse response of the GT filters reads:

$$g(t) = a^{-1} t^{n-1} e^{-2\pi b t} \cos(2\pi f_c t)\,. \tag{3.1}$$

It can be interpreted as a cosine function shaped by an envelope decaying with an exponential function and rising from zero with a power function. The specific factors and parameters define the filters' properties:

- a is the normalization factor constraining the time integral over $0 < t < \infty$ of the envelope to 1;
- b is the factor determining the decaying slope of the exponential factor and can be seen as the duration of the response. It is closely related to the filter width;
- n is determining the slope of the rising part of the envelope and is referred to as the order of the filter. A value of $n = 4$ was chosen in the current work, Glasberg and Moore (1990);

- f_c is the center frequency, i. e. the frequency at which the filter is peaking in the frequency domain.

By taking the Fourier Transform (FT) of $g(t)$ in Eq. (3.1), the Gamma function (Γ) is introduced (Slaney, 1993), thus explaining the name chosen for the filter. In the current case a set of 189 filters have been employed, whose center frequencies are equally spaced on the Equivalent Rectangular Bandwidth (ERB) scale and range from 100 to 4000 Hz, the Nyquist frequency of the audio files of the adopted corpus. Figure 3.3 shows an example of the processing of a speech signal consisting of a series of five digits (top panel). The second panel from the top represent the Internal Representation (IR) after passing the signal through the GT filterbank.

The filterbank output gives an illustration of how the frequency content of the signal varies with time, and with the spoken digits. The frequency representation can also be used to visually inspect some differences and similarities between the speech segments (e. g. the similar frequency distribution between the utterances of the digit "one" in the time interval ca. 1.5 to 2 s or the difference with the digit "zero", at time ca. 1 to 1.5 s). After passing the signal through the GT filterbank, the processing of the following steps will be applied in parallel on each one of the frequency channels.

3.1.2 *Hair cell transduction*

The multiple outputs from the auditory filters represent the information about the processed sound in a mechanical form. At this point, in the auditory system, signals representing mechanical vibrations are converted to a form that is able to be processed by the higher stages of the auditory pathway. Thus the place-dependent vibrations of the BM are converted into neural spikes traveling along the auditory nerve.

The movements of the BM cause the displacement of the Inner-Hair Cells (IHCs) tips, called *stereocilia*. This displacement, in turn, opens up the small gates on the top of the each stereocilium, causing an influx of positively charged potassium ions (K^+), Plack (2005). The positive charging of the IHC causes the cell depolarization and triggers the neurotransmitter release in the synaptic cleft between the IHC and the auditory nerve fiber. Accordingly, an action potentials in the auditory nerve is created.

The described transduction mechanism only occurs at certain phases of the BM's vibration. Thus, the process is often referred to as the phase-locking property of the inner ear, Plack (2005). Nevertheless, the inner ear coding is performed simultaneously by a great number of IHCs. Therefore, a combined informational coding can be achieved, meaning that if the single cell cannot trigger an action potential each time the basilar membrane vibration causes the opening of its gates (e. g. due to the spurs of a pure tone at a frequency f_0), the overall spiking pattern

of a bunch of cells can successfully follow the timing of the input signal (Smith, 1976; Westerman and Smith, 1984). An illustration of the concept can be found in Plack (2005, Fig. 4.18). Although considering the aforementioned mechanism, there is a natural limit to the highest frequency that can be coded by the IHCs. That is why for the high frequency content of audio signals, the auditory nerve fibers tend to phase-lock to the envelope of the signal (and not to the fine structure anymore).

In order to simulate the mechanical-to-neural signal transduction via basic signal processing operations, the frequency channels' contents are half-wave rectified (to mimic the mentioned phase-locking property) and low-pass filtered using a second order Butterworth filter with a cut-off frequency of 1000 Hz. Although the latter exhibits a rather slow roll-off, it reflects the limitation of the phase-locking phenomenon for frequencies above 1000 Hz. The output after IHC transduction is shown in the middle panel of Fig. 3.3. It can be seen how the half-wave rectification causes the only the positive parts of the frequency channels' time trajectories to be retained. The low-pass filtering determines an attenuation of the higher frequency components, i. e. the top part of the auditory spectrogram.

3.1.3 *Adaptation stage*

The following step, called *adaptation* in the block diagram, performs dynamic amplitude compression of the IR. As the name suggests, the compression is not performed statically (e. g. taking the logarithm of the amplitudes) but adaptively, meaning that the compressive function changes with the signal's characteristics. The stage is necessary to mimic the adaptive properties of the auditory periphery, Dau *et al.* (1996a), and it represent the first inclusion of temporal information within the model. The presence of this stage accounts for the twofold ability of the auditory system of being able to detect short gaps of a few milliseconds duration, as well as integrate the information over intervals of hundreds of ms.

The implementation consists of five consecutive nonlinear adaptation loops, each one formed by a divider and a low-pass filter whose cutoff frequency (and therefore the time constant) takes the values defined in Dau *et al.* (1996a). The values of such time constants in Dau *et al.* (1996a) were chosen to fit measured and simulated data in forward masking conditions. An important characteristic introduced by the adaptive loop consists in the application of a non-linear compression depending on the rate of change of the analyzed signal. If the fluctuations within the input signal are fast compared to the aforementioned time constants, these changes are processed almost linearly. Therefore, the model produces an emphasis (strictly speaking it does not perform any compression) of the dynamically changing parts

(i. e. onsets and offsets) of the signal. When the changes in the signal are slow compared to the time constants, like in the case of more stationary segments, a quasi-logarithmic[2] compression is performed.

The result of the adaptation loop can be examined from the IR in the second panel from the bottom of Fig. 3.3, illustrating the enhancement of the digits' onsets (except for the central ones which are not separated by silence) and the compression of some of the peaks spotted within some of the digits' utterances (e. g. the two peaks within the third digit, "zero").

For the reasons that will be listed in following chapters, this stage is to be considered of great importance for the results obtained in the current work.

3.1.4 *Modulation filtering*

Humans perception of modulation, i. e. the sensitivity to the changes in the signals' envelopes, has often been studied in the past employing the concept of Temporal Modulation Transfer Function (TMTF), introduced in Viemeister (1979). The TMTF is defined as the threshold (expressed by the minimal modulation depth, or modulation index) for detecting sinusoidally modulated noise carriers and measured as a function of the modulation frequency. Data from the threshold detection experiments were used to derive the low-pass model of human sensitivity to temporal modulations. In Viemeister (1979) the cutoff frequency was found to be approximately 64 Hz, associated to a time constant of 2.5 ms.

Thee low-pass behavior of the filter was also maintained in the Dau *et al.* (1996a) model, where the last step is given by a first order low-pass filter with cutoff frequency, f_{cut}, of 8 Hz, found to be the optimized parameter to simulate a series of psychoacoustical experiments. The filter operates in the modulation domain, meaning that it reduces the fast transitions within time trajectories of frequency channels contents. Fast modulations are attenuated, because experimental data suggest that they are less important than low modulation frequencies (Dau *et al.*, 1996b) and this is particularly true for speech perception (Drullman *et al.*, 1994a,b).

The attenuation of fast envelope fluctuations in each frequency channel, characterizing the IR of audio signals after the processing of the previous stages, can be seen from the panel on the bottom of Fig. 3.3, where the time trajectories of the frequency channels within the auditory spectrogram get smoothed in time.

The combination of the last two stages can be interpreted as a band-pass transfer function in the modulation domain, i. e. a Modulation

2 The actual relation between input I and output O is $O = \sqrt[2^n]{I}$, where n is the number of adaptation loops. In case of $n = 5$, as it is in Dau *et al.* (1997a), the function approaches a logarithmic behavior.

Transfer Function (MTF): the adaptation loops provides low modulation frequency attenuation whilst the low-pass filter introduces high modulation frequency attenuation. Due to the nonlinearity introduced by the adaptation stage, the MTF of the model is signal dependent, Dau *et al.* (1996a); therefore, a general form of the MTF cannot be found. However, both in Tchorz and Kollmeier (1999) and Kleinschmidt *et al.* (2001), where an adapted version of the Dau *et al.* (1996a) model for ASR was employed, the MTF was derived for a sinusoidally amplitude-modulated tone at 1 kHz. The IR was computed via the auditory model when such a stimulus was provided and the channel with the greatest response, i. e. the one centered in 1 kHz, was extracted as the output. The MTF was then calculated between these two signals.

The result was reproduced in the current work using the same procedure, even though the details about the actual calculation of the MTF were not provided in the referenced studies. Among the different procedure that have been proposed in literature to calculated the MTF (Goldsworthy and Greenberg, 2004), it was chosen to quantify the the modulation depths in the two signals and simply divide them. Such an approach is close to the method proposed in Houtgast and Steeneken (1985). Due to the onset enhancement caused by the adaptation stage, the estimation of the modulation depth on the output signal was performed after the onset response had died out.

The MTF was calculated for three different modulation low-pass cutoff frequencies: 4, 8 and 1000 Hz. As in Tchorz and Kollmeier (1999), a second order filter was used for the cutoff frequency in 4 Hz and a first order one for the remaining two conditions. Figure 3.2 shows the three MTFs. When f_{cut}= 1000 Hz, no attenuation from the low-pass is provided in the low-frequency range of interest. For the other two cases, the transfer function shows a band-pass behavior for the modulation frequencies around 4 Hz, which were found to be very important frequencies for speech perception as pointed out in Drullman *et al.* (1994a,b). In Chapter 6, the role of the MTF band-pass shape in the improvement of ASR experiment scores will be further discussed.

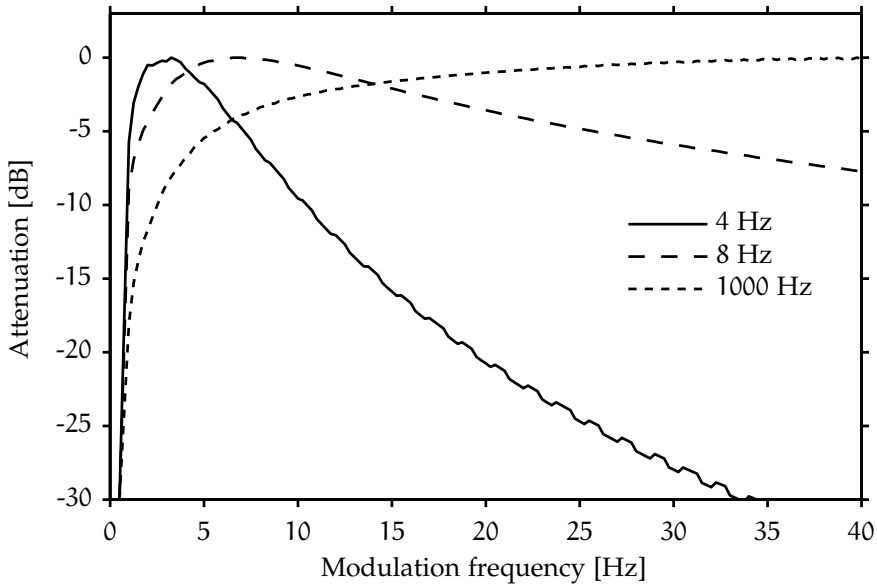

Figure 3.2: Modulation Transfer Function computed with the MLP model between the output of the channel, extracted from the IR, with center frequency of 1 kHz and a sinusoidally amplitude-modulated sinusoid at 1 kHz input. The result for three different modulation low-pass cutoff frequencies are shown (solid, dashed and dotted lines correspond, respectively, to 4, 8 and 1000 Hz).

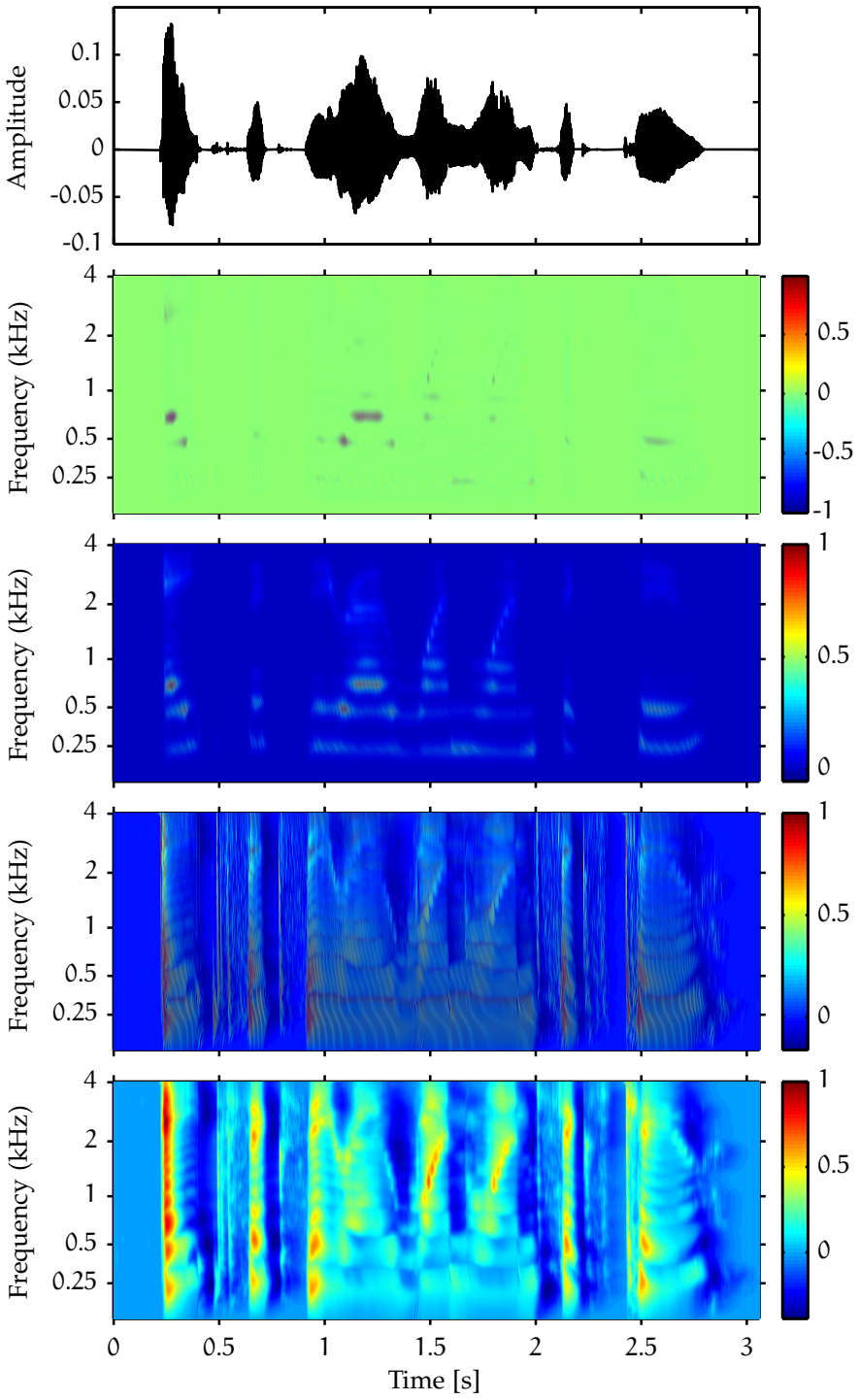

Figure 3.3: MLP model computation of the speech utterance of the digit sequence "8601162". From the top to the bottom: speech signal, output of the GT filtering, output of the IHC transduction, result of the adaptation stage and modulation low-pass filtering ($f_{cut} = 8$ Hz).

3.2 MODULATION FILTERBANK

The experimental framework that the Dau *et al.* (1996a) model was meant to simulate did not regard temporal-modulation related tasks, but other kind of psychoacoustical tasks such as simultaneous and forward masking. Therefore, in order to account for other aspects of the auditory signal processing related to modulations, a bank of modulation filters was introduced in Dau *et al.* (1997a). In this way, tasks such as modulation masking and detection with narrow-band carriers at high center frequencies, which would have not been correctly modeled by the previous approach, can be correctly simulated.

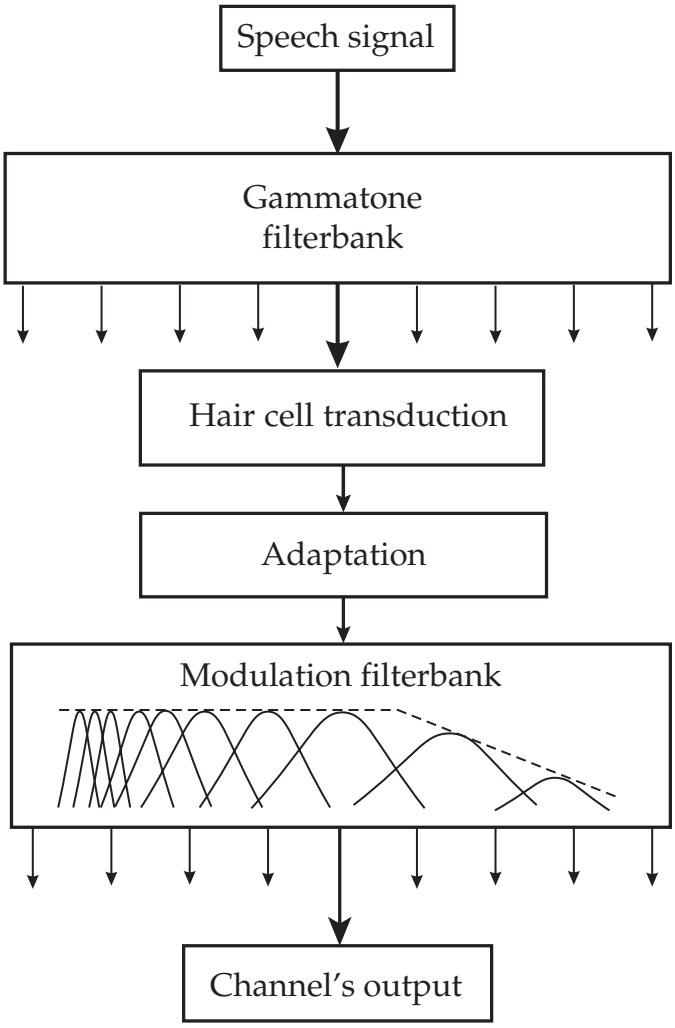

Figure 3.4: Block diagram of the MFB model.

The improvement was performed by substituting the single low-pass modulation filter with a modulation filterbank (formed by the low-pass itself and a series of band-pass filters). The steps to be performed before the modulation domain operations were retained, with some minor modifications, see Dau *et al.* (1996a, 1997a). In this way, the

		DauOCF	DauNCF	FQNCF
Low-pass	Order	2	3	3
	f_{cut} [Hz]	2.5	1	1
Band-pass	Type	Resonant	Resonant	Fixed-Q
	Order	1	1	2
	f_C [Hz]	5	2	2
		10	4	4
		16.67	8	8

Table 3.1: Different modulation filterbanks employed in the current study. In all the three cases the low-pass was a Butterworth filter with the listed characteristics.

updated model both maintains the capabilities of the former version and also succeeds in modeling the results of modulation experiments.

Moreover, evidence that the model behavior can be motivated by neurophysiological studies, mentioned for non-human data from Langner and Schreiner (1988) in Dau *et al.* (1997b), were found in following works for humans subjects in Giraud *et al.* (2000). These findings were provided by functional magnetic resonance images of five normal hearing test subjects, taken while stimuli similar to the ones in Dau *et al.* (1997a) were presented to the listeners. Giraud *et al.*'s study suggests the presence of a hierarchical filterbank distributed along the auditory pathway, composed by different brain regions sensitive to different modulation frequencies (i. e. a distributed spatial sensitivity of the brain regions to modulations).

As in the previous case, the model presented in Dau *et al.* (1997a) was slightly modified to be used in the current work, leaving out the optimal detector stage; an illustration is provided in Fig. 3.4. From now on the original filterbank presented in Dau *et al.* (1997a) will be referred to as DauOCF; Table 3.1 lists the characteristics of the DauOCF while a plot of the filterbank is shown in Fig. 3.5. The output of the MFB model with the first three modulation channels (i. e. the low-pass and the first two band-pass filters of DauOCF in Fig. 3.5) is illustrated in Fig. 3.6. The number of modulation channels, i. e. filters, reflects the number of 2-D auditory spectrogram (i. e. three in this case).

3.2.1 *Alternative filterbanks*

The center frequencies and the shapes of the filters derived in Dau *et al.* (1997a), were chosen to provide good data fitting, as well as a minimal computational load with the framework analyzed in the mentioned study. However, the experiments investigated with the

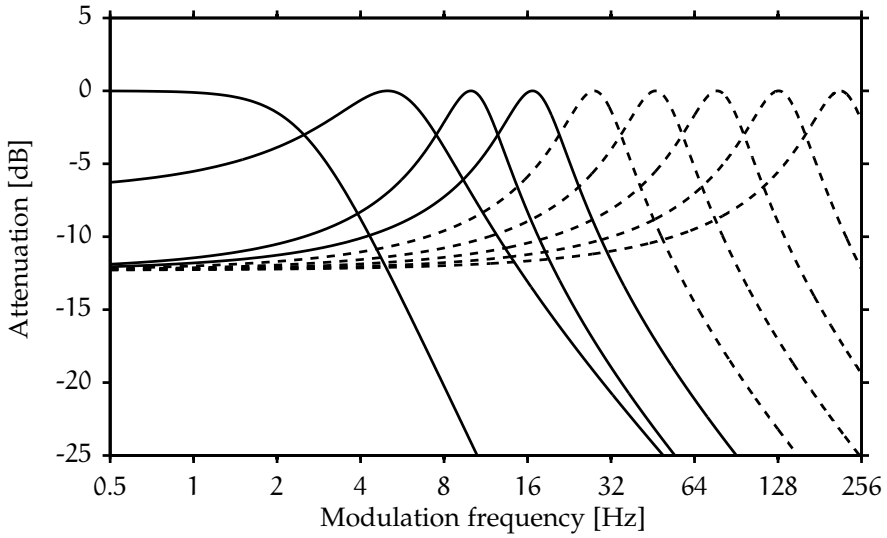

Figure 3.5: Modulation filterbank with the original central frequencies and filter bandwidths derived in Dau *et al.* (1997a). The dashed lines represent the filters of the Dau *et al.* (1997a) filterbank left out from DauOCF (which comprises only the first four filters and it is illustrated with solid lines).

mentioned model were not dealing with speech signals. Studies from perceptual data like Drullman *et al.* (1994a,b) indicated that the modulation frequencies with stronger importance are restricted to a much smaller interval — approximately 1 to 16 Hz — than the one taken into consideration in Dau *et al.* (1997a). Such high modulation frequencies provides cues when performing other kind of tasks but they seem to have only a minor importance in the human speech perception.

Therefore, after using the DauOCF filterbank for the first set of experiments, it was chosen to change it both introducing modifications in the filters' shapes and in the center frequencies to closely inspect the smaller modulation frequency range of interest. The center frequencies have been changed into a new set of values, separated from each other by one octave and listed in Table 3.1, defining the filterbank referred to as DauNCF.

Regarding the new filters' shapes, different strategies have been taken into consideration: instead of the resonant filters from the original model — which do not decay and approach the DC with a constant attenuation — symmetric filters were implemented, motivated by the work in Ewert and Dau (2000). Both Butterworth and fixed-Q bandpass filters were considered.

The digital transfer function of a fixed-Q Infinite Impulse Response (IIR) filter is given by, Oppenheim and Schafer (1975):

$$H_{fQ}(s) = \frac{1-\alpha}{2}\left[\frac{1-z^{-2}}{1-\beta\ (1+\alpha)\ z^{-1}+\alpha\, z^{-2}}\right] \tag{3.2}$$

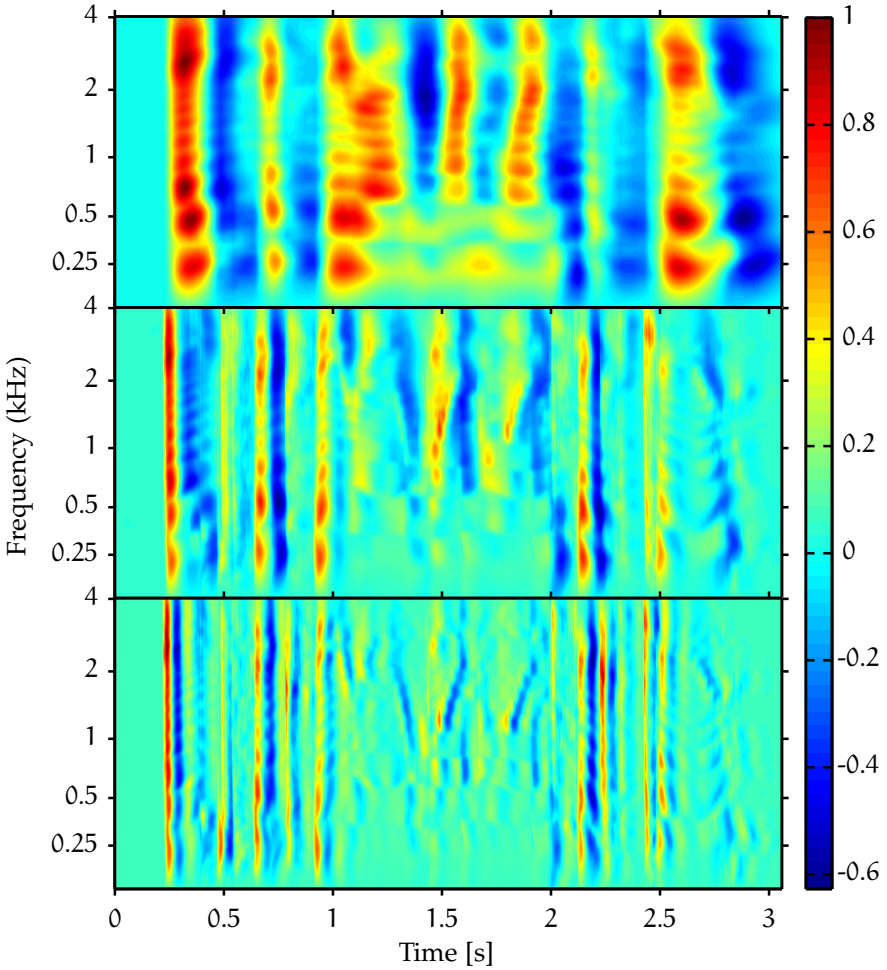

Figure 3.6: Output of the MFB model including the first three channels of the Dau *et al.* (1997a) filterbank for the speech utterance of the digit sequence "8601162". From the top to the bottom the auditory spectrograms refer, respectively, to the filters: low-pass with $f_{cut} =$ 2.5 Hz and resonant band-pass in 5 and 10 Hz.

where α and β are constants linked to bandwidth and center frequency of the filter. The frequency responses of the fixed-Q filterbank (referred to as FQNCF in Table 3.1) are compared with the resonant filters from Dau *et al.* (1997a) with new center frequencies (DauNCF) in Fig. 3.7. The low-pass filter in both the filterbanks had cutoff frequency of 1 Hz. It has been changed from the one in original case, centered at 2 Hz, to reduce the overlapping with the first resonant filter.

Due to problems involving the proper interface between the front-end and the back-end of the ASR system, in a subsequent series of experiments a set of independent 12^{th} order band-pass and low-pass Butterworth filters has been implemented. The processing was therefore carried out using a single filter at the time. Inspired by the work done in Kanedera *et al.* (1999), which proposes a very similar approach, this new set of filters was employed to confirm the evidence about

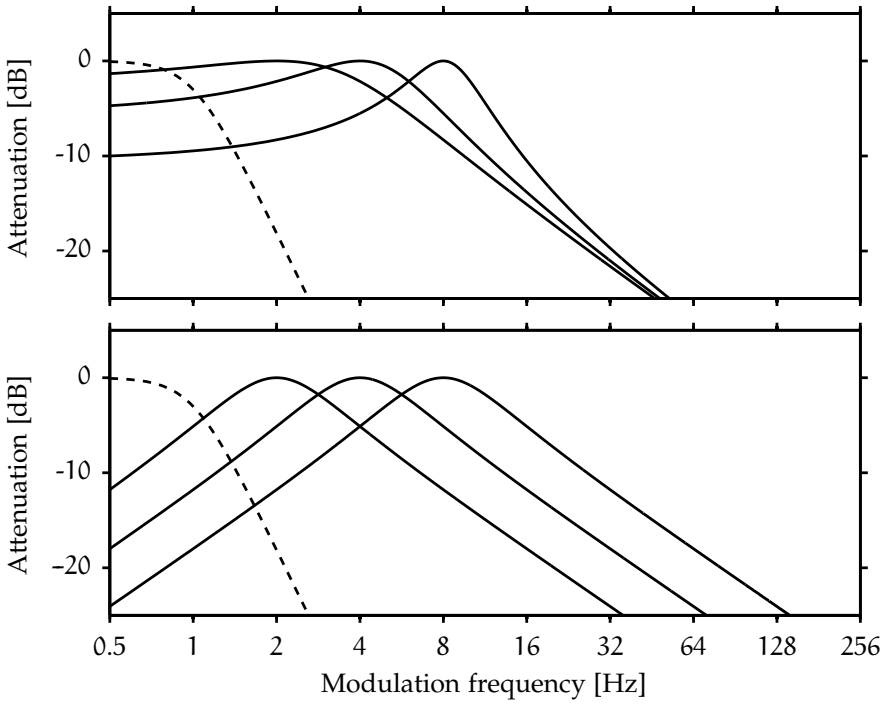

Figure 3.7: Comparison between the frequency responses of the filters from the new filterbanks. On the top panel, the DauNCF filterbank. On the lower panel, the FQNCF filterbank (see Table 3.1). The dashed line represents the third order Butterworth low-pass filter, f_{cut}= 1 Hz used in both the filterbanks.

the importance of low modulation frequencies for speech recognition linked to the perceptual results obtained in Drullman *et al.* (1994a,b). The filters were built from seven frequency values chosen to be related by an octave spacing: 0[3], 1, 2, 4, 8, 16 and 32 Hz. The lower (upper) cutoff frequency[4], defined by $f_{m,l}$ ($f_{m,u}$), were related to each of the seven frequencies by a factor $2^{-\frac{1}{6}}$ ($2^{\frac{1}{6}}$). For instance, the actual cutoff frequencies for the band $[2, 4]$ Hz were $[2 \cdot 2^{-\frac{1}{6}}, 4 \cdot 2^{\frac{1}{6}}]$ Hz. This choice was made in order to have the different filters overlapping at the seven octave spaced frequencies at approximately 0 dB (see Fig. 3.8).

All the permutations of the seven frequencies were used to determine the set of filters of the filterbank (provided that $f_{m,l} < f_{m,u}$). Thus, the total number of filters considered, given the $n_f = 7$ frequencies, was $n_{bins} = n_f \left(n_f - 1\right)/2 = 21$. When the lower cutoff frequency was 0 Hz, low-pass filters were implemented; for all the other combinations of $f_{m,l}$ and $f_{m,u}$, band-pass filters were implemented. Given the spacing between the chosen frequencies, the smallest filters in the considered set were approximately one octave wide while the broadest cutoff frequencies' combinations gave rise to filters with bandwidths

3 0 Hz is not linked to the other values using the octave relation, of course.
4 The cutoff frequencies were defined as the −3 dB attenuation points.

up to five octaves[5]. It was chosen to use Butterworth filters, to get the maximally flat response on the pass-band, even though the roll-off of such filters is not as steep as other kind of implementations, such as Chebyshev or Elliptic filters (Oppenheim and Schafer, 1975). However, a satisfactory compromise on the overlap between adjacent filters was reached at the implemented order with a small increase in the computational need.

In Fig. 3.8 is given an illustration of some of the filters employed (only the narrower of the filterbank, i. e. the ones between two subsequent octave spaced values).

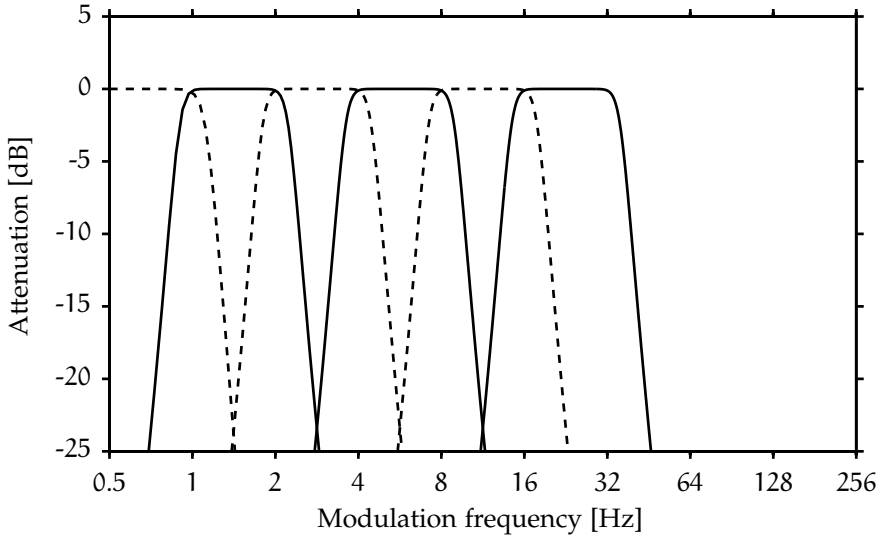

Figure 3.8: Filters from the filterbank employed in the BPE. Only the filters between two subsequent octave spaced values are shown and different line styles were used to distinguish the contiguous ones.

5 The five octaves wide filter has cutoffs $f_{m,l} = 1 \cdot 2^{-\frac{1}{6}}$ Hz and $f_{m,u} = 32 \cdot 2^{\frac{1}{6}}$ Hz. Again, the 0 Hz frequency is not included in this calculation.

4

METHODS

In this chapter, the methods employed to extract the feature vectors from the IRs computed via the auditory models will be presented and a brief introduction will be given of the *corpus*, i. e. the speech material, employed for the recognition experiments describing the kind of utterances, levels, noises and SNRs used throughout the simulations performed.

4.1 AUDITORY-MODELING-BASED FRONT-ENDS

One of the main reasons why auditory models have been employed as front-end for ASR systems is the idea that the speech signal is meant to be processed by the auditory system. It is plausible to argue that human speech has evolved in order to optimally exploit different characteristics of human auditory perception. Thus, it is sensible to develop a feature extraction strategy emulating the stages of the human auditory system (Hermansky, 1998).

Many studies have investigated these possibilities (e. g. Brown *et al.*, 2010; Holmberg *et al.*, 2006; Tchorz and Kollmeier, 1999 among others) and a common conclusion seems to be the increase of noise robustness of an auditory-oriented feature representation. Nevertheless, so far, most of the worldwide feature extraction paradigms for ASR do not employ the state of the art in auditory modeling research. According to Hermansky (1998), there are several reasons for this. Among others:

- the possibility that auditory-like features may not be completely suitable for the statistical models used as back-ends: the fact that they must be decorrelated in order to be fed to an HMMs based model, as it will be described later in this section, could be a limitation to the achievable model accuracy;
- some of the classical feature extraction's methods have been employed for a long time and in most of the cases fine parametrical tunings for given tasks have been developed; poorer scores sometimes obtained with auditory-based methods in certain experiments, could derive from the usage of models not tuned to the particular tasks;
- some of the stages within the different auditory models could be not strongly relevant for the recognition task or their implementation could be somehow unsuitable to represent speech in ASR; the inclusion of such features could, in principle, degradate the results;

- the, often, higher computational power needed to go through the feature encoding process in an auditory based framework compared to the classical strategies.

In the case of auditory-signal-processing-based features, the encoding strategy has some substantially different aspects from the previously discussed *classical* methods. However, some other aspects were implemented considering their counterpart in the MFCC procedure, in order to match the constraints imposed by the HMM-based back-end framework (e. g. the Discrete Cosine Transformation illustrated later).

The first step of the process consists in the computation of the IR of the speech signal using the auditory model. As previously described, the auditory model employed in this study emulates, to a certain extent, the functions of the human auditory system, accounting for different results observed in psychoacoustical tests.

The IR obtained in the last of the steps of the model calculation (shown in Fig. 3.3) is further processed in order to meet some requirements needed for the features to be used from the HTK.

Although the paradigm employed in the two cases is somehow similar, there are some notably differences in the way Modulation Low-Pass (MLP) and Modulation FilterBank (MFB) IRs were processed in the current study.

4.1.1 *Modulation Low-pass*

Two main facts have to be accounted for, in order to convert the feature vectors in a format suitable to be processed by the HTK.

Due to the high time and frequency resolutions of the IRs (respectively in the order of 10^4 and 10^2 samples in the considered work), a reduction in the number of samples for both the domains has to be performed. The reason of this data compaction is mainly due to computational power problems as well as poorer models generalization that would arise from high resolution IRs (Hermansky, 1998),

Additionally, the usage of overlapping filters within the auditory filterbank, returns correlated inter-channel frequency information (i. e. correlated features). Correlation is a property to be avoided for the features used in HMM-based ASR systems whether diagonal covariance matrixes are employed (see Section 2.2). In order to solve both the mentioned problems, two signal processing steps are implemented:

A. filtering and downsampling via averaging of overlapping windows was used to reduce the time resolution;

B. downsampling in the frequency domain and decorrelation were both achieved via Discrete Cosine Transform (DCT).

The reduction in the time resolution was simply performed by averaging the IRs within overlapping frames of the same dimensions

of the ones considered in the MFCC method: 25 ms long windows overlapping for the 40% (i. e. 10 ms). This operation decreases the sampling frequency to 100 Hz (the inverse of the overlap time) after low-pass filtering the IR by means of a moving average filter. The choice of the two parameters (as well as the averaging procedure) was the same performed in other studies (e. g. Brown *et al.*, 2010; Holmberg *et al.*, 2007; Jankowski *et al.*, 1995). Although the rather slow roll-off of the moving average filter, some mild attenuation is introduced on the low-frequency region considered in the different experiments performed in the current study (32 Hz being the higher limit in the BPE, see Section 3.2) but it can be considered to be negligible.

The remaining issues were solved employing the DCT operation. As mentioned in Appendix A.2, the DCT is an approximation of Principal Components Analysis (PCA); therefore, its computation on the IR returns a set of pseudo-uncorrelated time varying feature vectors. Due to the energy compaction introduced by the DCT (Khayam, 2003), the number of coefficients than can be used to describe the feature vectors is much smaller than in the frequency representation obtained with the auditory model. As for the MFCCs, 14 coefficients were retained excluding the energy terms. Additionally, 14 first- and second-order dynamic coefficients were calculated and appended to the feature vectors using an approach similar to the ones adopted in other studies (e. g. Holmberg *et al.*, 2007; Brown *et al.*, 2010).

The role of the DCT in auditory-features-based ASR systems has been investigated in a set of simulations where no transformation was applied. The accuracy results have been compared with the ones obtained when the DCT was correctly performed, as shown in Section 5.1.

Figure 4.1 illustrates the ASR-oriented processing of a speech signal, showing the IR computed via the MLP (middle panel) and the sequence of feature vectors after DCT-based decorrelation (bottom panel). It can be noticed how great part of the energy of each frame is concentrated at the beginning of the three segments of the feature vectors (i. e. coefficients 1, 14 and 28). Moreover, the temporal structure of the IR is somehow maintained, showing peaks in correspondence of the words onsets. Figure 4.2 illustrates the decorrelation property of the DCT. The correlation matrixes computed on an IR obtained with the MLP model before (top panel) and after (bottom panel) this operation show that in the second case high values are concentrated in a narrow area around the diagonal (i. e. less correlated variables).

A final clarification is needed about the IRs onsets. Due to the discussed properties of the adaptation stage, the model enhances the onsets of the speech signal. In case of utterances corrupted by noise (which is applied from the very beginning to the very end of the correspondent clean utterances), an onset emphatization is performed at the beginning of the IR due to the noise. To exclude this corrupted part of the model computation, for a first set of simulations of the

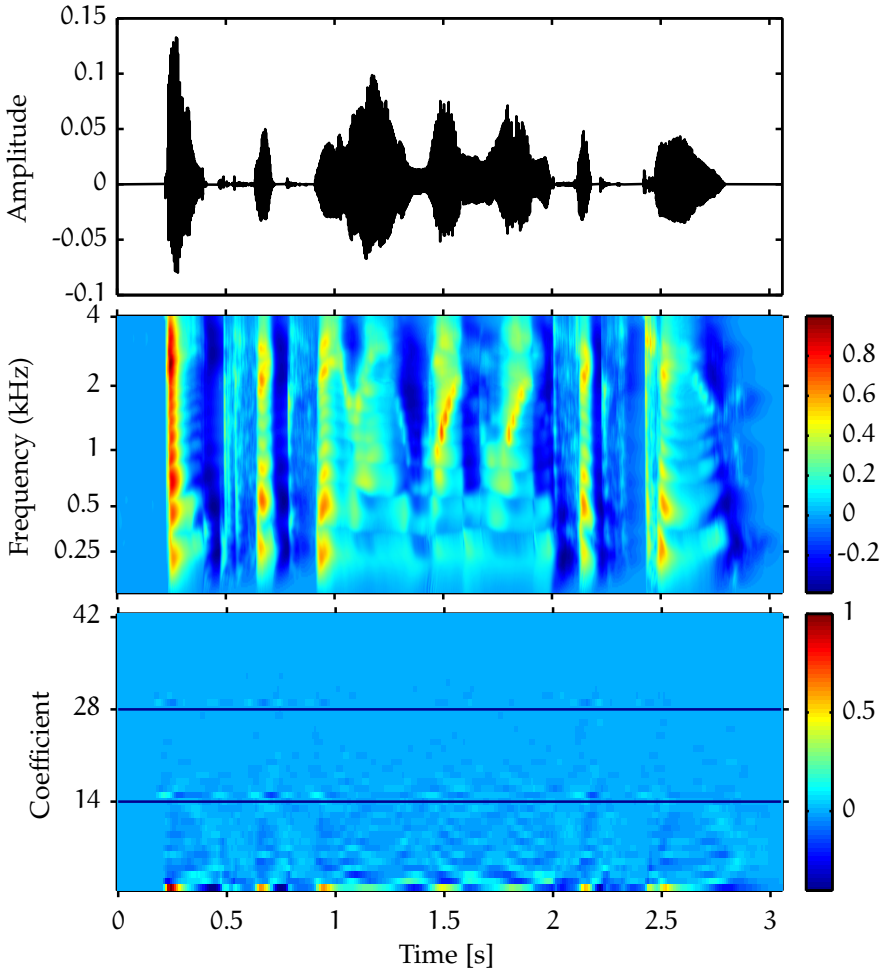

Figure 4.1: Feature extraction from a speech signal processed using the MLP model. The speech utterance is given by the digit sequence "8601162", as in Fig. 3.3. From the top to the bottom: speech signal, output of the MLP model ($f_{cut} = 8$ Hz) and features vectors extracted from the IR.

current work the initial 150 ms of the IRs were left out. However, the removal of the noise was shown to have a negligible effect on the results[1]. Thus, in subsequent simulations the onsets were simply left untouched in the encoded features vectors.

4.1.2 *Modulation Filterbank*

The process of features encoding from IRs computed using the MFB model introduced a more challenging problem. Essentially, providing additional information about the modulation domain is reflected in a

1 The reason of this could arise from the fact that in most of the cases the adaptation to the noise was achieved before the actual start of the spoken digits within the utterance (placed on average after 200 ms from the beginning of the recorded signals).

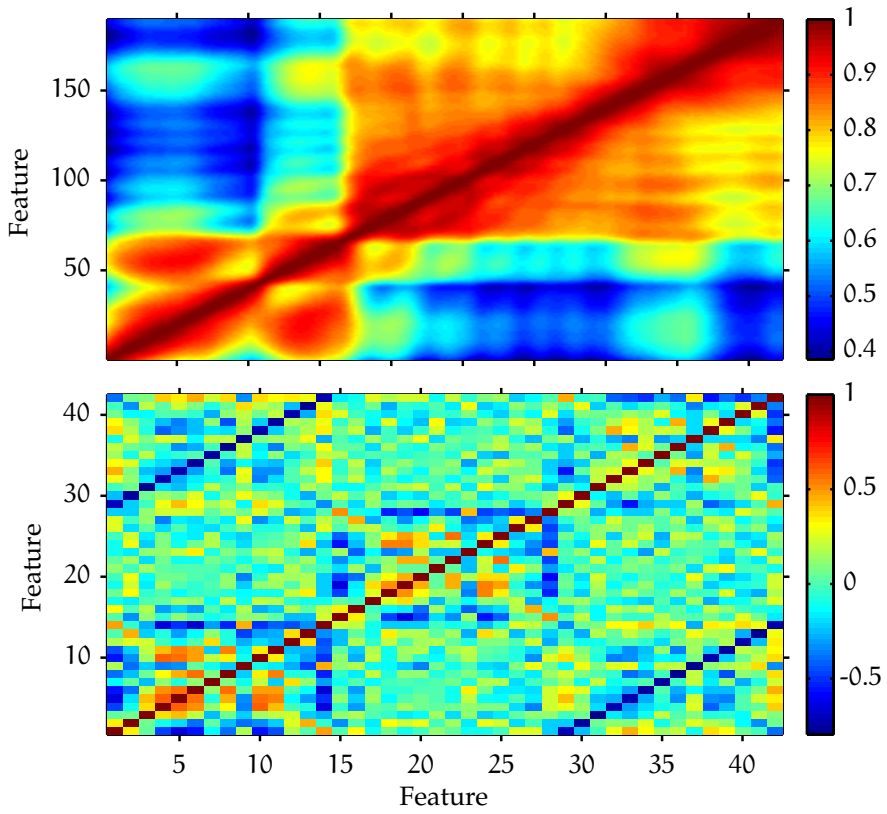

Figure 4.2: Correlation matrix of an IR obtained with the MLP model before (top) and after (bottom) DCT. The responses shown in the middle and bottom panel of Fig. 4.1 were used, respectively. The concentration of higher amplitude values along the diagonal reflects the fact that the features composing the transformed IR are less correlated than the samples of the untouched IR (which is strongly correlated as it can be seen by the off-diagonal high amplitude values on the top figure).

dimensionality increase of the IR, as shown in Fig. 3.6; in such a case the output varies with time, frequency and modulation frequency.

As for the MLP features encoding, a downsampling operation in the time domain can be performed to reduce the resolution of the time samples. However, the second step employed in the previous encoding strategy cannot be blindly applied. The problem arises for two main reasons:

1. like the filters in the auditory frequency domain, the filters composing the modulation filterbank are partially overlapped, thus introducing additional correlation;

2. a method successfully decorrelating the features in both the frequency domains, would anyway return a three dimensional signal which is not suitable to be analyzed by the statistical back-end chosen for the current work.

Different approaches have been tried to perform features encoding of MFB-derived IRs; however, the problem has not been completely solved. In a first attempt, it was chosen to simply apply the DCT singularly on the different channels and subsequently merge the information from the separate channels into a single vector. This encoding approach, succeeds in decorrelating the information within the single auditory frequency channels, but it does not take into consideration the modulation frequency domain. Because of this, the correlation problem is not solved. The top panel of Fig. 4.3 illustrates the content of the feature vectors from two different time frames extracted from the feature representation of a DCT-transformed IR (shown in the bottom panel of Fig. 3.6). The three channels (separated using the dashed lines) show rather similar trends between each other for both the observations. The middle and bottom panels of Fig. 4.3 show, respectively, the cross-correlation between the first channel of the two feature vectors and the cross-correlation between the two entire feature vectors. While in the first case the decorrelation is achieved, to a certain extent, in the second case a rather strong correlation is retained at lags corresponding to the integer multiple of channel's coefficients[2]. Thus, by placing multiple-channel information within single vectors, the correlation is reintroduced at the DCT-representation level. For these reasons, no simulation were carried out with such features. In the attempt of developing a method satisfying the HTK constraints, a different encoding strategy was considered.

In a second approach, referred to as *Method 1* (M1)[3], the decorrelation in both the frequency domains (auditory and modulation) was performed via a 2-D DCT applied on each time frame. However, the situation is very similar to the one previously discussed because correlation is reintroduced once the features from different channels are compacted together. Thus, the decorrelation seem not to be achieved via 2-D DCT. The problem could be due to the very limited number of modulation channels for which a redistribution of the energy in a more compacted form is not achiavable. Nevertheless, M1 has been employed to encode MFB features in some of the simulations (being aware of its limitations).

A third approach, referred to as *Method 2* (M2), was lastly implemented. A 2-D DCT was applied as in M1. As far as it concern the vectors encoding, it was chosen to compress the modulation frequency dimension along time, i. e. the 3-D IR represented as a matrix of size $T \times N \times M$ — with T time frames, N frequency samples and M modulation frequency channels — was resized as a new 2-D matrix of size $(T \times M) \times N$. Essentially, the result can be seen as the 2-D matrix obtained with M1 where, for a fixed time frame, the frequency

2 In the example at the lags $l = 42\,k$, $k \in [-2, 2]$.

3 The number notation of the methods only refers to the procedures actually employed in the simulations. Since the first encoding approach was not tested, it was not associated to a "method name".

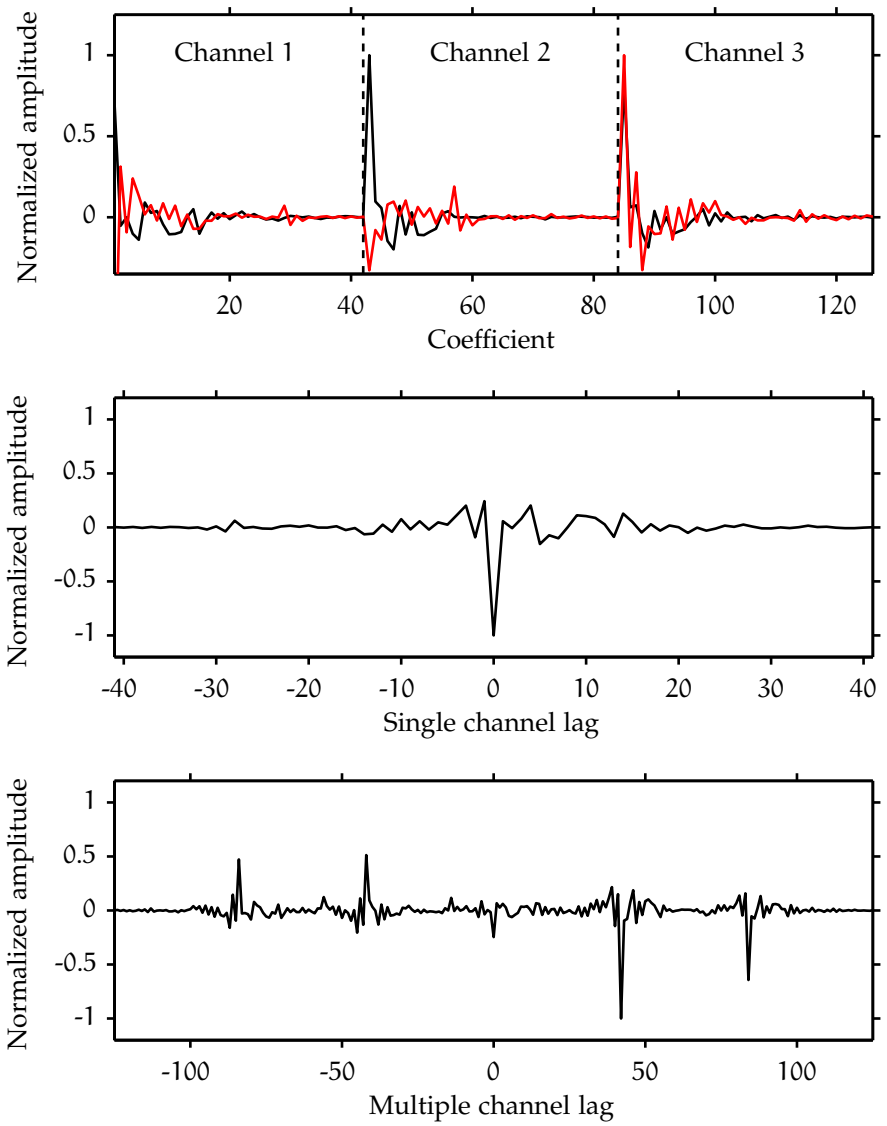

Figure 4.3: Top: content of the feature vectors from two different time frames of a DCT-transformed multi-channel IR. Middle: cross-correlation between the coefficients of the first channel of the two feature vectors. Bottom: cross correlation between the two feature vectors.

information of two modulation channels $m_{j,k'}$ and $m_{j,k''}$ ($j = 1 \dots N$, $k' \neq k''$) are placed one after the other in the time domain.

Although this encoding paradigm seemed to be suitable at first, it was subsequently observed that the use of such representation could lead to problems in the model characterization. The different nature of adjacent time frames in this approach (as they derive from different channels), should not be problematic for the HMM-based recognizer which assumes independence between adjacent observations. However, the application of a derivative-like operation on such features could no longer be suited due to the discontinuities between adjacent frames.

Figures 4.4 and 4.5 (top panels) show the result of the features encoding methods M1 and M2. The difference between the encoding methods can be noticed by comparing the number of frequency and time samples. In the proposed simulations $M = 3$ modulation channels are considered; therefore, the output of M1 consists of three sets of 42 coefficients per channel (i. e. a total of $42 \times 3 = 126$ features per frame), whilst the output of M2 is only made by 42 coefficients but a triple number of time frames. One can also notice the much more discontinuous fine structure of the second representation mentioned earlier. A measure of the degree of decorrelation introduced by the DCT in the two methods is given by the correlation matrix, see Appendix A.3, illustrated in the lower panels of Figs. 4.4 and 4.5.

Although both the methods have been used to perform some experiments, the feature correlation problems encountered in the encoding process of MFB features suggested that the back-end employed for this study, i. e. the HTK, was not completely suitable for the ideas to be investigated. Regarding the current study, it was decided to move to a different kind of experiment relying on the computation of single-channel IRs treatable in the same way as the MLP-derived IRs. Anyway, other approaches to properly encode 3-D IRs — involving multi-streams models (e. g. Zhao and Morgan, 2008; Zhao *et al.*, 2009) for used defined features, which HTK seems to only support partly — that could be employed are briefly discussed in Chapter 6.

4.2 SPEECH MATERIAL

In ASR certain kinds of speech materials or corpora (singular corpus) are used to train and test the recognizing system. Several different corpora have been developed and used in the field of ASR. There exist a number of aspects distinguishing corpora from one another, see Harrington (2010). Modern ASR systems are still quite dependent on the particular task they were built for. Therefore, the choice of the corpus should be made carefully, considering the kind of experiment one is working on. The structure of the speech material is one of the key parameters characterizing a speech corpus; amongst others, in ASR, one can distinguish corpora based on:

- syllables
- isolated words or digits
- sequence of words digits
- sentences

Some other constraints that can be used to tune the different ASR systems are, for instance, represented by:

- finite alphabet (e. g. only some categories of words are present in the corpus)

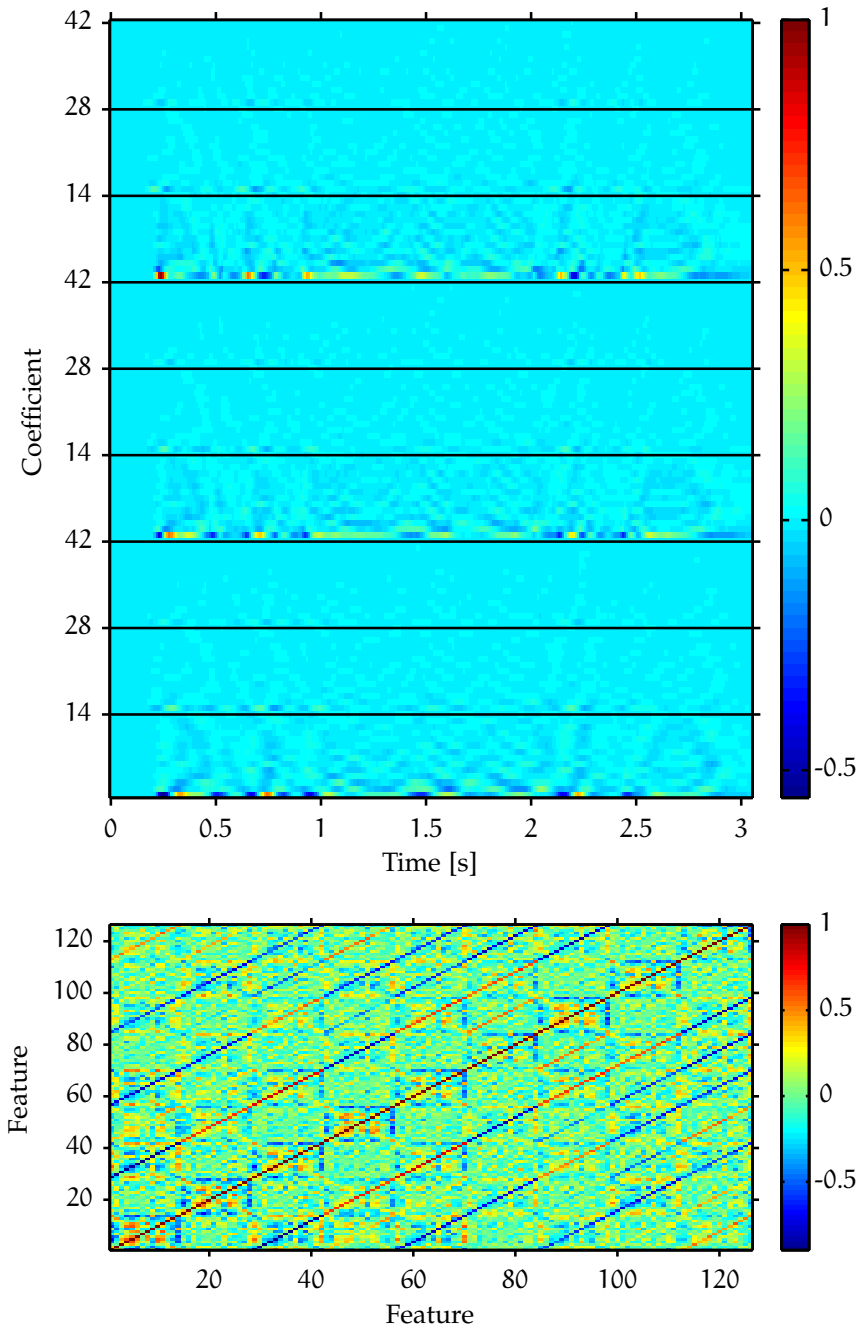

Figure 4.4: Top: feature extraction from a speech signal processed using the MFB model and M1. The speech utterance is "8601162", as in Fig. 3.3. The 3 modulation channels correspond to the first three filters of the Dau *et al.* (1997a) filterbank, i. e. a low-pass with $f_{cut} = 2.5$ Hz and resonant band-pass in 5 and 10 Hz. Bottom: correlation matrix of the encoded file, showing the strong correlations (given by the lines parallel to the diagonal) between features.

- defined grammar (e. g. the presence of a silence before and after each spoken word)

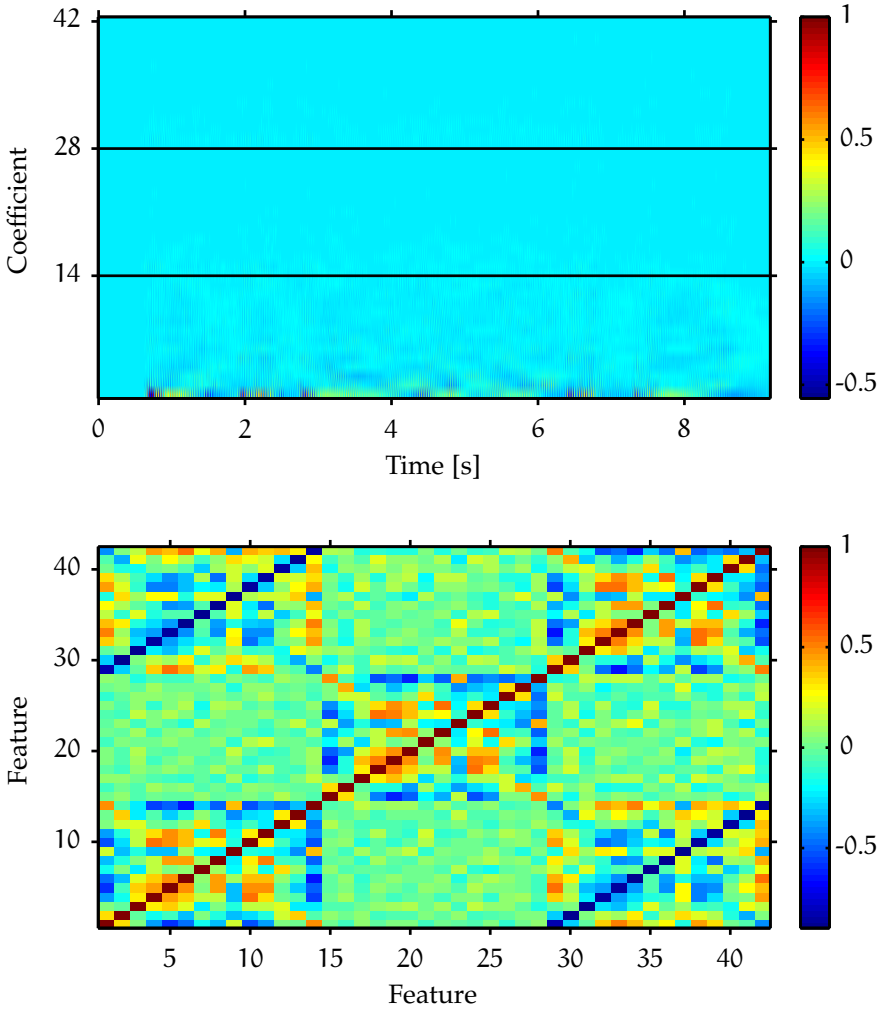

Figure 4.5: feature extraction from a speech signal processed using the MFB model and M2. The speech utterance is "8601162", as in Fig. 3.3. The correlation between features is lowered compared to the case in Fig. 4.4 and resembles more the structure obtained for the file converted using the HTK tool, HCopy, see Fig. 2.3.

In the current work, a simple digit-based corpus has been employed. The choice has been done to match the speech materials used in other works investigating ASR performance with auditory-based feature extraction techniques (i. e. Brown *et al.*, 2010; Holmberg *et al.*, 2007; Tchorz and Kollmeier, 1999) thus allowing a kind of comparison of the results.

4.2.1 AURORA 2.0

The corpus employed in the current study is the AURORA 2.0 Corpus, described in Pearce and Hirsch (2000). The corpus was designed in order to evaluate the performance of different speech recognition

algorithms both in clean and in noisy conditions. The clean speech samples were extracted from another well known corpus in the field of ASR: the *TIdigits* Corpus (Leonard, 1984) which consists of utterances, spoken by male and female US-American speakers, of isolated digits ("oh", "zero" and "one" to "nine") and sequences up to seven digits. An important feature of the corpus is its speaker-independency, since 52 male and 52 female speakers were employed to record the speech material. Therefore, an improvement in the recognition accuracy from a given features encoding strategy would suggest the robustness of the method to extract speaker-independent characteristics of the speech signals. The noisy speech was created by artificially adding the distortions given by 8 different real scenario noises: *suburban train*, *babble*, *car*, *exhibition hall*, *restaurant*, *street*, *airport* and *train station*. An additional signal-processing step introduced in the corpus, consists in the distortion of the signals by two standard frequency characteristics (G.712 and MIRS), mimicking realistic telecommunication-channel transfer functions (Pearce and Hirsch, 2000). For each of the noises, there exist seven SNRs conditions: clean speech (i. e. ∞ dB SNR) and from 20 to -5 dB SNR in 5 dB steps. The minimum SNR included in the corpus is somehow higher than average minimum SNRs used in perceptual tests, such as speech intelligibility tests, because lower levels will typically lead to extremely poor results Lippmann (1997); Sroka and Braida (2005).

Only two out of four parts of the database were employed in the current study: the *train* folder and the *testa* folder. The *train* folder was used for both clean and multi condition training, the latter referring to a training procedure for the statistical model performed on both clean and noisy speech. In the first case, 8440 utterances of the *clean* folder were used; in the second case, a total of 8440 utterances were summed up from 422 utterances for 20 different conditions: five SNRs (clean, 20, 15, 10 and 5 dB) for each of four noise types (subway, babble, car and exhibition), allowing to train the model using multi condition data. The *testa* folder consists of 28 conditions: all the seven SNRs for each of four noise types (subway, babble, car and exhibition) for a total 28028 utterances (1001 per condition).

It can be noticed that the same noise types are used both in training and testing the model (this is referred to as matched condition). Moreover, the same frequency characteristics (G.712) is employed for both the sets, see Pearce and Hirsch (2000). ASR with mismatched conditions (both regarding noise types and frequency characteristics) can be performed using the speech material in the folders *testb* and *testc* of the database, but these two parts were not used in the current work.

Two main modifications have been applied to the speech corpus: a level correction and the inclusion of an additional noise condition (white noise).

4.2.2 *White noise addition*

An additional noise condition was introduced in the *testa* set, employing white noise. This was done since white noise has been used in several psychoacoustics studies (e. g. French and Steinberg, 1945; Hawkins and Stevens, 1950; Festen and Plomp, 1990) as well as in different ASR works (e. g. Jankowski *et al.*, 1995) and represents a strong masker (e. g. French and Steinberg, 1945; Hawkins and Stevens, 1950), thus allowing to test the noise robustness of the feature extraction paradigm employed in the ASR task.

The noise addition was performed to the clean speech utterances of the *subway* noise subset using the same signal-processing steps employed by the AURORA 2.0 Corpus authors, by means of the freely available *C*-language script provided in the AURORA project website[4]. Therefore, both the noise addition and the application of the convolutional distortion (G.712) were carried out in the same way as in the original study.

The white noise sample was a 70 s long uniformly random distributed signal with a sampling frequency of 8 kHz and it was created using the MATLAB function `rand`.

4.2.3 *Level correction*

Another modification performed on the Corpus was related to the dB RMS levels of the speech data. The Root Mean Square (RMS) values were computed in dB SPL considering a finite signal, sampled at a uniform rate and using the convention stating that a pure tone at 100 dB SPL has an RMS value of 1.

The AURORA 2.0 Corpus dB RMS levels' distribution approaches a Gaussian, as it can be seen in Figure 4.6, and the mean is 92.7 dB RMS. In order to work with a level closer to the average spoken speech level (ANSI-S3.5, 1997) and to reflect levels used in other ASR studies (e. g. Brown *et al.*, 2010; Holmberg *et al.*, 2007) the mean level was shifted down to 65 dB RMS.

4 http://aurora.hsnr.de/download.html

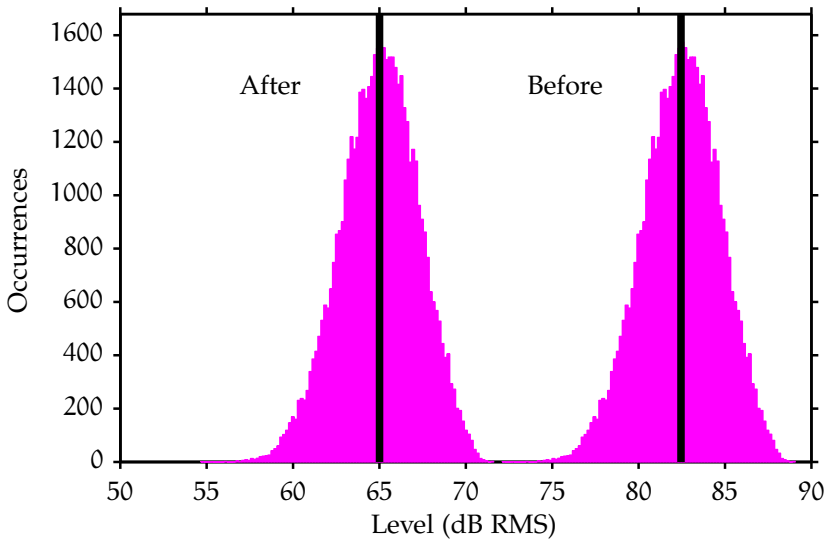

Figure 4.6: AURORA 2.0 Corpus level distribution.

5 RESULTS

Two different sets of experiments will be described in this chapter. The results of several feature extraction procedures involving both the variation of the Modulation Low-Pass (MLP) model cutoff frequency f_{cut} in the last stage of the auditory model and the introduction of the Modulation FilterBank (MFB) model with varying number of filters and encoding strategies are presented in the first part. In the second part, the results of an experiment inspired by the Kanedera *et al.* (1999) study are discussed. All the results presented in these sections were computed using the utterances from the AURORA 2.0 corpus with mean level shifted to 65 dB RMS and for all but the initial tested condition the training was performed both with clean and noisy speech. Moreover, no variations of the back-end were involved, i. e. no parameters of the HTK have been varied or fitted throughout all the tests.

5.1 STANDARD EXPERIMENT RESULTS

In order to compute the results of the experiments presented in this first part, the initial portions of the Internal Representations (IRs) have been left out due to the noisy onsets, see Section 4.2.1. For simplicity, the features sets will be referred to with the same name of the model used to encode them, e. g. MLP features indicates features extracted from the IRs obtained with the MLP model. Throughout all the experiments, a value of ∞ for the SNR indicates the clean speech testing condition.

5.1.1 *MLP and MFCC features in clean training conditions*

As a first experiment, the results obtained for the baseline (i. e. the MFCC features) were compared to the ones obtained using features encoded with the MLP model with the optimal cutoff frequency in the modulation stage found in Dau *et al.* (1996a), i. e. $f_{cut} = 8$ Hz. Five noise conditions were available for testing. In Fig. 5.1 are shown the recognition accuracies for the baseline and for the MLP model, respectively on the left and right panel.

A first thing to notice is the trend of the recognition accuracy as a function of the SNR. A decrease in the accuracy with the SNR is what expected in ASR experiments (e. g. Tchorz and Kollmeier, 1999; Holmberg *et al.*, 2006, 2007) and what will be observed throughout all

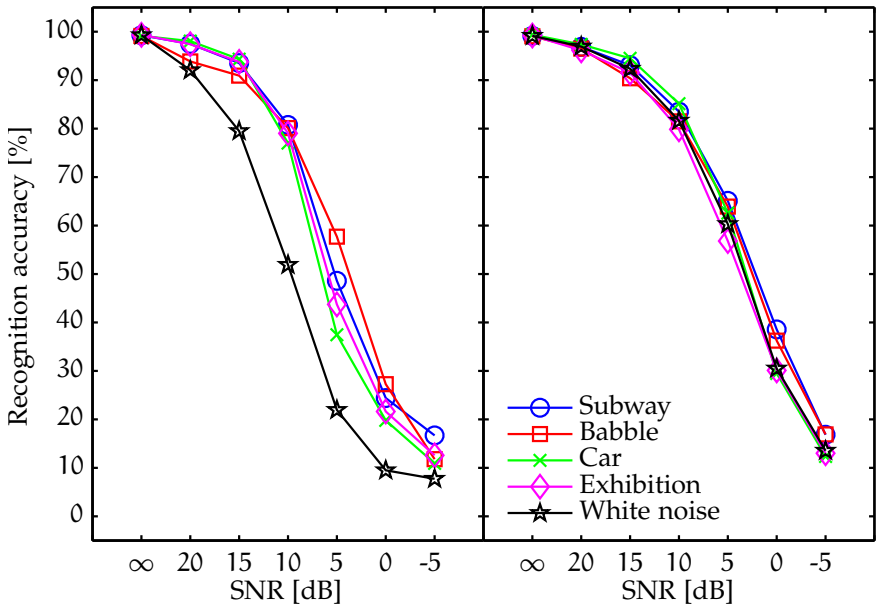

Figure 5.1: Clean condition training. Left panel: recognition accuracies obtained with MFCC features. Right panel: recognition accuracies obtained with the MLP model using $f_{cut} = 8$ Hz. In both the cases, five noise conditions were tested: *suburban train* (○), *babble* (□), *car* (×), *exhibition hall* (◇) and *white noise* (☆).

the following experiments[1]. It can be easily seen that the results varies in a smaller range for the MLP-based case as far as it concerns the four AURORA 2.0 noises. The white noise condition produces the greatest difference: the results are comparable in the case of the auditory model based processing whilst they get dramatically worse (up to approximately 30% at 10 and 5 dB SNR) when computed with the MFCC baseline. Due to the similarities of the results of the MLP features for the different noises, they can be characterized using an inter-noise mean value for each SNR. As mentioned in Section 4.1.1, the features from the IRs obtained with the MLP model were augmented with their velocities and accelerations. As noticed in other works (Tchorz and Kollmeier, 1999; Holmberg *et al.*, 2006) the auditory-modeling-based features provide a more noise-robust representation, thus leading to better results, especially at the mid-level SNRs of the corpus (i. e. 20 to 0 dB RMS).

5.1.2 *MLP and MFCC features in multi-conditions training*

In Fig. 5.2 it is shown the comparison between the same two conditions presented earlier (MFCC and MLP features) using acoustic models trained in multi-condition. The multi-condition training was intro-

1 In multi-condition training, small improvements were sometimes observed for models tested in mild SNR conditions, compared to tests in clean-condition.

duced to investigate the effect of this model adaptation on the MLP features. Besides the general improvements for MFCC features, expected considering results from other studies (among others Pearce and Hirsch, 2000 or Holmberg *et al.*, 2007), it can be seen that the similarities of inter-noise results for the baseline are extended to lower SNRs (from 10 to 5 dB) and that for the white noise condition the difference is narrowed down but still relevant (from 10% up to 35% at low SNRs).

Regarding the MLP results, an improvement can be noticed at all the considered SNRs, even though at the lowest values (0 and -5 dB) some inter-noise variability is introduced compared to the clean training case. It can be seen that the improvement introduced from the multi-condition training is smaller than for the MFCC features. In Fig. 5.3 the comparison of the accuracies for MFCC and MLP features averaged across noises, for clean-trained (left panel) and multi-condition-trained (right panel) models are shown, indicating the smaller gap between the two methods when training in multi-condition. This is somehow different from the more constant improvement in the two cases presented in Holmberg *et al.* (2007).

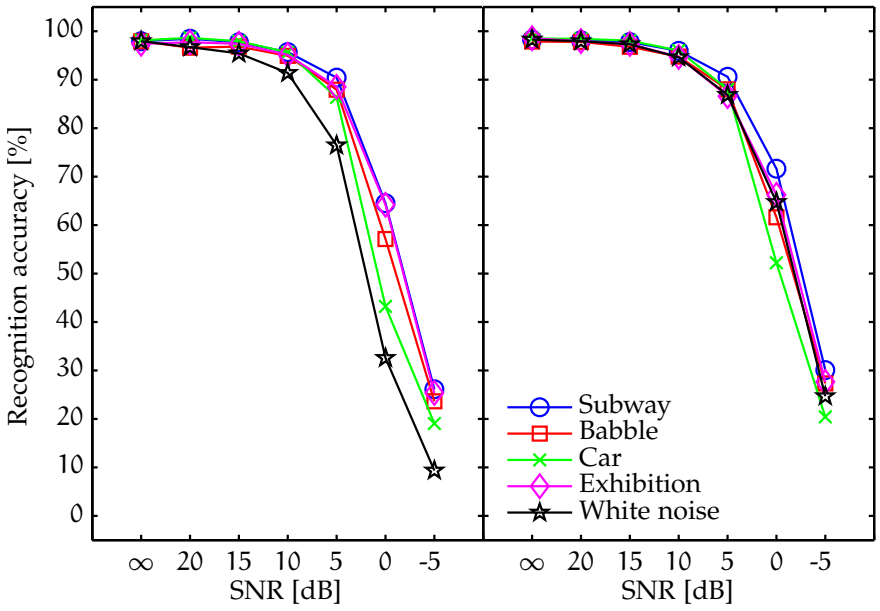

Figure 5.2: Multi-condition training. Left panel: recognition accuracies obtained with MFCC features. Right panel: recognition accuracies obtained with MLP features using $f_{cut} = 8$ Hz. In both the cases, 5 noise conditions were tested: *suburban train* (○), *babble* (□), *car* (×), *exhibition hall* (◇) and *white noise* (☆).

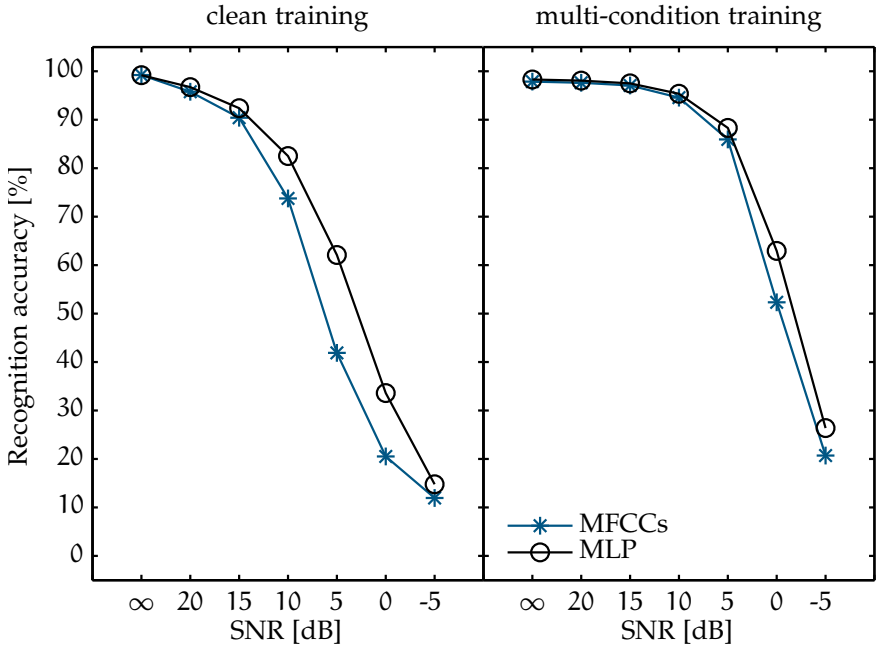

Figure 5.3: Left panel: recognition accuracies averaged across the five noises for each SNR value, obtained for clean-trained HMMs with MFCC (✳) and MLP (◯) features. Right panel: same as left panel with the acoustic models trained in multi-condition.

5.1.3 *MLP features encoded with and without performing the DCT*

The role of the DCT has been investigated by comparing results obtained with the auditory model (fixing $f_{cut} = 8$ Hz), in the cases where the mentioned transformation was either applied or not. Such a condition was considered to examine the importance of mapping the data to a different domain in order to fulfill the back-end's constraints, mainly the requirement of diagonal covariance matrixes (even though this process has a poor physical representation, Batlle *et al.*, 1998; Nadeu *et al.*, 1995).

The left panel of Fig. 5.4 shows that applying the DCT increases the accuracy scores (up to 15%), thus providing a better acoustics model for the different words. The two curves show approximately the same accuracies when tested in clean speech, but the scores degradates more rapidly with decreasing SNR when no transformation is applied, as hypothesized in Tchorz and Kollmeier (1999).

In the right panel of Fig. 5.4, the differences in the results with and without computing the DCT are shown for HMMs trained in multi-condition, exhibiting a similar trend of the ones obtained for clean-trained models although the gap between the two curves is narrower.

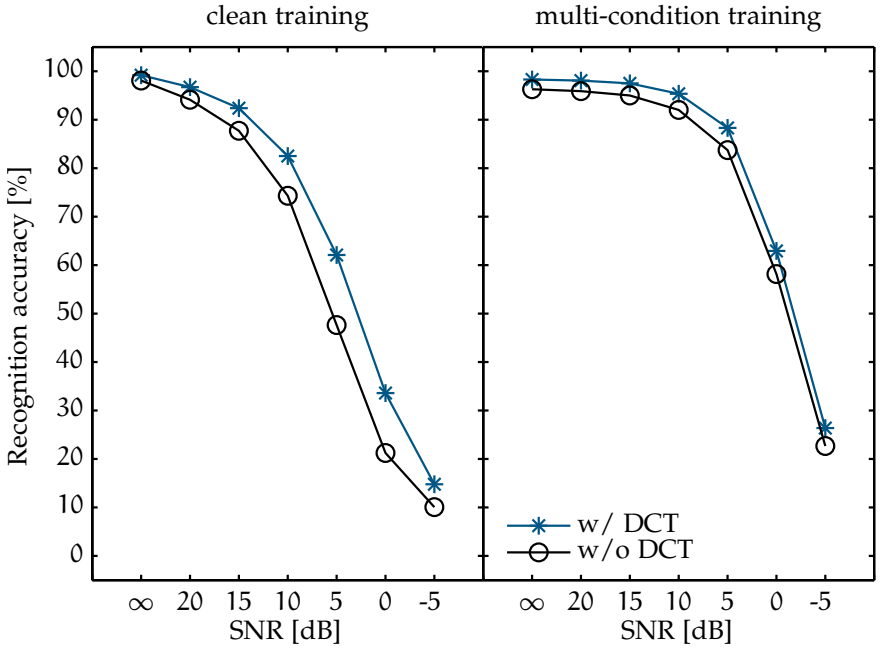

Figure 5.4: Left panel: recognition accuracies obtained for clean-trained HMMs with MLP features obtained with and without performing the DCT operation (○ and ✳, respectively). Right panel: same as left panel with the acoustic models trained in multi-condition. The results are averaged across noise conditions.

5.1.4 *MLP features with different cutoff frequencies*

Subsequently, the effect of changes in the cutoff frequency of the MLP model were investigated by comparing accuracy results obtained with MLP features with f_{cut} taking the values: 2.5, 4, 8 and 20 Hz. The experiment was carried out to determine the effect of the suppression of high modulation frequencies from the IR of a speech signal, wich should provide additional improvements (Tchorz and Kollmeier, 1999; Hermansky and Morgan, 1994).

The results for clean- and multi-condition-trained models are illustrated in Fig. 5.5 (left and right panel, respectively) where it can be seen that the deviation between the four cases is almost unnoticeable (all the differences in accuracy are below 3%), suggesting that the information important to perform recognition is retained in all the conditions. In Tchorz and Kollmeier (1999), a change in the cutoff was performed between 4 and 8 Hz, causing a greater difference than the one obtained in this study. However, in the mentioned study the order of the filter was changed too.

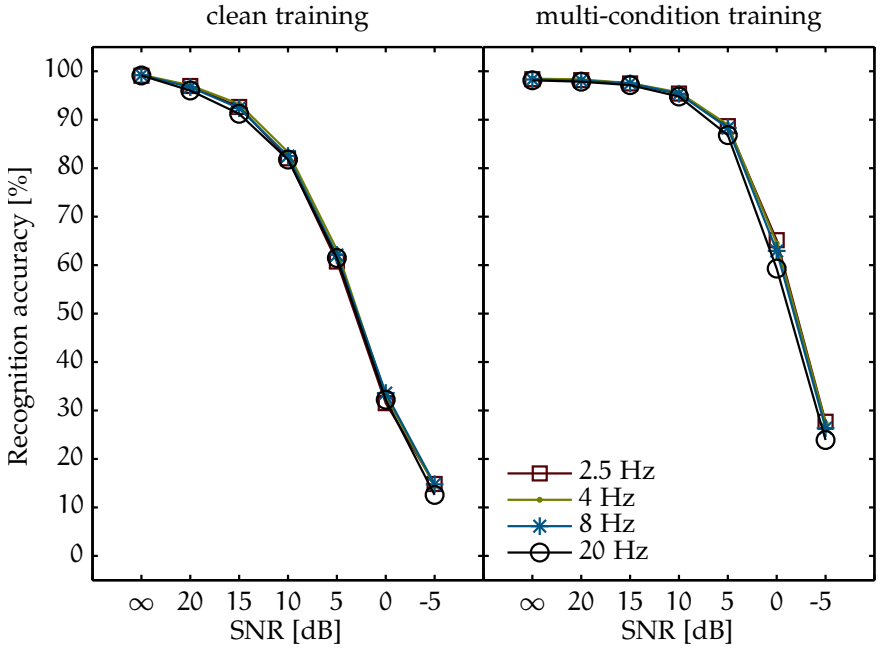

Figure 5.5: Left panel: recognition accuracies obtained for clean-trained HMMs with MLP features obtained for different cutoff frequencies: 2.5, 4, 8 and 20 Hz (□, •, ✻ and ○, respectively). Right panel: same as left panel with the acoustic models trained in multi-condition. The results are averaged across noise conditions.

5.1.5 *MLP features with different filter orders*

Due to the apparent invariance shown in the results with different cutoff frequencies, it was chosen to encode sets of features where the order of the modulation low-pass filter was changed. This was done so as to show how influential the removal of information in the modulation domain is for the results. This simulation was carried out in a later stage of the study and the cutoff value was not chosen to be part of the set of cutoffs frequencies used in the other experiment with MLP features. The cutoff frequency was fixed to 2 Hz since it was expected to reflect greater degradation in the accuracies as the filter's order was increased (from two to twelve in steps of two). Figure 5.6 shows how the results strongly deteriorate as the order gets higher.

5.1.6 *MLP and MFCC features with and without dynamic coefficients*

Figure 5.7 shows the comparison between the results for MLP and MFCC features obtained with and without appending the dynamic coefficients. The modulation low-pass filter of the MLP model was chosen to be $f_{cut} = 4$ Hz. The comparison was performed in order to assess the importance of dynamic coefficients — known to provide increased robustness with classical encoding methods such as MFCC

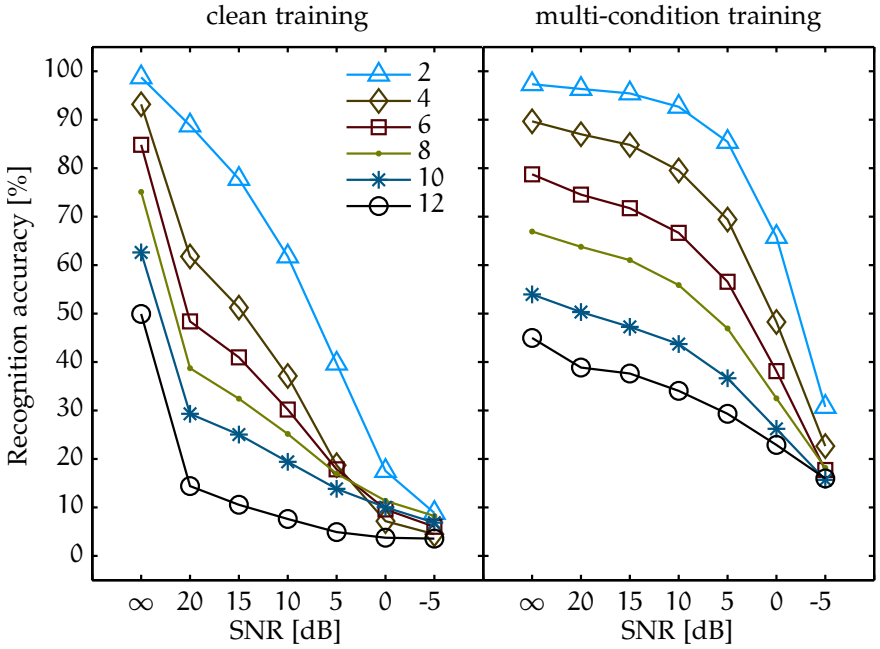

Figure 5.6: Left panel: recognition accuracies obtained for clean-trained HMMs with MLP features. The cutoff was fixed to 2 Hz and the filter's order was set to 2, 4, 6, 8, 10 and 12 (△, ◇, □, •, ✻ and ○, respectively). Right panel: same as left panel with the acoustic models trained in multi-condition. The results are averaged across noise conditions.

features (Furui, 1986) — when using auditory-signal-processing-based features. It can be seen that the introduction of the dynamic coefficients for MFCC features causes the expected great improvements, whereas it only induces a mild (though noticeable) improvement of the results for the MLP features. The outcome of such an experiment supports the idea that the additional processing performed when calculating the dynamic coefficients somehow resembles part of the processing within the MLP model, thus causing a smaller improvement when deltas and accelerations are used to augment MLP features compared to when they are combined with MFCC features.

For MFCC features the increase ranges approximately from 5% up to almost 40% in clean-condition training and from 5% to 15% in multi-condition training. For MLP features the improvement is more contained, ranging approximately from 2% to 5% for the clean-trained models and from 1% to 8% for multi-condition-trained models.

5.1.7 *MFB features with different numbers of filters*

The first set of simulations performed after the replacement of the modulation low-pass filter stage were carried out introducing of the modulation filterbank presented in Dau *et al.* (1997a), i. e. DauOCF.

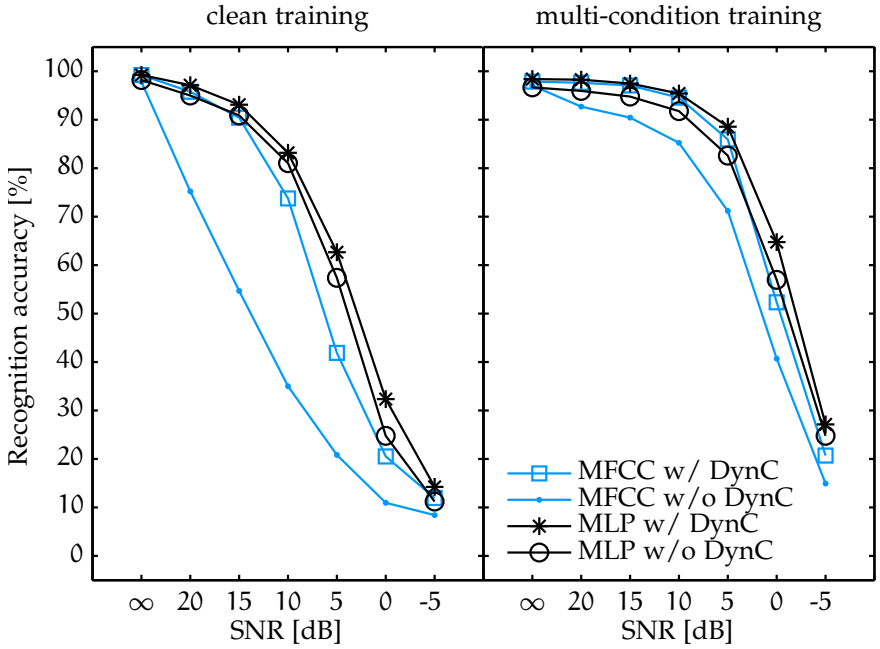

Figure 5.7: Left panel: recognition accuracies obtained for clean-trained HMMs with MFCC features including (□) and not including (•) dynamic coefficients and MLP features including (⁎) and not including (○) dynamic coefficients. The low-pass modulation filter in the MLP model had cutoff $f_{cut} = 4$ Hz. Right panel: same as left panel with the acoustic models trained in multi-condition. The results are averaged across noise conditions.

It was chosen to encode the features using an increasing number of filters to see the differences in recognition accuracy given by the progressive introduction of other channels information. The features were encoded from the IRs obtained with one up to four modulation filters, where the first case represents the single low-pass filtering at 2.5 Hz (from the MLP model) and the last case represents the encoding of the first four channels using the Dau *et al.* (1997a) modulation filterbank (i. e. low-pass fitler and band-pass filters with center frequencies in 5, 10 and 16.67 Hz, see Table 3.1). The results are shown in Fig. 5.8 for both clean and multi-condition training. Almost no changes are introduced using different numbers of filters. Regarding the features encoding procedure, M1 (described in Section 4.1.2) was employed. Like for the comparison between MLP features obtained using different cutoff frequencies, the similarity between the results is somehow unexpected considering that a greater amount of information is introduced. Some of the causes that might be responsible for this results invariance are discussed in the following chapter.

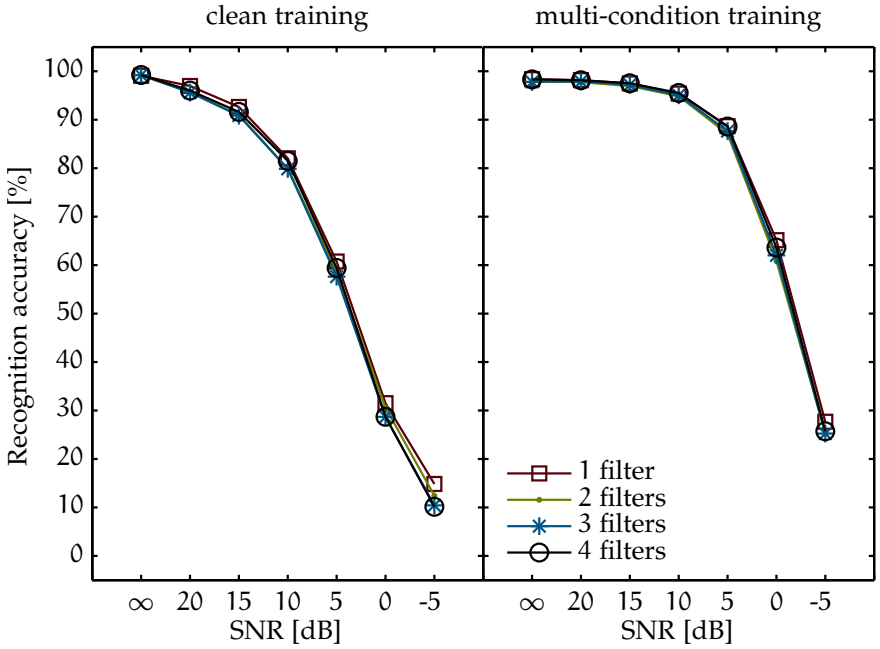

Figure 5.8: Left panel: recognition accuracies obtained for clean-trained HMMs with features encoded using the MFB model, Dau *et al.* (1997a) filterbank and the approach M1 with a number of filters ranging from 1 to 4 (□, •, ⁎ and ○ symbols, respectively). Right panel: same as left panel with the acoustic models trained in multi-condition.

5.1.8 *MFB features with different center frequencies and encoding methods*

The last two conditions investigated changes that have been applied on both the filterbank and on the encoding methods. In a first experiment the results for MFB features obtained using the filterbank with resonant filters and new center frequencies, i. e. DauNCF, were compared with the results for MFB features obtained using the original filterbank from Dau *et al.* (1997a), i. e. DauOCF. Only the first two filters of the filterbanks were employed. Also the encoding methods were different in the two cases: M1 (i. e. multiple channels coefficients stacked in the same feature vector for each time frame) was employed for the features encoded with the DauOCF and M2 (i. e. multiple channels coefficients stacked in separate time frames) was used for the features encoded with the DauNCF. The introduction of M2 was performed to investigate how changes in the encoding methods would have affected the accuracy results. The change of the filterbank was performed to give more importance to lower modulation-frequency regions than with the original Dau *et al.* (1997a) filterbank, keeping in mind the perceptual results from Drullman *et al.* (1994a,b). The results are illustrated in Figure 5.9, showing that no difference is basically introduced by the described modification other than a small difference for the testing condition in clean speech for multi-condition-trained models

where the two curves are separated by approximately 5%. The gap gets smaller at 20 dB SNR causing the recognition accuracy to increase with worse SNR conditions.

In a second tested case, features encoded using the DauNCF filterbank were compared with features encoded using the FQNCF (i. e. a Butterworth low-pass and fixed-Q band-pass filters). The features were encoded using M2 even though M1 returned slightly better results, since it was the role of the filter shape, more than the absolute score that was investigated. Such a change should point out the differences in recognition accuracy due to the use of a filterbank where the band-pass filter does not retain the low-modulation-frequency energy in the same way as the resonant filters in the Dau *et al.* (1997a) filterbank (constant at low frequencies), but strongly attenuates the DC. Figure 5.10 shows that the changes are very mild but the results are worse with the new filterbank. Also in this case, an increase in the accuracy is notice between test in clean speech and in noisy speech at 20 dB SNR.

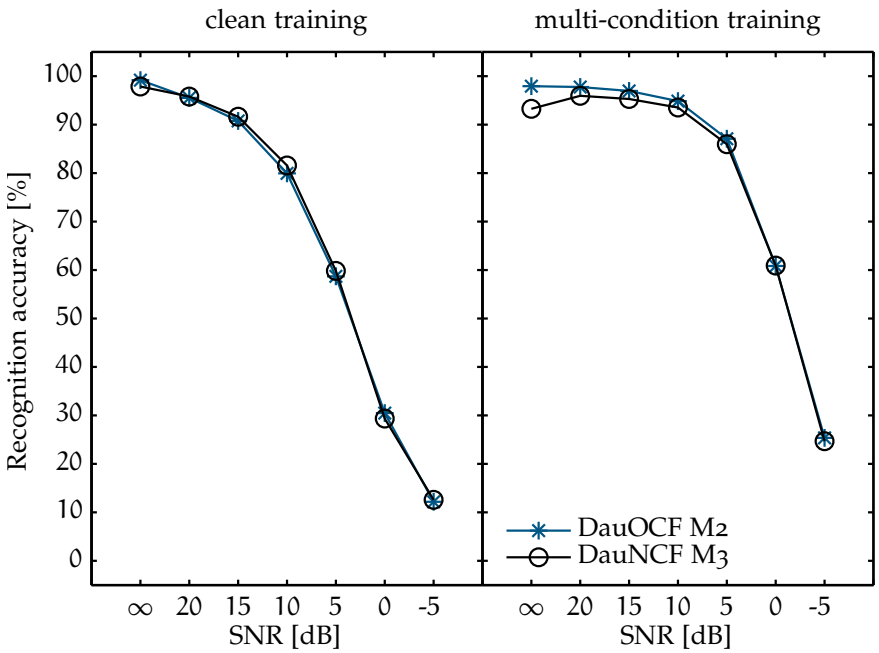

Figure 5.9: Left panel: recognition accuracies obtained for clean-trained HMMs with two different setups. The curve marked with ✻ refers to features encoded using the MFB model, DauOCF filterbank and the approach M1. The curve marked with ○ refers to features encoded using the MFB model, DauNCF filterbank and the approach M2. Right panel: same as left panel with the acoustic models trained in multi-condition.

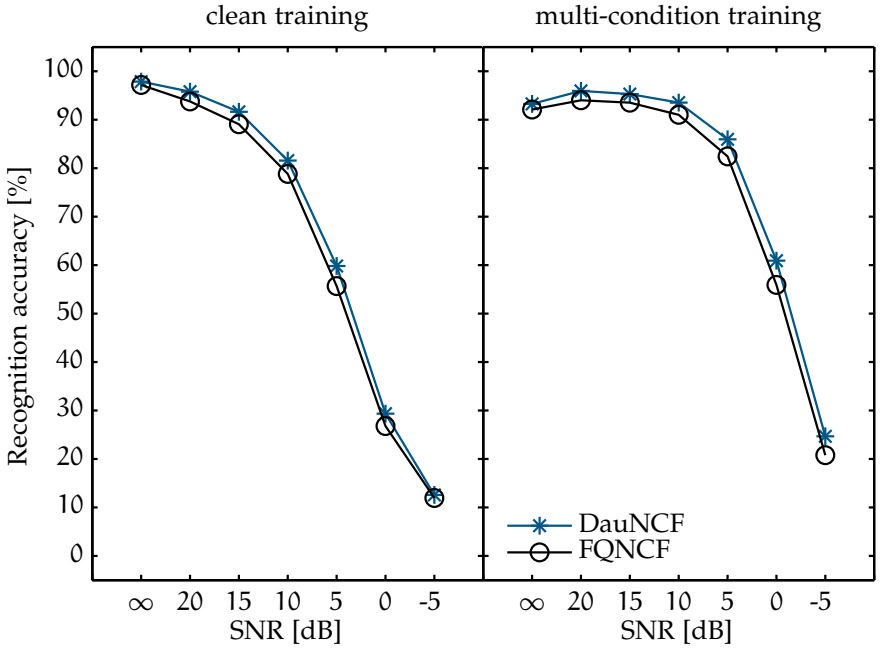

Figure 5.10: Left panel: recognition accuracies obtained for clean-trained HMMs with features encoded using the MFB model, two filters from the DauNCF filterbank and two filters from the FQNCF filterbank (✻ and ○ symbols, respectively). The features have been encoded using the approach M2 in both cases. Right panel: same as left panel with the acoustic models trained in multi-condition.

5.2 BAND PASS EXPERIMENT RESULTS

The results obtained in the Band Pass Experiment (BPE) are reported in this section. Figure 5.11 shows an example of the results obtained from this set of simulations. Only a single SNR condition is shown, 20 dB SNR, trained in clean condition. It was chosen for comparison, since it is the closest condition to the one tested in Kanedera *et al.* (1999), where an SNR of 23.7 dB was considered. Moreover, this particular condition returns a strong non-monotonic behavior for the recognition accuracy as the band of the filters gets narrower (see e. g. the peaks for the bands $[1, 4]$ Hz and $[2, 8]$ Hz), which will be motivated later.

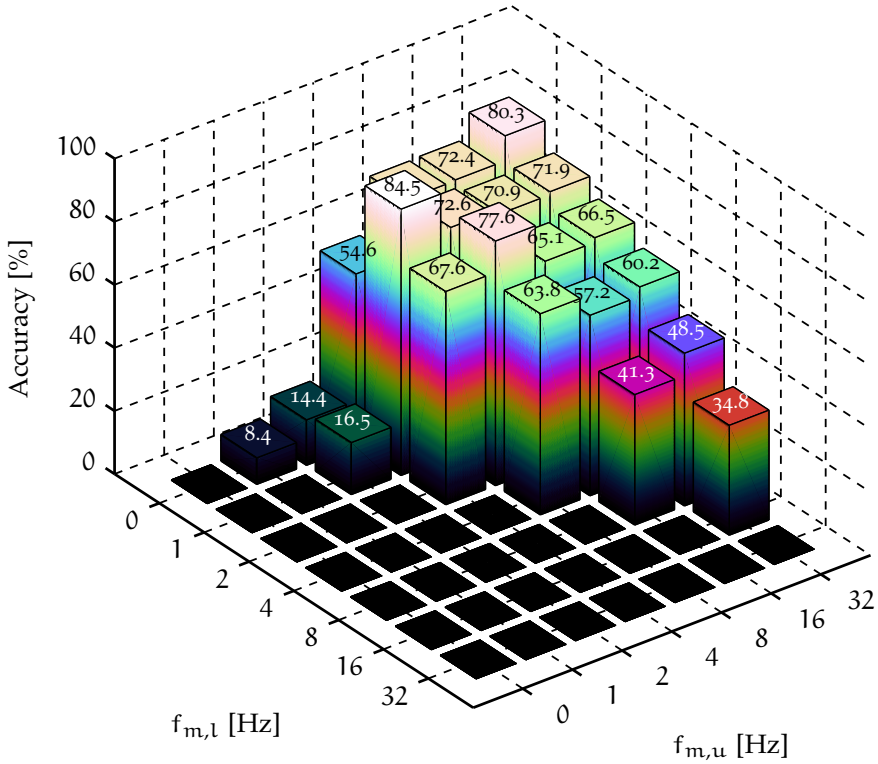

Figure 5.11: Recognition accuracies of the BPE for clean-trained HMMs and an SNR value of 20 dB, averaged across all the noise conditions.

The BPE has been carried out for all the noise conditions and for both clean- and multi-condition training. It was preferred to plot the whole set of results in a different way than Fig. 5.11, to simplify the understanding of the illustrations. For such a purpose, the results were chosen to be shown by progressively fixing one of the two variables ($f_{m,l}$ and $f_{m,u}$, respectively), thus allowing only the other cutoff frequency to be varied (i. e. $f_{m,u}$ and $f_{m,l}$, respectively).

Figures 5.12 and 5.13 show the results obtained varying $f_{m,u}$ for six fixed values of $f_{m,l}$ (0, 1, 2, 4, 8 and 16 Hz). The mentioned behavior regarding Fig. 5.11, can be seen as a non-monotonic increase of the recognition accuracy. For instance, in the curve obtained fixing $f_{m,l} = 1$ Hz (denoted by the $\diamond$ symbol) a maximum is present for $f_{m,u} = 4$ Hz in most of the left panels of Figs. 5.12 and 5.13.

On the other hand, Figs. 5.14 and 5.15 show the results obtained varying $f_{m,l}$ for six fixed values of $f_{m,u}$ (1, 2, 4, 8, 16 and 32 Hz). In this case the non-monotonic decrease can be seen when fixing $f_{m,u} = 4$ Hz (denoted by the $\square$ symbol) which has a maximum for $f_{m,l} = 1$ Hz in most of the left panels of Figs. 5.14 and 5.15, i. e. the same conditions described for the other figures.

It is noticed that such effect appears only for the clean-trained models cases and it is more pronounced at higher SNRs, e. g. panels *C* and *E* of Figs. 5.12 and 5.14. For the multi-condition training case, the behavior seen in clean training is not observed anymore. The accuracy simply gets higher as more information from the modulation domain is retained, giving the highest results for the widest considered band (the low-pass filter with cutoff 32 Hz).

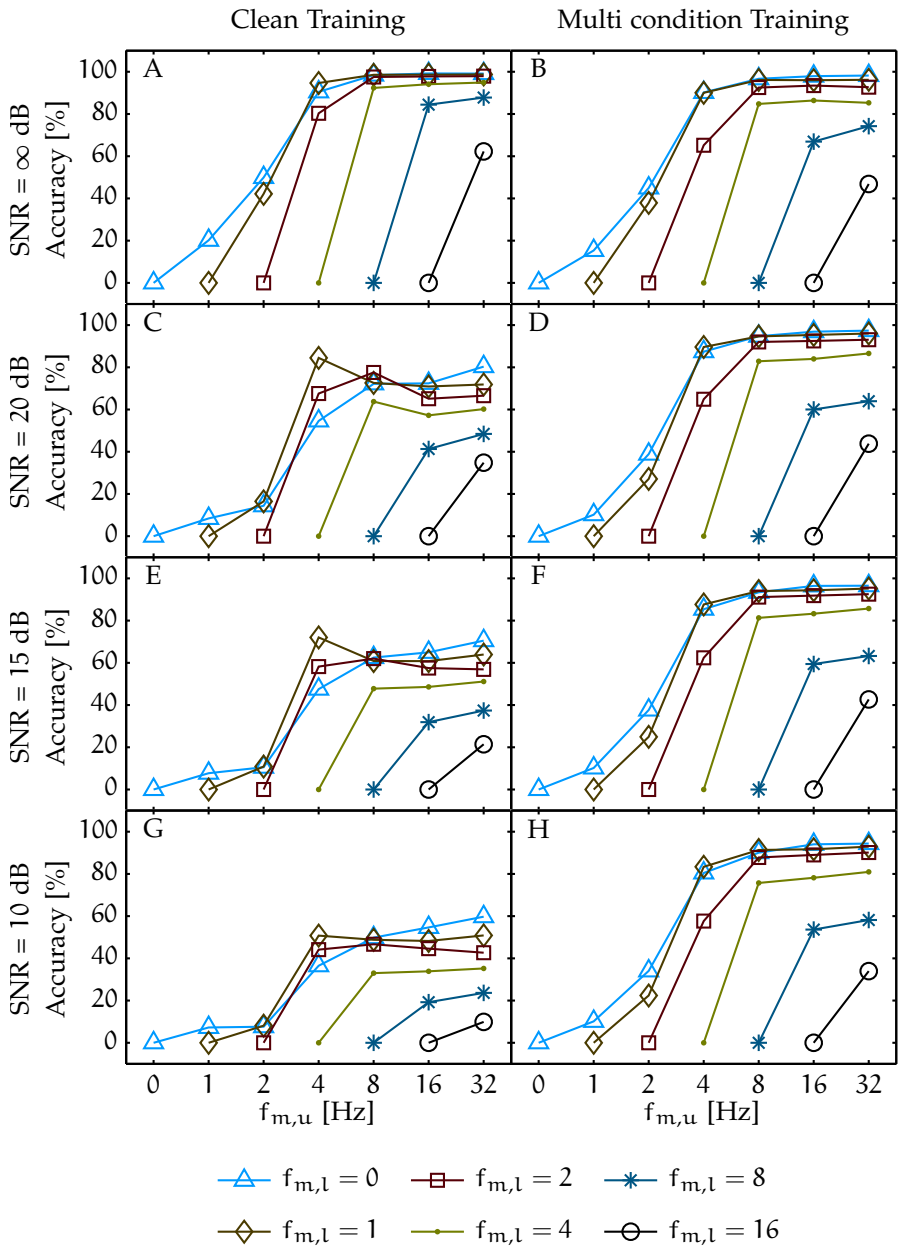

Figure 5.12: Recognition accuracies of the BPE as a function of the higher cutoff frequencies ($f_{m,u}$) parameterized in the lower cutoff frequency ($f_{m,l}$). Four conditions (from the top to the bottom, clean speech, SNR of 20, 15 and 10 dB) tested with HMMs trained in clean- (left panels) and multi-conditions (right panels) are considered. The values of the parameter are $f_{m,l} \in \{0, 1, 2, 4, 8, 16\}$, corresponding to the symbols △, ◇, □, •, ✻ and ○, respectively. The scores are averaged across noises.

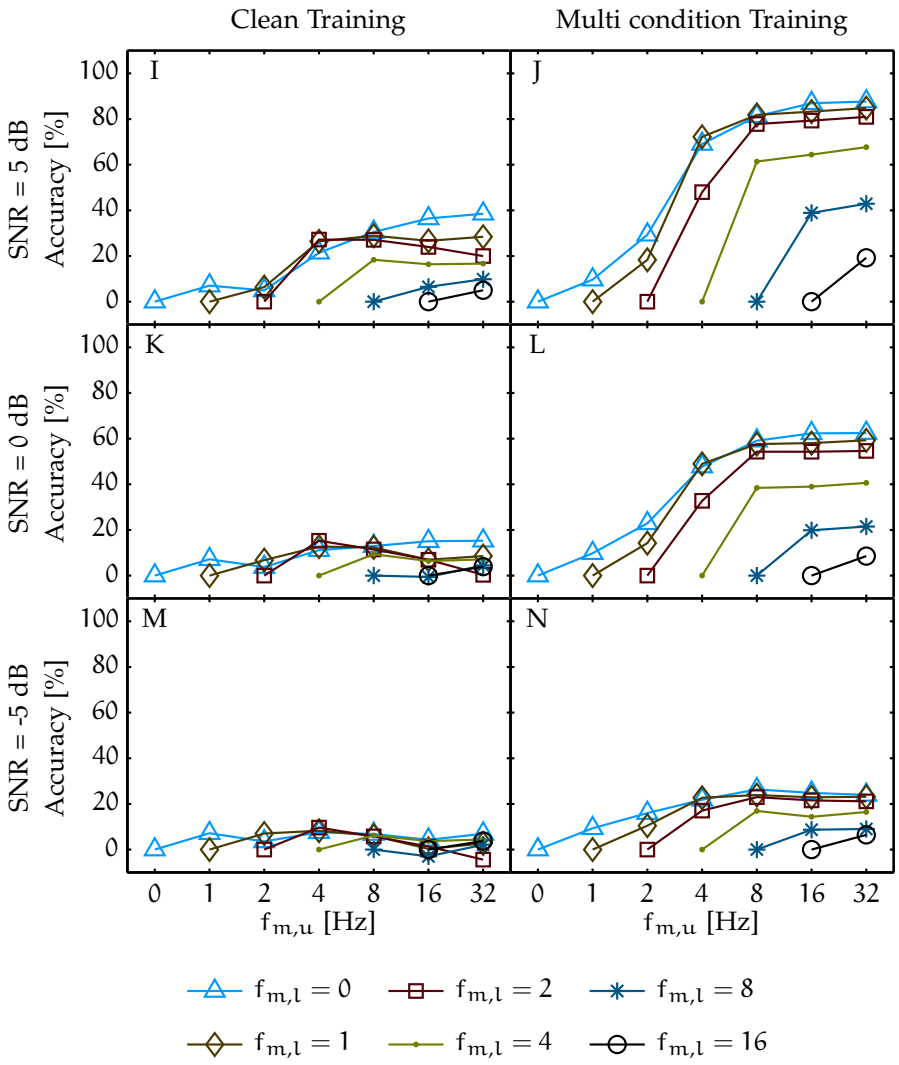

Figure 5.13: Recognition accuracies of the BPE as a function of the higher cutoff frequencies ($f_{m,u}$) parameterized in the lower cutoff frequency ($f_{m,l}$). Three conditions (from the top to the bottom, SNR of 5, 0 and -5 dB) tested with HMMs trained in clean- (left panels) and multi-conditions (right panels) are considered. The values of the parameter are $f_{m,l} \in \{0, 1, 2, 4, 8, 16\}$, corresponding to the symbols △, ◇, □, •, ✻ and ○, respectively. The scores are averaged across noises.

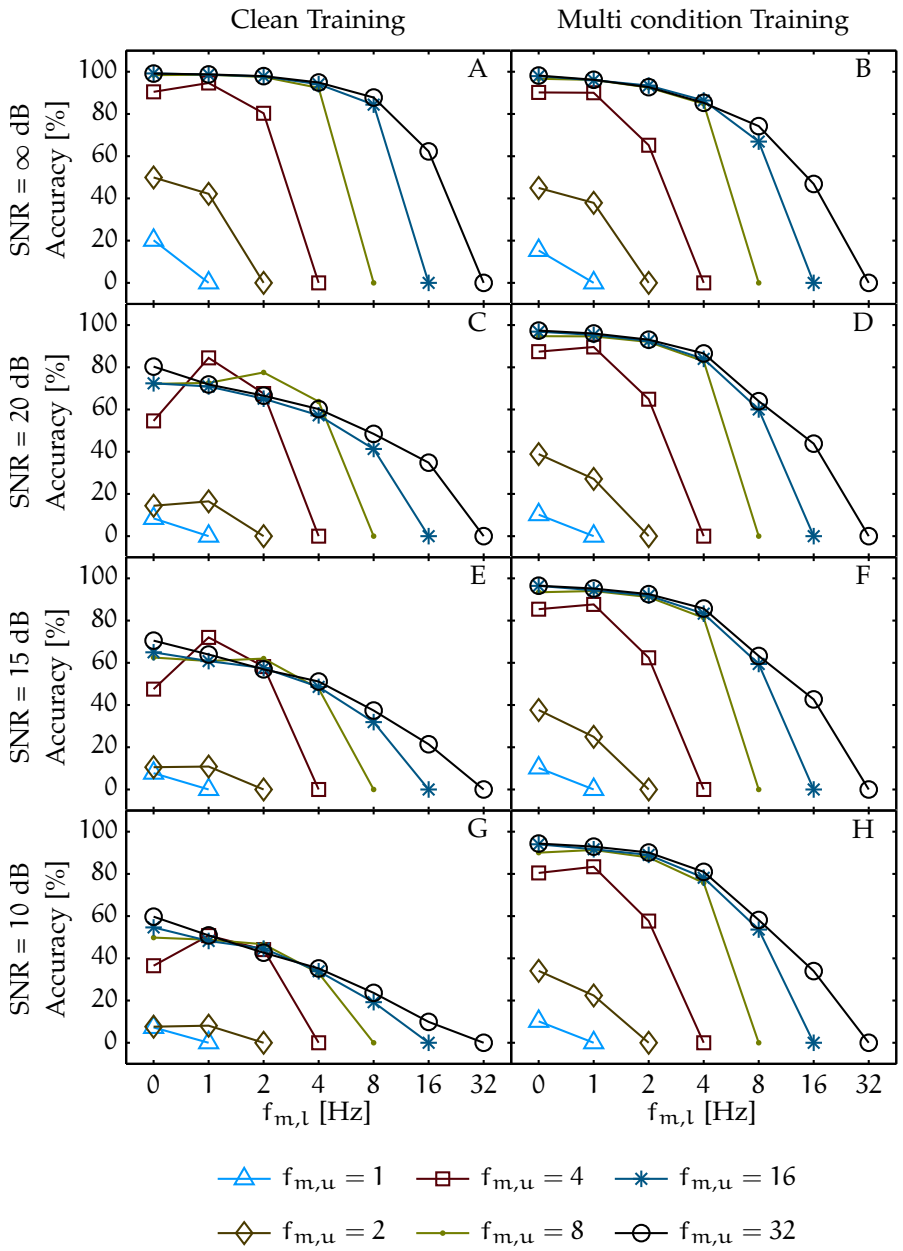

Figure 5.14: Recognition accuracies of the BPE as a function of the lower cutoff frequencies ($f_{m,l}$) parameterized in the higher cutoff frequency ($f_{m,u}$). Four conditions (from the top to the bottom, clean speech, SNR of 20, 15 and 10 dB) tested with HMMs trained in clean- (left panels) and multi-conditions (right panels) are considered. The values of the parameter are $f_{m,u} \in \{1, 2, 4, 8, 16, 32\}$, corresponding to the symbols △, ◇, □, •, ✻ and ○, respectively. The scores are averaged across noises.

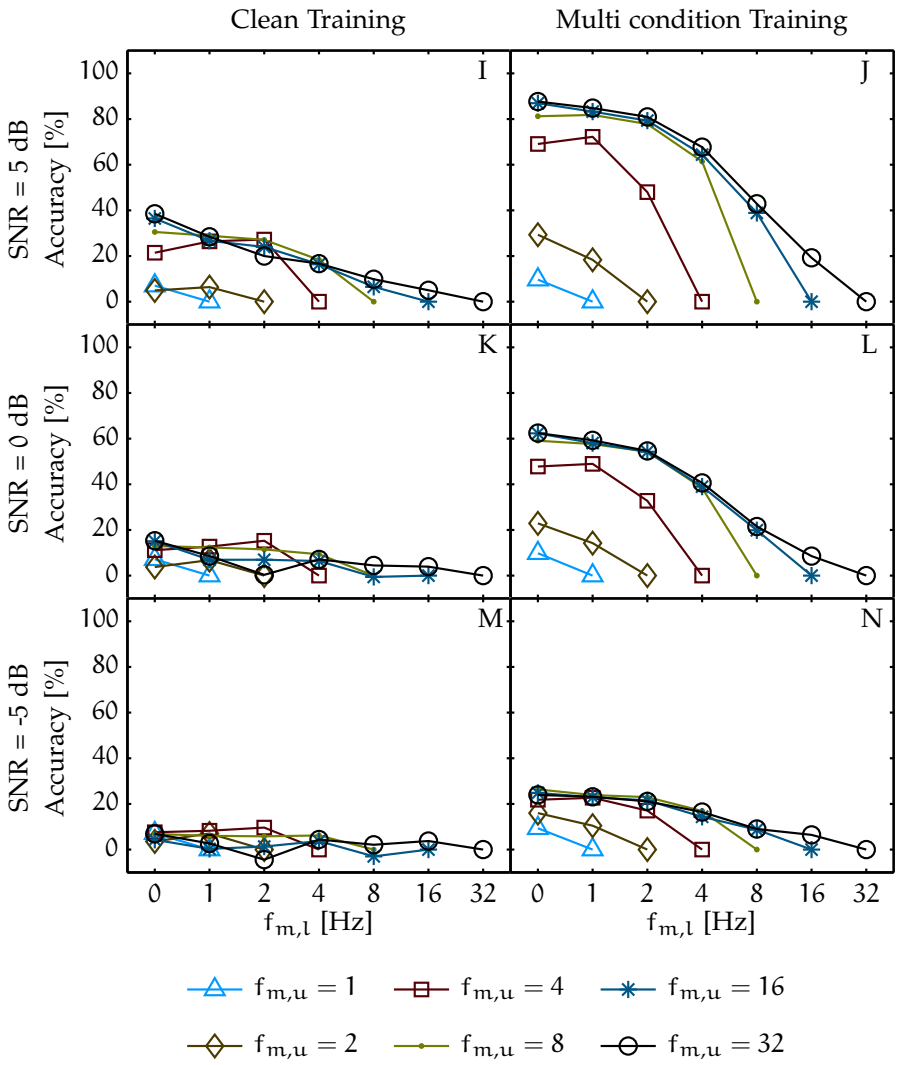

Figure 5.15: Recognition accuracies of the BPE as a function of the lower cutoff frequencies ($f_{m,l}$) parameterized in the upper cutoff frequency ($f_{m,u}$). Three conditions (from the top to the bottom, SNR of 5, 0 and -5 dB) tested with HMMs trained in clean- (left panels) and multi-conditions (right panels) are considered. The values of the parameter are $f_{m,u} \in \{1, 2, 4, 8, 16, 32\}$, corresponding to the symbols △, ◇, □, •, ✻ and ○, respectively. The scores are averaged across noises.

6 DISCUSSION

The results collected in this study will hereby be discussed and compared to already existing studies. These results can be used to give a rational explanation toward some of the reasons that characterize the auditory system as a robust speech recognizer.

Although at the beginning of the current project some parametric tuning was performed in order to enable absolute result comparisons (i. e. in the actual accuracy scores obtained) with the literature, it was soon found to be is quite unfeasible. This was due to the great number of difficult to control parameters playing a role in the final results of the complex framework of ASR systems. Nevertheless, general behaviors of the results from previous works, such as relative improvement between two tested conditions, can be a matter of comparison.

6.1 NOISE ROBUSTNESS IN AUDITORY-MODEL-BASED AUTOMATIC SPEECH RECOGNITION

The main advantage introduced by auditory-model-based feature encoding is the increased noise robustness of the encoded speech representations. The ASR experiments carried out in the current work were performed on speech material corrupted by additive noise (comprising different real life noise types) as well as by convolutional distortions introduced by filtering the speech material with a transfer function simulating a transmission channel. Thus, two of the most common forms of disturbances that can corrupt speech signals in real life were taken into account.

The comparison between results obtained with Mel-Frequency Cepstral Coefficient (MFCC) features and auditory-model-based features (Figs. 5.1 to 5.3) shows that the latter provides an improvement both in clean and in multi-condition training. In the second case, the improvements introduced by the auditory-based features computation are less prominent, suggesting that the role of the adaptation of the statistical models (not to be confused with the adaptation stage within the auditory model) given by the multi-condition training is somehow dominant compared to the improvements introduced by the auditory model.

Nonetheless, the gap in the results for the clean training condition between MFCC and MLP features is more than just noticeable. The reason of the improvement introduced by auditory-like processing can be associated to the modulation frequency filtering provided by the adaptation stage and the modulation low-pass filter of the

auditory model (Tchorz and Kollmeier, 1999) and this concept can be understood more easily considering other signal processing techniques usually employ to deal with distortions in ASR.

For instance, Cepstral Mean Subtraction (CMS), Atal (1974), is often employed (e. g. Hermansky and Morgan, 1994; Holmberg *et al.*, 2006) when the removal of convolutional distortions is required; as mentioned in Appendix A.1, such operation can be interpreted as a (linear) high-pass filtering of the modulation frequencies (i. e. the time trajectories of the short-term cepstral representation of the speech signal). The importance of such high-pass filtering can be accounted for, by noticing that the temporal evolution of several kinds of noise is slow compared to speech (Nadeu *et al.*, 2001; Hermansky and Morgan, 1994; Bourlard *et al.*, 1996). Hence, the slowly varying envelopes of these disturbances can be associated with the high energy concentration in the low-frequency part of the modulation spectrum and these frequencies can be rejected by the high-pass filter provided by the CMS technique (Kanedera *et al.*, 1999; Atal, 1974).

The same behavior in the modulation domain can be associated to the calculation of dynamic coefficients (Hermansky and Morgan, 1994), whereas it is further improved by the RASTA filtering, which combines the high-pass filter with an additional low-pass filter to reduce the high-frequency envelope fluctuations (see Section 2.1.3).

6.1.1 *Adaptation stage contribution*

Similarly to RASTA, the features computation via the Modulation Low-Pass (MLP) model provides a band-pass shaped Modulation Transfer Function (MTF), as seen in Section 3.2. For these reasons it can be argued that the high-pass portion of the MTF obtained with the MLP model, which is a direct consequence of the adaptation loops, is responsible for the rejection of convolutional disturbances.

As far as additive noise concerns, low-modulation-frequencies attenuation by the adaptation stage is again playing an important role. In Section 3.1.3, it has been described the twofold behavior of the adaptation stage, performing a quasi-logarithmic transformation in case of slowly varying signal changes and a quasi-linear transformation in case of fast signal fluctuations. In the first case the additive relation between noise and speech signal is maintained[1] in the frequency domain (i. e. after the auditory filterbank), but it is subsequently lost in the following step due to the nonlinearity introduced by the log-transformation of the two element sum (i. e. noise and speech signal). In other words, if both signal and noise can be considered to be stationary within the considered time frame, a distinction in statistical terms cannot be performed.

1 Due to linearity of the Discrete Fourier Transform (DFT).

In the second mentioned case, provided a slowly varying additive noise, the additive relation is maintained due to the linear transformation applied. Thus, the segregation via high-pass filtering will again be possible.

6.1.2 *Low-pass modulation filter contribution*

The robustness of the model discussed so far, was linked with the processing introduced by the adaptation stage. According to the results reported in Tchorz and Kollmeier (1999) the modulation low-pass filtering should also play a role in the final recognition scores since it should provide a smoother representation of the spectral changes, thus limiting the short-spectra estimation artifacts, Hermansky and Morgan (1994). However, the results presented in Fig. 5.5 show a rather small variability in the accuracy scores obtained with the MLP model using different cutoff frequencies. This could be interpreted as some form of independence of the model to the amount of high modulation frequencies suppressed by the filter, but this is known to be untrue considering the data in literature regarding speech perception (e. g. Drullman *et al.*, 1994a,b).

For all the cutoff frequencies tested with the MLP model (i. e. 2.5, 4, 8 and 20 Hz), it must be noticed that a second order low-pass Butterworth filter was used. The 12 dB/octave roll-off of such a filter could be shallow enough to retain most of the important information present in the higher part of the modulation spectrum, thus leading to very similar accuracy scores. The behavior observed in Tchorz and Kollmeier (1999) between the cases with cutoffs 4 and 8 Hz show a greater variability which can be accounted for by considering that the filter's order is also varied. This hypothesis is strengthened by the results presented in Fig. 5.6, where the recognition accuracies strongly decrease as the filter's order is increased for a fixed cutoff frequency ($f_{cut} = 2$ Hz) due to the rejection of important information. These data seem to point in the same direction of the conclusions drawn in Kanedera *et al.* (1999), where it was shown that in an ASR framework one can also confirm the perceptual findings about the importance of the modulation frequencies around the 4-6 Hz band for speech communications, reported e. g. in Drullman *et al.* (1994a,b) and Drullman (1995).

6.1.3 *Temporal analysis in ASR*

A final aspect to notice, is that the analysis performed on time intervals with duration of approximately one order of magnitude greater than the time frames lengths (150-200 ms against 10-20 ms) seems to be the most important factor in the processing carried out within the ASR oriented framework based model. These time constants roughly

correspond to those adopted in RASTA, dynamical features and CMS (which can both be somehow seen as special cases of RASTA, Hermansky and Morgan, 1994) as well as in the adaptation stage of the considered auditory model. Therefore, it is sensible that a model built to simulate temporal effects in the auditory system, such as forward masking and temporal integration, can also be applied in the field of ASR because the properties of the real system modeled by such temporal-oriented approaches seem to be strongly related to those accounting for robustness in Human Speech Recognition (HSR) processes (and ASR, consequently). It is interesting to notice the duality of this effect referring to a study in which RASTA was used to partially model some aspects of forward masking experiments (Hermansky and Pavel, 1998).

Moreover, fundamental speech units like syllables (or syllable-like units) — which are thought to be very important for speech perception and understanding (Greenberg, 1996) — can be described in time intervals of similar lengths. Therefore, considering the idea mentioned earlier about the evolution of the auditory system in order to exploit the different characteristics of speech (Hermansky, 1998), the comparability of such time quantities strengthens the assumption of using longer temporal intervals to extract unique features from speech signals.

6.2 ROBUSTNESS INCREASE BY DYNAMIC COEFFICIENTS AND DCT COMPUTATION

Figure 5.7 shows the changes in the recognition accuracies for MFCC and MLP features where the dynamic coefficients are either computed or not. The derivative-like operation performed to evaluate dynamic features can be considered as a high-pass filtering in the modulation domain, Hermansky and Morgan (1994). Thus, the small improvement given by the inclusion of these coefficients in the MLP case could be interpreted considering that an additional robustness is introduced by the fixed (i. e. signal-independent) modulation high-pass filter. Due to the nonlinearity introduced by the adaptation stage of the auditory model, the obtained MTF is not fixed and its characteristics, e. g. its steepness at low frequency, changes with different input signals. For certain signals it could happen that the high-pass section of the model's MTF is not completely suited to remove the low frequency attenuation corrupting the speech signal. Thus, the additional high-pass filter could help rejecting the unwanted components. The fact that the adaptation stage and the dynamic features possess similar behaviors could also motivate why the improvement obtained when introducing dynamic coefficients in auditory-based features is smaller than the improvements shown for MFCC features. In the latter case, temporal information are not coded within the preprocessing operations and

the inclusion of the dynamics of the coefficients strongly influences the results.

Finally, the role of the DCT can be recognized by observing the results illustrated in Fig. 5.7. The hypothesis of uncorrelated features must be fulfilled when using the framework employed in the current study; the DCT applied on the IRs provides a tool to guarantee such a property. The improvement of the recognition accuracies supports this fact. The necessity to perform a transformation of the IRs is one of the reasons suggesting that diagonal-covariance-HMMs might not be the ideal models to deal with such features (Tchorz and Kollmeier, 1999); this is because of the unclear physical meaning of the transformed auditory features (Batlle *et al.*, 1998; Nadeu *et al.*, 1995).

6.3 MULTIPLE CHANNEL FEATURE ENCODING

The results obtained introducing the MFB features are now discussed. Figure 5.8 shows the results for an increasing number of filters in the filterbank and essentially no difference is introduced by this change. This behavior was not expected considering the usefulness of the modulation filterbank concept to describe different temporal aspects of the auditory system (Dau *et al.*, 1997a; Ewert and Dau, 2000). One of the causes of this invariance, could be again the too shallow slope of the filters employed in the filterbank which would retain very similar information across channels although the center frequencies are changed. Another cause for this invariance was considered to be the unfulfillment of the requirement necessary in order to use diagonal covariance matrixes. In Section 4.1 the encoding procedure M1 has been described and it was pointed out how, after successfully achieving feature decorrelation via DCT in the single channels, the compaction of multiple channel coefficients in single feature vectors reintroduced features correlation. Thus, performing HMM-based ASR tasks with such feature vectors introduces a systematic error. However, it is not simple to quantify the entity of such an error since several parameters could play a role in it.

The following comparison, presented in Fig. 5.9, shows that similar results are obtained when both the center frequencies of the filterbank and the feature encoding method are changed. Thus, both the modifications seem to be unnecessary to achieve better results. The argument about the filter's slopes mentioned earlier for the data about different MFB filters number could again motivate the similarity in results when changing the center frequencies. Regarding the encoding method, it is difficult to estimate how relevant is the improper dynamic-coefficients encoding (i. e. the problem of computing derivative-like features on the coefficients obtained with M2) compared to the gain that could have been achieved by the new frame distribution.

In the last comparison (Fig. 5.10), from the results obtained with feature encoded using different filter shapes (DauNCF and FQNCF), better results are obtained with resonant filters suggesting that they are preferred to symmetric filters in this framework. The degradation in recognition accuracy when using fixed-Q filters is very contained and could be due to the additional attenuation introduced for very low-modulation frequencies in the range $[1, 2]$ Hz, which was also shown to be relevant in speech perception (Drullman *et al.*, 1994a,b; Kanedera *et al.*, 1999).

After all the experiments were concluded and the whole set of data analyzed, a different cause that accounts for the similarity of the results for MFB features was considered. It was postulated that the problem could derive from the intrinsic error made in the attempt of encoding (and treating) the features from each frame as *single* feature vectors. In such a way, the amount of information when considering multiple filters is increased but it is treated as a single quantity without making distinctions between the different channels. Therefore, the multi-channel information is somehow "integrated", without exploiting the possibility to treat the cues that could be enhanced by each channel separately. For the same reason, it becomes unfeasible to investigate the contributions of the different channels to the total recognition accuracy once the results are obtained.

As a consequence, the results from the experiments carried out with MFB features can be compared to the results of other MFB experiments, but they are not easily comparable with the results of experiments based on MLP features earlier described. On the other hand, it could be expected that using a framework allowing to work with both MLP and MFB features in the same way, also the change of these parameters would lead to different results. Different approaches that would allow to independently consider the information from the different modulation channels when training the statistical models were investigated in literature (e. g. Ellis, 2000; Zhao and Morgan, 2008; Zhao *et al.*, 2009), but due to the limited amount of time and the necessity to employ a different recognizer for such a purpose they were not implemented in this project.

6.4 BAND-PASS EXPERIMENT RESULTS

Although the processing steps implemented in this work were slightly different than the ones employed in Kanedera *et al.* (1999) due to the auditory processing, the results of the referenced study were expected to be confirmed since both the approaches exhibited capabilities toward modeling perceptual-like characteristics.

The results obtained for the BPE for the clean training case, partially reflect the data in Kanedera *et al.* (1999, Fig. 2). From Fig. 5.11, or alternatively from the C panels of Figs. 5.12 and 5.14, it can be seen how

including only part of the information from the modulation domain, returns better results than using the whole modulation spectrum in the subsequent processing.

Again, the bands concurring in this effect are those including the modulation frequencies that were found to be more perceptually important. In the considered case the peaks in recognition accuracy were found for $f_m \in [1,4]$ Hz and $f_m \in [2,8]$ Hz. However, unlike the results in the mentioned study, the accuracies obtained when processing the signals with broader filters (i. e. higher $f_{m,u}$ in Fig. 5.11) than the two mentioned bands show a lower accuracy. Thus, the inclusion of higher modulation frequency information destructively contributes to the ASR score. The higher impact of the $[1,4]$ Hz and $[2,8]$ Hz bands is maintained also at lower SNR, with a less pronounced effect. The different behavior of the described simulations compared to Kanedera *et al.*'s results could arise from three main reasons:

1. the low-pass filtering introduced when downsampling the IRs by averaging (moving average filter);
2. the different preprocessing performed before modulation filtering (auditory model as opposed to the J-RASTA technique);
3. the use of a different recognizer.

The low-pass filtering is introducing some additional attenuation in the higher part of the modulation spectrum considered in the experiments (up to 32 Hz). However, such attenuation is rather small and cannot be accounted for all the introduced changes.

Regarding the second point, it is recalled that the adaptation stage introduces signal-dependent changes in the MTF; this could account for a more auditory-based discrimination of the important temporal modulation bands characterizing the speech signals and therefore cause the differences between the two cases.

Lastly, in Kanedera *et al.* (1999) both a DTW- and a HMM-based system were used. Only part of the data is available for the HMM recognizer, but they seem to show a similar behavior to the results obtained in the current study.

An attempt to relate the contributions of the sub-bands which make up the wider bands (e. g. the relation between the accuracies obtained from the bands $[1,2]$ Hz, $[2,4]$ Hz and the result from the $[1,4]$ Hz band) has been made, but only a general behavior could be observed. Aside from the constructive contribution of the narrowest contiguous bands (i. e. the 1-octave-wide filters) and the somehow stable values for results extracted from broader bands (i. e. the top-right corner of Fig. 5.11), it has not been possible to define a model describing the details shown in the figure.

6.5 LIMITATIONS

One of the main limitations of the framework employed to carry out the described ASR experiment was given by the impossibility to perform a feature encoding process completely suitable with the subsequent back-end. This required to introduce different encoding approaches which simply brings in more parameters and uncertainty to consider.

Another practical limitation consisted of the high computational time and load needed to encode the whole set of files from the two test sets of the AURORA 2.0 corpus (approximately 36000) using the auditory model. The processing time of each single file was depending on the encoding procedure, but it was on average between 1 and 3 s with the computer employed in this work (approximately between 10 to 15 s in a normal workstation). Thus, more than one day was required to encode each single feature set. The higher computational need is one of the main drawbacks encountered when working with auditory models to extract features for ASR experiments (Jankowski *et al.*, 1995; Hermansky, 1998), therefore limiting their widespread in applications that require real time computation.

Moreover, as pointed out in Hermansky (1998), it can be sometimes very confusing to deal with all the parameters that can play a role in a complete ASR system, leading to a difficulty interpreting the results and comparing with other studies. In addition to this, it must be noticed that, although at the present day HMMs represent by far the most popular technique employed to model speech signal in ASR (Morgan *et al.*, 2004), many constraint of this approach have been reported (e. g. Hermansky, 1998; Bourlard *et al.*, 1996). This has to be kept in mind especially when the encoding methods adopted in the front-end could be not perfectly suited to an HMM-based back-end.

6.6 OUTLOOK

As repeatedly mentioned in the previous sections, a modification of the statistical model to properly deal with multi-channel IRs could be done by adopting a back-end designed to perform multi-stream recognition. Retrospectively, an additional toolkit could have been employed together with the HTK, which has limited support for multi-stream recognition when combined with non-standard features[2], to work with MFB features. For instance, the RESPITE CASA Toolkit (CTK) decoder[3] could have been used[4]. In such a case it would be possible to recognize utterances in parallel by means of different models trained

2 Standard features are features encoded using algorithms not provided by the HTK, like the auditory features computed by means of external models.

3 http://spandh.dcs.shef.ac.uk/projects/respite/ctk/notes/releases.html

4 As suggested by Guy J. Brown in a personal communication.

on the multiple modulation channels. Subsequently, the obtained posterior probabilities could be used to investigate both the contribution in recognition from each channel and the combination of multi-stream information to give the resultant (i. e. recognized using all the channels information) word sequence. Furthermore, dynamically weighted bands, depending on the type of noise background, could be used.

Also the results of the BPE provide additional ideas that could be applied whether a multi-stream recognition paradigm was employed. If the behavior of the curves obtained fixing one of the two parameters $f_{m,l}$ and $f_{m,u}$ (i. e. the curves in Figs. 5.12 to 5.15) was modeled, the results in each modulation frequency band could be seen as an "importance factor" and be used to somehow weigh — in other words, introducing additional "a priori" knowledge — the multi-stream HMM models. In such a case, it should also be investigated whether the behavior of these "importance factors" of the different frequency bands is substantially changed in different ASR tasks, e. g. by reproducing the experiments with different speech materials.

Some final considerations regard the alternative choices of the speech material. First of all, it could be interesting to introduce a more systematic way of testing different disturbances. By using speech corrupted by single types of noise, e. g. an additive noise or a convolutional distortion, it would be easier to assess the improvements relatively to single noise conditions. Moreover, testing the signal with large word vocabularies (thus relying on phonemes classification tasks) would be more meaningful for the sake of the comparison with HSR experiments.

7 CONCLUSIONS

An auditory-signal-processing-based approach to extract features for ASR experiments was described in the current study. Several results were computed in an attempt to validate previous studies investigating the use of auditory-like features in ASR as well as to link these results to perceptual data. Although part of the parameter tuning of the system was performed to match those employed in other studies, in order to be able to compare the results obtained, it soon turned out that such task is more challenging than expected due to a the great number of parameters to deal with when working with an ASR system. It was therefore considered to perform the comparisons on the general behaviors of different encoding strategies, focusing on the relative changes in the results.

The main property investigated results obtained with different auditory-like sets of features from different modifications performed on the last stage of the models presented in Dau *et al.* (1996a, 1997a), concerning the filtering of temporal modulations. An increased feature robustness was achieved by such encoding strategies, compared to a classical feature extraction technique (i. e. MFCCs), especially in noise-corrupted speech. The improvement was greater for clean-trained HMMs compared to multi-condition-trained models.

In a subsequent series of experiments, the influence of changes in the cutoff frequency of the low-pass modulation filter was investigated and found to have almost no relevance. This behavior was hypothesized to be due to the shallow slope of the filters employed in the described experiments. Evidence of this were also found in a following set of experiment showing a severe performance degradation obtained by a progressive increase of the filter's order.

The computation of the DCT before passing the features to the recognizer was found to improve the recognition accuracies, as expected since the features decorrelation was achieved using this operation. A mild improvement was also shown for MLP features with the introduction of first- and second-order dynamic coefficients as opposed to the great increase in accuracy when dynamic coefficients are used with MFCC features. This was justified considering that in the auditory-based approaches the temporal information is already coded in the IRs from the steps involving temporal modulation operations.

Experiments including features from multiple modulation channels were performed and very similar results were obtained as the number of these channels was varied. This was explained by both an inadequate choice of the filters' characteristics employed (which

were thought to be too shallow toward high frequencies) and the usage of feature encoding procedures not completely suitable with the statistical back-end. Regarding the problems of interfacing front-end and back-end, both the issue deriving from the correlation of the features employed and the problem caused from the impossibility of treating the information from different channels separately were discussed. The results for MFB features obtained using filterbanks with different characteristics, i. e. filters' center frequencies and shapes, and encoding procedures also showed a very low variability. Hence, the filters' parameters seemed not to be as crucial as they were expected to be, within the framework employed to perform ASR experiments in the current project. It was noticed, however, that when using a system correctly treating the MFB features, the results could behave differently, with respect to the parameters mentioned.

Subsequently, an experiment inspired by the work of Kanedera *et al.* (1999), to assess the importance of single modulation frequency bands in the low-modulation-frequency spectrum, was reproduced. The relevance of the information contained in frequency bands that were found in literature to be more perceptually significative (Drullman *et al.*, 1994a,b; Drullman, 1995) was partially confirmed. This information was hypothesized to be applicable when performing a perceptually-oriented weighting of the different frequency bands.

Finally, some ideas that could be developed to deal with the limitations encountered using the described framework were briefly introduced.

A

APPENDIX

A.1 HOMOMORPHIC SIGNAL PROCESSING AND REMOVAL OF CONVOLUTIONAL DISTURBANCES

Homomorphic processing consists in the usage of some strategies to convert a given non-linear system into a linear one, in order to be able to use standard analysis techniques and handle different problems in a well known and understood way, Oppenheim and Schafer (1975).

Cepstral domain transformations represent an example of this, Proakis and Manolakis (2007), employing the logarithmic operator to transform signals (in the considered case speech, but cepstrum has other areas of interests too) in the log domain where some kind of operations can be more easily performed (e. g. removal of convolutional disturbances). The cepstrum $c_f(\tau)$ of a signal $f(t)$ is an operation introduced in the 60s to analyze the rate of change of a signal's spectrum, Bogert *et al.* (1963). It results from the (nonlinear) transformation defined by:

$$c_f(\tau) = \mathcal{F}\left\{\log\left(|\mathcal{F}\{f(t)\}|^2\right)\right\} \tag{A.1}$$

where $\mathcal{F}$ denotes the Fourier Transform (FT) operator, substituted by the Discrete Fourier Transform (DFT) operator in the case of digital signals. Dealing with audio signals, i. e. real valued signals, the external FT of Eq. (A.1) reduces to Discrete Cosine Transform (DCT), i. e. :

$$c_f(\tau) = \mathcal{DCT}\left\{\log\left(|\mathcal{F}\{f(t)\}|^2\right)\right\}. \tag{A.2}$$

Due to its properties, some of which will be briefly introduced in this section, cepstral techniques can be used to separate (i. e. deconvolve) different components within speech signals, e. g. speech from convolutional distortions or vocal tract transfer function from glottal excitatory signals Gudnason and Brookes (2008). For instance, given a convolutional disturbance, $h(t)$ acting on a signal $x(t)$ to give the corresponding degradated signal $y(t)$, it can be written Kolossa (2007):

$$\begin{aligned} y(t) &= x(t) \star h(t) \\ &\updownarrow \\ Y(\omega) &= X(\omega) \cdot H(\omega) \end{aligned} \tag{A.3}$$

$$\begin{aligned} &\Downarrow \quad \log||\cdot||^2 \\ Y^l(\omega) &= X^l(\omega) + H^l(\omega) \end{aligned} \tag{A.4}$$

where $\star$ denotes the convolution integral operator[1]. By now taking the DCT of the new obtained relation, we have (using the linearity property of the DCT):

$$\begin{aligned} \mathcal{DCT}\left[Y^l(\omega)\right] &= \mathcal{DCT}\left[X^l(\omega)\right] + \mathcal{DCT}\left[H^l(\omega)\right] \\ c_y(\tau) &= c_x(\tau) + c_h(\tau). \end{aligned} \tag{A.5}$$

One of the main advantages of using cepstral analysis consists in the possibility of separating nonstationary zero-mean sources (e. g. speech) from a stationary source (e. g. the distortion), see e. g. Kanedera *et al.* (1999); Hermansky and Morgan (1994); the stationarity of the distortion can be hypothesized if its statistics vary much slower than the speech's ones. In such a case $E[c_h(\tau)] = c_h(\tau)$, i. e. c_h is stationary in the considered time range, and due to the zero-mean characteristics of the speech $E[c_x(\tau)] = 0$. Thus:

$$\begin{aligned} E\left[c_y(\tau)\right] &= E[c_x(\tau)] + E[c_h(\tau)] \\ &= c_h(\tau) \\ &\Downarrow \\ c_y(\tau) - E\left[c_y(\tau)\right] &= c_x(\tau). \end{aligned} \tag{A.6}$$

Equation (A.6) is usually referred to as Cepstral Mean Subtraction (CMS), Atal (1974), and such a technique is broadly used to remove convolutional disturbances from corrupted signals (e. g. Holmberg *et al.*, 2006), somehow representing a way to perform an operation inverse to convolution. As the name suggests, the main feature of the CMS consists in the removal of the DC component from the cepstrum and it can therefore be seen as a high-pass filtering operation, Hermansky and Morgan (1994). In a frame oriented framework, the quantity in the LHT of Eq. (A.6) can be redefined as:

$$\begin{aligned} c_y^{cms}(\tau, k) &= c_y(\tau) - E\left[c_y(\tau)\right] \\ &= c_y(\tau, k) - \bar{c}_y(\tau) \\ &= c_y(\tau, k) - \frac{1}{N}\sum_{n=1}^{N} c_y(\tau, n). \end{aligned} \tag{A.7}$$

The summation in Eq. (A.7) calculates the mean of the whole processed signal, but it is usually substituted by an adaptive estimate of the cepstral mean in real time implementation, Kolossa (2007). In such a case, the latter represents a moving average filter (i. e. a low-pass filter); thus, by subtracting the summation value from $c_y(\tau, k)$ the result

1 Strictly speaking, the step defined in Eq. (A.3) is not mathematically correct for the considered case, since it neglects the signal windowing applied to obtain the frames. However, the assumption would become more accurate if the window length is chosen to be long enough compared to the impulse response of the distortion, Stockham *et al.* (1975). A way to correctly express Eq. (A.3) is given in Avendano and Hermansky (1997)

is a high pass filter, which can be used to reject the low frequency part of the modulation spectrum.

An operation similar to the CMS can be defined by exploiting the homomorphic signal processing steps just described but leaving out the DCT (see e. g. Hermansky and Morgan, 1994), i. e. to define Eq. (A.7) by replacing $c_y(\tau)$, $c_x(\tau)$ and $c_h(\tau)$ with $Y^l(\omega)$, $X^l(\omega)$ and $H^l(\omega)$, respectively. However, such an operation is not completely suitable for diagonal covariance HMM-based ASR systems, see Appendix A.3, since the DCT is needed to introduce the desired features decorrelation necessary in the mentioned statistical approach.

A.2 DISCRETE COSINE TRANSFORM

The Discrete Cosine Transform (DCT), Ahmed *et al.* (1974), is a mathematical operation related to the Discrete Fourier Transform (DFT) providing the decomposition of a function as a sum of a finite number of cosine components oscillating at different frequencies. The usage of cosine functions instead of complex exponential functions as in the DFT provides that real signals are mapped into real signals (e. g. Proakis and Manolakis (2007)). The DCT is employed in many different areas, e. g. video and image coding as well as audio related applications (Khayam, 2003), because it offers several useful properties.

Mainly, it provides good decorrelation and energy compaction of the signals to which it is applied (e. g. Khayam, 2003; Ahmed *et al.*, 1974). These two properties are linked since the removal of redundancies within the signal, i. e. the decorrelation, allows to characterize it using a smaller number of samples without having a big loss in the information. Moreover, decorrelation is required for the features vectors in order to be further processed by the HMMs-based back-end in the case the parametric estimation is done considering diagonal covariance matrices (see Section 2.2).

Additionally, the DCT shares linearity, separability, symmetry and orthogonality with the DFT. However, unlike the latter, the DCT does not request to retain the information of both magnitude and phase since it provides real outputs, Khayam (2003), and it is thus preferred to DFT in many situations.

It is recalled that the DCT can be seen as an approximation of the Karhunen-Loève Transform (KLT), also known as Principal Components Analysis (PCA), a linear transformation capable of extracting the so called principal components, i. e. the directions of maximal variance, out of a set of multivariate observations (Batlle *et al.*, 1998; Khayam, 2003; Hunt, 1999). As the length of the considered signal tends to infinity the DCT approaches the KLT. PCA is often exploited to perform dimensionality reduction, for its optimal property of energy compaction and thus it is suitable for the purposes of ASR. However, it is computationally much more expensive than the DCT due to its signal

dependence. Thus, given the possibility of implementing FFT-based fast algorithms computing the DCT and the consideration about the approached optimality in energy compaction and decorrelation, the DCT seems to be a rather convenient choice in an ASR framework.

There exist several versions and as many definitions of DCT (eight according to Sanchez *et al.*, 1995). A common definition used in ASR literature, and the one used in the current work, is referred to as DCT-II and reads (Young *et al.*, 2006; Ahmed *et al.*, 1974; Khayam, 2003):

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=0}^{N-1} f_j \cos\left[\frac{\pi i}{N}(j - 0.5)\right] \tag{A.8}$$

where f_j is an N samples 1-D discrete signal.

In some of the experiment it has been necessary to use the DCT on two dimensional signals. The 1-D formula given in Eq. (A.8) can be extended to the 2-D DCT definition (Khayam, 2003):

$$c_{i,k} = \sqrt{\frac{2}{N}}\sqrt{\frac{2}{M}} \sum_{j=0}^{N} \sum_{l=0}^{M} f_{j,l} \cos\left[\frac{\pi i}{N}(j - 0.5)\right] \cos\left[\frac{\pi k}{N}(l - 0.5)\right] \tag{A.9}$$

where $f_{j,l}$ is an $N \times M$ samples 2-D discrete signal.

A.3 FEATURES CORRELATION

As mentioned in Section 2.2, the distributions used to model the observation data, giving raise to the emission probabilities, in most of the HMM-based systems are chosen to be multivariate Gaussian mixtures of uncorrelated variables. Figure A.1 shows an example of a mixture of two bivariate distributions of the uncorrelated variables X and Y (i. e. the covariance matrixes of both the components of the mixture were diagonal as it can be notice from the directions of maximal variance parallel to the axes). This constraint allows to lower the number of parameters defining the mixtures distributions and, as a consequence, to properly train the HMMs using a smaller number of training data (Young *et al.*, 2006). Furthermore, a reduction of the computational load and time is achieved.

A measure of the correlation between features is given by the correlation coefficient. According to the definition, the sample correlation coefficient (also known as Pearson's *r*, Sheskin, 2004) between two variables X and Y (in this case representing two different features of the feature vector, with sample mean value $\overline{X}$ and $\overline{Y}$ respectively) is given by (Sheskin, 2004):

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} \tag{A.10}$$

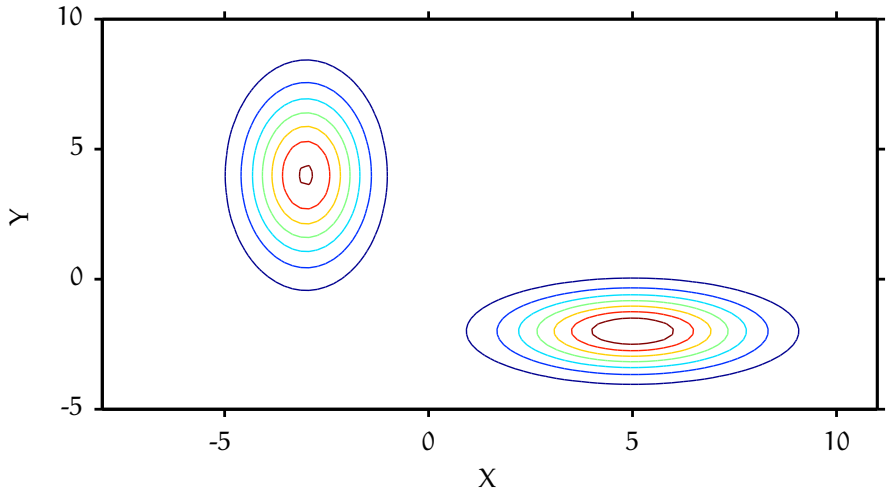

Figure A.1: Example of a mixture of two bivariate Gaussian distribution of the uncorrelated variables X and Y. Diagonal covariance matrixes implies that the directions of maximal variance are orientated along the axis.

In the case of features vectors with dimension larger than two, like in the case of ASR, Eq. (A.10) is evaluated for all the combination between variables (i. e. features). By plotting the matrix containing the correlation coefficients, one can get a visual description of the variables' uncorrelatedness. An uncorrelated set of variables returns a correlation matrix with ones along the diagonal (i. e. maximum autocorrelation) and zeros in all the other positions, see Fig. A.2.

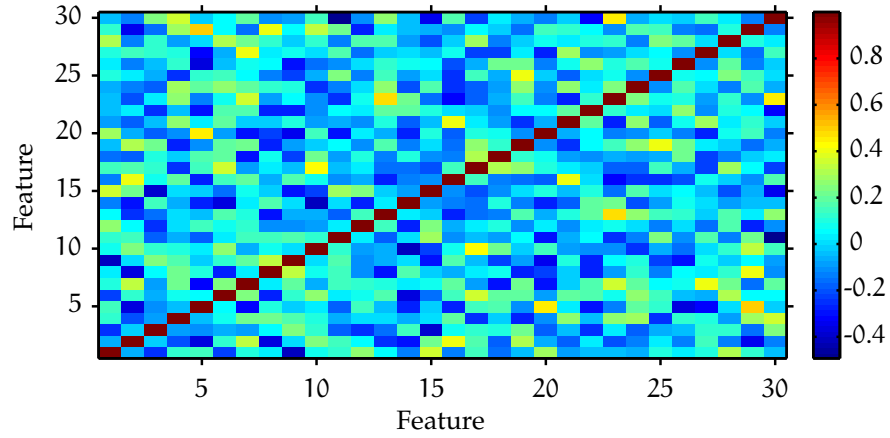

Figure A.2: Correlation matrix from a set of uncorrelated variables. The non-zero values in the off-diagonal positions derives from the small number of samples used to obtain the matrix.

BIBLIOGRAPHY

Ahmed, N., Natarajan, T., and Rao, K. (**jan. 1974**), "Discrete Cosine Transfom," Computers, IEEE Transactions on **C-23**(1), 90 – 93.

ANSI-S3.5 (**1997**), *ANSI-S3.5-1997. American National Standard: Methods for Calculation of the Speech Intelligibility Index*.

Atal, B. S. (**1974**), "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," The Journal of the Acoustical Society of America **55**(6), 1304–1312, URL http://link.aip.org/link/?JAS/55/1304/1.

Avendano, C. and Hermansky, H. (**jul 1997**), "On the effects of short-term spectrum smoothing in channel normalization," Speech and Audio Processing, IEEE Transactions on **5**(4), 372 –374.

Batlle, E., Nadeu, C., and Fonollosa, J. A. R. (**1998**), "Feature Decorrelation Methods in Speech Recognition. A Comparative Study," in *Fifth International Conference on Spoken Language Processing*.

Békésy, G. v. (**1949**), "On the Resonance Curve and the Decay Period at Various Points on the Cochlear Partition," The Journal of the Acoustical Society of America **21**(3), 245–254, URL http://link.aip.org/link/?JAS/21/245/1.

Bogert, B., Healy, M., and Tukey, J. (**1963**), "The quefrency alanysis of time series for echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking," in *Proc. Symp. on Time Series Analysis*, pp. 209–243.

Bourlard, H., Hermansky, H., and Morgan, N. (**1996**), "Towards increasing speech recognition error rates," Speech Communication **18**, 205–231.

Brown, G. J., Ferry, R. T., and Meddis, R. (**2010**), "A computer model of auditory efferent suppression: Implications for the recognition of speech in noise," The Journal of the Acoustical Society of America **127**(2), 943–954, URL http://link.aip.org/link/?JAS/127/943/1.

Cui, X. and Alwan, A. (**November 2005**), "Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR," Speech and Audio Processing, IEEE Transactions on **13**(6), 1161 – 1172.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**a), "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," The Journal of the Acoustical Society of

America **102**(5), 2892–2905, URL http://link.aip.org/link/?JAS/102/2892/1.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**b), "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," The Journal of the Acoustical Society of America **102**(5), 2906–2919, URL http://link.aip.org/link/?JAS/102/2906/1.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**a), "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," The Journal of the Acoustical Society of America **99**(6), 3615–3622, URL http://link.aip.org/link/?JAS/99/3615/1.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**b), "A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements," The Journal of the Acoustical Society of America **99**(6), 3623–3631, URL http://link.aip.org/link/?JAS/99/3623/1.

Davis, S. B. and Mermelstein, P. (**aug 1980**), "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," Acoustics, Speech and Signal Processing, IEEE Transactions on **28**(4), 357 – 366.

Drullman, R. (**1995**), "Temporal envelope and fine structure cues for speech intelligibility," The Journal of the Acoustical Society of America **97**(1), 585–592, URL http://link.aip.org/link/?JAS/97/585/1.

Drullman, R., Festen, J. M., and Plomp, R. (**1994**a), "Effect of reducing slow temporal modulations on speech reception," The Journal of the Acoustical Society of America **95**(5), 2670–2680, URL http://link.aip.org/link/?JAS/95/2670/1.

Drullman, R., Festen, J. M., and Plomp, R. (**1994**b), "Effect of temporal envelope smearing on speech reception," Journal of the Acoustical Society of America **95**(2), 1053–1064.

Ellis, D. P. (**2000**), "Stream combination before and/or after the acoustic model," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, pp. 1635–1638.

Ewert, S. D. and Dau, T. (**2000**), "Characterizing frequency selectivity for envelope fluctuations," The Journal of the Acoustical Society of America **108**(3), 1181–1196, URL http://link.aip.org/link/?JAS/108/1181/1.

Festen, J. M. and Plomp, R. (**1990**), "Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing," The Journal of the Acoustical Society of

America **88**(4), 1725–1736, URL http://dx.doi.org.globalproxy.cvt.dk/doi/10.1121/1.400247.

French, N. R. and Steinberg, J. C. (**1945**), "Factors Governing the Intelligibility of Speech Sounds," The Journal of the Acoustical Society of America **17**(1), 103–103, URL http://dx.doi.org.globalproxy.cvt.dk/doi/10.1121/1.1902408.

Furui, S. (**feb 1986**), "Speaker-independent isolated word recognition using dynamic features of speech spectrum," Acoustics, Speech and Signal Processing, IEEE Transactions on **34**(1), 52 – 59.

Gales, M. and Young, S. (**January 2007**), "The application of hidden Markov models in speech recognition," Found. Trends Signal Process. **1**, 195–304, URL http://portal.acm.org/citation.cfm?id=1373536.1373537.

Giraud, A.-L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., and Kleinschmidt, A. (**2000**), "Representation of the Temporal Envelope of Sounds in the Human Brain," Journal of Neurophysiology **84**(3), 1588–1598, URL http://jn.physiology.org/content/84/3/1588.abstract.

Glasberg, B. R. and Moore, B. C. (**1990**), "Derivation of auditory filter shapes from notched-noise data," Hearing Research **47**(1-2), 103 – 138, URL http://www.sciencedirect.com/science/article/pii/037859559090170T.

Goldsworthy, R. L. and Greenberg, J. E. (**2004**), "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," The Journal of the Acoustical Society of America **116**(6), 3679–3689, URL http://dx.doi.org/doi/10.1121/1.1804628.

Greenberg, S. (**1996**), "Understanding Speech Understanding: Towards A Unified Theory Of Speech Perception," in *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, pp. 1–8.

Gudnason, J. and Brookes, M. (**31 2008-april 4 2008**), "Voice source cepstrum coefficients for speaker identification," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4821 –4824.

Harrington, J. (**2010**), *Phonetic Analysis of Speech Corpora* (Wiley Publishing).

Hawkins, J. E. and Stevens, S. S. (**1950**), "The Masking of Pure Tones and of Speech by White Noise," The Journal of the Acoustical Society of America **22**(1), 6–13, URL http://dx.doi.org.globalproxy.cvt.dk/doi/10.1121/1.1906581.

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

Hermansky, H. (**1997**), "The modulation spectrum in the automatic recognition of speech," 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings , 140–147URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=658998.

Hermansky, H. (**1998**), "Should recognizers have ears?" Speech Communication **25**(1-3), 3–27.

Hermansky, H. and Morgan, N. (**oct 1994**), "RASTA processing of speech," Speech and Audio Processing, IEEE Transactions on **2**(4), 578 –589.

Hermansky, H. and Pavel, M. (**1998**), "RASTA model and forward masking," in *Proc. NATO/ASI Conference on Computational Hearing* (Il Ciocco, Italy), pp. 157–162.

Holmberg, M., Gelbart, D., and Hemmert, W. (**January 2006**), "Automatic speech recognition with an adaptation model motivated by auditory processing," IEEE Transactions on Audio, Speech and Language Processing **14**(1), 43–49, URL http://dx.doi.org/10.1109/TSA.2005.860349.

Holmberg, M., Gelbart, D., and Hemmert, W. (**2007**), "Speech encoding in a model of peripheral auditory processing: Quantitative assessment by means of automatic speech recognition," Speech Communication **49**(12), 917 – 932, URL http://www.sciencedirect.com/science/article/B6V1C-4NXRMDX-1/2/ff71e7cc5c37319428c21738e7342440.

Hon, H.-W. and Wang, K. (**2000**), "Unified frame and segment based models for automatic speech recognition," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, vol. 2, pp. II1017 –II1020 vol.2.

Houtgast, T. and Steeneken, H. J. M. (**1985**), "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," The Journal of the Acoustical Society of America **77**(3), 1069–1077, URL http://link.aip.org/link/?JAS/77/1069/1.

Huang, C., Chen, T., Li, S. Z., Chang, E., and Zhou, J.-L. (**2001**), "Analysis of speaker variability," in *INTERSPEECH*, edited by P. Dalsgaard, B. Lindberg, H. Benner, and Z.-H. Tan (ISCA), pp. 1377–1380.

Hunt, M. J. (**1999**), "Spectral Signal Processing for ASR," in *IEEE Workshop on Automatic Speech Recognition and Understanding*.

Jankowski, J., Charles R., Vo, H.-D. H., and Lippmann, R. P. (**Jul. 1995**), "A comparison of signal processing front ends for automatic word recognition," Speech and Audio Processing, IEEE Transactions on **3**(4), 286 –293.

Jürgens, T. and Brand, T. (**2009**), "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," The Journal of the Acoustical Society of America **126**(5), 2635–2648, URL http://link.aip.org/link/?JAS/126/2635/1.

Kanedera, N., Arai, T., Hermansky, H., and Pavel, M. (**1999**), "On the relative importance of various components of the modulation spectrum for automatic speech recognition," Speech Communication **28**(1), 43–55.

Khayam, S. A. (**2003**), "The discrete cosine transform DCT: theory and application," Tech. rep., Michigan State University.

Kleinschmidt, M., Tchorz, J., and Kollmeier, B. (**2001**), "Combining speech enhancement and auditory feature extraction for robust speech recognition," Speech Communication **34**(1-2), 75–91.

Kolossa, D. (**2007**), "Independent component analysis for environmentally robust speech recognition," Ph.D. thesis, Technischen Universität Berlin.

Ladefoged, P. (**2005**), *Vowels and consonants: an introduction to the sounds of languages*, no. v. 1 in Vowels and Consonants: An Introduction to the Sounds of Languages (Blackwell Pub.), URL http://books.google.com/books?id=cO9yDkqS1Y0C.

Langner, G. and Schreiner, C. E. (**1988**), "Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms," Journal of Neurophysiology **60**(6), 1799.

Leonard, G. R. (**Mar. 1984**), "A database for speaker-independent digit recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.*, vol. 9, pp. 328 – 331.

Lippmann, R. P. (**July 1997**), "Speech recognition by machines and humans," Speech Communication **22**(1), 1–15, URL http://dx.doi.org/10.1016/S0167-6393(97)00021-6.

Markel, J. D. and Gray, A. H., Jr. (**1976**), *Linear Prediction of Speech Signals* (Springer, Berlin).

Morgan, N., Bourlard, H., and Hermansky, H. (**2004**), "Automatic Speech Recognition: An Auditory Perspective," in *Speech Processing in the Auditory System*, edited by R. R. Fay and A. N. Popper (Springer New York), vol. 18 of *Springer Handbook of Auditory Research*, pp. 309–338, URL http://dx.doi.org/10.1007/0-387-21575-1_6.

Morgan, N. and Hermansky, H. (**10-13 November 1992**), "RASTA extensions: Robustness to additive and convolutional noise," in *ETRW on Speech Processing in Adverse Conditions* (Cannes-Mandelieu, France).

Nadeu, C., Hernando, J., and Gorricho, M. (**1995**), "On the decorrelation of filter-bank energies in speech recognition," in *Eurospeech'1995*, pp. 1381–1384.

Nadeu, C., Macho, D., and Hernando, J. (**2001**), "Time and frequency filtering of filter-bank energies for robust HMM speech recognition," Speech Communication **34**(1-2), 93 – 114, URL http://www.sciencedirect.com/science/article/pii/S0167639300000480, noise Robust ASR.

Oppenheim, A. V. and Schafer, R. W. (**1975**), *Digital signal processing / Alan V. Oppenheim Ronald W. Schafer* (Prentice-Hall, Englewood Cliffs, N.J. :).

Paliwal, K. K. (**1999**), "Decorrelated and Liftered Filter-Bank Energies for Robust Speech Recognition," in *EUROSPEECH'99* (Budapest, Hungary).

Palomäki, K. J., Brown, G. J., and Barker, J. R. (**2004**), "Techniques for handling convolutional distortion with [] missing data'automatic speech recognition," Speech communication **43**(1-2), 123–142.

Palomäki, K. J., Brown, G. J., and Barker, J. R. (**2006**), "Recognition of reverberant speech using full cepstral features and spectral missing data," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on* (IEEE), vol. 1, pp. I–I.

Patterson, R., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1988**), "An efficient auditory filterbank based on the gammatone function," APU report **2341**.

Pearce, D. and Hirsch, H.-G. (**16-20 October 2000**), "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," in *Sixth International Conference on Spoken Language Processing (ICSLP 2000)* (Beijing, China), pp. 29–32.

Picone, J. W. (**sep 1993**), "Signal modeling techniques in speech recognition," Proceedings of the IEEE **81**(9), 1215 –1247.

Plack, C. J. (**2005**), *The sense of hearing* (Lawrence Erlbaum Associates), URL http://books.google.dk/books?id=wpYSS8o0PeoC.

Proakis, J. G. and Manolakis, D. G. (**2007**), *Digital signal processing* (Pearson Prentice Hall), URL http://books.google.com/books?id=H_5SAAAAMAAJ.

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

Rabiner, L. R., Wilpon, J. G., and Soong, F. K. (**apr 1988**), "High performance connected digit recognition, using hidden Markov models," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pp. 119 –122 vol.1.

Sakoe, H. and Chiba, S. (**1978**), "Dynamic programming algorithm optimization for spoken word recognition," IEEE Transactions on Acoustics, Speech and Signal Processing **26**(1), 43–49, URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1163055.

Sanchez, V., García, P., Peinado, A. M., Segura, J. C., and Rubio, A. J. (**nov 1995**), "Diagonalizing properties of the discrete cosine transforms," Signal Processing, IEEE Transactions on **43**(11), 2631 –2641.

Sheskin, D. (**2004**), *Handbook of parametric and nonparametric statistical procedures* (Chapman & Hall/CRC), URL http://books.google.com/books?id=bmwhcJqq01cC.

Slaney, M. (**1993**), "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," in *USENIX Technical Conference*.

Smith, R. L. (**1976**), "Short-term adaptation in single auditory-nerve fibers-some poststimulatory effects," The Journal of the Acoustical Society of America **59**(S1), S16–S16, URL http://link.aip.org/link/?JAS/59/S16/6.

Sroka, J. J. and Braida, L. D. (**2005**), "Human and machine consonant recognition," Speech Communication **45**(4), 401 – 423, URL http://www.sciencedirect.com/science/article/B6V1C-4F7VNF5-1/2/34573b357646087ddf2edb0a34b2de70.

Stevens, S. S., Volkmann, J., and Newman, E. B. (**1937**), "A Scale for the Measurement of the Psychological Magnitude Pitch," The Journal of the Acoustical Society of America **8**(3), 185–190, URL http://link.aip.org/link/?JAS/8/185/1.

Stockham, J., Thomas G., Cannon, T. M., and Ingebretsen, R. B. (**april 1975**), "Blind deconvolution through digital signal processing," Proceedings of the IEEE **63**(4), 678 – 692.

Tchorz, J. and Kollmeier, B. (**1999**), "A model of auditory perception as front end for automatic speech recognition," The Journal of the Acoustical Society of America **106**(4), 2040–2050, URL http://link.aip.org/link/?JAS/106/2040/1.

Viemeister, N. F. (**1979**), "Temporal modulation transfer functions based upon modulation thresholds," The Journal of the Acoustical

Society of America **66**(5), 1364–1380, URL http://link.aip.org/link/?JAS/66/1364/1.

Westerman, L. A. and Smith, R. L. (**1984**), "Rapid and short-term adaptation in auditory nerve responses," Hearing Research **15**(3), 249 – 260, URL http://www.sciencedirect.com/science/article/pii/0378595584900327.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (**2006**), *The HTK Book, version 3.4* (Cambridge University Engineering Department, Cambridge, UK).

Zhao, S. Y. and Morgan, N. (**2008**), "Multi-stream spectro-temporal features for robust speech recognition," in *Ninth Annual Conference of the International Speech Communication Association*.

Zhao, S. Y., Ravuri, S., and Morgan, N. (**2009**), "Multi-stream to many-stream: Using spectro-temporal features for asr," in *Tenth Annual Conference of the International Speech Communication Association*.