Speaker Verification
in a
Time-Feature Space

Sarel van Vuuren
M. Eng., Computer Engineering, University of Pretoria,
Pretoria, South Africa, 1994
B. Eng. Electronic Engineering, University of Pretoria,
Pretoria, South Africa, 1991

A dissertation submitted to the faculty of the
Oregon Graduate Institute of Science and Technology
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Electrical and Computer Engineering

March 1999

The dissertation "Speaker Verification in a Time-Feature Space" by Sarel van Vuuren has been examined and approved by the following Examination Committee:

Dr. Hynek Hermansky
Professor
Thesis Research Adviser

Dr. Chin-Hui Lee
Department Head
Dialogue Systems Research Department
Bell Laboratories, Lucent Technologies

Dr. Douglas Reynolds
Senior Member of Technical Staff
Information Systems Technology Group
MIT Lincoln Laboratory

Dr. Michael Macon
Assistant Professor

Dedication

Wei Wei

Acknowledgments

I would like to express my foremost thanks to my thesis advisor, Dr. Hynek Hermansky, for taking me on as his student, and for instilling in me the idea for a medium-term processing of speech. Hynek's valuable suggestions and ideas have been a great source of inspiration for this dissertation. I would like to express my gratitude to Dr. C-H. Lee, Dr. M. Macon and Dr. D. Reynolds for serving on my thesis committee and for reviewing this dissertation. I appreciate their valuable comments and suggestions and wish to thank them for their setting a high standard.

I would like to thank many others as well. I would like to thank Dr. E. Barnard for his helpful suggestions and guidance in the early stages of this thesis and for bringing my attention to the field of speaker verification. I would like to extend my thanks also to Dr. T. Leen, Dr. M. Pavel and Dr. B. Yegnanarayana, for all their valuable inputs over the years. And I would like to thank the CSLU toolkit group: J. de Villiers, Dr. M. Fanty, J. Schalkwyk, and Dr. P. Vermeulen. It's been fun working with you all and thank you for all your help and advice. I'd like to thank N. Jain, who in the early years shared an office with me and whose optimism was very motivational. And of course, as with any large undertaking, I would like to thank the members of the Anthropic Signal Processing Lab and CSLU, with a special thanks to Dr. T. Arai, Dr. C. Avendano and Dr. H. Yang, for your collaborations, inspiration and helpful suggestions.

My deepest gratitude goes to Dr. W. Wei. She has been the source of unwavering support, strength and inspiration that has helped to bring this thesis to fruition. She has contributed many an hour encouraging and helping me in this endeavor and it is to her that I dedicate this dissertation.

Special thanks go to my parents for their support and encouragement through the many years of this and previous endeavors. Finally, I would like to thank the sponsors and organizations that have helped to support my graduate studies, my family and friends, my fellow students and faculty – both here and abroad.

SVV, January 1999.

Contents

List of Tables

List of Figures

Abstract

Speaker Verification
in a
Time-Feature Space

Sarel van Vuuren

Supervising Professor: Dr. Hynek Hermansky

The goal of this dissertation is to determine the relative importance of components of the modulation spectrum for automatic speaker verification and to use this knowledge to improve the performance of an automatic speaker verification system. It is proposed that the power spectrum of a *time sequence* of logarithmic energy, called the *modulation spectrum*, provide information that may be used to reduce the effects of adverse environments. The proposed strategy is to attenuate spectral components that are not particularly useful for speaker verification. The aim is to reduce system sensitivity to telephone handset variability without reducing verification accuracy.

By computing the effect of carbon-button and electret microphone transducers on the modulation spectrum of telephone speech, it is found that handset transducer variability accounts for a substantial portion of the total variability at moderate to high modulation frequencies. This is shown to be the case also at very low modulation frequencies, where variability is ascribed to the effect of a convolutional channel. This result is substantiated with verification results on the Switchboard corpora as used in 1997-1998 NIST speaker recognition evaluations. The main conclusion is that components of the modulation spectrum between 0.1 Hz and 10 Hz contain the most useful information for speaker verification.

To deal with adverse environments, it is proposed that the time sequences of logarithmic energy be lowpass filtered. When compared to other filtering techniques such as cepstral mean subtraction that may retain components up to 50 Hz or RASTA processing that retains components between 1 Hz and 13 Hz, lowpass filtering to 10 Hz is found to significantly reduce verification error in conditions where handset transducers differ between training and testing. It is furthermore proposed that the feature stream be sampled down from a 100 Hz sampling rate to as low as a 25 Hz sampling rate after lowpass filtering. Using this processing, a relative reduction in error of about 10% is shown for the 1997 and 1998 NIST speaker recognition evaluations. Additional contributions of the dissertation include the design and implementation of a modular, high-performance speaker recognition toolkit.

Chapter 1

Introduction

Speech conveys information on several levels. It contains a message generically expressed as a sequence of words, information specific to the speaker that produced the speech, and information about the environment in which the speech was produced and transmitted. Speaker specific information include the identity of the speaker, the gender of the speaker, the language or dialect of the speaker and possibly the physical and emotional condition of the speaker. With this richness of information it comes as no surprise that, with the advent of computers, speech has found wide-spread application in human-computer communication. In particular, *automatic speech recognition* is the process of extracting the underlying message and *automatic speaker recognition* is the process of verifying the identity of the speaker. Applications range from using voice commands over the telephone to control financial transactions and verifying the identity of the speaker, to continuous dictation and speaker detection in multi-party dialogues. The application generally dictates the types of information in the speech signal that are useful. For example, for the purpose of extracting the underlying message in automatic speech recognition, the presence of speaker and environmental information may actually lead to confusions and degrade system accuracy. Similarly, message and environmental information may degrade speaker recognition accuracy. For an application to be successful, an accurate modeling of the desired type of information is therefore important.

1.1 Speaker Verification

Speaker verification can be considered within the wider context of *speaker recognition*. Speaker recognition collectively describes the tasks of extracting or verifying the identity of the speaker [4, 20]. In *speaker identification*, the task is to use a speech sample to select the identity of the person that produced the speech from among a set of candidate identities, or population of speakers. This task involves classification from N-possibilities, where $N > 1$ is the population of speakers. In *speaker verification*, the task is to use a speech sample to test whether a person who claims to have produced the speech did in fact do so. This task involves a two-way classification which is a test of whether the claim is correct or not. In speaker identification the number of possible choices are the number of speakers in the population, whereas in speaker verification the outcome is limited to one of two choices. *Closed-set* speaker identification is the task where every speaker in the population is known to the system at the time of use. *Open-set* identification is the task where some speakers in the population are unknown to the system at the time of use and hence must be rejected on the basis of being unknown. Open-set identification is therefore a combination of closed-set identification and speaker verification. An example where speaker identification has found use is audio indexing, which involves the automatic detection and tagging of speakers in a small multi-party dialogue. In this dissertation the focus will be on the task of speaker verification, but it should be understood that the techniques investigated here can be readily applied to speaker identification.

Taking a broader view, speaker identification and verification themselves can be placed in the field of *biometric* identification and verification [14], where the goal is to use any of a number of person-specific cues to classify that person. Examples of commonly used cues are as diverse as a facial image [96], iris pattern, finger print, genetic material or even keyboard typing pattern. The advantage of using a biometric cue for access control is that it is always accessible, unlike a key or password that can be misplaced, forgotten or stolen.

Using a speaker recognition system is usually a two-step process [27]. The user first *enrolls* by providing the system (computer) with one or more representative samples of his or her speech. These *training* samples are then used by the system to *train* (construct) a *model* for the user. In the second step the user provides a *test* sample that is used by the system to *test* the similarity of the speech to the model(s) of the user(s) and provide the required service. In this second step the speaker associated with the model that is being tested is termed the *target speaker* or *claimant* [60].

In speaker verification, when the person is constrained to speak the same text during both training and testing the task is *text-dependent* [27]. For example, the verification phrase may be a unique password or a fixed string of digits. Applications requiring access control, such as voice-mail, telephone banking and credit card transactions have successfully used this type of verification [14, 11]. A similar system using fixed phrases is currently being tested at a US border crossing at Otay Mesa, in San Diego, California, that would allow frequent travelers to gain clearance by speaking into a hand-held computer inside the car. While text-dependent verification potentially requires only a small amount of speech it requires the user to faithfully produce the required text. As such it requires a *cooperative* user and a structured interaction between the user and system [14]. When the person is not constrained to speak the same text during training and testing the task is *text-independent* [27]. This is required in many applications where the user may be uncooperative or applications where speaker recognition occurs as a secondary process unknown to the speaker as in audio indexing. For example, a forensic application may require verifying the identity of a speaker based on speech from a recorded telephone conversation and the speaker may not actually be aware of this process. In both text-dependent and text-independent modes of operation the verification decision can be sequentially refined as more speech is input until a desired significance level is reached [55, 27, 25]. The word "authentication" has sometimes been used for "verification" and "talker" or "voice" for "speaker". Similarly, "text-free" has been used for "text-independent" and "fixed-text" for "text-dependent" [27].

A block diagram of the major stages in a speaker verification system is shown in Fig. 1.1. First is the acquisition stage, where the speech produced by the speaker is converted from a sound pressure waveform into an electrical signal using a transducer. This *acoustic signal* is digitized and sampled at a suitable rate. Second is the signal processing

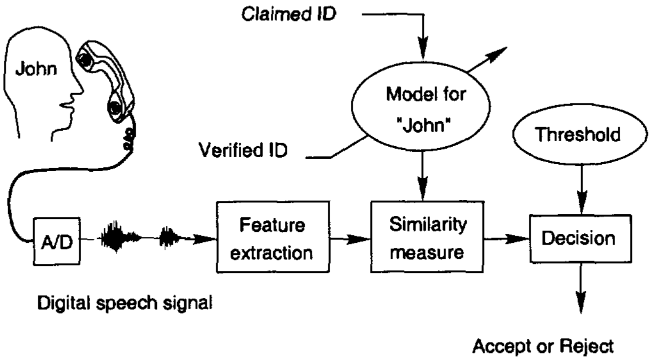


Figure 1.1: Block diagram of the major processes in a speaker verification system.

and *feature extraction* stage, where salient parameters conveying speaker identity are extracted from the acoustic speech signal. Design of the feature extraction stage is based on the existing body of knowledge of the speech process – such as models of the articulatory and auditory systems [67, 37], theory of linguistics and phonetics [46], perceptual cues used by listeners [102, 22], transmission process [76], and application specific requirements. The third stage involves computing a similarity measure [25] between the information retrieved from the speech of the current speaker and a previously constructed model representing the person the speaker claims to be. The model training (construction) forms a major component of the speaker verification system. It determines storage cost and computation and dictates accuracy of the similarity measure. The fourth and final stage is to compare the similarity measure to a predetermined value or threshold and decide whether to accept or reject the claimed identity of the speaker. In this last stage for example, if the model of the claimed speaker is deemed to represent the information retrieved from the acoustic signal accurately, i.e. the two are similar, then the decision is to accept the claim made by the speaker.

There has been, and continues to be, a great deal of interest in speaker verification with a vast number of speaker specific cues, feature extraction techniques, modeling techniques, and evaluation measures proposed. These are covered in a number of tutorial papers [5, 84, 20, 29, 27, 14, 21, 48]. Recently a number of speaker verification systems have also been deployed commercially. Examples include systems from ITT, Lernout & Hauspie, T-NETIX, Veritel, Texas Instruments, Voice Control Systems and Nuance Corporation [14].

A speaker verification system has to have certain characteristics to be useful. Obviously, for a specified mode of operation it is desirable that the system be accurate and consistent in its performance. An important characteristic is that the system should be relatively insensitive or *robust* to adverse environmental disturbances such as distortions introduced by the transmission channel. Furthermore, a system that can make accurate decisions based on a small sample of speech would be preferable to a system requiring a large sample of speech, since acquiring a large sample may be annoying to the user. As discussed previously, depending on the application, another useful characteristic is that of text-independent operation. Other useful characteristics from a practical point of view are that the system should be fast, operate in real time, be extendible (e.g. allow improvements) and be scalable (e.g. allow new users to be added at any time).

In the important case of speech having been spoken into a telephone handset and transmitted over a telephone network, robustness to environmental changes becomes an important issue [20]. The term *environment* will be used rather liberally here to collectively refer to effects specific to the environment in which the speech was produced – such as ambient noise and the lombard effect, and to effects specific to the transmission of the speech – such as contributed by handset and channel. Robustness to environmental changes are important since a call from a cellular telephone instead of an office telephone, for example, may cause a machine to falsely reject a speaker.

1.2 Analysis of Speech in a Time-Feature Space

To better understand the effect of the environment it is necessary to first consider the nature of the acoustic speech signal. The acoustic speech signal is produced by exciting the vocal tract system of the speaker with a wide-band excitation. The vocal tract changes shape relatively slowly with time and thus can be modeled as a *slowly time-varying filter* that imposes its frequency response on the spectrum of the excitation. For the time-varying filter, fixed (stationary) properties over a time interval of 10-40 ms can be assumed [4, 76]. Over this short time interval the vocal tract shape can be characterized by its natural frequencies (called formants) which correspond to resonances in its *frequency response*.

The acoustic speech signal, which is a measure of the changes in acoustic pressure at the mouth opening, can then be understood to reflect the excitation and shape of the vocal tract due to the movement of the speech articulators (such as the tongue and lips).

The short-term assumption can be used to analyze the speech signal in a *time-feature* space. An example of a short-term analysis is the well-known behavior of a graphic equalizer found in some sound systems. At a given time instant the graphic equalizer may display the energy for different frequency components in the speech signal as vertical bars. Over time the lengths of these bars change, reflecting the change in energy for that frequency component and the *non-stationary* nature of speech.

In the *short-term analysis* of speech, the speech signal is segmented into short segments that are individually analyzed and/or modeled. A segment is usually represented or decomposed in terms of its frequency components or *spectrum.* This short-term analysis of speech has been used successfully in a large number of automatic speech and speaker recognition systems as a basic feature extraction step [14]. In the case of a spectral representation, the short-term analysis produces a two-dimensional signal in time and frequency, where the time dimension refers to the segment that is being analyzed and the frequency dimension to its spectral components. This is commonly displayed as a *spectrogram.* Thus the two-dimensional signal can be viewed as a sequence of *frames* or *feature vectors* with each feature vector indexed by the time dimension and formed by the spectral components of the signal at that particular point in time – see Fig. 1.2 (a). The sequence of feature vectors is sometimes referred to as a *feature stream.* Each individual spectral component or feature in the feature stream can then be seen to describe a one-dimensional signal in time, or *time sequence* as it will be called – see Fig. 1.2 (b). *Medium-term analysis*, which is the analysis of each of these time sequences over an interval of time extending beyond that of short-term analysis, forms the basis of this dissertation. Time sequences of a number of different feature representations will be considered but the focus will be mainly on *time sequences of logarithmic spectral energy.* In general, since the representation will be clear from the context, these representations will sometimes also be referred to as *time sequences of spectral features*, *time sequences of energy*, *time sequences*, or simply *sequences.*

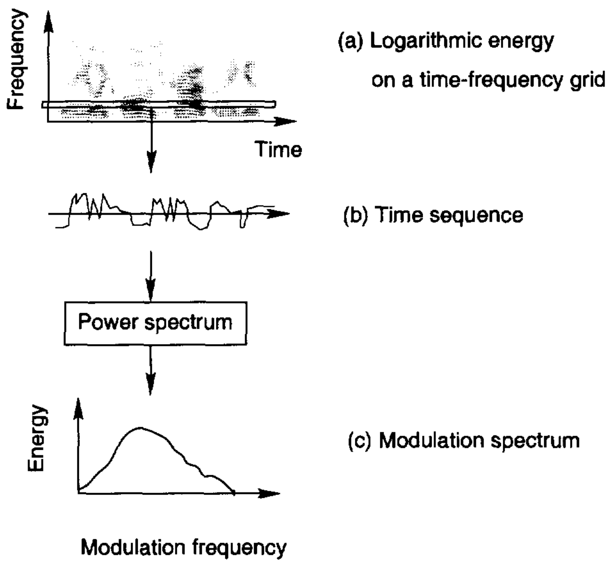


Figure 1.2: Representing speech in a time-feature space.

The *power spectrum* of each time sequence – see Fig. 1.2 (c) – is known as its *modulation spectrum* [41] and is considered to convey important characteristics of speech [41, 22, 2, 36]. For example, dominant components in the modulation spectrum of speech have been associated with average syllabic and phonetic rates [22, 2, 36].

1.3 Adverse Environments

It is well known that adverse environments, such as present with the use of different telephone handset transducers, affect the time sequences of the speech signal. For example, assuming that the environment acts like a time-invariant filter, it has an approximately constant multiplicative effect on the short-term frequency response [4, 76, 26]. In general however, the environment may be non-linear, time-varying, noisy and not well modeled [7].

Given that the environment affects the time sequences, one way to gain an understanding of the effects is to analyze the environment in terms of its modulation spectrum and compare this to the modulation spectrum of speech. In this dissertation, the strategy will be to determine the relative importance of the components in the modulation spectrum for speaker verification. The view will be that *attenuation* of less important components, such as components that are overly affected by the environment or that do not actually convey useful speaker information may improve performance both in terms of verification accuracy and system speed. The motivation for this view stems from the following argument [36]. Human speech communication is a highly specialized process and constrained by the organs that are involved. The process involves a source (organs of speech production), a transmission channel (environment), and a receiver (organs of speech perception). For optimal communication, these components have to be in tune with each other. It is likely that nature may have designed the speech communication process in a way that alleviates or avoids the variability inherent in the transmission channel. If, for example, evidence exists that certain modulation frequency components are more important than others for perception, then this knowledge should guide system design. Conversely, if the transmission channel can be implicated in contributing highly and variably to certain modulation frequency components, compared to the contribution of the speech production process, then the attenuation or perhaps even removal of those modulation frequency components may be warranted and lead to improved performance.

1.4 Dealing with Adverse Environments

In the previous section it was proposed that a possible strategy for dealing with adverse environments may be to attenuate or deemphasize the redundant and overly noisy information in the speech signal. This strategy can be compared to some alternative strategies [75] that deal with adverse environments.

In ASR for example, when the adverse environment includes speaker variability, one popular strategy is to *adapt* to the speaker and environment[1]. An example is the so-called *stochastic matching* technique where the idea is to adapt the models or features to the test environment and thus reduce mismatch that may have existed between the training and test environments. In this technique the models are transformed by maximizing the data likelihood [95]. The maximization is used to find the parameters of a transformation function that describes the environmental disturbance. Linear transformation

[1] Adaptation techniques fall outside of the scope of this dissertation and will be reviewed only briefly in this section.

functions have been popular [50] and used successfully, while non-linear transformation functions have also been investigated [95, 77]. In general the adaptation techniques require that the transformation function matches the environmental disturbance and that the transformation will not map different models to each other. The latter requirement is necessary to preserve model uniqueness and discriminability. Adaptation to the transmission channel using a maximum likelihood linear regression (MLLR) [50] has been tried for text-independent speaker verification [57], but was reported to be unsuccessful.

We speculate why this may be the case. In an analysis of variance (ANOVA) decomposition of high-quality speech from the TIMIT corpus [94], it has been observed that while intra- and inter-phonetic variability may account for as much as 60% of the total variability in the speech, the speaker variability (including that due to dialect and gender) accounts for only about 10% of the total variability[2]. The variability (differences) between the models for two speakers may therefore be small relative to adverse sources of variability, which, in the case of text-independent speaker verification, would include phonetic and environmental variability. It has also been observed that dominant speaker and environmental variations may actually be quite similar. For example, it is known that the long term average spectrum of speech contains speaker information, but also that this average may be influenced by the transmission channel. These observations imply that the requirement that the transformation will not map different models to each other, may not be met in the case of speaker verification.

In contrast with an adaptation strategy, where values of parameters for the adverse environment have to be estimated from the test data [36], the attenuation or deemphasis strategy attempts to localize and contain the environmental degradation, but not to measure it. This suggests a possible advantage for the attenuation strategy in dealing with unknown variability.

The attenuation or deemphasis of redundant information as a strategy to improve performance when there is a mismatch of training and testing environments, such as with the use of different telephone handsets, may also be understood as a particular form of

[2] We observed similar contributions in other corpora such as the OGI-TS (stories) corpus of continuous telephone speech and the NTIMIT corpus of telephone quality speech.

regularization. Regularization [83] is motivated from a Bayesian point of view [25, 23] and deals with the issue of controlling feature and modeling complexity. Regularization is known to improve system performance or *generalization* ability when there is a mismatch between training and testing environments (see [98] for an analysis and discussion). The improvement results from a suitable choice of a *prior* probability distribution function for the features that deemphasizes aspects of the features that may be deemed unimportant while emphasizing important aspects, such as smoothness. As an extreme case of this regularization, the prior could be chosen to effectively *remove* certain aspects of the features which may be considered redundant or noisy.

1.5 Outline

The dissertation is organized into three parts. The first part reviews, analyzes and motivates techniques for the processing of speech by characterizing different sources of variability in telephone speech in a time-feature space. This part of the dissertation presents a rather general treatment of telephone handset variability in speech and as such does not specifically deal with speaker variability. It does serve however to indirectly motivate and guide the development of a proposed linear filtering of the time sequences of logarithmic energy that would attenuate unwanted variability in the speech signal. Whereas the first part was concerned with the effect of telephone handset variability in speech in general, the second and third parts narrow the focus to the speaker verification task specifically. The second part covers the motivation, design and specification of a text-independent speaker verification system that incorporates the proposed filtering. The third part presents a systematic investigation of the relative importance of the components of the modulation spectrum for speaker verification followed by an exploration for the usefulness of the proposed lowpass filtering for speaker verification. The aim is to find a filter or filters that, when applied to the time sequences of logarithmic energy to generate features, would improve speaker verification performance in terms of verification accuracy and/or computational cost.

1.5.1 Outline by Chapter

Chapter 2 covers acoustic feature extraction and processing in a time-feature space. The main aspect of this processing is a linear filtering of the time sequences of spectral energy. In the chapter, short-term acoustic features are first motivated based on perceptual and physiological considerations. Next, the theory of short-term analysis of the speech signal is reviewed along with common feature representations used in ASR and speaker verification. The *modulation spectral domain* is then defined and introduced as a domain in which to study and manipulate these short-term features. Various practical and theoretical issues of the analysis are examined. The problem of *acoustic mismatch* in automatic speaker verification is then examined and existing methods for its alleviation reviewed. As a general strategy, it is proposed that filtering of the short-term features be employed as a processing technique for alleviating acoustic mismatch in adverse environments.

Chapter 3 explores the characteristics of the short-term features in the modulation spectral domain. As expected from a convolutional model for the transmission channel, it is shown that telephone handset variability severely contaminates the DC-modulation component. Importantly, it is also shown that the moderate to high modulation frequency components are severely contaminated by handset variability. The result is obtained by computing the variability in speech due to carbon-button and electret microphone transducers and comparing it to the overall variability in speech. The computation is based on an analysis-of-variance model (ANOVA). Speaker specific characteristics are not explored in this chapter, but rather handset variability is contrasted to the overall speech variability to obtain an indication of where and how handset variability may be affecting the recorded speech. Whether the observed variability is actually relevant to speaker verification in particular, is tested later in Chapter 5.

Chapter 4 describes the feature extraction, modeling and evaluation measures used for speaker verification in this dissertation. Speaker verification is formulated as a problem in statistical hypothesis testing and a test statistic based on two probability density distribution functions (pdfs) defined. The decision of whether to accept or reject the claim of a speaker is made by comparing the test statistic to a global threshold. One pdf describes *speaker independent* (SI) features and the other describes *speaker dependent* (SD) features. A Gaussian mixture modeling approach is adopted based on statistical considerations of the features and a review of existing modeling approaches. The well-known Expectation-Maximization algorithm is used to estimate the parameters in the SI model and Bayesian maximum aposteriori (MAP) adaptation of the SI model is used to derive the SD models. Various results related to optimizations of the feature and modeling parameters are presented. Speech data and various training and testing conditions similar to recent NIST Speaker Recognition Evaluations (NIST-SRE) are used. Descriptions of the NIST-SREs and evaluation plans can be found in [72, 73, 60] and NIST's URL at `http://www.nist.gov/speech`. Appendix A presents a detailed description of the setup used in this dissertation.

Chapter 5 presents a further systematic investigation of the relative importance of the components of the modulation spectrum for speaker verification. This investigation for speaker verification specifically, is to be contrasted to the more general investigation speech and handset variability that is presented in Chapter 3. In Chapter 5, an analysis of the error surface is proposed to confirm the observation that higher modulation frequencies are less important for speaker verification. The approach is to measure and analyze the effect on the speaker verification error for various filters designed in the modulation spectral domain and applied in the time-feature space. The choice of filters and effect of down sampling of the time sequences of spectral features are further investigated, based on a finding that these time sequences can be lowpass filtered without degradation in performance. The findings are supported with results from the official 1998 NIST-SRE [59].

Chapter 6 summarizes the major results, conclusions and contributions of this dissertation and ends with suggested directions for future research.

1.5.2 Outline by Original Contribution

Chapter 3 presents a novel framework for the study and characterization of handset transducers in the modulation spectral domain. The framework incorporates an analysis-of-variance (ANOVA) that was modified to allow an interpretation at different modulation frequencies, and allows different sources of variability to be modeled in the speech signal.

Chapter 4, provides an optimization study of the salient parameters in a state-of-the-art speaker verification system.

Chapter 5 provides a systematic investigation of the relative importance of the components of the modulation spectrum for speaker verification as well as a processing strategy of lowpass filtering for alleviating the effects of environmental mismatch. To the best of our knowledge, the modulation spectrum has not been used before to characterize speaker verification performance in a time-feature space as is done here. The analysis contributes to an understanding of the effects and usefulness of contemporary processing techniques such as CMS and RASTA. Importantly, the chapter includes also the proposal for a reduction of the frame rate – from a traditional 100 Hz to as low as 25 Hz. The benefits of such processing for speaker verification have not been demonstrated before.

Appendix C provides a discussion and application of McNemar's significance test [28] that as far as we know is not commonly used in speaker verification.

Appendix D describes a modular and efficient speaker recognition toolkit build around a script language that facilitates rapid prototyping. This toolkit has contributed substantially to the speaker verification and ASR research effort in our laboratory and elsewhere. The toolkit and parts of it have been used by IIT Madras and CSLU among others.

Appendix E describes the original use of linear discriminant analysis (LDA) in the automatic derivation of FIR filters that optimizes phoneme discriminability for ASR.

Chapter 2

Feature Extraction in a Time-Feature Space

The purpose of this chapter is to review and examine acoustic feature extraction and processing in a time-feature space. The main aspect of this processing is a linear filtering of the time sequences of spectral features. The acoustic feature extraction is considered for its usefulness in adverse environments. In Section 2.1, short-term acoustic features are first motivated based on perceptual, physiological and acoustic considerations. Short-term analysis of the speech signal is then reviewed and discussed in Section 2.2, followed by a review of common feature representations used in ASR and speaker verification in Section 2.3. Section 2.4 extends the short-term analysis to a medium-term analysis. The concepts of *modulation frequency* and *modulation spectrum* are defined and introduced in terms of their usefulness for the study and manipulation of the short-term features. The effects of the length of the short-term analysis window, analysis sampling rate and transmission channel on the modulation spectrum of speech is subsequently examined. The usefulness of the modulation spectrum becomes apparent in Section 2.5 where the problem of *acoustic mismatch* is considered. This problem is examined and existing methods for its alleviation reviewed. The acoustic mismatch is considered as a degradation of the speech signal in an adverse environment and compensated for by filtering of the short-term features. Results from a small experimental study are described that highlight the problem of acoustic mismatch in speaker verification.

2.1 Perceptual and Physiological Bases

Before considering various feature extraction techniques in this chapter, and various modeling techniques in Chapter 4, it is appropriate to review the perceptual and physiological bases for speaker recognition. The purpose is to use perceptual, physiological and acoustic considerations to guide the choice of features and to use human listener performance as a guide-line to the performance of an automatic system.

One way to determine which acoustic features might successfully represent speaker identity is to examine correlates between the acoustic signal and perceptual cues used by listeners [102, 22]. Another way is to examine the speech production system for possible speaker dependent attributes that may be manifested in the speech signal [67, 86].

2.1.1 High- and low-level cues

Speaker-specific information can be categorized in two levels [86]. *High-level cues* are linguistic, semantic and to a lesser extent prosodic in nature and derive from word usage, pronunciation, habit and manner. *Low-level cues* are acoustical in nature and describe for example loudness, speaking rate, clarity, nasality and breathiness. The high-level cues are associated with behavioral characteristics of the speaker and, on account of being difficult to quantify, generally not well understood [86]. The low-level cues are associated with physical and structural characteristics of the speaker and comparatively easier to quantify. Most studies and approaches have focused on the low-level cues [5].

2.1.2 Physiological attributes

The human speech production system is depicted in Fig. 2.1. Speech is driven by an excitation source, generating airflow from the lungs, through the trachea and through the *vocal folds.* Phonation occurs when the vocal folds modulate the airflow by alternatingly opening and closing under tension and air pressure. This results in a stream of air pulses that excites the *vocal tract*, causing it to resonate at its characteristic frequencies. This happens in the production of *voiced* sounds such as *vowels*. The characteristic frequencies are often referred to as *formant* frequencies. The formant frequencies can be modified by

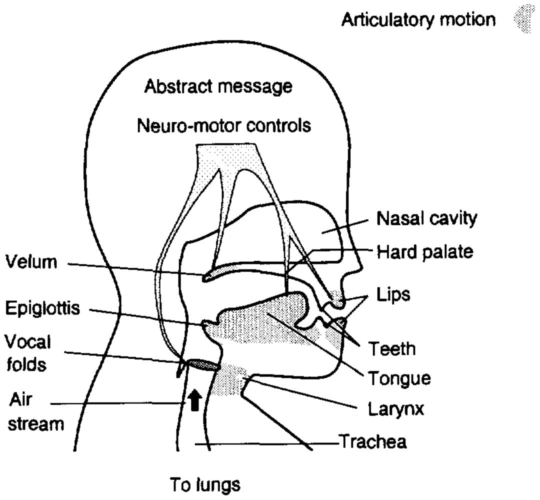


Figure 2.1: Human speech production system.

changing the configuration of the vocal tract by the process of *articulation*. This happens when any of the articulators (shaded in the figure) such as the lips, tongue and velum are moved. *Fricative* sounds are generated when the vocal tract is constricted in some location. The frication may be *voiced* or *unvoiced*, depending on whether the vocal folds modulated the airflow. An unmodulated airflow results in a wide-band excitation as air rushes through the vocal tract. *Plosive* or *affricate* sounds are formed if the constriction is suddenly or gradually released. Moving the velum to allow airflow to the nasal cavity, produces *nasalized* sounds.

A number of physiological attributes have been correlated with speaker identity and associated with the acoustic speech signal:

- Variations in the size of the vocal tract cavities produce differences in the spectrum of the speech signal. The length of the vocal tract affects the overall spectrum [4].

- Variations in the size of the vocal folds are associated with changes in the average pitch or fundamental frequency of speech [3].

- Variations in velum and size of nasal cavities produce spectral differences in nasalized speech sounds [86, 64].

- The configuration of the teeth and palate affects frication [64, 46].

- Behavioral traits such as speaking rate [64], breathing, nasalization and dialect affect the speech signal.

Pitch has proven to be susceptible to mimicry [84], but spectral features associated with the vocal tract have been more successful in automatic speaker verification [14]. The temporal characteristics of such spectral features are investigated in this dissertation.

When sounds such as phonemes and syllables are articulated, the vocal tract changes shape relatively slowly with time and thus can be modeled as a *slowly time-varying filter* that imposes its frequency response properties on the spectrum of the excitation. For the time-varying filter, fixed (stationary) properties over a time-interval of 10-40 ms can be assumed [4, 76]. Spectral analysis of a speech signal over a short time interval can therefore provide features that point to the shape of the vocal tract. The acoustic speech signal, which quantifies the changes in acoustic pressure at the mouth opening, can then be understood to reflect the excitation, shape of the vocal tract and movement (rate and position) of the speech articulators (such as the tongue and lips), albeit in a complex way.

2.1.3 Perceptual cues

In the past, several studies have been aimed at identifying perceptual cues used by listeners. In [102], the approach was to ask 32 listeners to classify speakers according to a set of 49 differential factors such as "clear-hazy" and "fast-slow". Significant features were determined using a factor analysis. The most significant features were "clarity", "roughness", magnitude" and "animation". However, the features were not mapped to acoustic cues and given the difficulty of associating abstract features with the physical acoustic signal, the usefulness of the study remains questionable [20]. Later approaches have taken a more direct approach. In [103], desirable properties for cues, such as measurability, were delineated and investigated for discriminability using the F-ratio. Fundamental frequency, vowel spectra, glottal source spectrum slope, and word duration were found to be useful cues. Importantly, in this and other studies [4, 78] the use of spectral features among others have been advanced. Spectral features appear to be statistical in nature and not easily quantifiable. In [10], interesting observations are presented on the difficulties that humans have in visually verifying speakers from spectrograms of their speech.

A number of studies have examined temporal features and their importance for perception as well as their auditory relevance. Exploring the time-feature space generated by the time sequences of spectral features, [22] and [2] reported perceptual speech recognition results using the concept of the *modulation spectrum* [41]. The modulation spectrum may be viewed as the power spectrum of the time sequence of a short-term spectral feature. In the usual short-term analysis paradigm, where the short-term analysis may be performed every 10 ms, the highest frequency component in the time sequence of spectral features is 50 Hz according to the Nyquist sampling criterion [65]. In the perceptual studies, it was reported that the ability of listeners to recognize sounds such as consonant-vowel-consonant sequences are relatively little impaired if the speech signal is constrained to have modulation frequencies only in the range of 1 to 15 Hz. Employing a sensitivity analysis over modulation frequency, it was further shown that the components in the 2-6 Hz range may be the most important for *speech recognition.* In fact, components around 4 Hz have been associated with the syllabic and temporal structure of speech and taken as indicative of the dominant rate of speech and associated rate of change of the vocal tract shape [36]. This supports earlier studies. Hermansky et al. [37] refer to early experiments in [82, 105] and [31], which indicate a greater sensitivity of human hearing to modulation frequencies around 4 Hz than to lower (or higher) modulation frequencies.

These results have since been confirmed by a number of automatic speech recognition (ASR) experiments. In [40] it was shown that a highpass filtering of the time sequences improved performance in the presence of mismatched channels. In [37] and [44] it was shown that a bandpass filtering of the time sequences to a modulation frequency range of 1 to 15 Hz improved ASR performance in the presence of mismatched channels. In [37], the so-called RASTA-processing technique was designed as an engineering solution to improve speech recognition in adverse environments. The motivation for filtering the spectral features was that the rate of extra-linguistic changes may be outside of the typical rate of change of linguistic components.

It has been reported that such bandpass filtering may be justified in that it emulates properties of human hearing, such as temporal masking [67, 15, 37, 36]. In [44], recognizer performance was measured using a sensitivity analysis similar to the one used in [2], with findings in agreement with the perceptual studies. In [100], FIR filters to be applied to the time sequences of spectral features were designed directly from the data using a linear discriminant analysis (LDA). The filters that were designed on a telephone speech corpus of conversational speech to maximize phoneme discriminability exhibited bandpass frequency responses in close agreement with that suggested by the perceptual studies. Features derived from the filters outperformed a number of conventional features in recognition experiments on a telephone speech corpus of connected digits in matched conditions. While these experiments in the temporal domain were performed for the task of speech recognition, it is of interest to know whether the results pertain to speaker recognition as well. Whether the time sequences may indeed be constrained to contain only certain spectral components is the main topic of this dissertation.

2.1.4 Human performance

In a recent perceptual study using short speech segments of nominally 3 seconds duration, it was reported [89] that human listeners can perform the same or better than current state-of-the-art speaker verification systems. Although limited to small speaker populations due to a working memory capacity of about 7 items, humans were able to outperform machines in adverse conditions. Human performance was not very consistent however – performance varied greatly among individuals. The significance of this study is that it used speech from the 1998 NIST Speaker recognition evaluation corpus[1], speech that is also used for speaker verification in this dissertation. Human performance in adverse conditions was also reviewed in [20], where it was reported that human listeners are adept at using various cues to verify speakers in the presence of acoustic mismatch. Human performance must however, be interpreted with caution. In general, in human performance studies it is difficult to control all extraneous factors. This makes comparison to an automatic system difficult. The main conclusion to be drawn from the fact that humans can verify speakers well in an adverse environment, is that the human speech process may serve as a useful guide-line to improve the performance of machines in those conditions.

[1] For more information the reader may refer to NIST's URL at `http://www.nist.gov/speech`.

2.1.5 Sources of error

To complete the discussion on perceptual and physiological bases for speaker verification, it is appropriate to consider sources of speaker verification error. Typical sources of error manifest themselves over multiple sessions of use and include [14] the following:

- acoustic mismatch (e.g. different microphones for enrollment and verification, noise corruption, varying channels);
- microphone placement (time-varying intra- and inter-session);
- room acoustics (e.g. multi-path and noise);
- emotional state (e.g. stress, excitement and interest);
- sickness (e.g. head colds);
- aging (e.g. change in vocal tract);
- amount (duration) of test speech; and
- amount (duration and number of sessions) of enrollment speech.

Acoustic mismatch will be investigated in this dissertation.

2.2 Short-term Analysis of Speech

The *short-term Fourier transform* (STFT) has been used successfully in the analysis of speech [4, 65, 76]. The key idea is to treat the speech signal as quasi-stationary and examine spectra of the underlying time-varying process over short observation intervals of the signal using the Fourier Transform (FT). Given a speech signal $s(m)$ that is sampled at times t_m (i.e. the sequence is indexed by m), and a windowing sequence $w(m)$, centered at consecutive observation times t_n (indexed by n), the STFT is defined as

$$S(n,\omega) = \mathrm{FT}_m[w(n-m)s(m)] = \sum_{m=-\infty}^{\infty} w(n-m)s(m)e^{-j\omega m}. \tag{2.1}$$

The windowing sequence is commonly referred to as the *analysis window*, the windowed speech segment as a *frame*, and $S(n,\omega)$ as the short-term spectrum. The subscript m in

FT_m is used to denote the FT variable. The FT is used here for generality – in practice, the STFT is computed using the Fast Fourier Transform (FFT), i.e. ω is sampled at frequencies $2\pi k/K$, $1 \leq k \leq K$. The spectral vector resulting from the FFT, or a similar representation, is generated for each *frame* so that the sampling rate of n is commonly referred to as the *frame rate*. Components of the spectral vector or transformations thereof, are sometimes referred to as *features*. For the FT to exist, $w(n-m)s(m)$ must be absolutely summable [65] – a requirement that is met if the window is of finite duration. In the STFT, the *analysis frequency* in radians is $\omega = 2\pi f t_s$, where f is the analysis frequency in Hertz (Hz), $t_s = 1/f_s$ is the *sampling period* and f_s is the *sampling frequency*.

Equation (2.1) can be rewritten in terms of the FT of the unwindowed signal $s(n)$ and analysis window. Denoting the respective FTs of $s(m)$ and $w(m)$ as

$$S(\omega) = \sum_{m=-\infty}^{\infty} s(m)e^{-j\omega m}, \quad W(\omega) = \sum_{m=-\infty}^{\infty} w(m)e^{-j\omega m},$$

with the understanding that $S(\omega)$ pertains to the segment of the short-term analysis and that outside the analysis window it is either zero or periodic, it follows from the *modulation* or *windowing theorem* [65] that

$$\begin{aligned} S(n,\omega) &= \mathrm{FT}_m[w(n-m)s(m)] && (2.2) \\ &= \mathrm{FT}_m[w(n-m)] *_\omega \mathrm{FT}_m[s(m)] \\ &= \frac{1}{\pi}\int_{-\pi}^{\pi} W(\theta)e^{j\theta n}S(\omega+\theta)d\theta, && (2.3) \end{aligned}$$

where the $*_\omega$ operator represents linear convolution with respect to the analysis frequency ω. Equation (2.3) exemplifies the *frequency domain* interpretation of short-term analysis [76]. The FT of the speech signal is convolved with the FT of the analysis window, resulting in the STFT. The estimated spectrum is seen to provide an estimate for the speech spectrum, with the frequency resolution of the estimate limited by the bandwidth of the analysis window.

As is popular in speech processing, the Hamming window is used as the analysis window

$$w(m) = \begin{cases} 0.54 - 0.46\cos(2\pi m/(l_w - 1)), & 0 \leq m \leq l_w - 1 \\ 0 & \text{otherwise.} \end{cases} \qquad (2.4)$$

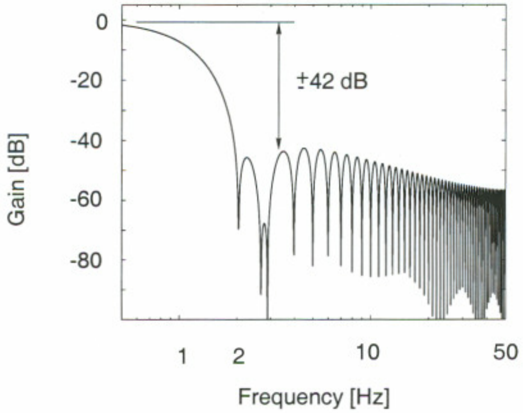


Figure 2.2: Frequency response of a 100-point Hamming window at a 100 Hz sampling rate.

Properties of the Hamming window and other analysis windows have been explored extensively [65] and will be of importance in latter chapters. The frequency response of a 100-point Hamming window at a 100 Hz sampling rate is depicted in Fig. 2.2. The bandwidth (in Hertz) of an l_w-point Hamming window, as defined by the location of the first zero, is

$$B = 2f_s/l_w, \tag{2.5}$$

which is twice that of a rectangular window. The Hamming window however offers the advantage of a lower side lobe (about 42 dB down from the main lobe) compared to the rectangular window (about 13 dB down) [65]. For a sampling rate of 8 kHz and window length of 20 ms or 160 samples, as is common for telephone speech, the frequency resolution is 100 Hz. Similarly, for a frame rate of 100 Hz and window length of 1 second or 100 samples, as will be used to estimate the modulation spectrum in the next chapter, the frequency resolution is 2 Hz.

2.3 Short-term Feature Representations

Many different representations of short-term spectral features have been used for speaker verification. The aim with short-term features is to reduce dimensionality of the representation and to smooth out spurious information. Common representations include linear predictive coefficients (LPC) and various transformations such as PARCOR coefficients, reflection coefficients, and cepstral coefficients [4, 76, 75]. (The LPC representation can be derived from the short-term auto-correlation coefficients by use of the short-term power spectrum.) A drawback of the LPC representation is that it is based on an all-pole model that is strictly valid only for vowels. Another drawback is that without additional processing it is sensitive to noise. Other representations include line spectral pairs (LSP) [14] (also derived losslessly from LPC's), and filter bank energies and their cepstral representation [20, 16, 78]. Filter bank features can be derived from the FFT or LPC envelope. Filter bank features derived from the FFT have proven to be particularly useful for speaker verification [58], and will be used in this dissertation. The specific parameterization used will be detailed in Chapter 4.

The computation and parameterization of filter bank features have been influenced by a number of additional considerations and findings:

- In [30] an RMS distance on the *logarithmic spectral energies* was reported to be meaningful for speech processing. That a logarithmic representation may be suitable is also supported by statistical modeling considerations [14] and a model of convolutional distortion for the environment [76, 37, 7].

- Based on auditory considerations [43, 16, 67, 34], a log-like (Mel or Bark) *warping* of the frequency axis (to allow higher resolution at low frequencies) may be warranted.

- Similarly, based on the theory of auditory masking, which holds that frequency resolution is limited to *critical bands* [67, 34], a suitable limiting of *spectral resolution* appears useful. (This is usually done by weighting and averaging of adjacent power spectral components.)

- A telephone channel [20] has a fairly narrow passband (about 300-3300 Hz, but variable). The result is that spectral energies below 300 Hz and above 3300 Hz tend to be inconsistent and have a relatively high noise to speech content. Ignoring such energies as a post-processing step may improve robustness to acoustic mismatch [20, 78, 29].

Short-term spectral features have been reported to have certain shortcomings [36] if used without contextual information. There is a growing body of evidence that information may be spread across syllable length (about 200 ms) portions of the speech signal, due to articulatory effects [36, 100, 104]. Furthermore, the short-term spectral features may be overly specific and sensitive to adverse environments [20] and may require *context* for improved estimates. This context may be build either directly into the feature representation or provided in some other form to the follow-on classifier. A framework for the analysis of contextual information is provided in the next section.

2.4 Medium-term Analysis of Speech

Equation (2.1) suggests a medium-term[2] analysis of speech. This follows from a *filter bank* interpretation of short-term analysis, where the time sequences of spectral features are obtained by linear filtering. The short-term spectrum $S(n,\omega)$ may be obtained by demodulating the speech signal and convolving it with the analysis window. This follows directly from Equation (2.1) by viewing it as a function of the observation time n, for a fixed analysis frequency ω

$$S(n,\omega) = \sum_{m=-\infty}^{\infty} w(n-m)s(m)e^{-j\omega m} = s(n)e^{-j\omega n} *_n w(n). \tag{2.6}$$

Here the $*_n$ operator denotes linear convolution with respect to the time index n. The interpretation is that the time sequence of a particular spectral component in the short-term spectrum can be obtained by demodulation of the original signal $s(n)$ and application of a lowpass filter $w(n)$ to attenuate the aliased components [65]. This is a basic demodulation step in many communication systems [93]. As discussed before, the continuous analysis frequency ω may be sampled on a fixed grid ω_k, $1 \leq k \leq K$, using a finite number of filter bank *bands* to give outputs $S(n,k) = S(n,\omega_k)$. Fig. 2.3 depicts the filter bank interpretation for the k-th filter band.

It is also informative to interpret the filter bank in terms of Fourier transforms with respect to the observation index n while the analysis frequency ω remains fixed. For this

[2] *Medium-term* refers to time-spans that are longer than the frame length (typically 20 ms) up to time-spans on the order of the utterance duration (in this dissertation typically 3-60 seconds).

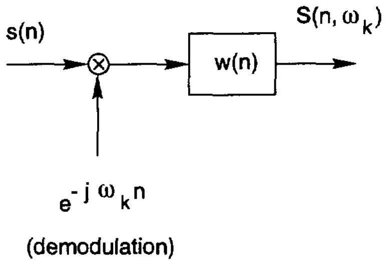


Figure 2.3: Filter bank interpretation of STFT.

purpose, since ω is considered a constant, it is convenient to denote[3] the FT of $s(n)$ as $S(\theta)$. Then as a result of the demodulation process the FT of the input to the filter $w(n)$ is $S(\omega+\theta)$. Thus the spectrum of $s(n)$ at frequency ω is shifted to zero frequency. The FT at the output of the filter is then $S(\omega+\theta)W(\theta)$. At the output, applying the inverse FT gives the time sequence that is generated by the ω frequency component of the short-term spectrum

$$\begin{aligned} S(n,\omega) &= \mathrm{IFT}_\theta\left[S(\omega+\theta)W(\theta)\right], \\ &= \frac{1}{\pi}\int_{-\pi}^{\pi} W(\theta)e^{j\theta n}S(\omega+\theta)d\theta, \end{aligned} \tag{2.7}$$

where IFT_θ denotes the inverse FT with respect to the frequency variable θ. From Equation (2.7) it can be seen that if $W(\theta)$ is a lowpass filter with a very narrow passband (i.e. $w(m)$ is almost constant) then up to a constant, the STFT spectrum $S(n,\omega)$ becomes a good approximation for the actual FT spectrum $S(\omega)$.

2.4.1 Modulation Frequency

Equation (2.7) introduces the concept of *modulation frequency* [41, 22]. The frequency variable θ that is related to the spectrum of the time sequence generated by a particular frequency component of the short-term spectrum,

$$S(\theta,\omega) = S(\omega+\theta)W(\theta), \tag{2.8}$$

[3] The *analysis frequency* ω should not be confused with the *modulation frequency* θ.

is known as the modulation frequency. The modulation spectral domain therefore follows from Equation (2.2) as the two-dimensional FT of the windowed speech signal

$$S(\theta, \omega) = \mathrm{FT}_n \left[S(n, \omega) \right] = \mathrm{FT}_{n,m} \left[w(n-m) s(m) \right]. \tag{2.9}$$

2.4.2 Modulation Spectrum

The concept of the *modulation spectrum* derives from the well-known modulation transfer function (MTF) used in the evaluation of optical systems. In the transmission of sound, the MTF refers to the transfer function that characterizes a system in terms of changes in the modulation depth of a temporally sine-modulated test signal [90]. The reduction in modulation depth as a function of modulation frequency constitutes the MTF. The MTF tends to be different for different systems and provide a good characterization of the properties of the system. Typically the MTF shows a lowpass character, reflecting the limited resolution power of the system in preserving finer details.

In [41], the modulation spectrum was defined as the power spectrum of the octave-band energy envelope of speech. In this definition, before computation of the power spectrum, the average energy of the octave-band energy envelope was first normalized. This was done because the average energy was considered to be of limited interest since it depends on the recording conditions and is highly variable. The definition of the modulation spectrum[4] was later extended in [7] and elsewhere to apply to non-linear transformations of the energy sub-band envelope. In particular, in [22, 37, 2, 44, 7] and [100] the usefulness of the modulation spectrum of *logarithmically* transformed energy has been demonstrated for analysis, recognition and enhancement of speech. (Note that normalization of the average energy of the sub-band envelope can be achieved in this domain by simply normalizing the mean of the logarithmic sub-band envelope.) In [7] it was shown that processing in a non-linear domain, such as the logarithmic domain, cannot be implemented using a time-domain filter. That the logarithmic energy domain may be the most extensively used domain in speech processing for ASR and speaker recognition, may be due to the fact that this domain allows for compensation of the effect of a convolutional channel

[4] The term *modulation spectrum* should be understood to refer to the modulation *power spectrum*.

distortion through the process of *blind deconvolution* [75]. Given these observations, the approach used in this dissertation will be to compute, analyze and modify the modulation spectrum of the time sequences of *logarithmic* energy.

Logarithmic energies can be computed directly from the STFT by ignoring the phase component:

$$X_1(n,\omega) = \log(|S(n,\omega)|^2) = \log(|\mathrm{FT}_m\,[w(n-m)s(m)]|^2). \tag{2.10}$$

Each time sequence can be subsequently modified by linear filtering to yield

$$X_2(n,\omega_k) = \sum_{r=-\infty}^{\infty} f(n-r)X_1(r,\omega_k) = f(n) *_n X_1(n,\omega_k), \tag{2.11}$$

where $f(n)$ is the impulse response of an FIR filter. The proposed processing is a convolution in the time dimension of the STFT and a multiplication in the frequency dimension.

From Equations (2.9), (2.10) and the preceding discussion, the modulation (power) spectrum of an unfiltered time sequence is formally defined as

$$G_1(\theta,\omega) = |\mathrm{FT}_n\,[X_1(n,\omega)]|^2\,, \tag{2.12}$$

with a similar definition for the filtered time sequence $X_2(n,\omega)$. An in depth discussion of the modulation spectrum and requirements for resynthesis of the speech signal can be found in [7].

In the interpretation of the characteristics of the short-term *logarithmic* energies, it is informative to investigate the effect of the analysis window length, frame rate, and environment in the modulation spectral domain.

2.4.3 Sampling Considerations

Recall from the filter bank interpretation of short-term analysis, that the analysis window acts as a filter applied to the demodulated speech signal (refer Equation (2.6)). Accordingly, the bandwidth of $S(n,\omega_k)$ is therefore limited by the bandwidth of the analysis window. Fig. 2.4 depicts the theoretical band-limiting effect for a Hamming analysis window (refer Equation (2.5)). The figure shows the frequency responses of Hamming windows of different lengths. The responses are shown in the modulation spectral domain

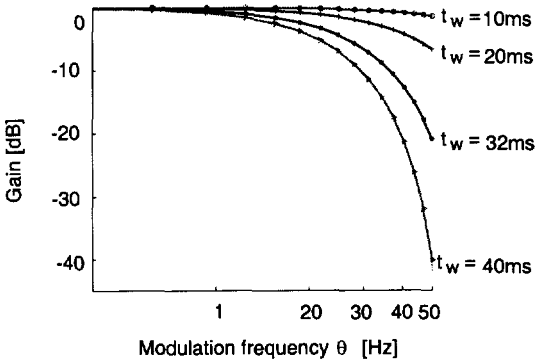


Figure 2.4: Theoretical band-limiting effect of Hamming analysis windows of different lengths t_w.

Table 2.1: Minimum sampling rate θ_s to avoid aliasing for different Hamming analysis window lengths l_w.

l_w (samples)	80	160	256	320
t_w (ms)	10	20	32	40
B (Hz)	200	100	64	50
θ_s (Hz)	400	200	128	100

assuming a frame rate of 100 Hz. In the figure it can be seen that the analysis window acts as a lowpass filter in the modulation spectral domain. Using the Nyquist sampling criterion [65], it follows that to avoid aliasing, it is necessary to choose a frame rate $\theta_s \geq 2B$ that is equal to or larger than two times the bandwidth of the window.

Table 2.1 lists values for the minimum frame rate θ_s that avoids aliasing. The frame rate and bandwidth B of the Hamming window are shown as a function of short-term analysis window length l_w in samples and t_w in milliseconds. The results apply to a speech sample rate of 8000 Hz. The table shows that a 20 ms long Hamming window requires a frame rate $\theta_s \geq 200$ Hz. In ASR and speaker verification it is actually quite common to use a 20 ms long Hamming window and sample at $\theta_s = 100$ Hz. To avoid aliasing at the 100 Hz sampling rate, a 40 ms long analysis window would be needed. A suitable choice of analysis window length and sampling rate, as well as an understanding of the potential for aliasing, is necessary in the next chapter, where modulation spectra of speech and telephone handsets are examined.

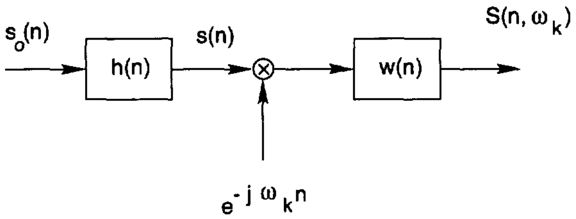


Figure 2.5: Model for convolutional channel distortion.

2.5 Medium-term Feature Processing

The strategy proposed in Equation (2.11) was to obtain useful features by filtering the logarithmic energies defined over points in time $t = ..., t_{n-1}, t_n, ...$ for each particular short-term analysis frequency ω_k. This section elaborates on the usefulness of such filtering.

2.5.1 Convolutional Distortion

Sources of extraneous/non-speaker variability, such as a changing transmission channel, affect the modulation spectrum of speech. For example, it is generally accepted that a linear time-invariant transmission channel affects the DC components in the time sequences of the *logarithmic* energies (see [4, 26, 7] and references therein). A model for the degradation is shown in Fig. 2.5. The degraded speech signal is denoted as $s(n)$, the original undegraded speech signal as $s_o(n)$ and the impulse response of the transmission channel as $h(n)$. For the convolutional transmission channel it follows that

$$S(n, \omega) = \mathrm{IFT}_\theta \left[S_o(\omega + \theta) H(\omega + \theta) W(\theta) \right]. \tag{2.13}$$

Then an argument similar to the one following Equation (2.7) is that if $W(\theta)$ is a lowpass filter with a very narrow passband so that $W(\theta) \to \mathit{const} \times \delta(\theta)$, i.e. $w(m)$ is almost constant, then the STFT spectrum $S(n, \omega)$ approximates the actual FT spectrum $S_o(\omega)H(\omega)$ up to a constant factor. To get a reasonable estimate of the actual channel it is necessary that the impulse response of the transmission channel $h(n)$ be shorter than the length of the analysis window $w(n)$ as implied by the convolution operation and windowing (refer Equation (2.6)). Requirements for the approximation are discussed in more detail in [7].

With a suitable choice of analysis window the short-term logarithmic spectrum of the degraded speech signal $s(n) = s_o(n) * h(n)$ is modeled in the n-th analysis segment as

$$\log[|S(n,\omega)|^2] \approx \log[|H(n,\omega)|^2] + \log[|S_o(n,\omega)|^2], \tag{2.14}$$

where $H(n,\omega)$ is the short-term frequency response of the transmission channel and $S_o(n,\omega)$ is the short-term frequency response of the speech signal before passing through the transmission channel. Assuming that the transmission channel is time-invariant, with a relatively short impulse response, so that $H(n,\omega)$ changes little with time, taking the expectation over time may approximate the channel term well

$$\mathrm{E}_n\left\{\log[|S(n,\omega)|^2]\right\} \approx \log[|H(\omega)|^2] + \mathrm{E}_n\left\{\log[|S_o(n,\omega)|^2]\right\}, \tag{2.15}$$

and is seen to affect the DC component of the modulation spectrum.

This simple model of course only accounts for the convolutional effect of a linear time-invariant transmission channel and does not model any DC contribution due to the term $\log[|S_o(n,\omega)|^2$ that describes the undegraded speech signal [4]. While it is clear that the transmission channel affects the DC component of the modulation spectrum, to the extent that the convolutional distortion model is only an approximation of one type of distortion, it would be of interest to determine the effect of distortions such as for example changes in telephone handset on the DC component as well as other components of the modulation spectrum.

2.5.2 Additive Noise

The presence of additive or ambient noise is common in adverse environments. Fig. 2.6 shows an additive noise source $v(n)$ within the filter bank interpretation of short-term analysis of speech. Following the discussion in Section 2.5.1, it is informative to decompose the degraded speech signal $s(n) = s_o(n) * h(n) + v(n)$ in terms of the original undegraded speech signal. Assume that the additive noise is uncorrelated with the original speech signal $s_o(n)$ and stationary. For this idealized model, the decomposition in the power spectral domain is

$$|S(\omega)|^2 = |S_o(\omega)|^2|H(\omega)|^2 + |V(\omega)|^2. \tag{2.16}$$

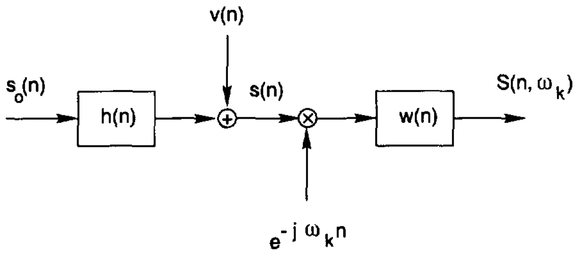


Figure 2.6: Model for convolutional channel distortion and additive noise.

Let $\tilde{S} = |S(\omega)|^2$ denote the power spectrum of the degraded speech, with similar notation $\tilde{H}$, $\tilde{S}_o$ and $\tilde{V}$ for the other terms. In the log-spectral domain (see for example [77]), it follows[5] that

$$\log \tilde{S} = \log \tilde{S}_o + \log \tilde{H} + \log(1 + \frac{\tilde{V}}{\tilde{S}_o \tilde{H}}). \tag{2.17}$$

Equation (2.17) has two important implications. First, with the presence of additive noise, a convolutional distortion can not be completely isolated in the log-spectral domain (compare Equation (2.14)). Conversely, additive noise cannot be completely isolated in the spectral domain if a convolutional distortion is also present. Second, while the transmission channel may be linear and time-invariant (with a fixed convolutional distortion and stationary additive noise) the effects of the acoustic disturbances may be time-varying in the log-spectral domain. This can be seen from the third term in Equation (2.17) which is a non-linear function of the time-varying undegraded speech term and noise term. This time-varying nature of the acoustic disturbances will be important in the analysis of the time sequences of logarithmic energy in the next chapter.

2.5.3 Compensating for Distortions and Noise by Filtering

Various filters have been proposed to compensate for convolutional channel distortions and for extracting useful features from the logarithmic energies. A common filtering technique is that of cepstral mean subtraction (CMS) [26, 4]. CMS suppresses convolutional noise (such as the frequency characteristic of a transmission channel which is additive in

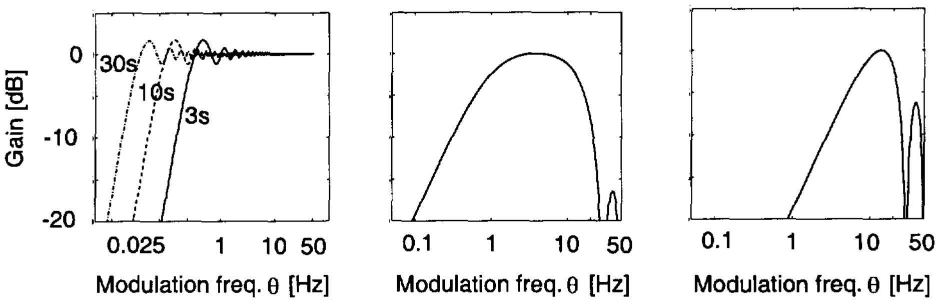
[5] The relation follows directly by factoring out the first term from Equation (2.16), so that $\tilde{S} = \tilde{S}_o\tilde{H}(1 + \tilde{V}\tilde{S}_o^{-1}\tilde{H}^{-1})$, and manipulating the resulting second term.

the logarithmic spectrum or cepstrum) by suppressing the DC component in each time sequence of cepstral coefficients.[6] CMS therefore performs highpass filtering. In this dissertation, since logarithmic energies are not projected onto a cosine basis, attenuation or removal of the mean will be referred to as *mean subtraction* (MS) to indicate that it is the logarithmic energies that are filtered and not the cepstral coefficients.

The temporal structure of speech was explored in [26] with the use of so-called "dynamic cepstral features". These features are estimates of the first and second temporal derivatives of the time sequences of cepstral coefficients. The derivatives are usually computed as the first and second order orthogonal polynomial expansion of 50 to 90 ms long segments of the time sequence. These first and second order derivative features, also called *delta* and *double delta* features, are common in speech and speaker recognition systems and the benefits of these representations are widely recognized. While usually viewed as an expansion on the polynomial basis, the expansion can also be treated as a linear filtering operation and the characteristics of the polynomial filters examined using the modulation spectrum. An important characteristic of the derivative features is that they attenuate the DC component in the modulation spectrum and thus are relatively insensitive to convolutional channel distortions.

Another technique [37] that has been shown to reduce the effect of the transmission channel by limiting the frequencies present in the time sequences of logarithmic energy is RelAtive SpecTrAl Processing (RASTA). The RASTA filter is a passband filter with a spectral zero at zero modulation frequency and a relatively narrow passband. It is implemented as an IIR filter with the same filter used for all frequency bands. The relative successes reported for the CMS and RASTA techniques can be interpreted from the perspective of removing or attenuating redundant and possibly noisy information. Both techniques achieve a measure of channel independence by attenuating specific components in the modulation spectrum. Regardless of what the transmission channel does to these components, their removal leads to more consistent features.

[6] Cepstral coefficients are obtained by applying the Discrete Cosine transform (DCT) to the logarithmic spectrum.



(a) MS (b) RASTA (c) Delta

Figure 2.7: Frequency responses of various filters in the modulation spectral domain.

Fig. 2.7 illustrates some of the differences between these filters in terms of their frequency responses for a sampling rate of 100 Hz. The MS filter has a highpass frequency response with a -3 dB cut-off frequency[7] that depends on the length of the averaging window - here 0.025, 0.075 and 0.25 Hz respectively for window lengths of 30, 10 and 3 seconds. The RASTA filter has a -3 dB passband[8] of about 1 to 13 Hz. The delta polynomial $(-0.2, -0.1, 0, 0.1, 0.2)$ computed in a 50 ms window of speech has a -3 dB passband of about 7 to 21 Hz. Clearly the effect of these filters on the modulation spectrum of the speech will differ substantially and in general will depend on the modulation spectrum of the speech itself. For speaker verification the filtering should of course be used to enhance speaker specific information while suppressing non-informative and possibly confusing information. This suggests an analysis of the relative importance of the components of the modulation spectrum for speaker verification. The processing strategy would be to filter the logarithmic energies to extract salient and consistent features. The objective would be to improve performance in adverse environments.

2.5.4 Experimental study

This section highlights a small experimental study that introduces the effects of acoustic mismatch in adverse environments and examines the performance of the popular MS

[7] The cut-off frequency is defined as the frequency at which the squared magnitude frequency response of the filter is 3 dB below the nominal passband value.

[8] The passband is defined as the range of frequencies delimited by the lowest and highest frequency for which the squared magnitude frequency response of the filter is 3 dB below the peak value.

and RASTA filtering techniques. For the purpose here it is sufficient to give the main results. The speaker verification system is described in Chapter 4, the setup is explained in Appendix A, and the study is detailed in Chapter 5.

Speaker verification is performed on a population of 500 speakers obtained from the telephone speech corpus used in the 1997 NIST speaker recognition evaluation [72, 58]. Three systems that use features based on the short-term logarithmic spectral energies are compared. In the baseline system, the time sequences of spectral energy are used unmodified. In the 'MS'-system, the mean of each time sequence is set to zero (a highpass filtering). In the 'RASTA'-system, each time sequence is filtered with a RASTA filter (a bandpass filtering). The frequency responses of the filters were shown in Fig. 2.7.

A model for each speaker is trained in one environment and tested in the same environment (matched condition) and different environment (mismatched condition). The matched or mismatched condition refers to whether the phone number and handset type are the same. The segments of test speech are nominally 3 or 30 seconds in length. Verification performance is assessed separately for males and females using the equal error rate[9] (EER), and the errors averaged to give a gender independent error.[10]

Table 2.2 lists the average EER in percent for the matched and mismatched conditions. It can be seen that errors in the matched condition are much lower than errors in the mismatched condition. In 30 second segments with MS filtering, the EER increases by about a factor of 5 from the matched condition to the mismatch condition. This shows that acoustic mismatch indeed affects performance deleteriously. For the mismatched condition it can be seen that MS and RASTA filtering reduces the EER considerably

[9] The EER is the point where false rejection and false acceptance errors are equal.

[10] The errors for males and females are about the same, with similar tendencies. As a result it was decided to average them and report only a single number. This greatly simplifies the representation without loss of generality. It was decided not to pool scores for males and females before computing the error, as is done in the NIST speaker recognition evaluation. Motivation for this decision came from the observation that the statistics of male and female scores differ to the extent that pooling the scores and comparing each score to the same global threshold would lead to suboptimal performance and a higher error compared to the case where, either the scores for males and the scores for females were first normalized to have the same statistics before pooling, or where the error was computed separately for males and females without a pooling of the scores. For the purpose intended here which was to compare the effect of processing it did not matter whether the scores were normalized and pooled or whether the scores were not pooled and the errors averaged except that the latter greatly simplified the comparison.

Table 2.2: Equal error rate in percent for speaker verification using 3 and 30 second test segments in (a) matched and (b) mismatched conditions.

	TEST SPEECH DURATION	
	3 seconds	30 seconds
Baseline	10.2	6.7
Baseline + MS	13.1	4.7
Baseline + RASTA	14.7	6.3

(a) Matched

	TEST SPEECH DURATION	
	3 seconds	30 seconds
Baseline	32.5	31.4
Baseline + MS	26.8	21.8
Baseline + RASTA	28.4	22.5

(b) Mismatched

with both giving a similar reduction in EER over the baseline (on average about 30% for the 30 second segments and about 15% for the 3 second segments). In contrast, this reduction is less in the matched condition (on average a reduction of about a 17% for the 30 second segments and actually an increase of 36% for the 3 second segments). MS filtering is seen to perform slightly better than RASTA filtering overall, especially in the matched condition. It can also be seen that both types of filtering become less effective as the length of the speech segment decreases. This suggests that benefits of both types of filtering may come from their effect on the very low modulation frequency components – components which are only present in the longer segments.

While it can be concluded from the results that a filtering of the time sequences can be beneficial, it is not clear exactly what type of filtering gives the best performance or may improve performance. Indeed, MS filtering was motivated assuming an ideal convolutional disturbance, and RASTA filtering was motivated from similar considerations as well as experiments in ASR [37]. It is clear though, that the filtering should be used to enhance speaker specific information while suppressing non-informative and possibly confusing information. This idea will be explored in depth in the next chapter, where the modulation spectrum of speech will be compared to a modulation spectrum that describes variability among telephone handset transducers. Recalling that the MS and RASTA filters have different frequency responses, a study of environmental variability in the modulation spectral domain may help to motivate the choice of a suitable filter.

2.6 Summary

This chapter motivated feature extraction in a time-feature space. Based on physiological, auditory and perceptual considerations, features for speaker verification were derived from a short-term analysis of speech. A medium-term analysis was motivated as necessary for the analysis of the characteristics of the resulting time sequences of spectral features. The modulation spectral domain was defined as a suitable domain for the analysis of the time sequences and for the characterization of various filters that manipulate these time sequences. The effects of analysis considerations such as window length were emphasized. The effects of convolutional channel distortions and additive noise were studied and various compensation techniques reviewed. An experimental study demonstrated the deleterious effect of acoustic mismatch that is present in adverse environments and suggested that filtering of the time sequences of spectral features may help to alleviate this mismatch. It was suggested that an analysis of the relative importance of the components of the modulation spectrum for speaker verification may provide insight as to what type of filter may improve verification performance.

Chapter 3

Handset Variability

In this chapter it is proposed that the modulation spectrum provide information that may be used to understand and reduce the effects of telephone handset variability. The chapter presents a rather general characterization of telephone handset variability in speech and as such does not specifically deal with speaker variability. It does serve however to indirectly motivate and guide the development of a processing strategy adopted in latter chapters that attenuates unwanted variability in the speech signal. The approach will be to compare modulation spectra of speech from different speakers that were recorded using handsets with different microphone transducer types – specifically, carbon-button and electret transducers. As expected from a convolutional model for the transmission channel, it is found that telephone handset variability severely contaminates the DC-modulation component. Importantly, this is also found to be the case for the higher (above about 10 Hz) modulation frequencies. The result is obtained by computing the variability in speech due to carbon-button and electret transducers and comparing it to the overall variability in speech. The computation is based on an analysis-of-variance model.

The outline of the chapter is as follows. Section 3.1 provides a brief review of the problem of handset variability. Previous approaches to compensate for this variability are categorized as having considered variability in terms of the short-term spectrum or *frequency* dimension of the feature space. It is then suggested to consider variability in terms of the time sequences of logarithmic energy or *time* dimension of the feature space. Section 3.2 next describes a database that will be used to investigate handset variability in this time dimension. Section 3.3 then describes a model that can be used to decompose the speech signal into different sources of variability. The section starts by outlining the key ideas behind the decomposition and then proceeds with a decomposition based on an analysis-of-variance model. The variability is interpreted in the modulation spectral domain. Section 3.4 describes various results related to the decomposition. Section 3.5 investigates limitations of the technique and Section 3.6 provides additional results.

3.1 Variability in Time and Frequency

Handset variability has been a major source of verification errors in the recent NIST Speaker Recognition Evaluations [58, 59, 72, 73]. One strategy to compensate for handset variability has been to normalize model scores conditioned on *handset type* [58, 80]. Another strategy has been to improve the feature processing. In [74] the approach has been to try to classify telephone handsets according to *electret-* and *carbon-button* transducer type. The one handset type was then *mapped* to the other using a memoryless polynomial nonlinearity so as to match short-term *spectral magnitude.* An interesting observation from this work was that among other things, the effect of the handset transducer, and in particular the carbon-button transducer, was to create so-called "phantom-formants" in the spectrum of the speech signal. These phantom-formants were shown to occur as peaks in the spectrum at multiples and sums of the original formant frequencies. A possible explanation for this effect was that non-linearities in the transducer would tend to emphasize certain harmonics of the formants of the original (undegraded) speech signal.

The degradation model depicted in Fig. 2.6 will be used in this chapter but it will be assumed that the handset may impose a possibly time-varying convolutional distortion on the undegraded speech signal as well as introduce additive noise. While sources other than the handset may also introduce convolutional distortions and additive noise, it will be convenient to assign all degradations as being due to the "handset". From the discussion in Section 2.5.2 it can be argued that the effects of handset variability will be time-varying, since the speech signal itself is time-varying. Instead of investigating the effects of the transducers in terms of short-term spectral magnitude, the approach followed in this chapter will be to analyze the effects in terms of the modulation spectrum of the time sequences of logarithmic energy. Motivation for this approach is for instance that "phantom-formants" in the spectrum of the speech signal, if only occasionally present, would show up in the higher modulation frequency components. Whereas the work described in [74] was concerned with the frequency dimension of the feature space, this chapter deals with the time dimension of the feature space. As far as we know, a characterization of handset transducers along this time dimension has not been done before.

3.2 Handset Data

To study handset variability, the HTIMIT corpus which consists of speech from 192 males and 192 females transmitted over different telephone handsets is used. This corpus was collected and kindly made available by Douglas Reynolds [81] and is available through the LDC [52]. The HTIMIT corpus contains ten utterances per speaker, each containing different spoken text, which originate from the TIMIT corpus[1]. The HTIMIT utterances are the original TIMIT utterances of a speaker which were transmitted through different electret- and carbon-button transducers to simulate the effect of different telephone handsets. We used three different electret and three different carbon-button transducers.

We randomly selected 15 male speakers for the study presented here. Utterances were chosen so that the same text strings were spoken by each speaker: "She had your dark suit in greasy wash water all year" and "Don't ask me to carry an oily rag like that". We aligned the recordings for a particular speaker speaking the same text over different handsets using a waveform-based correlation and verified visually that alignment errors are on the order of 3 ms or less. The alignment errors are therefore well within a typical short-term analysis frame spacing of 10 ms.

Fig. 3.1 illustrates typical handset variability. The figure shows one second of logarithmic energy with mean removed for the same speech from the same speaker transmitted over electret- and carbon-button transducers. Small differences can be seen between the two time sequences. These differences are examined in the remainder of this chapter using the modulation spectrum that was defined in Section 2.4.2.

[1] The TIMIT utterances were recorded in a sound booth using a high quality microphone.

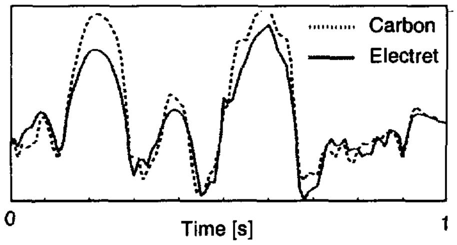


Figure 3.1: Time sequences $X(n,k)$ from the $f_k = 1$ kHz filter bank band for speech from a speaker transmitted over an electret and a carbon-button transducer.

3.3 Analysis-of-Variance Model

An analysis of variance (ANOVA) [91] provides a convenient way to estimate the effect of handset variability. In the following, the decomposition is performed on a time sequence of logarithmic energy $X(n,k)$ and the contributions of the individual factors analyzed using the modulation spectrum. This section is structured as follows. We will first describe the procedure for estimating the modulation spectrum of a time sequence. Then we will provide an outline of an algorithm for the analysis of variance. The outline will serve to highlight some of the assumptions and issues involved in the analysis. Finally we will introduce the ANOVA model.

3.3.1 Estimating the Modulation Spectrum

The modulation spectrum was defined as the power spectrum of a time sequence of logarithmic energy in Chapter 2. Estimates for the power spectrum were computed using Welch's averaged periodogram method [65]. This computation is similar to the short-term Fourier Transform technique described in Section 2.2, except that a medium-term analysis is performed on a sequence of logarithmic energy using a long analysis window. Since the speech utterances and thus time sequences are nominally 2.5 seconds in duration, it was decided to use a 1 second long Hamming analysis window advanced in 1 second steps. To interpret the results it will be useful to recall (refer Section 2.2) that the bandwidth, and thus modulation frequency resolution, of a 1 second long Hamming analysis window is 2 Hz. For the STFT computation of the time sequences of logarithmic energy, an analysis window with a length t_w of 40 ms and a frame rate θ_s of either 100 Hz or 200 Hz will be used. Since results differ little for either frame rate, the default 100 Hz frame rate used in the latter part of this dissertation will be used in most of the results reported here. When other window lengths and frame rates are used, this will be noted in the text. While results will be reported for time sequences of logarithmic energy from the $f_k = 1$ kHz analysis frequency the main conclusions are fairly general and can be shown to pertain to time sequences from other analysis frequencies as well. Accordingly, while the frequency index k will be dropped from the notation in the following, results should be treated as being with respect to a particular analysis frequency.

3.3.2 Outline of Algorithm for the Analysis of Variance

This section provides an outline of an algorithm for analyzing the different sources of variability in the speech signal in the modulation spectral domain. The section is intended to introduce the key ideas behind an analysis-of-variance decomposition that will be described in the next section. It is assumed that the effects of the handset transducers on the logarithmic energies are independent and identically distributed and additive. The latter assumption is justified for a convolutional degradation which is approximately additive in the logarithmic energy domain as discussed in Section 2.5.1.

The goal is to analyze the different sources of variability in terms of the time sequences of logarithmic energy that were computed for the speech signal. Table 3.1 outlines an algorithm for estimating the variability due to variations among handsets specifically ("handset variability"), and variations among handsets, speakers, and utterances in general ("total variability"). The handset variability as a function of modulation frequency will be denoted as $G_H(\theta)$ and called the *handset variation* and the total variability as a function of modulation frequency will be denoted as $G_X(\theta)$ and called the *total variation*. The part of the total variability that is not attributable to the handset variability will be denoted as $G_S(\theta)$ and called the *speech variation*. Several comments related to the algorithm in Table 3.1 are in order:

- Step 1 is a global energy normalization for the time sequences of logarithmic energy. This is necessary to allow a comparison of $G_H(\theta)$ and $G_X(\theta)$, since for $G_H(\theta)$ the

Table 3.1: An algorithm for computing handset variation $G_H(\theta)$ and total variation $G_X(\theta)$.

1. For all time sequences of logarithmic energy $X_{l,j}(n)$, spoken by speaker l using handset j, compute the average over all speakers and handsets to get $\bar{X}(n)$.

2. (a) Let c be a constant. Take any two time-aligned sequences of logarithmic energy, each from identical speech spoken by speaker l but recorded over different handsets i and j and compute the difference sequence $E_{i,j,l}(n) = c\,[X_{l,i}(n) - X_{l,j}(n)]$

   (b) Compute the power spectrum $P_{i,j,l}(\theta)$ for the difference sequence $E_{i,j,l}(n)$.

   (c) Repeat steps 2a and 2b for all combinations of handsets i, j and all speakers l and average the estimates $P_{i,j,l}(\theta)$ to get $G_H(\theta)$.

3. (a) Take a sequence of logarithmic energy spoken by speaker l and recorded over handset i and subtract the global average $\bar{X}(n)$ computed in step 1.

   (b) Compute the power spectrum $P_{l,i}(\theta)$ for the difference sequence $X_{l,i}(n) - \bar{X}(n)$.

   (c) Repeat steps 3a and 3b for all handsets i and all speakers l and average the estimates $P_{l,i}(\theta)$ to get $G_X(\theta)$.

energy is normalized by the difference computation in step 2.

- The constant scale factor c in step 2 is necessary to normalize for correlation of the two time sequences. For example, assume that $X_{l,i}(n) \sim \mathcal{N}(0, \sigma^2)$, and that $X_{l,i}(n)$ and $X_{l,j}(n)$, $i \neq j$ are *not* correlated. In this extreme case then, it follows that $\text{Var}\{X_{l,i}(n) - X_{l,j}(n) | i \neq j\} = 2\sigma^2$. Then a choice of $c = 0.5$ is necessary for $E(n)$ to give an estimate of $\text{Var}\{X_{l,i}(n)\}$.

- Since the only difference between the two time sequences of logarithmic energy is due to the handset, $G_H(\theta)$ reflects the variations among handset transducers.

- $G_X(\theta)$ reflects the variations among handsets, speakers, and utterances. In addition to the handset variation $G_H(\theta)$ it therefore includes variations attributable to differences between speakers and differences between the spoken text and manner of speaking. In general, $G_X(\theta)$ reflects the total variability present in speech.

It is important to note that while the definition of $G_X(\theta)$ closely matches the classical definition of the modulation spectrum of speech, it generalizes and differs from this definition in the following way. In the classical definition [41], the modulation spectrum was defined as the power spectrum of a time sequence from which the long term average (computed over n) has been subtracted. Here, in the generalized definition, the *ensemble average* (computed over time sequences), and not the long term average is subtracted from the time sequence. The algorithm can of course be modified so that the long term average of each time sequence is removed before computing the power spectrum, but not doing so allows a more general analysis. Not removing the long term average means that variability that gets reflected in the zero modulation frequency component can also be studied. The differences and similarities between the two definitions will be discussed in Section 3.6.2.

3.3.3 Nested Analysis of Variance

The outline given previously relates to a nested analysis of variance [91] for the time sequences of logarithmic energy, with interpretation of the variances in the modulation

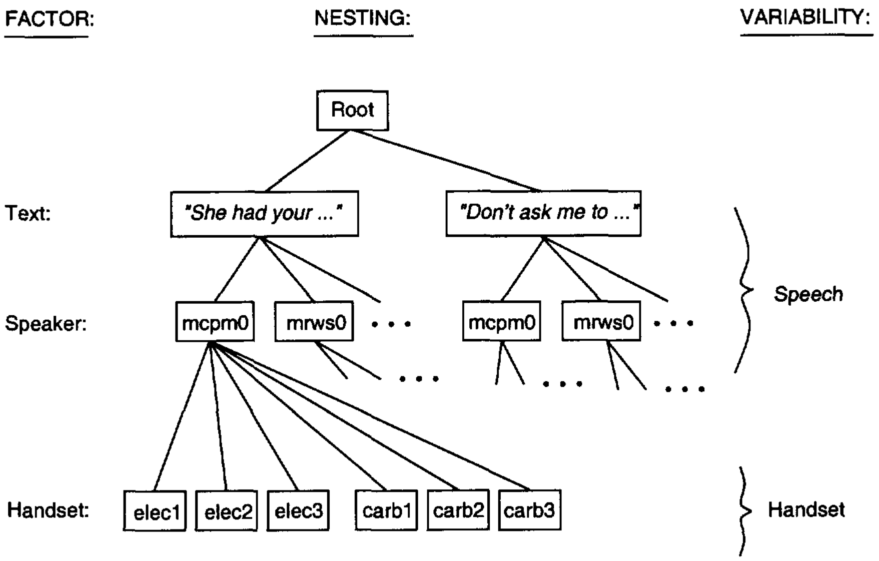


Figure 3.2: Nesting of factors for analysis of variance.

spectral domain. The ANOVA is nested on the factors *handset*[2] (H), *speaker* (S) and *text* (T). Fig. 3.2 shows the nesting of the factors. Variations due to different handsets are taken to reflect handset variability, while variations due to different speakers and spoken text are taken to reflect speech variability.

The speech from different speakers speaking the same text will in general not be finely aligned due to variations in rate of speech and variations in prosody, as well as phoneme, syllable and word insertions, deletions and substitutions. This suggests that variations due to different speakers and text should not be interpreted separately as it would be difficult to decide to what extent the speaker variation is due to textual or speaker specific variations. It may also not be possible to compensate for this ambiguity, since trying to do so may impose unnatural constraints on the speech signal and introduce artificial variability[3]. The handset factor, though, is well defined at the speaker level. This is due to the explicit association of handset transducer label with speech signal and careful alignment of all recorded speech signals that originated from a particular speech signal spoken by a particular speaker. The total variability is also well defined since it is

[2] More specifically *handset transducer*.

[3] This would happen for instance if a technique such as dynamic time-warping were to be applied to the signals.

simply the *total* variation among all speech signals. The approach taken here is therefore to estimate the handset and total variability with the understanding that the difference between the total and handset variability represents the speech variability.

In the analysis, all time sequences are truncated to the same length and treated as vectors in the time dimension. The time sequence of logarithmic energy $X_1(n,k)$ for a particular analysis frequency f_k with $1 \leq n \leq N$ is therefore treated as an N-dimensional vector in the time index n. In the following we drop the subscript 1 that was used to indicate logarithmic energy and also drop the frequency index k from the notation, but results should be understood to apply to a time sequence of logarithmic energy for a particular analysis frequency. Unless it is noted otherwise, results will be reported for the $f_k = 1$ kHz analysis frequency.

The observed response at the j-th level of T, l-th level of S and i-th level of H is then

$$\begin{aligned} X(n) = \bar{X}(n) + X_j(n) + X_{l|j}(n) + X_{i|lj}(n) + \epsilon_{ilj}, \\ \text{for } j = 1 \ldots J,\ \ l = 1 \ldots L(j),\ \ i = 1 \ldots I(j,l), \end{aligned} \qquad (3.1)$$

where $\bar{X}(n)$ is the average response at time n. Here $\epsilon_{ilj} \sim \mathcal{N}(0,\sigma^2)$ is the experimental error which we assume to be zero for the purposes of this study[4]. The *total* variation can be obtained in terms of the sums of squares (SS) as

$$SS_{total}(n) = SS_X(n) + \sigma^2 = SS_T(n) + SS_S(n) + SS_H(n) + \sigma^2. \qquad (3.2)$$

Letting $L_t(j) = \sum_{l=1}^{L(j)} I(j,l)$, the individual terms can be computed as

$$\begin{aligned} SS_X(n) &= \sum_{j=1}^{J} \sum_{l=1}^{L(j)} \sum_{i=1}^{I(j,l)} \left| X_{ilj}(n) - \bar{X}_{...}(n) \right|^2, \\ SS_H(n) &= \sum_{j=1}^{J} \sum_{l=1}^{L(j)} \sum_{i=1}^{I(j,l)} \left| X_{ilj}(n) - \bar{X}_{.lj}(n) \right|^2, \\ SS_S(n) &= \sum_{j=1}^{J} \sum_{l=1}^{L(j)} I(j,l) \left| X_{.lj}(n) - \bar{X}_{..j}(n) \right|^2, \\ SS_T(n) &= \sum_{j=1}^{J} L_t(j) \left| X_{..j}(n) - \bar{X}_{...}(n) \right|^2. \end{aligned} \qquad (3.3)$$

[4] Changes in a transducer over time, such as related to temperature, as well as sampling errors are ignored in this study, but can be envisaged to be contributing to an experimental error.

The dot notation, as in $\bar{X}_{..j}$, is used to indicate the average value of X computed over the "dotted" factors – in this example, i and l.

In the previous discussion, it was noted that it would be difficult to decide to what extent the speaker variation $SS_S(n)$ is due to textual or speaker specific variations. Accordingly, the speech variation $SS_{S'}(n)$ is substituted for the speaker and text variations

$$SS_{S'}(n) = SS_S(n) + SS_T(n) = SS_X(n) - SS_H(n). \tag{3.4}$$

To interpret the variations in the modulation spectral domain it is necessary to modify the computation. For example, the term $SS_X(n)$ for the total variation is modified as follows[5]

$$G_X(\theta) = \sum_{j=1}^{J} \sum_{l=1}^{L(j)} \sum_{i=1}^{I(j,l)} \left|\mathrm{FT}_n \left\{X_{ilj}(n) - \bar{X}_{...}(n)\right\}\right|^2, \tag{3.5}$$

where FT_n denotes the Fourier Transform with respect to the time index n. The other terms $SS_S(n)$ and $SS_H(n)$ are modified in the same way to obtain the variations[6] in the modulation spectral domain $G_S(\theta)$ and $G_H(\theta)$. The variations can be scaled by a constant factor without loss of generality. In the following we normalize G_H, G_S and G_X so that the total variation G_X is approximately 1 at $\omega = 0$.

3.4 Handset Variability

Fig. 3.3 depicts the total variation and handset variation on a logarithmic scale for time sequences from the $f_k = 1$ kHz analysis frequency obtained at a frame rate of $\theta_s = 200$ Hz using a short-term analysis window of length $t_w = 20$ ms. The 1 second long Hamming analysis window that was used to estimate the modulation spectrum had the effect of smearing the energy in the DC component into components up to about 2 Hz. The Nyquist frequency for the analysis is 100 Hz.

[5] Note that the modification preserves the overall energy. This can be seen by applying Parseval's Theorem [65]

$$\sum_{n=-\infty}^{\infty} |X(n)|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X(\theta)|^2 d\theta$$

and by rearranging the summation.

[6] We will use the term *variation* to mean variability as a function of modulation frequency θ.

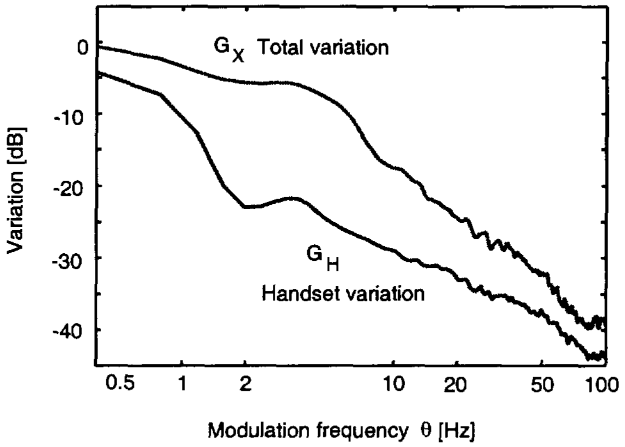
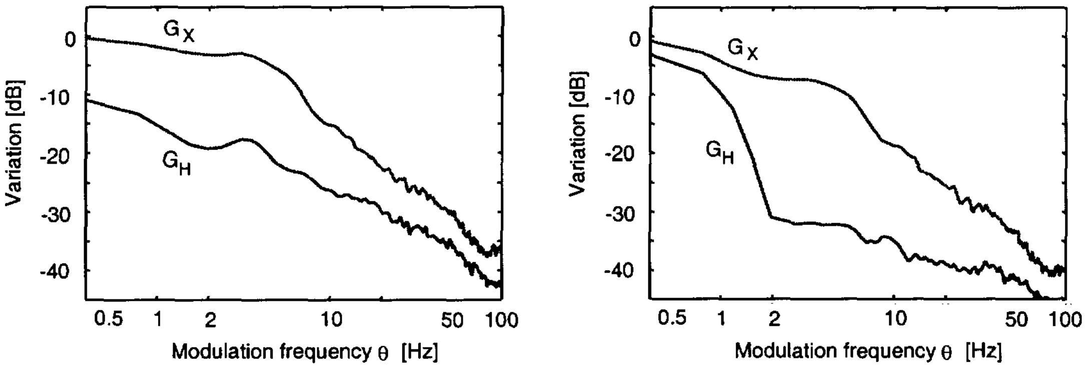


Figure 3.3: Total variability and handset variability as a function of modulation frequency θ. ($\theta_s = 200$ Hz and $t_w = 20$ ms.)

In the figure, the upper curve is the total variation $G_X(\theta)$ (or generalized modulation spectrum of speech) as computed over 15 speakers in the HTIMIT corpus. The lower curve is the handset variation $G_H(\theta)$ as computed over the three electret- and three carbon-button transducers *jointly*. For modulation frequencies $\theta > 5$ Hz, it can be seen that the total variation falls off by about 7 dB per octave, whereas the handset variation falls off by about 3.5 dB per octave.

It can be seen that handset variability contributes considerably to the total variability at very low modulation frequencies ($\theta < 1$ Hz), and at moderate to high modulation frequencies ($\theta > 10$ Hz). The high handset variability at the very low modulation frequencies suggests a strongly varying DC component in the time sequences of the logarithmic energies. This result conforms with the model for a convolutional degradation that was described in Section 2.5.1, where it was concluded that a linear time-invariant transmission channel will add a bias to the time sequences of logarithmic energy.

Fig. 3.4 depicts the total and handset variations as computed over the two types of microphone transducer (carbon-button and electret) *individually*. Fig. 3.4 (a) depicts the case where $G_H(\theta)$ was estimated over the three carbon-button transducers only, while Fig. 3.4 (b) depicts the case where $G_H(\theta)$ was estimated over the three electret transducers. The variation among carbon-button transducers is clearly different from the variation



(a) Intra carbon-button

(b) Intra electret

Figure 3.4: Total variability and handset variability as a function of modulation frequency θ. (a) Depicts variations among carbon-button transducers. (b) Depicts variations among electret transducers. ($\theta_s = 200$ Hz and $t_w = 20$ ms.)

among electret transducers. Interestingly, it can be seen that the variability of electret transducers is less in the middle range of modulation frequencies than the variability of carbon-button transducers. Why this is the case is not entirely clear, but we will offer a possible explanation at the end of the chapter.

3.5 Limitations of the Analysis

It is informative to consider the limitations of the analysis of variability in the modulation spectral domain. In this section the effects of frequency smearing, aliasing, and alignment of the speech signals are investigated in more detail.

3.5.1 Frequency Smearing

As was shown in Section 2.2, the 1 second long Hamming window used to compute the modulation spectra limits the frequency resolution to 2 Hz. Fig. 3.3 showed that the handset variation $G_H(\theta)$ exhibits a strong peak below about 2 Hz. We suggest that the actual peak may be well below 2 Hz – close to DC, and that the observed peak is the result of a smearing of this DC component into higher frequencies – refer Equation (2.3). Fig. 3.5 supports this suggestion. In the figure the handset variation $G_H(\theta)$ computed using a 2 second long analysis window (b) is compared to that computed when using a

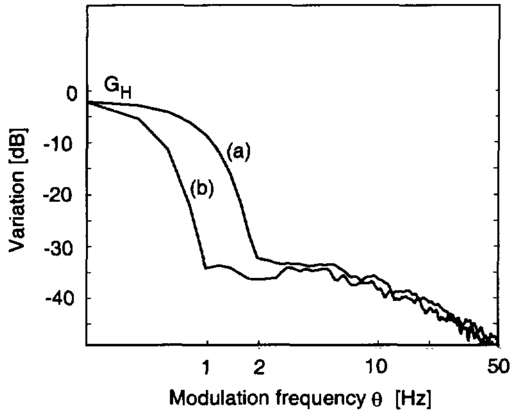


Figure 3.5: Handset variability as a function of modulation frequency θ for medium-term analysis Hamming window lengths of (a) 1 second, and (b) 2 seconds. ($\theta_s = 100$ Hz and $t_w = 40$ ms.)

1 second long analysis window (a). For the 2 second long analysis window, where the frequency resolution is limited to 1 Hz, the peak moves down to below 1 Hz.

Fig. 3.5 also indicates that the choice of medium-term analysis window is important. It was remarked in Section 2.2 that the main side lobe of a Hamming window is about 42 dB below the main lobe, whereas this is only about 13 dB for a rectangular window. A rectangular window would therefore be inappropriate for the medium-term analysis since it would not be able to resolve the sharp 30 dB change in handset variation that can be observed in the figure. It can therefore be concluded that the Hamming window would be the better choice for the medium-term analysis window. It is important to note though that even the Hamming window would not be able to resolve sharp changes of more than about 40 dB.

3.5.2 Aliasing .

In Section 2.4.3 it was noted that to prevent aliasing in the modulation spectral domain it is necessary to sample with a frame rate that is higher than twice the bandwidth of the analysis window that is used in the short-term analysis. Table 2.1 listed typical minimum values for the frame rate that avoids aliasing when using a short-term analysis Hamming window of a certain length. Window lengths t_w of 20, 32 and 40 ms respectively, required frame rates θ_s of 200, 128 and 100 Hz to avoid aliasing. In this subsection the effects of

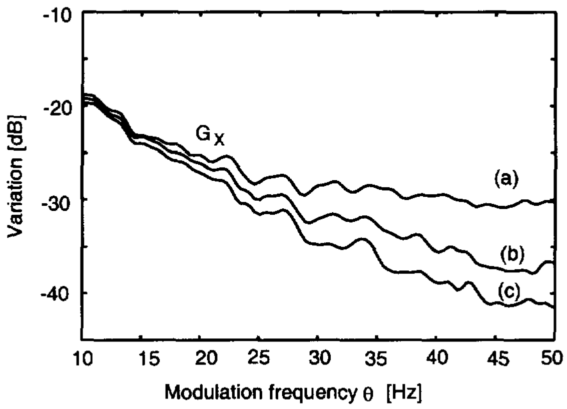


Figure 3.6: Total variability as a function of modulation frequency θ for a frame rate $\theta_s = 100$ Hz and short-term analysis Hamming window length t_w of (a) 20 ms, (b) 32 ms, and (c) 40 ms.

aliasing when these conditions are not met are examined using the modulation spectrum.

Fig. 3.6 shows the total variation $G_X(\theta)$ for different lengths of the Hamming window. The frame rate is $\theta_s = 100$ Hz and the Nyquist frequency is 50 Hz. Cases (a) and (b) in the figure are to be compared to case (c). For case (c) where $t_w = 40$ ms, no aliasing occurs, whereas for case (a) with $t_w = 20$ ms aliasing theoretically occurs above 25 Hz and for case (b) with $t_w = 32$ ms aliasing theoretically occurs above 36 Hz. While it is important to consider the potential for aliasing in the interpretation of the results presented here, it appears that aliasing will not be a considerable problem due to the band-limited nature of the modulation spectrum of speech. In Fig. 3.3 $G_X(\theta)$ was seen to fall off rather rapidly at about 7 dB/octave so that at $\theta = 36$ Hz $G_X(\theta)$ is already about 25 dB down from its the maximum value. Indeed, in Chapter 4 it will be shown that speaker verification performance is optimum for choices of analysis window length anywhere in the range 20 to 64 ms. Given these observations, computational efficiency suggests using a 100 Hz frame rate and a 20 – 32 ms long analysis window (i.e. not more than 256 samples). A 32 ms long analysis window is the default choice in this dissertation.

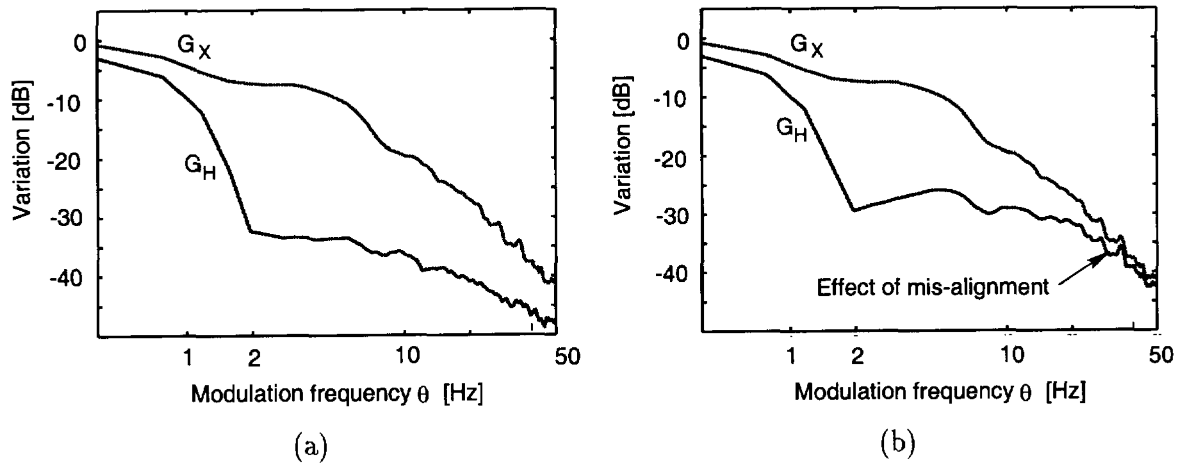


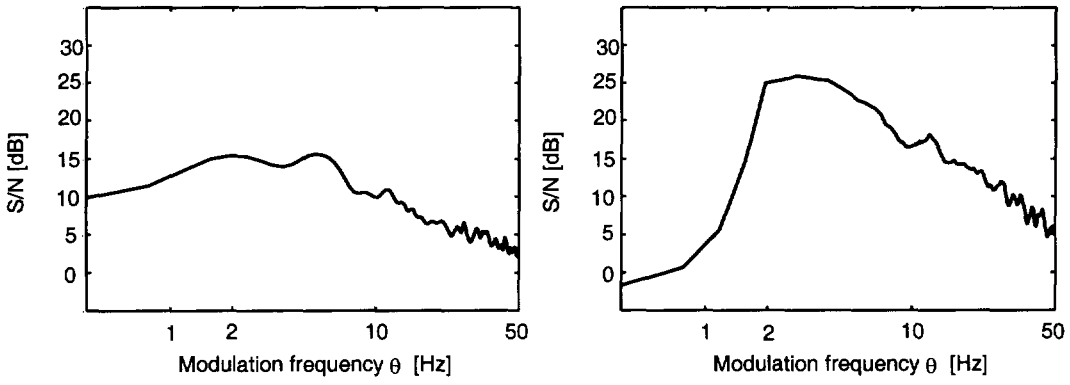
Figure 3.7: Total variability and handset variability as a function of modulation frequency θ. (a) Time sequences aligned, (b) time sequences randomly shifted by one frame. (Electret speech, $\theta_s = 100$ Hz and $t_w = 40$ ms.)

3.5.3 Time Alignment

The results in this chapter are based on time-aligned recordings for a particular speaker speaking the same text over different handsets. It was reported in Section 3.2 that the alignment errors are on the order of 3 ms or less. The alignment can be tested by comparing variations for time sequences that were randomly shifted by one frame to the variations of the original time sequences. This is shown in Fig. 3.7. Fig. 3.7 (a) shows the variations computed using the original time sequences and Fig. 3.7 (b) shows the variations computed using the randomly shifted sequences. In the latter it can be seen that handset variability, that reflects the alignment variability, becomes the dominant variability at high modulation frequencies. This suggests that shifting the time sequences by a single frame, increases the alignment error and suggests that the original alignment is optimal.

3.6 Additional Results

Additional results related to handset variability are presented in this section. Speech-to-handset signal-to-noise ratios are studied individually for electret and carbon-button transducers and as a function of short-term analysis frequency in Section 3.6.1. Differences



(a) Carbon-button

(b) Electret

Figure 3.8: SNR for (a) carbon-button and (b) electret transducer variability as a function of modulation frequency θ. ($\theta_s = 100$ Hz and $t_w = 40$ ms.)

between the modulation spectrum resulting from the classical definition and the modulation spectrum resulting from the generalized definition are examined in Section 3.6.2. Finally, the effects of additive noise and signal gain are investigated in Section 3.6.3.

3.6.1 Signal to Noise Ratios

Comparing speech and handset variations may provide an indication of which components of the modulation spectrum may be useful for ASR and speaker verification. A higher SNR for a particular modulation frequency implies that the speech signal is affected less by handset variability at that frequency. Define a signal-to-noise ratio (SNR) at a particular modulation frequency as the ratio of the speech variation and handset variation

$$[S/N](\theta) = \frac{G_{S'}(\theta)}{G_H(\theta)} = \frac{G_X(\theta) - G_H(\theta)}{G_H(\theta)}. \tag{3.6}$$

Fig 3.8 depicts the variations in terms of the SNR for handsets with (a) carbon-button and (b) electret transducers. It is can be seen that the SNR is highest in a band of modulation frequencies centered around about 3 Hz. This band extends down to 1 Hz and up to around 10 or 20 Hz. For electret transducers the SNR is higher than 15 dB in a band from 1 to 10 Hz. The SNR is somewhat lower for carbon-button transducers. Possible reasons for this will be suggested in the next section.

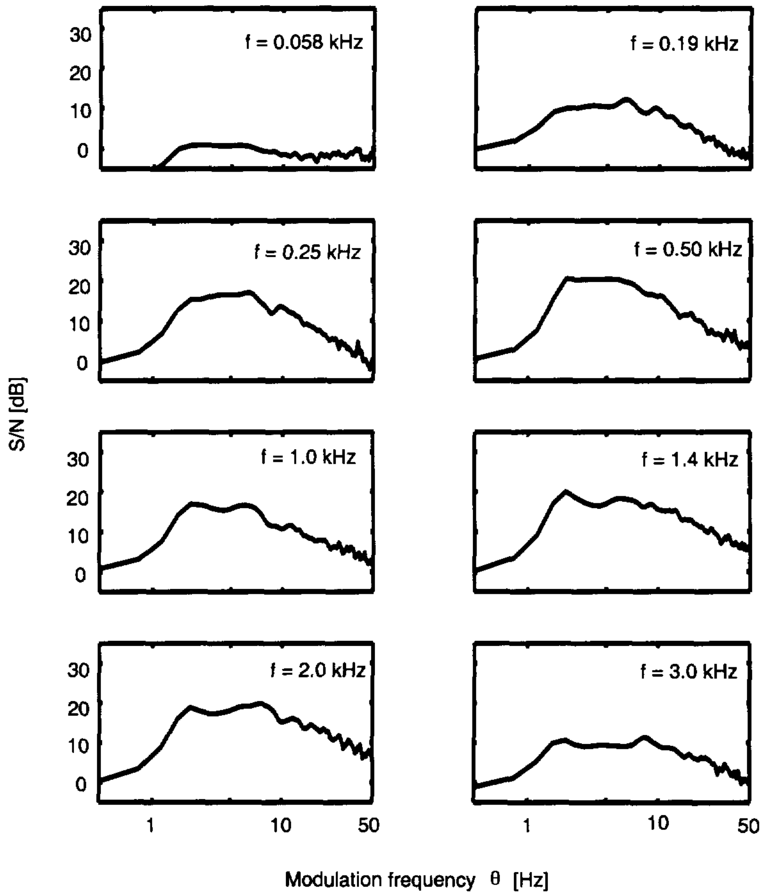


Figure 3.9: SNR as a function of modulation frequency θ for various short-term analysis frequencies f. ($\theta_s = 100$ Hz and $t_w = 40$ ms.)

Fig. 3.9 depicts the SNR as a function of short-term analysis frequency f. Results are shown for filter bank bands with center frequencies f spanning the telephone band. The SNRs shown were computed using handsets with both carbon-button and electret transducers. The same conclusions as before hold for the low and high modulation frequencies. Here though, it can be seen that the SNR also depends on the short-term analysis frequency, and is highest for a short-term analysis frequency in a range of 0.3-3 kHz. The dominant range of frequencies agrees with the typical bandwidth of telephone speech [20] as can be seen in Fig. 3.10 which shows the SNRs as a function of short-term analysis frequency f at a modulation frequency $\theta = 4$ Hz.

The signal-to-noise-ratios in the figures suggest that moderate to high and very low

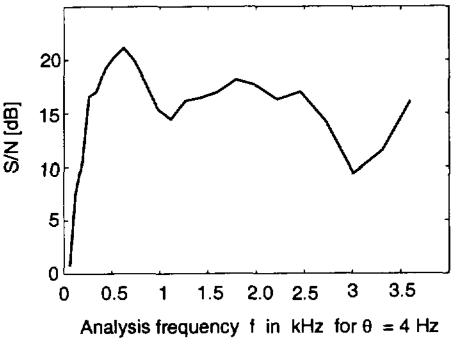


Figure 3.10: SNR as a function of short-term analysis frequency f for the case where $\theta = 4$ Hz. ($\theta_s = 100$ Hz and $t_w = 40$ ms.)

modulation frequencies are susceptible to handset variability and that attenuating them may reduce unwanted variability in the acoustic features. This observation will be investigated in more detail in Chapter 5.

3.6.2 Comment on the Use of Long-term or Ensemble Average

As was noted in Section 3.3.2, in its classical definition [41, 7], the modulation spectrum is defined as the power spectrum of a time sequence of logarithmic energy from which the *long-term* average has been subtracted. The definition was generalized in this chapter to define the modulation spectrum as the power spectrum of a time sequence of logarithmic energy from which the *ensemble average* has been subtracted. Fig. 3.11 shows the modulation spectra that result from the two definitions. The solid lines show the modulation

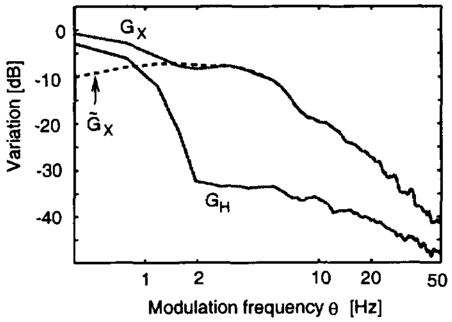


Figure 3.11: Comparison between two definitions for modulation spectra. See text for details. (Electret speech, $\theta_s = 100$ Hz and $t_w = 40$ ms.)

spectrum or total variation G_X and handset variation G_H that comes from the generalized definition. Compared to these, the dashed line shows the modulation spectrum $\tilde{G}_X$ computed using the classical definition. For the latter, the long-term average of each time sequence was normalized before computing its power spectrum.

The difference between the two definitions for the modulation spectrum can be seen to be in the DC-component. In the classical definition, the DC-component is attenuated[7] and essentially ignored, whereas this is not the case in the generalized definition.

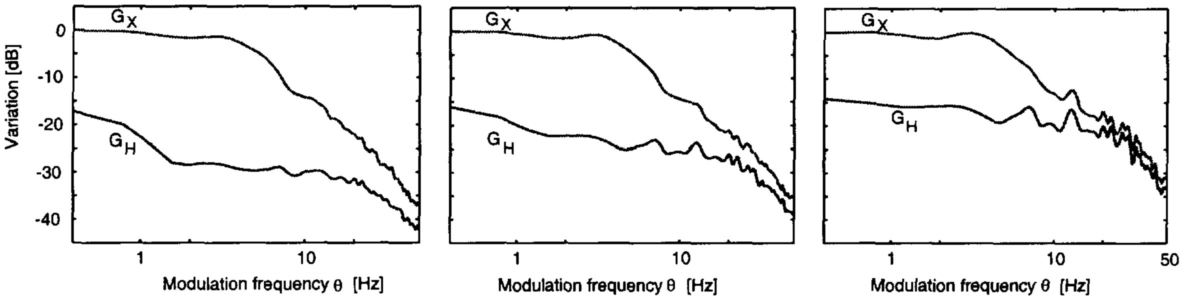
3.6.3 Additive Noise

In previous sections it was seen that the handset variability ascribed to variations among carbon-button transducers is different from the handset variability ascribed to variations among electret transducers. A possible explanation is offered in this section.

First of all, notice that since the DC component in the logarithmic domain reflects the energy in the speech signal, it may be affected among other things by a normalization of the speech signal. This would be the case when changing the gain of the transmission channel. The observed differences in DC variation for the carbon-button and noisy electret handset may therefore depend on whether the noise was added to the speech signals before or after a change in the broad-band speech signal energy. To compensate for this effect the broad-band energy of each speech signal will be normalized. This only affects the DC component.

Now consider the effect of additive noise. It is proposed that additive noise that may be present in the carbon-button speech signal may be in part responsible for the observed difference between transducer types. This can be modeled by artificially adding white noise to the electret speech signal. The goal is to try to make the handset variation among electret transducers look more like the handset variation among carbon-button transducers. In the following, noise will be added in two different ways.

[7] Note that the DC component is not exactly zero as may be expected. This is because in estimating the modulation spectrum, the analysis window was applied to the time sequence *after* the long-term average was removed from it. The non-zero DC component reflects the fact that the long-term average of the windowed time sequence is not exactly zero.



(a) SNR = 30 dB (b) SNR = 20 dB (c) SNR = 10 dB

Figure 3.12: Total variability and handset variability as a function of modulation frequency θ. The effect of adding noise to speech signals recorded using electret transducers is shown for noise levels at an SNR of (a) 30 dB, (b) 20 dB and (c) 10 dB. See text for details. ($\theta_s = 100$ Hz and $t_w = 40$ ms.)

Fixed SNR

Suppose that white noise[8] is added at a fixed SNR to *each* electret speech signal. The SNR is measured in decibels as the ratio of average energy in the speech signal relative to the average energy of the noise signal. Fig. 3.12 depicts the resultant total and handset variations for different SNR levels of additive noise.

Variable SNR

Suppose instead that white noise is added at SNR levels that vary randomly from 10 – 30 dB among the electret speech signals. Fig. 3.13 depicts the resultant total and handset variations.

Discussion

Several important conclusions can be drawn from the analysis. Variations in broad-band energy contribute considerably to the DC variability. This can be seen by comparing the handset variations shown in the current section, with the handset variations shown in previous sections. In the current section the broad-band energy of each speech signal was normalized, whereas in the previous sections the broad-band energy of each speech signal

[8] White noise from the NOISEX-92 corpus is used.

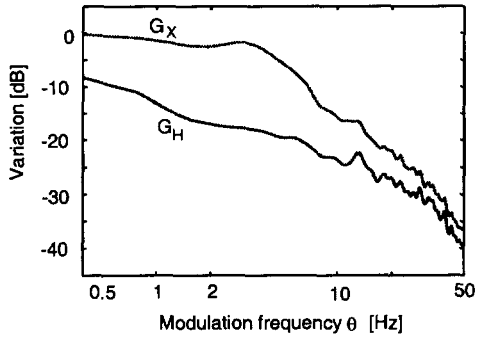


Figure 3.13: Total variability and handset variability as a function of modulation frequency θ. The effect of adding noise to the speech signals recorded using electret transducers is shown. Noise is added at SNRs that vary from 10 to 30 dB. ($\theta_s = 100$ Hz and $t_w = 40$ ms.)

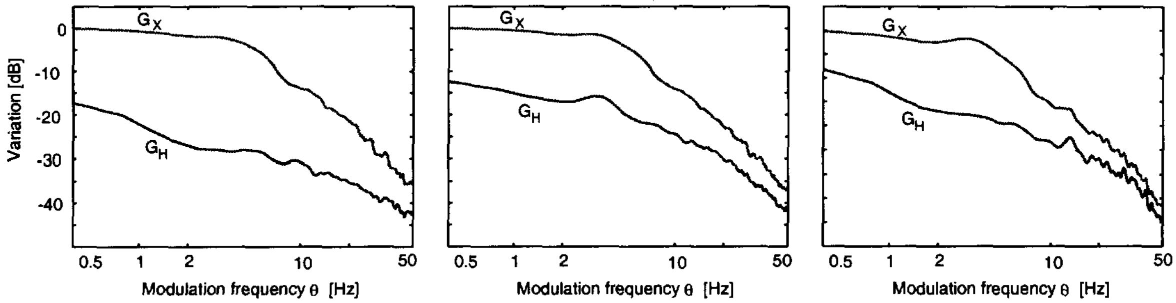
was not normalized. Normalizing the broad-band energy can be seen to reduce handset variability for the DC modulation component.

It can be seen that adding noise to the electret speech affects the SNR at all modulation frequencies. The behavior agrees with the conclusion at the end of Section 2.5.2 that the effects of additive noise in the log-spectral domain is time-varying.

The intra-electret and intra-carbon-button handset variations can be compared to the noisy electret handset variation. The variations are shown in Fig 3.14. In Fig. 3.14 (c) the variations were computed after adding white noise to the electret speech signals at SNR levels that varied randomly from $10 - 30$ dB. Fig. 3.14 (c) is the same as Fig. 3.13.

It can be concluded that additive noise may indeed be partly responsible for the observed difference in carbon-button and electret handset variation. The intra electret handset variation in the presence of additive noise is quite similar to the intra carbon-button handset variation. This supports the assumption that additive noise may be responsible for the observed differences in variation between the transducer types.

Importantly, it can be seen that increasing the additive noise leads to an increase in the handset variation at all frequencies with the effect that handset variation contributes substantially to the total variation at high modulation frequencies. This can be seen best in Figs. 3.12 (b-c). The additive noise has basically a 'pinching' effect. When the signal to noise ratio for the additive noise is around 10-20 dB it can be seen that frequency



(a) Electret (b) Carbon-button (c) Noisy electret

Figure 3.14: Total variability and handset variability as a function of modulation frequency θ. (a) Intra electret, (b) intra carbon-button and (c) intra noisy electret transducer variability. ($\theta_s = 100$ Hz and $t_w = 40$ ms.)

components in the modulation spectrum of speech that are higher than about 10 Hz contain relatively little speech variability compared to the noise variability. It can be concluded that attenuating the higher modulation frequency components may make the features less sensitive to such changes and perhaps increase system robustness.

Notice also from the figure that additive noise has the effect of increasing the total variation at higher modulation frequencies relative to the total variation at lower modulation frequencies. This suggests that aliasing, if allowed to occur, may become a problem at high levels of additive noise and should not be ignored.

3.7 Summary

In this chapter it was proposed that the modulation spectrum provides information that may be used to understand and reduce the effects of telephone handset variability. The approach that was followed was different from previous approaches in that characteristics of handset types were analyzed along the time dimension of the feature space and not along the frequency dimension. A framework for this analysis was proposed based on a nested ANOVA model. The model generalized the concept of modulation frequency and allowed the variations to be interpreted in the modulation spectral domain.

Several results followed and led to a number of important conclusions which we summarize.

- Handset variability constitutes a large portion of the overall variability in the speech signal at very low (DC) and moderate to high (> 10 Hz) modulation frequencies. Gain in particular, was seen to affect the DC modulation frequencies.

- The ratio of speech variation to handset variation is highest in a band of modulation frequencies centered around 3 Hz. This band extends down to 1 Hz and up to around 10 or 20 Hz.

- Carbon-button transducers exhibited a lower SNR than electret transducers.

- It was shown that the SNR also depends on the short-term analysis frequency, and is highest for a short-term analysis frequency in a range of 0.3-3 kHz. This observation agreed with a bandwidth for telephone speech in the same range. Irrespective of short-term analysis frequency though, it was observed that very low and moderate to high modulation frequencies consistently have lower SNRs.

- It was shown that variations in the broad-band energy of the speech signal contributes substantially to the DC variability and that a normalization is appropriate.

- It was seen that additive noise affects the modulation spectrum across the range of modulation frequencies.

- By adding noise to the speech that was recorded using the electret transducer, it was concluded that additive noise may indeed be partly responsible for the observed difference in carbon-button and electret handset variation and that adding such noise caused the electret handset variation to better approximate the carbon-button handset variation.

- When additive noise is present in the speech signal, it was seen that frequency components in the modulation spectrum of speech that are higher than about 10 Hz convey relatively little speech information. This was seen to be the case for additive noise at SNR levels as low as 20 dB.

This chapter presented a rather general characterization of telephone handset variability in speech and did not specifically deal with speaker variability. Handset variation was contrasted to the total and speech variations. It was concluded that attenuating the higher modulation frequencies may make the features less sensitive to adverse environments and likely increase the robustness of an ASR or speaker verification system. We will test this conclusion in Chapter 5 by performing a systematic investigation of the relative importance of the components of the modulation spectrum for speaker verification specifically. By using the speaker verification system that we describe in the next chapter, the approach will be to measure and analyze the effect on the speaker verification error for various filters designed in the modulation spectral domain.

Chapter 4

Speaker Verification

The purpose of this chapter is to describe and motivate the system that will used to evaluate text-independent speaker verification performance in this dissertation. The extraction and parameterization of suitable acoustic features, and choice, optimization and assessment of a statistical model for these features are described.

The chapter is organized as follows. Section 4.1 describes the specific processing and parameterization used for feature extraction. Section 4.2 formulates speaker verification as a problem in statistical hypothesis testing. This formulation motivates the use of a likelihood ratio test. A likelihood ratio test statistic is introduced that uses two probability density distribution functions (pdfs). The one describes a *speaker independent* (SI) distribution of the features and the other a *speaker dependent* (SD) distribution. Performance of the speaker verification system is measured using different evaluation measures. These measures are motivated by the requirements of recent speaker recognition evaluations run by the National Institute of Standards and Technology. Section 4.3 provides further details of the statistical modeling. To decide on a suitable choice for parameterizing the pdfs, Section 4.3.1 reviews past and current approaches to modeling the acoustic features. A Gaussian mixture modeling approach is then motivated and described. The well-known Expectation-Maximization (EM) algorithm is used to estimate the parameters in the SI model and a Bayesian maximum aposteriori (MAP) adaptation of the SI model is used to derive the SD models. Finally, Section 4.4 presents results related to optimization of the various parameters.

4.1 Feature Extraction and Parameterization

Features derived from a short-term analysis using a filter bank were motivated in Chapter 2 in terms of their suitability for speaker verification. This section details the feature extraction and parameterization used in the remainder of the dissertation. Following the common approach [16], unless where otherwise noted, acoustic features for the speaker verification experiments are derived from a short-term analysis of the speech signal as follows. A 32 ms long Hamming window is advanced in $t_n - t_{n-1} = 10$ ms steps. For each time point t_n, the logarithmic energies $X_1(n, k) = \log |S(n, w_k)|^2$ at frequencies w_k, where $k = 1, ...K$, are computed from the squared magnitude FFT using a triangular integration window in a manner similar to that of the computation of Mel-frequency cepstral coefficients [16]. Fig. 4.2 depicts the filter bank. In this computation the center frequencies of the filters are warped according to the Mel scale $f' = 2595 \log_{10}(1 + f/700)$. Since the purpose is to process telephone speech, only the $K = 19$ spectral energies falling within the range $[f_l, f_h] = [200, 3500]$ Hz are retained. This process is depicted in Fig. 4.2. In this dissertation specifically, the time sequences of logarithmic energies are further processed by one or more FIR filters as depicted in the figure. The effects of various choices for these filters are analyzed in Chapter 5. After filtering, speech frames with energies more than 3 dB below a floating noise floor in the signal are labeled as silence (about 40% of the samples). The noise floor is estimated as the cross-over point of a two-component

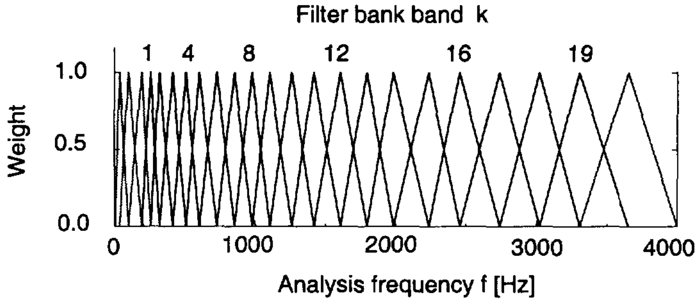


Figure 4.1: Filter bank used in deriving short-term acoustic features. The integration window for each filter bank is shown. The filter bank bands falling between 200 and 3500 Hz are shown as solid lines.

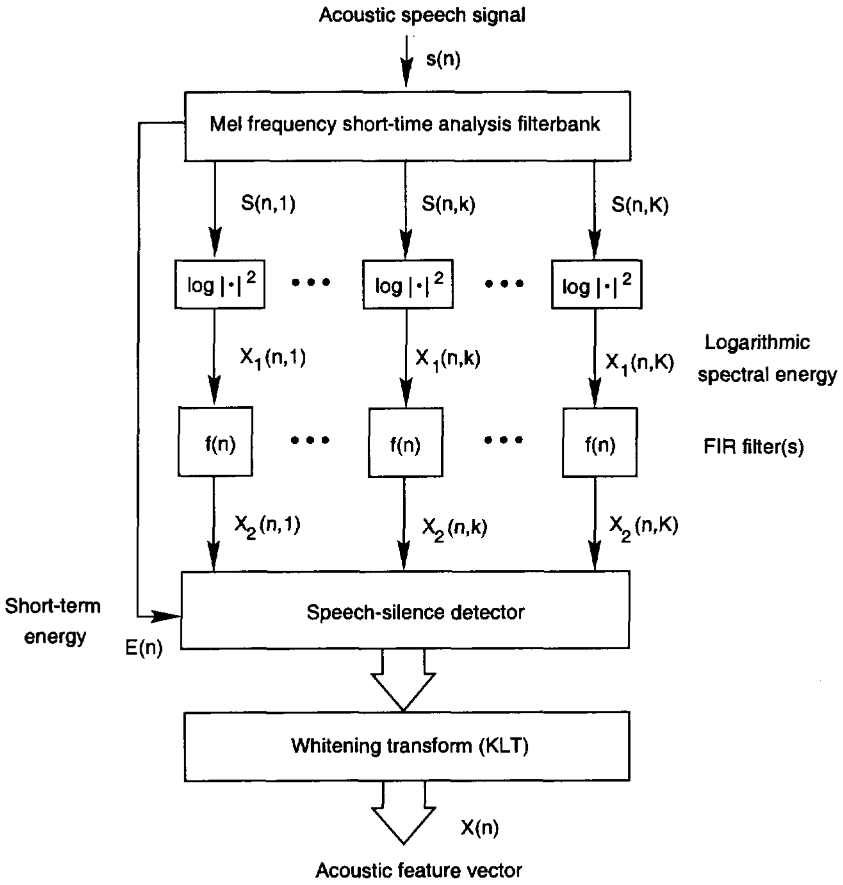


Figure 4.2: Acoustic feature processing.

Gaussian mixture fit [97] to the broad-band short-term logarithmic energy that is estimated within a sliding two second long window. The estimate is updated two times per second. Silence samples are discarded and the remaining vectors decorrelated using the optimal Karhunen-Loéve transformation (KLT). The KLT basis functions are estimated from the training data [25] to form the acoustic features $\boldsymbol{X} = [X(n)]$. In contrast to other approaches [78], cepstral coefficients are not computed. Parameterizations were optimized on data from the 1996 NIST Speaker Recognition Evaluation (NIST-SRE) [57]. An optimization study is included at the end of the chapter.

4.2 Statistical Hypothesis Testing and Likelihood Ratio Test

In the verification task the system has to decide whether to accept or reject the identity of a speaker claiming to be the source of the speech signal. This suggests testing the hypothesis that the claimant speaker is the true speaker. The following development closely follows that outlined in [47, 48] where speaker verification was formulated as a *statistical hypothesis* testing problem.

Given acoustic features $\boldsymbol{X}$, test the *null hypothesis*, H_0, against the *alternative hypothesis*, H_1, where H_0 assumes that $\boldsymbol{X}$ was generated by speaker Y_0 and H_1 assumes that $\boldsymbol{X}$ was generated by another speaker Y_1. Usually the other speaker is not specified. In this case the alternative hypothesis is that $\boldsymbol{X}$ was *not* generated by speaker Y_0 which makes H_1 a composite hypothesis. If the probability density functions (pdfs) of the acoustic features $\boldsymbol{X}$ under the two hypotheses are known exactly and fall into a specific class of distributions such as the exponential family, techniques are available in the literature for designing an optimal test. In general however, given the fact that H_1 is composite, and that the two pdfs $p(\boldsymbol{X}|H_0)$ and $p(\boldsymbol{X}|H_1)$ are not known, but have to be estimated from a limited number of data samples, no known optimal testing procedure exists (see [48] for a discussion).

Good performance can however be achieved by adopting the likelihood ratio statistic [69, 53, 80, 48]. The test statistic is then given by

$$T(\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\Theta_0)}{p(\boldsymbol{X}|\Theta_1)} \geq \lambda^* \tag{4.1}$$

where given acoustic features $\boldsymbol{X}$, the null hypothesis is accepted if the statistic is larger than a test threshold λ^*. In this formulation Θ_0 and Θ_1 are model parameters characterizing H_0 and H_1 respectively and $p(\boldsymbol{X}|\Theta_0)$ and $p(\boldsymbol{X}|\Theta_1)$ are the likelihoods that the acoustic features $\boldsymbol{X}$ were generated under the two competing hypotheses. Since the number of possible sources for $\boldsymbol{X}$ can be large under the alternative hypothesis that $\boldsymbol{X}$ was *not* generated by speaker Y_0, it may be assumed that $p(\boldsymbol{X}|H_1) \approx p(\boldsymbol{X})$ so that $p(\boldsymbol{X}|\Theta_1)$ can be chosen to approximate $p(\boldsymbol{X})$.

Given the preceding discussion then, the log-likelihood ratio statistic is based on a *speaker dependent* (SD) pdf $p(X|\Theta_0)$ and a *speaker independent* (SI) pdf $p(X|\Theta_1)$. It is convenient to use the log-likelihood ratio instead of the likelihood ratio directly. It suffices to accept the hypothesis that the speaker associated with the target model Θ_0 generated $\boldsymbol{X}$ if

$$\log \frac{p(\boldsymbol{X}|\Theta_0)}{p(\boldsymbol{X}|\Theta_1)} \geq \lambda. \tag{4.2}$$

If it is assumed that the acoustic features are random vectors $X(1), X(2), ..., X(n), ..., X(N)$ drawn from the same distribution, and that the vectors are independent and identically distributed the log-likelihood ratio statistic becomes

$$t(\boldsymbol{X}) = \log \prod_{n=1}^{N} \frac{p(X(n)|\Theta_0)}{p(X(n)|\Theta_1)} = \sum_{n=1}^{N} \log p(X(n)|\Theta_0) - \log p(X(n)|\Theta_1). \tag{4.3}$$

Since the $t(\boldsymbol{X})$'s are also independent and identically distributed [25], with

$$\begin{aligned} E[t(\boldsymbol{X})|H_i] &= NE[t(X)|H_i], \\ \text{Var}[t(\boldsymbol{X})|H_i] &= N\text{Var}[t(X)|H_i], \end{aligned}$$

normalizing by N gives a statistic that does not directly depend on the length of the utterance. That is, given a target and acoustic features $\boldsymbol{X} = [X(n)]$, $n = 1...N$, the following statistic or *model score* is compared to the threshold

$$\frac{1}{N} t(\boldsymbol{X}) = \frac{1}{N} \sum_{n=1}^{N} \log p(X(n)|\Theta_0) - \log p(X(n)|\Theta_1) \geq \lambda. \tag{4.4}$$

The speaker that generated $\boldsymbol{X}$ is called the *true speaker* and a speaker that did not generate $\boldsymbol{X}$, but claims to have generated $\boldsymbol{X}$, is called an *imposter*. The speaker associated with the claim is called the *target*. Two types of error can occur. A *false rejection* error, often referred to as type I error, occurs when a true speaker is falsely rejected for being an imposter. A *false acceptance* error, often referred to as type II error, occurs when an imposter is accepted as a true speaker. An evaluation measure E can be constructed as a combination of the type I and II errors

$$\begin{aligned} E &= E_{\text{fr}} + E_{\text{fa}} \\ &= C_{\text{fr}}\, p(fr|H_0)\, P(H_0) + C_{\text{fa}}\, p(fa|H_1)\, P(H_1) \end{aligned} \tag{4.5}$$

where C_{fr} and C_{fa} are the costs associated with making the errors.

In this dissertation verification performance is evaluated using the equal error rate (EER) and a minimum decision error (MDE). These two evaluation measures weighs the type I and II errors differently. The EER weighs the false rejection error probability $p(fr|H_0)$ and the false acceptance error probability $p(fa|H_1)$ equally and is defined as the minimum error for which $C_{\text{fa}} = C_{\text{fr}} = 1$ and $P(H_0) = P(H_1) = 0.5$. The MDE weighs false acceptance errors about 10 times more strongly than false rejection errors. For the MDE[1] $C_{\text{fr}} = 10$, $C_{\text{fa}} = 1$, $P(H_0) = 0.01$ and $P(H_1) = 1 - P(H_0) = 0.99$.

It is informative to plot the tradeoff of false acceptance and false rejection errors. This can be done conveniently using a detection error tradeoff (DET) plot [56]. Fig. 4.3 shows an example of a DET plot with EER point, MDE point and HDE point (see following discussion) indicated.

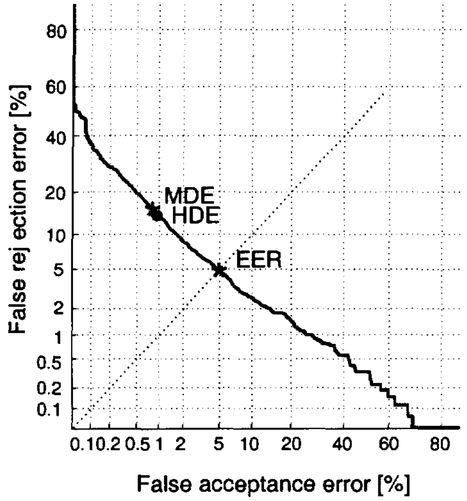


Figure 4.3: DET plot with EER, MDE and HDE points. (See text for details.)

Errors are estimated using *test* data. The test data refers to data obtained during actual use of the system and is to be contrasted to independent *training* and *development* data that were obtained previously during an enrollment stage. The training and development data are used to construct models for the target speakers whereas the test data

[1] The values for the costs and prior probabilities originated with the Speaker Recognition Evaluations run by NIST [60]. The MDE used here is the same as the *minimum detection cost* (min C_{DET}) that was defined and used in the NIST-SRE.

are used to assess the performance of the system.

A target speaker is accepted or rejected based on a comparison of the target's model score and a predetermined threshold. The threshold is estimated as the threshold that minimizes the EER or MDE evaluation measure. Whether to use the development or test data when determining the threshold is examined next.

In a real application, it is not possible to determine an optimum threshold using the test data as it is still being collected and not yet verified. A common approach is to determine a threshold that is optimum for previously collected development data and then apply that threshold to the incoming test data. In this scenario the threshold is not guaranteed to be optimal for the test data, since the statistics of the development and test speech may be different. To prevent the incorrect setting of a threshold on one data set to affect the conclusions reached for selecting useful features based on performance on another data set, the approach followed here will be to assume that the test data are available for computation of the threshold. Accordingly, unless otherwise specified, for both the EER and MDE, results will be reported at the optimum threshold as determined on the test data.

In Appendix C, the threshold that minimizes the MDE on development data will be used to assess performance on the test data. The resulting error will be called a "hard decision error" (HDE). Fig. 4.3 showed an example of a DET plot with EER point, MDE point and HDE point indicated.

4.3 Statistical Model

This section describes the statistical model used for speaker verification for the experiments in this dissertation. Derivation of a test statistic for speaker verification suggested use of a speaker independent and speaker dependent model. To guide and motivate a suitable parameterization, this section includes a discussion and review of existing modeling approaches for speaker verification.

4.3.1 Existing Approaches – Discussion and Review

Many different statistical models have been used for speaker verification. These include:

- the long-term average of spectral features, first proposed in [71] and used extensively in [4];
- the second order statistics [8, 88];
- vector quantization (VQ) of the features, first proposed for speaker recognition in [51] and investigated in [61];
- neural networks [63, 24];
- nearest neighbor classification [38];
- divergence and Battacharyya distance measures [14];
- a Gaussian mixture model (GMM), first proposed as a 5-state ergodic, autoregressive HMM in [70] and studied extensively in [78, 79]; and
- a hidden Markov model (HMM), first proposed for speaker recognition in [70] and used extensively in [62, 85, 61, 59].

The long-term average of spectral features was shown early on to contain speaker specific information [71]. In later work however, it was found to be rather sensitive to the transmission channel. A strategy to make the modeling less sensitive to perturbations of the long term average was proposed in [29]. The idea was to reduce the effect of an unreliable average, by a suitable weighting of a likelihood term for the sample average and a likelihood term for the sample covariance.

Temporal dynamics have also been modeled. In [26] it was proposed to model the dynamics with a projection onto orthogonal polynomials. Time spans of around 90 ms were reported to give good modeling results. The so-called delta features (see Sections 2.5.1 and 5.4) have found wide spread use in contemporary systems [59] where the features are usually appended to a static feature representation (such as the STFT energies) and modeled statistically. Dynamic time warping (DTW) [76] has been used in text-dependent speaker verification systems and in [39] dynamics of speech in 80 ms intervals were reported to be important and modeled using so-called "elastic" filler templates.

Normalizations for the effects of adverse environments and/or systematic deviations from the assumed statistical model have played an important role in speaker verification systems. Two types of normalization have been prevalent, namely parameter domain normalizations and similarity domain (score) normalizations. Removing the long term average of the features [26] is an example of a parameter domain normalization. The weighted mean and covariance model [29] discussed previously is an example of a similarity domain normalization. Other score normalizations have been proposed. In [80, 58] it was proposed that scores be normalized with respect to individual speaker (z-norm) and handset type (h-norm) categories. In this dissertation the focus is on parameter domain normalizations. Score normalizations of the type mentioned here are not performed. A number of approaches have also considered the removal of outliers [29, 61].

4.3.2 Proposed Approach – Motivation

In this dissertation a Gaussian mixture model (GMM) is used to model the SI and SD distributions. Motivation for this choice is as follows.

Feature distribution In [104] it was shown that logarithmic energies have strongly non-Gaussian distributions. In the current work the KLT is applied to the logarithmic energies to derive the features that are to be modeled statistically. However, it is expected that this transform will not be able to adequately compensate for the non-Gaussianity of the logarithmic energies. The implication is that the first and second order statistics may not be adequate to model the features properly. In informal experiments we found that modeling a particular time-sequence of logarithmic energy $X_1(n, k)$ directly with histograms with on the order of 25 bins, reflected the results and tendencies obtained with GMMs, although exhibiting a higher overall error rate. The conclusion is that a mixture of Gaussians or other suitable non-Gaussian distribution would be needed for accurate modeling.

Complexity HMM-based systems (of which GMMs are a special case) have been reported to outperform DTW [62] and VQ systems [86]. Limited evidence exists though to support a performance advantage for a more complex HMM model over a simpler GMM model. In [85] it was reported that an HMM with a left-right topology that is trained on sub-word units[2] gave better performance than VQ and an ergodic[3] HMM on a data base of isolated digit utterances of 100 speakers. On the other hand recently it has been reported [61, 58, 59] that HMM-based and GMM-based systems perform rather similarly. It has also been reported that there exists a tradeoff in performance between the number of mixture components per state K_1 and the number of states K_2 with reasonable performance achieved for a fixed product $K_1 \times K_2$ [61, 78, 59]. The implication is that a major modeling issue for speaker verification is the spectral and temporal resolution [85] of the model, perhaps more so than the exact state topology. A possible advantage of a GMM approach over an HMM approach is that a GMM does not require marked training data as does an HMM in the initial training phase.

Scalability The GMM approach that will be adopted here is more scalable than a neural network or similar approach which requires discriminative[4] training. Scalability was a requirement in the NIST-SRE.

The GMM described next, performed competitively in the 1998 NIST-SRE [101, 59].

4.3.3 Speaker Independent Model

The SI pdf is modeled using a mixture of Gaussians

$$p(X|\Theta_1) = \sum_{k=1}^{K} \omega_k p(X|\mu_k, \Sigma_k), \tag{4.6}$$

[2] Phone-like units (PLU) and acoustic segment units (ASU) were modeled. Similar performances were reported for both types of units.

[3] An HMM where every state is allowed to transition to every other state.

[4] If desired, the minimum classification error (MCE) could be used to make the GMM training discriminative. See for example [54, 53, 47]. It should be noted that interesting parallels exist between the MCE measure and a test statistic based on SD and SI pdfs.

with mixing weights $\omega_k \geq 0$, so that $\sum_{k=1}^{K} \omega_k = 1$, and with each Gaussian parameterized by a d-dimensional mean μ_k and a variance Σ_k as

$$p(X|\mu_k, \Sigma_k) = (2\pi)^{-d/2}|\Sigma_k|^{-1/2} \exp[-\tfrac{1}{2}(X - \mu_k)\Sigma_k{}^{-1}(X - \mu_k)^t)]. \tag{4.7}$$

While the main experiments in this dissertation use either $K = 128$ or $K = 256$ components, which represent a reasonable balance between modeling accuracy and computational complexity, some results will be reported for $K = 20$ to 1024 as well. The SI model is trained using the Expectation Maximization (EM) algorithm [19] on a population of male and female speakers different from those used in SD training and testing.

The mean of each component in the SI model is initialized to a centroid derived from the LBG algorithm [13] with centroids obtained by successive splitting starting from a single global centroid. The variance of each component is initialized to a scaled global covariance for the data and constrained to be diagonal. This constraint offers a computational advantage and does not appear to affect performance deleteriously [79]. The training procedure is detailed in Appendix B.

4.3.4 Speaker Dependent Models

The SD model of a speaker is obtained by updating the SI model using the training data of that speaker. Update rules were derived by unsupervised maximum aposteriori (MAP) reestimation of the parameters in the SI model [45, 18, 25]. This training is similar to that proposed in [80] for speaker verification and in [49] for ASR. Writing the posterior probability for the n-th training sample X_n being generated by the k-th mixture component as

$$p_k^n = \frac{\omega_k p(X_n|\mu_k, R_k)}{\sum_{j=1}^{K} \omega_j p(X_n|\mu_j, R_j)}, \tag{4.8}$$

the update rules are:

$$\text{weights:} \qquad \omega'_k = \frac{\sum_{n=1}^{N} p_k^n + \nu_k}{N + \sum_{k=1}^{K} \nu_k}, \tag{4.9}$$

$$\text{means:} \qquad \mu'_k = \frac{\sum_{n=1}^{N} p_k^n X_n + \eta \hat{\mu}_k}{\sum_{n=1}^{N} p_k^n + \eta}, \tag{4.10}$$

$$\text{variances:} \qquad \Sigma'_k = \frac{\sum_{n=1}^{N} p_k^n (X_n - \mu_k)(X_n - \mu_k)^t + \eta(\mu_k - \hat{\mu}_k)(\mu_k - \hat{\mu}_k)^t + \rho \hat{\Sigma}_k}{\sum_{n=1}^{N} p_k^n + \rho}, \tag{4.11}$$

where the meta-parameters ν_k, η and ρ reflect the respective confidence in the parameter estimates ω_k, μ_k and Σ_k.

During testing, the likelihoods for the SD and SI models are accumulated using only the five best scoring components identified for the SI model on the test vectors [80]. This is motivated in that only a few of the components in the mixture contributes to the likelihood for any one given acoustic vector and the SD model is adapted from the SI model. To decide whether to accept a claim or not, the likelihood ratio between the SD and SI models is compared to a speaker independent threshold[5].

4.4 Parameter Optimizations

Suitable values for the parameters described previously are determined in a series of optimization experiments in this section. The optimizations are based on the EER and MDE evaluation measures proposed in Section 4.2. Optimizations related to the feature extraction process are described in Section 4.4.1 and optimizations related to the statistical modeling process are described in Section 4.4.2.

Since it is prohibitive to search the whole parameter space exhaustively, the strategy adopted here is to perform line searches for a single parameter at a time, searching for a reasonable optimum for that parameter. This search is clearly sub-optimal, but does afford an interpretation of the effect of different parameter values on the overall performance of the system. Each line search was initiated starting from default values for all the parameters. Values for the parameters related to the default condition will be denoted

[5] Speaker or handset specific score normalizations are not used in this dissertation.

Table 4.1: Default values for the parameters in the speaker verification system related to experiments in this and subsequent sections.

Parameter or process			
Type	Description	Name	Value
Features	Frame rate	θ_s	100 Hz
	Low frequency cut-off	f_l	200 Hz
	High frequency cut-off	f_h	3500 Hz
	# of filter bank bands		24
	Analysis window length		
	(i) Section 4.4.1	t_w	20 ms
	(ii) Section 4.4.2 and Chapter 5	t_w	32 ms
	Preemphasis coefficient	a	0.97
	Mean subtraction	MS	Yes
	Use static features (cf Fig. 5.7)	A or C	Yes
	Use dynamic features (cf Fig. 5.7)	B or D	No
Models	# of mixture components		
	(i) Sections 4.4.1 and 4.4.2	K	128
	(ii) Chapter 5	K	128 or 256
	# of EM-iterations		5
	Variance regularization coefficient	ϵ	0.2
	Adaptation confidence parameters		
	(i) Weights	ν_k	16
	(ii) Means	η	16
	(iii) Variances	ρ	16
	Number of GMM components evaluated	N-best	5

with an asterisk (*) in the figures to follow. Default values for the system parameters used in this section are shown in Table 4.1.

Parameters of the feature extraction process are optimized using the telephone speech corpus of the 1996 NIST-SRE [60], while parameters of the statistical modeling process are optimized using the telephone speech corpus from the 1997 NIST-SRE [60]. The corpora and experimental setup are discussed in Appendix A. Results are computed for male and female speakers separately and averaged.

The EER and MDE results presented here relate to *matched* and *mismatched* training and testing conditions. The possibility for mismatch occurs when the telephone *number* and/or *handset type* differs between training and testing. Handset types are categorized into *electret* transducer and *carbon-button* transducer types[6]. For the feature extraction parameters, the conditions of interest are *same number* (SN) and *different number* (DN). For the statistical modeling parameters, the conditions are *same number* and *same handset type* (SNST) and *different number* and *different handset type* (DNDT). The conditions are described in more detail in Appendix A.

4.4.1 Feature Extraction

Various results related to the optimization of the feature extraction parameters are presented in this section. Parameters include the frequency cut-offs for telephone band-limiting, the length of the analysis window and the number of filter bank bands. The effectiveness of mean subtraction and use of dynamic features are also investigated. The 1996 NIST-SRE corpus is used for the experiments in this section. Results are computed separately for four available conditions of interest namely males and females, and 3 and 10 second long segments of test speech. To simplify interpretation, the results of these four conditions are averaged and reported here.

Length of Analysis Window

Various results in Chapters 2 and 3 indicated that to avoid aliasing, a good choice for the length of the short-term analysis window is longer than 20 ms. It was also remarked that to match the quasi-stationary nature of the speech signal, this length should be shorter than about 40 ms. Fig. 4.4 shows EER and MDE values for the matched and mismatched conditions as a function of length t_w of the analysis window.

The errors are seen to be relatively insensitive to the length of the analysis window as long as it is not shorter than about 20 ms or longer than about 64 ms. This suggests that aliasing is not much of a problem, even at 20 ms lengths. It also suggests that the quasi-stationary nature of the signal may extend up to 64 ms or that the spectral estimate benefits from the smoothing obtained with a longer analysis window. A default length of 20 ms is used for the experiments in this section and seems reasonable. In subsequent

[6] These distinctions were assessed using telephone number and handset type labels distributed by NIST.

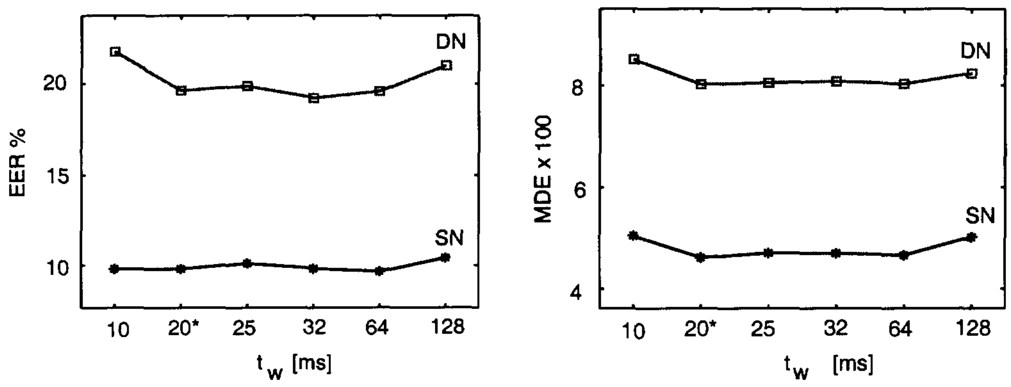


Figure 4.4: EER and MDE as a function of short-term analysis window length t_w (in milliseconds). 1996 NIST-SRE corpus.

sections however, a default length of 32 ms will be used. Reasons for this choice are the following. In Chapter 3 it was observed that lowpassed features may be less sensitive to the effects of environmental mismatch. In Chapter 2, it was shown that the longer the analysis window, the narrower is its bandwidth and that this is a way to obtain such features. A choice of 32 ms thus seemed appropriate.

Number of Filter Bank Bands

Fig. 4.5 shows that using a filter bank with around 20 to 26 bands is optimal. Given this result and so as to make the processing used here consistent with that used in other work (see e.g. [78]) it was decided to use 24 bands. This choice implies a 90 Mel spacing

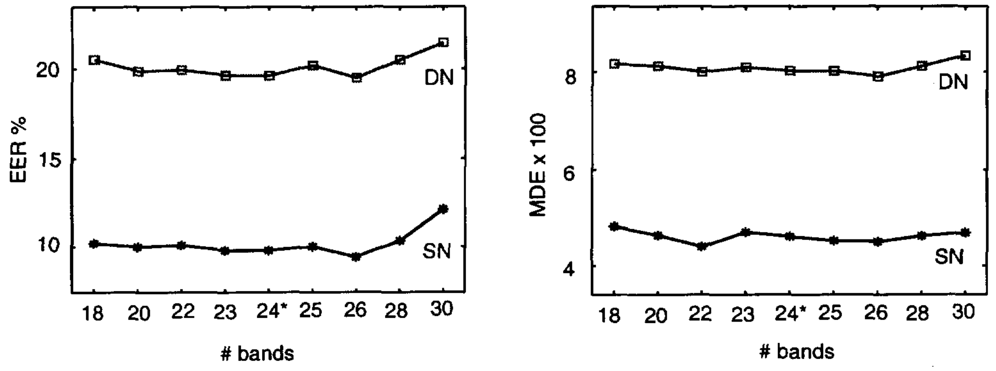


Figure 4.5: EER and MDE as a function of number of filter bank bands between 0-4 kHz. 1996 NIST-SRE corpus.

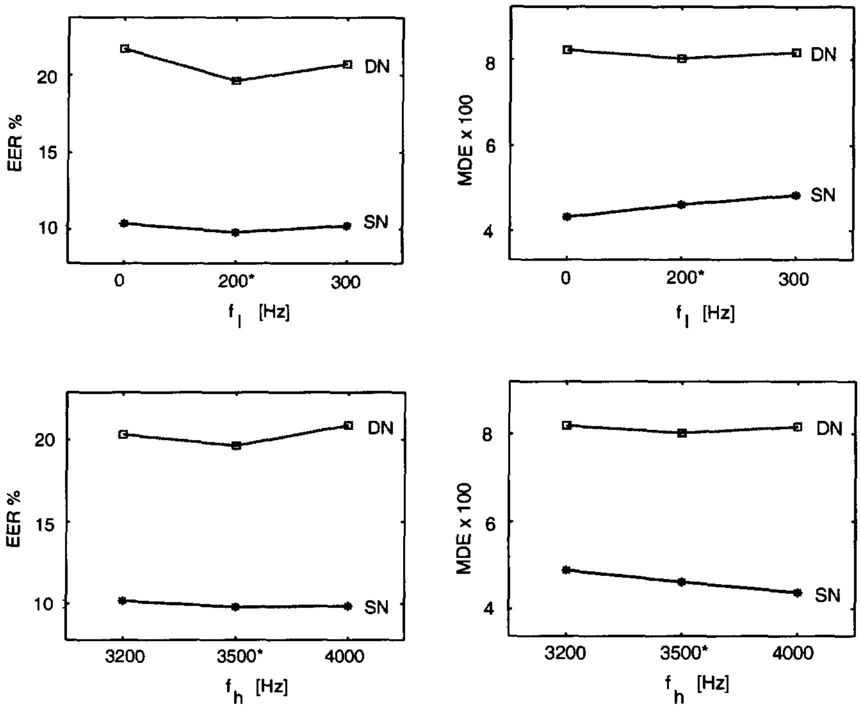


Figure 4.6: EER and MDE as a function of lower cut-off frequency f_l and as a function of higher cut-off frequency f_h . 1996 NIST-SRE corpus.

between the center frequencies of adjacent filter bands. Based on the results presented in Section 4.4.1, only the center 19 bands will be used.

Band-limiting

In Section 4.1 it was suggested that the speech signal be band-limited to a range $[f_l, f_h]$ = $[200, 3500]$ Hz. Fig. 4.6 show the EER and MDE as a function of the lower cut-off frequency f_l and higher cut-off frequency f_h. In the mismatched condition, the EER and MDE are minimum for both parameters at the proposed values. In the matched condition, band-limiting the speech signal increases the MDE. For robustness to environmental degradations though, it appears that the choice is reasonable.

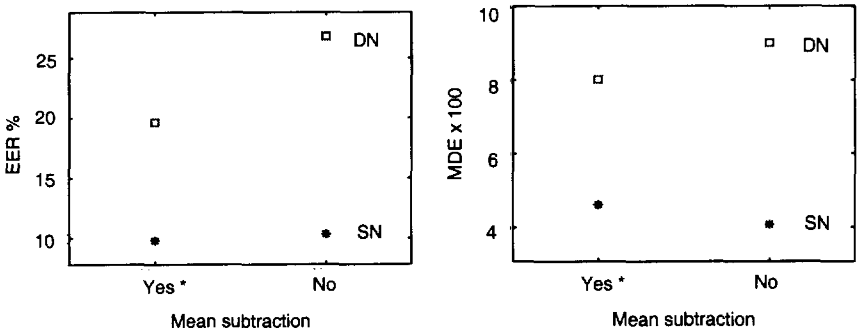


Figure 4.7: Effect of mean subtraction on the EER and MDE. 1996 NIST-SRE corpus.

Mean Subtraction and Preemphasis

Fig. 4.7 shows the effect of mean subtraction (MS). MS clearly reduces the EER and MDE in the mismatched condition. In the matched condition, MS is seen to slightly worsen the MDE. This suggests that a simple MS model for dealing with a convolutional degradation is suboptimal. This is not surprising since MS removes the long-term average which may be useful for speaker verification when not corrupted. In the matched condition the channel may actually help to verify or identify the speaker. In Section 2.5.2 it was suggested that a non-linear compensation strategy is necessary if additive noise is present.

Fig. 4.8 confirms however, that the system with mean subtraction is invariant to the effects of a linear time-invariant channel. The figure shows that preemphasis does not affect the EER and MDE. The value of the preemphasis coefficient a was set to either 0 or 0.97 to obtain the preemphasized speech signal $s'(n) = s(n) - as(n)$.

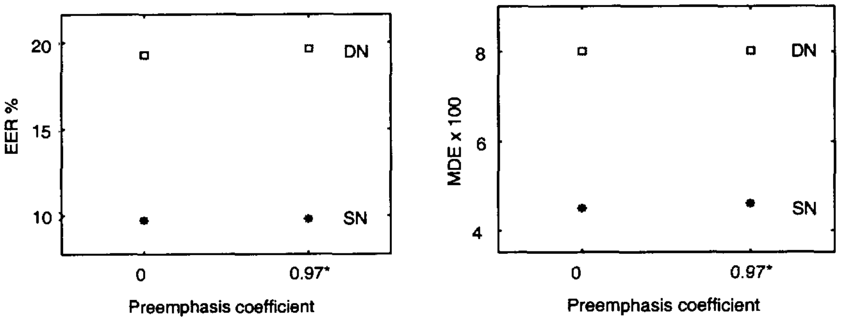


Figure 4.8: EER and MDE as a function of preemphasis coefficient showing invariance to a convolutional transmission channel. 1996 NIST-SRE corpus.

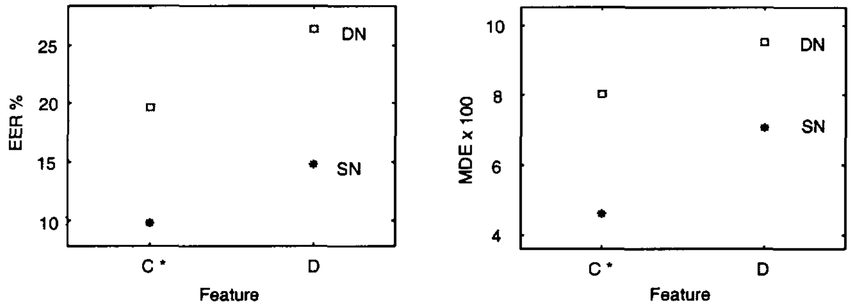


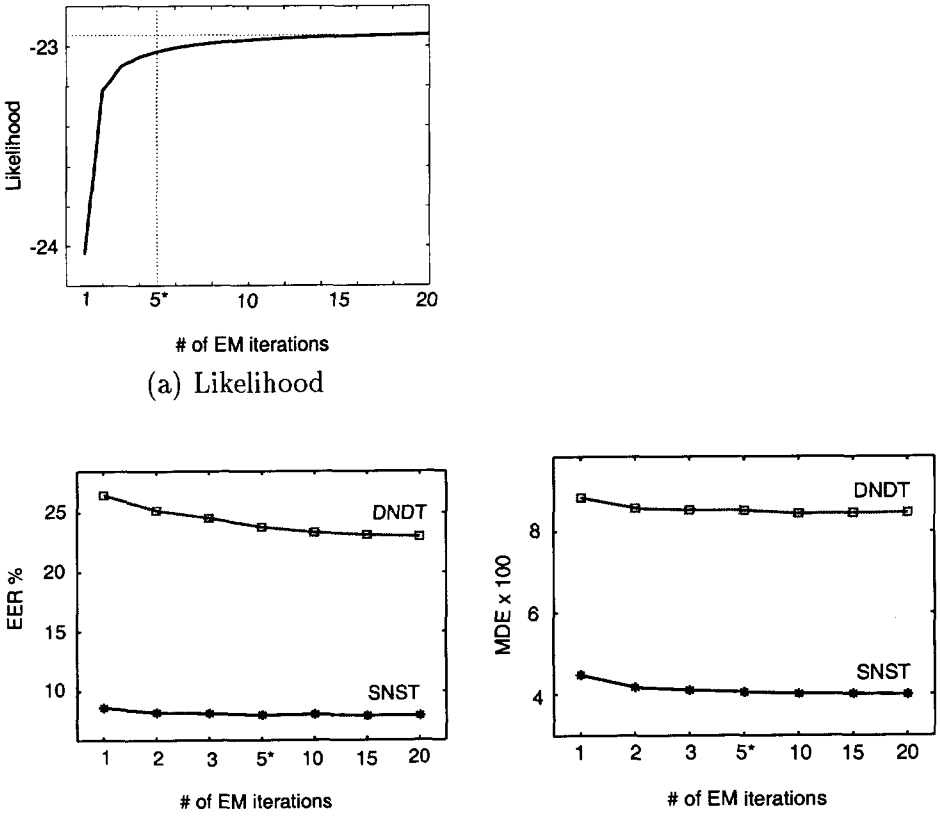
Figure 4.9: EER and MDE for static features (C) versus dynamic (delta) features (D). 1996 NIST-SRE corpus.

Dynamic Features

Fig. 4.9 contrasts the EER and MDE for a system using the default "static" features (C) to that of a system using "dynamic" features (D). These dynamic features are obtained by using the so-called delta polynomial fit to the time sequence of logarithmic energy. The polynomial used here spans 5 frames of speech (refer Section 2.5.3). When used individually, static features are seen to contain more speaker information than dynamic features. However, at the end of this chapter and in Chapter 5, it will be shown that using static and dynamic features together leads to a further reduction in error.

4.4.2 Statistical Modeling

Various results related to the optimization of the modeling parameters are presented in this section. These parameters are described in detail in Appendix B. Parameters include the number of iterations used in the EM-algorithm for training the SI model, the regularization parameter used for the covariances, the number of components in the GMM, the number of components that are evaluated in the GMM during scoring, and the confidence parameters used in the MAP adaptation. The effectiveness of using both static and dynamic features is also investigated. The 1997 NIST-SRE corpus is used for experiments in this section. Results are computed separately for six available conditions of interest namely males and females, and 3, 10 and 30 second long segments of test speech. To simplify interpretation, the results of these six conditions are averaged and reported here.



(a) Likelihood

(b) EER

(c) MDE

Figure 4.10: EER, MDE and likelihood as a function of the number of EM-iterations used for training the SI-model. 1997 NIST-SRE corpus.

Number of EM Iterations for SI model

Fig. 4.10 shows the EER, MDE and data-likelihood as a function of the number of iterations used in the EM-algorithm. For implementation details of the EM-algorithm refer to Appendix B. It can be seen that the EM-algorithm is close to convergence after about 5 iterations and has effectively converged after 10 iterations.

Covariance Regularization Parameter

In Appendix B it is proposed that the covariance matrices be regularized (see for example [66]) to prevent the occurrence of singularities in the likelihood function that is maximized by the EM-algorithm. Singularities can occur when the dimensionality of the

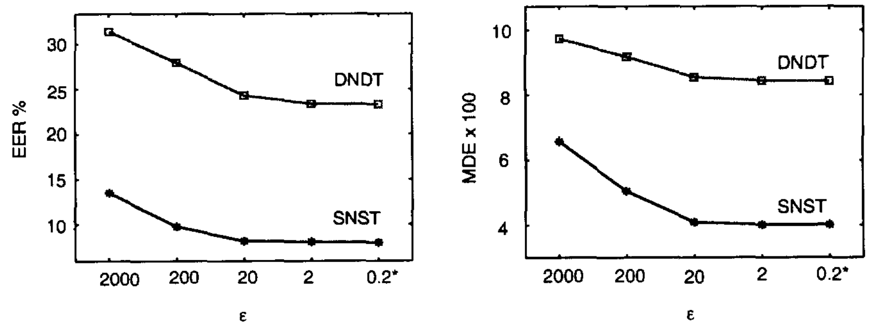


Figure 4.11: EER and MDE as a function of ϵ parameter used to regularize the covariances during training of the SI-model. 1997 NIST-SRE corpus.

data is high and the number of training samples is small so that some of the covariance matrices may become singular. As is evidenced by Fig. 4.11, singularities do not appear to be a problem here. This may be because the number of training samples is fairly large (10^6 samples were used here) and the samples were equally distributed among the mixture components during the VQ-training step.

N-Best GMM Components to be Evaluated

Fig. 4.12 shows the EER and MDE as a function of the number of best scoring N components evaluated in the SD and SI models during scoring. It can be seen that evaluating as little as one component does not affect performance deleteriously. In general, for experiments in this dissertation, the top 5 components will be evaluated.

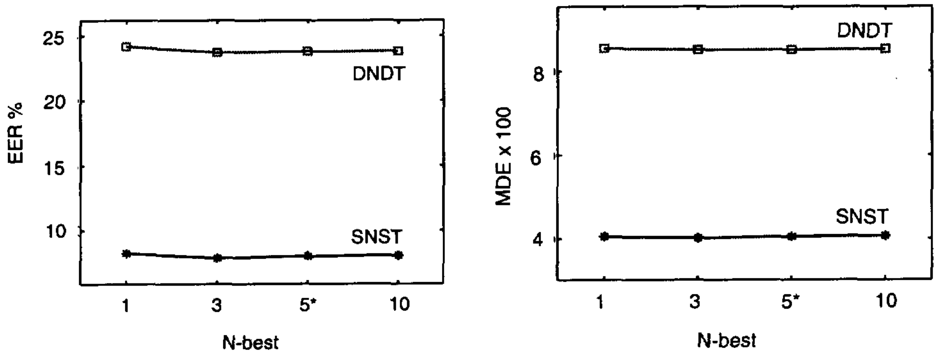


Figure 4.12: EER and MDE as a function of N-best components evaluated in the SD and SI models during scoring. 1997 NIST-SRE corpus.

Confidence Parameters Used in MAP Adaptation

In Appendix B it is shown that the general MAP update rules for the weights, means and variances take the form

$$\Theta_{SD} = \lambda\Theta_{\boldsymbol{X}} + (1-\lambda)\Theta_{SI}, \tag{4.12}$$

where $\Theta_{\boldsymbol{X}}$ is an estimate for the parameter based on the SD adaptation data $\boldsymbol{X}$ and Θ_{SI} is the previous estimate obtained for the SI data. The adaptation parameter λ can be seen to form the new parameter estimate from the weighted combination of the current and previous estimates. For the mean of a single component GMM for example, the adaptation parameter λ can be related to the confidence parameter by

$$\lambda = \frac{N}{N+\eta}, \tag{4.13}$$

where N is the number of samples in the adaptation data that is available for estimating the mean. The value of a confidence parameter is essentially a measure of how many samples are assumed to have been used to obtain the old parameter estimate, and is to be contrasted to the number of samples that is to be used in obtaining the current estimate. In this example, if the confidence parameter η is very large relative to N, $\lambda \rightarrow 0$ and the new estimate for the mean is equal to the old estimate, so that no adaptation is performed. Conversely, if the confidence parameter is very small relative to N, $\lambda \rightarrow 1$ and the new estimate for the mean is equal to the current estimate derived from the adaptation data.

Table 4.2 lists optimum values for the confidence parameters (see also Appendix B). The EER and MDE exhibit a minimum for a broad range of values for the parameters. The table shows that good performance is obtained for values of the confidence parameters set to $\nu_k = 16$, $\eta = 8$ and $\rho = 64$. Good performance is also obtained for the default values. It can be concluded that adaptation of the component means are the most important, with adaptation of the covariances ranking second. The higher value for the covariance confidence parameter ρ is to be expected given that more samples are needed to estimate the covariances than to estimate the component means.

Table 4.2: EER and MDE for various values of the MAP confidence parameters ν_k, η and ρ. 1997 NIST-SRE corpus.

PARAMETERS	ERROR				COMMENTS
	MDE x 100		EER %		
ν_k,η,ρ	SNST	DNDT	SNST	DNDT	
1024,1024,1024	7.8	9.5	22.5	36.7	Little adaptation
1,1,1	4.1	8.5	8.8	24.3	Full adaptation
16,8,64	**4.0**	**8.5**	**7.6**	**23.3**	**Best performance**
16,16,16*	4.1	8.5	8.0	23.8	Default system *
1024,1024,1	6.7	9.6	13.6	31.7	Adapt covariances only
1024,1,1024	4.4	8.7	8.2	23.8	Adapt means only
1,1024,1024	7.7	9.5	17.2	29.3	Adapt weights only

Number of Components in the GMM

Fig. 4.13 shows the effect of increasing the number K of components in the Gaussian mixture model and appending the vector of static features (C) with a vector of dynamic features (D) to form a new feature vector (C,D). The solid lines in the figure indicate the errors when using static features only. The dotted lines indicate the errors when using the static and dynamic features together.

Increasing K beyond 512 increases the error and suggests that a model with more than 512 components has too many free parameters. The result is that the model begins to fit irrelevant details in the training data – the well-known phenomenon of over fitting the training data. It appears that models with as few as 128 components give reasonable performance.

Appending dynamic features to the static features leads to a lower error rate, particularly so in the mismatched condition. This suggests that the dynamic features provide additional robustness to environmental degradations.

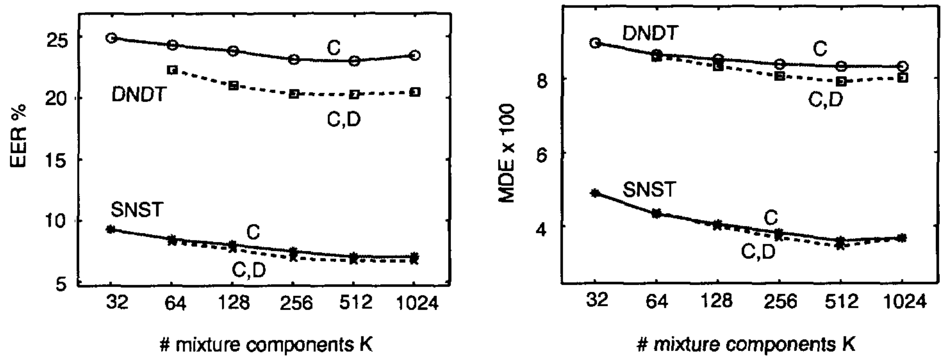


Figure 4.13: EER and MDE as a function of number of mixture components for static features (C) versus static and dynamic features (C,D). 1997 NIST-SRE corpus.

4.5 Summary

The speaker verification system that will used to evaluate text-independent speaker verification performance in this dissertation was described. Details of the feature extraction process were first described. Two different evaluation measures were subsequently defined. Derivation of a test statistic for speaker verification lead to the use of a speaker independent pdf as well as speaker dependent pdfs for modeling the speakers statistically. Based on previous work and observations about the non-Gaussianity of the speech data, the pdfs were parameterized using a mixture of Gaussians. Rules for deriving the parameters of these GMMs were then described. The last section of the chapter motivated specific choices for the values of the parameters.

In the optimization experiments it was observed that lengths from 20-64 ms for the analysis window give reasonable performance. It was seen that band-limiting the speech signal to the telephone bandwidth improves performance in the mismatched condition. Using a short-term analysis filter bank with between 20 and 26 bands gave good performance. Mean subtraction and the inclusion of dynamic features in addition to static features were seen to improve performance considerably in the mismatched condition. However, dynamic features used on their own performed less well. For the statistical modeling, a GMM with 256 to 512 components in the mixture performed best, although a GMM with as few as 128 components performed reasonably well.

The experimental setup that will be used in the remainder of this dissertation will be described next, before revisiting the performance of the speaker verification system in Chapter 5. There, by analyzing the errors of the system under various conditions, the approach will be to systematically investigate the relative importance of the components of the modulation spectrum for speaker verification.

Chapter 5

Speaker Verification in a Time-Feature Space

Whereas Chapter 3 was concerned with the effect of handset variability in speech in general, this chapter narrows the focus to the speaker verification task specifically. The aim is to use filtering in the time-feature space to measure the importance of the components of the modulation spectrum for speaker verification. The approach will be to design a number of filters in the modulation spectral domain, apply them to the time sequences of logarithmic energy, and measure and analyze their effect on the speaker verification error.

The outline of the chapter is as follows. In Section 5.1, a methodology is described for the measurement of the relative importance of the components of the modulation spectrum for speaker verification. Importance is measured by applying a specific bandpass filter to the time sequences, measuring the speaker verification error on the 1997 NIST-SRE corpus, and comparing the errors resulting from the application of different bandpass filters. Sections 5.2 and 5.3 then detail the effects of lowpass and highpass filtering of the time sequences of logarithmic energy. Section 5.4 reexamines the choice of dynamic features, using the results from the filtering experiments to guide the choice. By proposing a lowpass filtering of the time sequences, the potential for a subsequent down sampling of the filtered time sequences is examined in Section 5.5, where, using the suggested processing, a reduction in verification error is reported for the 1998 NIST-SRE corpus. The 1998 NIST-SRE corpus is used as a formal test set. Unless noted otherwise, results are computed for the test speech segments of 30 second duration only. This allows an analysis of modulation frequencies down to the centi-Hertz range.

5.1 Relative Importance of Components of the Modulation Spectrum

This section presents a methodology for the measurement of the importance of the components of the modulation spectrum for speaker verification. This is done by applying a specific bandpass filter to the time sequences, measuring the speaker verification error on the 1997 NIST-SRE corpus, and comparing the errors resulting from the application of different bandpass filters. The relative importance of these components is then measured and results detailing the relative importance presented and discussed.

5.1.1 Methodology

Verification error rates related to specific components of the modulation spectrum are obtained by bandpass filtering in the modulation spectral domain with different low frequency cut-offs θ_l and different high frequency cut-offs θ_h ranging from 0 to 50 Hz on a logarithmically spaced grid[1]. In each case the equal error rate $e(\theta_l, \theta_h)$ as a function of low and high cut-offs is computed using features with modulation spectral components from that particular passband only $\theta_l \leq \theta < \theta_h$. As an example, Fig. 5.1 depicts the grid generated by cut-offs at 2, 4, 8, 16, and 32 Hz. To derive charts that would summarize the relative importance of the spectral components, normalized differentials with respect to each of the low and high cut-off frequencies for the surface described by the error function $e(\theta_l, \theta_h)$ are averaged[2]. The average relative importance $R(\theta_1, \theta_2)$ of the spectral components between θ_1 and θ_2 is then estimated as the average of the normalized partial differentials evaluated in the band $\theta_1 \leq \theta < \theta_2$. The computation is set up so that a positive value for the average relative importance of components of the modulation spectrum within a given band reflects a relative reduction in verification error due to the inclusion into the acoustic features of those components.

[1] Filters were designed to have frequency responses with similar shape on the grid with sharp cross-overs and 50 dB attenuation in the stop band.

[2] This procedure is similar to one described in [44] except for the normalization applied here.

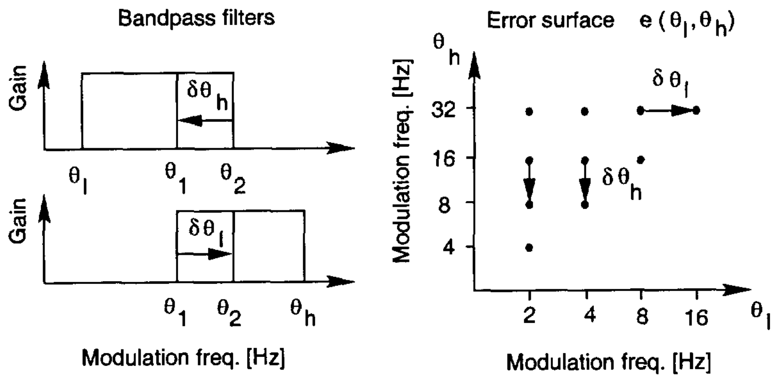


Figure 5.1: Grid for evaluating the importance of components of the modulation spectrum for speaker verification.

The relative importance is defined and computed as follows. For a given high cut-off $\theta_h > \theta_2$, let

$$R_l(\theta_1, \theta_2|\theta_h) = \frac{e(\theta_2, \theta_h) - e(\theta_1, \theta_h)}{e(\theta_1, \theta_h)} \tag{5.1}$$

denote the relative change in error associated with a change $\partial\theta_l = \theta_2 - \theta_1$ in the low cut-off. Similarly, for a given low cut-off $\theta_l < \theta_1$, let

$$R_h(\theta_1, \theta_2|\theta_l) = \frac{e(\theta_l, \theta_1) - e(\theta_l, \theta_2)}{e(\theta_l, \theta_2)} \tag{5.2}$$

denote the relative change in error associated with a change $\partial\theta_h = \theta_1 - \theta_2$ in the high cut-off. Then, when averaged over n_h possible high cut-offs and n_l possible low cut-offs, the average relative change in error for the inclusion of the band $\theta_1 \leq \theta < \theta_2$ is

$$R(\theta_1, \theta_2) = \frac{1}{(n_l + n_h)} \left[\sum_{\theta_h > \theta_2} R_l(\theta_1, \theta_2|\theta_h) + \sum_{\theta_l < \theta_1} R_h(\theta_1, \theta_2|\theta_l) \right]. \tag{5.3}$$

This gives an estimate of the relative importance of the spectral components between θ_1 and θ_2. Fig. 5.1 depicts this computation for $R(8, 16)$.

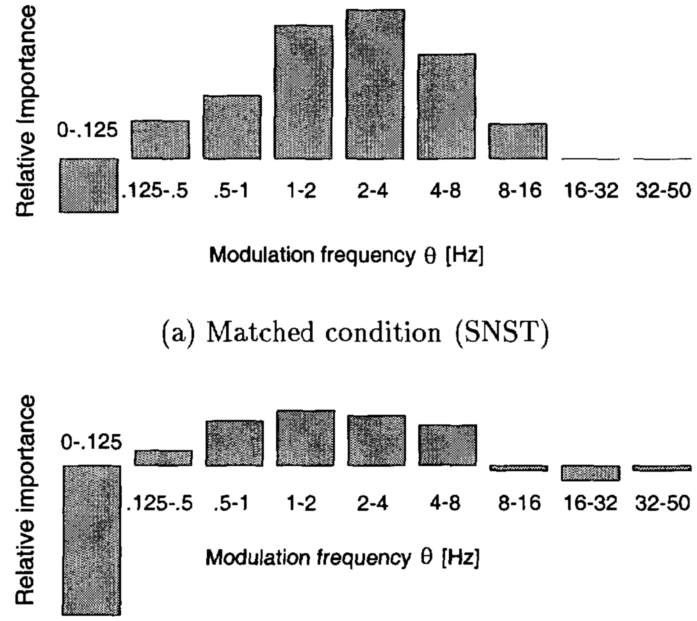
Note that the computation amounts to estimating an averaged gradient for the logarithmic error surface. For instance, for a change $\partial\theta_l = \theta_2 - \theta_1$ in the low cut-off, it follows that

$$\left.\frac{\partial \log e(\theta)}{\partial\theta}\right|_{\theta=\theta_1} = \left.\frac{1}{e(\theta)}\frac{\partial e(\theta)}{\partial\theta}\right|_{\theta=\theta_1} \approx \frac{1}{(\theta_2-\theta_1)} \cdot \frac{e(\theta_2)-e(\theta_1)}{e(\theta_1)}. \tag{5.4}$$

A similar interpretation holds for a change $\partial\theta_h = \theta_1 - \theta_2$ in the high cut-off. This relationship aids interpretation of the relative importance of the components of the modulation spectrum for speaker verification since it relates the relative importance to a minimization of the logarithmic error surface. It should be noted that the measure $R(\theta_1, \theta_2)$ used here only provides an indication of the importance of a band and as such does not provide information on the inter-dependence (e.g. correlation) of different modulation spectral components for speaker verification.

5.1.2 Results

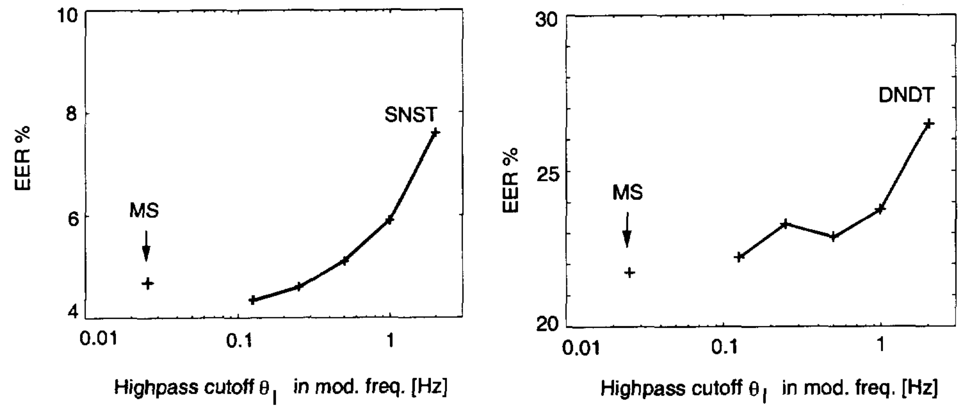
Fig. 5.2 depicts the relative importance of different components of the modulation spectrum for speaker verification on the 1997 NIST-SRE corpus. A positive value for a band reflects a relative reduction in verification error due to the inclusion into the acoustic features of the components of the modulation spectrum within that band. Components around 2 to 4 Hz are seen to be relatively more important for reducing the error rate. This finding agrees with the observations of Chapter 3 where it was seen that speech may be least affected by extraneous sources of variability in this range of modulation frequencies. This finding is also consistent with dominant rates of change in the logarithmic power spectrum of speech that have been estimated at around 2 to 4 times per second [100, 44].



(a) Matched condition (SNST)

(b) Mismatched condition (DNDT)

Figure 5.2: Relative importance R of components of the modulation spectrum. Positive values indicate a decrease in verification error due to the inclusion of a particular modulation spectral band in the acoustic features. Results were derived on 30 second test segments (male and female) from the 1997 NIST-SRE corpus.

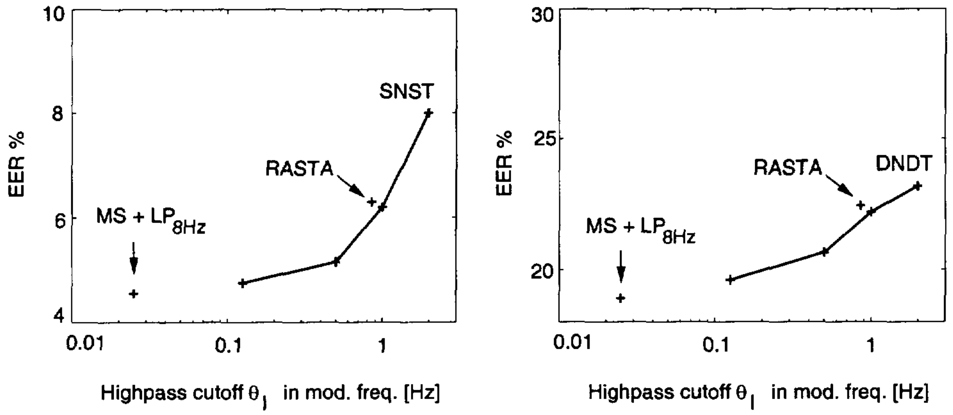


(a) Matched condition (SNST) (b) Mismatched condition (DNDT)

Figure 5.3: EER versus highpass cut-off for verification of 30 second test segments from the 1997 NIST-SRE corpus. θ_h=50 Hz.

5.2 Effect of Highpass Filtering

This section examines the effects of highpass filtering of the time sequences of logarithmic energy. It was seen in Fig. 5.2 that inclusion of components of the modulation spectrum below about 0.125 Hz increases error rate in the mismatched condition, while inclusion of components above about 0.125 Hz leads to a decrease in error rate. This suggests that the highpass cut-off of 1 Hz (see Fig. 2.7) used with RASTA filtering in ASR should be lowered for speaker verification. Figures 5.3 and 5.4 confirm this observation. Fig. 5.3 shows EER as a function of cut-off frequency for highpass filtering ($[\theta_l, 50]$ Hz). The figure shows that reducing the highpass cut-off generally reduces the error rate, with an optimum reached at a cut-off frequency close to that of MS. Fig. 5.4 shows EER as a function of cut-off frequency for a bandpass filtering ($[\theta_l, 8]$ Hz). The choice of an 8 Hz cut-off comes from the logarithmic grid defined earlier and as the results will show, is close enough to the 13 Hz cut-off of the RASTA filter ($[1, 13]$ Hz) to allow a comparison (see also Fig. 5.5.) It can be seen that the EER obtained with the RASTA filter closely matches the EER obtained with a bandpass filter ($[1, 8]$ Hz) and suggests that the 1 Hz highpass cut-off of the RASTA filter should be lowered for speaker verification. This result explains the observed differences reported in the experimental study of Section 2.5.4 where slightly better performance was achieved with MS filtering, than was achieved with RASTA filtering.



(a) Matched condition (SNST) (b) Mismatched condition (DNDT)

Figure 5.4: EER versus highpass cut-off for verification of 30 second test segments from the 1997 NIST-SRE corpus. θ_h=8 Hz.

5.3 Effect of Lowpass Filtering

This section examines the effects of lowpass filtering of the time sequences of logarithmic energy. Fig. 5.2 showed that in the matched condition, inclusion of components above 16 Hz does not affect the error rate, while in the mismatched condition inclusion of components as low as 8 Hz increases the error rate. Fig. 5.5 details these observations with bandpass results for a system with $[\theta_l, \theta_h] = [0.025, \theta_h]$. The figure shows the EER for matched and mismatched conditions and for testing with 30 second segments from the

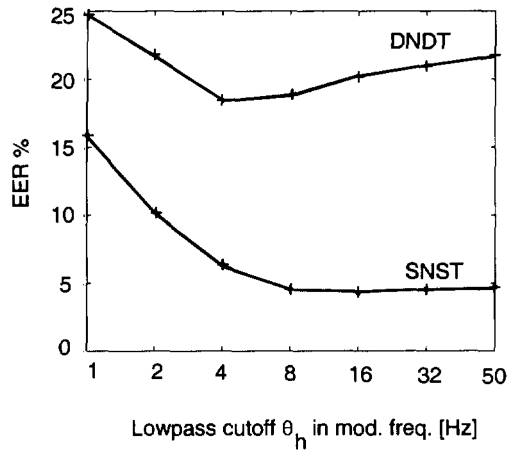


Figure 5.5: EER versus lowpass cut-off for verification of 30 second test segments from the 1997 NIST-SRE corpus. $\theta_l = 0.025$.

Table 5.1: EER in percent at a lowpass cut-off of 10 Hz (MS+LP10) and without lowpass filtering (MS) in matched (SNST) and mismatched (DNDT) conditions. Results are for verification of test segments (male and female) from the 1997 NIST-SRE corpus.

SYSTEM	CONDITION	DURATION (sec)		
		3	10	30
MS	DNDT	26.75	22.50	21.75
MS+LP10	DNDT	25.90	20.80	19.05
MS	SNST	13.10	6.50	4.70
MS+LP10	SNST	13.20	6.85	4.35

1997 NIST-SRE. Based on these results it can be concluded that higher modulation frequencies are not important for speaker verification and that removing them will improve performance.

It appears that a reasonable performance trade-off between matched and mismatched conditions may be obtained with a lowpass cut-off of around 10 Hz. Table 5.1 compares the EERs from a lowpass system with $\theta_h = 10$ and MS and a baseline system with $\theta_h = 50$ and MS. The table shows results for matched and mismatched conditions and different durations of 3, 10 and 30 seconds for the test segments.

For 30 second segments in the mismatched condition the lowpass system results in a relative reduction in EER of more than 14% while in the matched condition the lowpass system results in a relative reduction in EER of more than 8%. The lowpass filtering may be less effective for shorter segments because of the 1 second length of the FIR filter that was used and because of difficulties with initializing the filter at the beginning of the segment. It can be seen that lowpass filtering to about 10 Hz reduces acoustic mismatch. This reduction occurs in addition to the reduction of convolutional mismatch afforded by MS.

5.4 Temporal Features from Orthogonal Polynomials

Up to now only the modulation frequency characteristics of the filters have been specified. Temporal characteristics are examined in this section. The idea is to find a good combination of static and dynamic features. The static and dynamic features are defined as follows. If two types of features are used, then the distinction between a static and a dynamic feature is that the static feature exhibits higher energy at lower modulation frequencies than the dynamic feature.

In seminal work by Furui [26] superior performance was indicated for a system using static features and dynamic features obtained from a first order orthogonal polynomial expansion of 90 ms long segments of the time sequence. Neither the static or dynamic features were lowpass filtered however, and it is not known whether such lowpass filtering may lead to even better performance. It is the aim of this section to investigate this question.

5.4.1 Technique

Static and dynamic features can be obtained by projecting the time-sequences onto an orthogonal polynomial basis [26, 33]. The first three orthogonal polynomials in the basis are

$$\begin{aligned} p_{0,L}(n) &= 1, \\ p_{1,L}(n) &= n, \\ p_{2,L}(n) &= n^2 - (L^2-1)/12, \end{aligned} \tag{5.5}$$

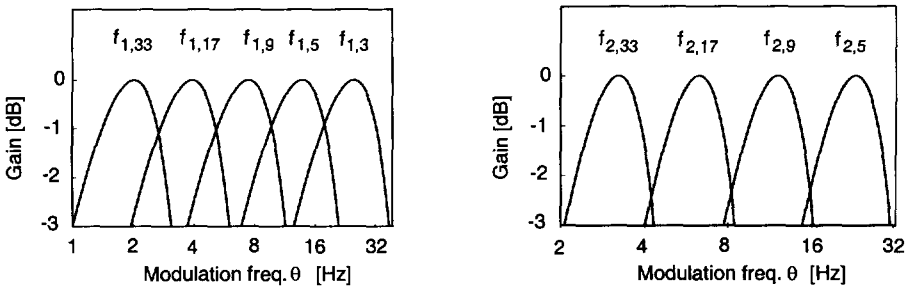
where $-(L-1)/2 \leq n \leq (L-1)/2$ and L is the length in frames spanned by the polynomial. Then the filter coefficients are

$$f_{i,L}(n) = p_{i,L}(n) / \sum_{j=-(L-1)/2}^{(L-1)/2} [p_{i,L}(j)]^2 , \quad i = 0, 1, 2, ... \tag{5.6}$$

By comparing polynomials of length L, it follows that static and dynamic features can be obtained by filtering respectively with $f_{i,L}(n)$ and $f_{j,L}(n)$, where $j > i$. In the remainder of this chapter, we will obtain static features using the allpass filter $f_{0,1} = \delta(0)$.

We will obtain dynamic features by filtering the static features using a first or second order polynomial filter $f_{1,L}(n)$ or $f_{2,L}(n)$.

Fig. 5.6 (a) shows the normalized frequency responses of the first order orthogonal polynomial filters: $f_{1,L}$, $L \in \{3, 5, 9, 17, 33\}$ and Fig. 5.6 (b) shows the normalized frequency responses of the second order orthogonal polynomial filters: $f_{2,L}$, $L \in \{5, 9, 17, 33\}$. The peak responses for the first order filters are at 25.0, 13.8, 7.5, 3.9 and 2.0 Hz. The peak responses for the second order filters are at 23.0, 12.1, 6.3 and 3.2 Hz.



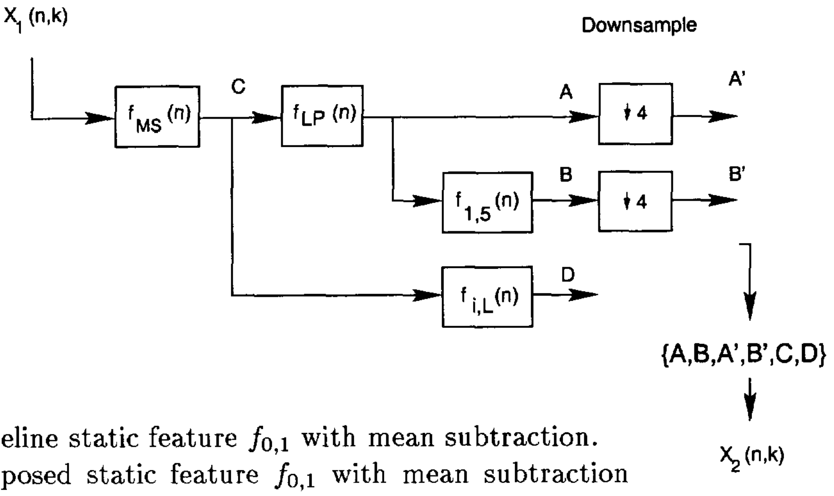
(a) 1st order

(b) 2nd order

Figure 5.6: Normalized frequency responses of the orthogonal polynomial filters.

The usefulness of different static and dynamic features can be evaluated by comparing speaker verification performance for various combinations of the features. Feature vectors are obtained by appending a dynamic feature vector to a static feature vector. Fig. 5.7 shows a block diagram for the extraction of different static and dynamic features. The features are obtained by filtering the time sequences of logarithmic energy $X_1(n, k)$. Note that the time sequences are sampled at a rate of $\theta_s = 100$ Hz. Features are obtained by MS filtering with f_{MS} followed by optional filtering with either f_{LP}, $f_{1,L}$ or $f_{2,L}$. The lowpass filter f_{LP} is a 101-tap FIR filter with cut-off at 10 Hz. Fig. 5.8 shows the composite frequency responses at outputs A, B, C and D of the different filters. They are to be compared to the frequency responses in Fig. 2.7.

In the previous sections it was observed that components of the modulation spectrum at relatively low frequencies (down to at least 0.125 Hz and as low as 0.025 Hz in the case of MS) contain useful speaker information and should not be attenuated. It was also observed that the removal of higher frequencies (above about 10 Hz) reduces verification error in the



Legend:

- C Baseline static feature $f_{0,1}$ with mean subtraction.
- A Proposed static feature $f_{0,1}$ with mean subtraction and lowpass filtering f_{LP} to 10 Hz.
- A' Same as A but with sampling rate reduction from $\theta_s = 100$ Hz to $\theta_s = 25$ Hz.
- $D_{i,L}$ Baseline dynamic feature $f_{1,L}$ with mean subtraction.
- $B_{1,5}$ Proposed dynamic feature $f_{1,5}$ with mean subtraction and lowpass filtering f_{LP} to 10 Hz.
- $B'_{1,5}$ Same as B but with sampling rate reduction from $\theta_s = 100$ Hz to $\theta_s = 25$ Hz.

Figure 5.7: Block diagram of system using polynomial filters for deriving dynamic acoustic feature vectors from logarithmic energies.

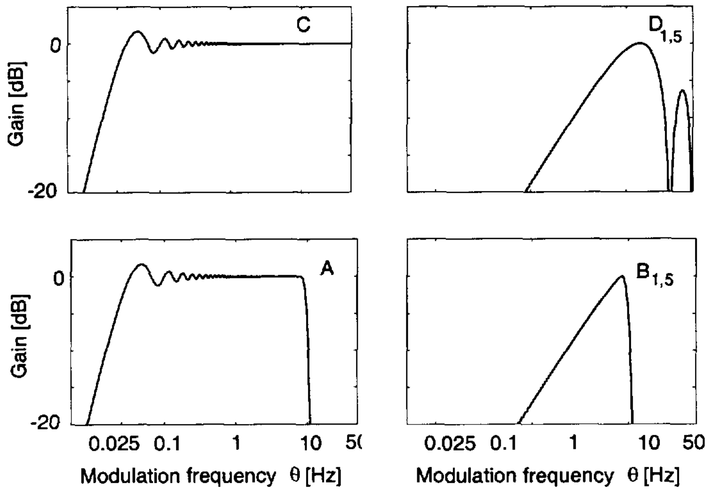


Figure 5.8: Effective filter frequency responses for deriving acoustic feature vectors from logarithmic energies.

mismatched condition. Based on these observations it appears reasonable to process the logarithmic energies with a bandpass filter that preserves modulation frequencies between 0.025 and 10 Hz. The usefulness of such filtering is investigated here.

It is proposed that a reasonable combination of features are a MS- and lowpass-filtered (to 10 Hz) static feature combined with a similarly filtered dynamic feature. The system using these features will be denoted as $(A, B_{1,5})$. Since these features have been lowpass filtered, it is possible to reduce the sampling rate. The system using features that were obtained by reducing the sampling rate from 100 Hz to 25 Hz will be denoted as $(A', B'_{1,5})$. Systems $(A, B_{1,5})$ and $(A', B'_{1,5})$ are the systems proposed in this dissertation. The performances of these proposed systems are to be compared to that of systems using other combinations of features.

5.4.2 Dynamic Features Based on First Order Polynomials

Fig. 5.9 shows the EER for various combinations of static and dynamic features obtained from the *first* order polynomial filters. While it is not shown here, it should be noted that the same trends were observed for the MDE. In the figures, from left to right, errors are shown for systems based on four major types of features. Table 5.2 lists the systems and their associated features.

Table 5.2: Systems and features related to Fig. 5.9.

1. Baseline system (A) with static features that are lowpass filtered with 10 Hz cut-off.
2. Proposed systems $(A, B_{1,5})$ and $(A', B'_{1,5})$ (down sampled) with static and dynamic features that are both lowpass filtered.
3. Baseline system $(C, D_{1,L})$ with static and dynamic features that are *not* lowpass filtered.
4. Comparison system $(A, D_{1,L})$ with static features that are lowpass filtered and dynamic features that are *not* lowpass filtered.

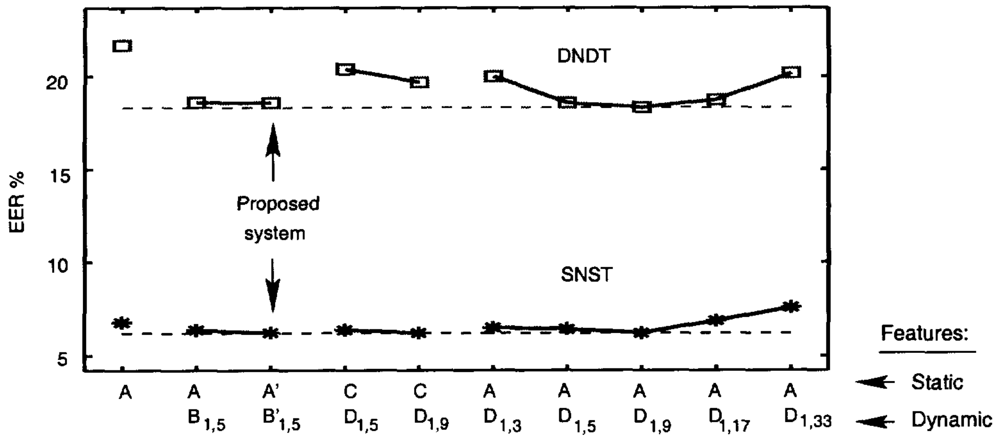


Figure 5.9: EER in percent for various combinations of static and dynamic $f_{1,L}$ features. Errors were averaged for males and females and 3, 10 and 30 second test conditions. 1997 NIST-SRE corpus.

The main observations to be reached from the figure are the following:

- Inclusion of dynamic features dramatically improves performance in mismatched conditions. This can be seen by comparing system (A) to the other systems.
- When dynamic features are appended to static features, lowpass filtering of the static features dramatically improves performance in mismatched conditions. Performance is unaffected in matched conditions. This can be seen by comparing systems $(C, D_{1,L})$ to systems $(A, D_{1,L})$ and $(A, B_{1,5})$. The same conclusion was reached in Section 5.3 where static features were lowpass filtered and dynamic features were not used.
- A first order polynomial filter spanning 5 to 9 frames gives the best dynamic feature.
- Down sampling does not affect the performance deleteriously as can be seen by comparing system $(A', B'_{1,5})$ to system $(A, B_{1,5})$.

Table 5.3: Systems and features related to Fig. 5.10.

1. Proposed system $(A', B'_{1,5})$ with static and dynamic features that are both lowpass filtered and down sampled.

2. Comparison system $(A, D_{1,5})$ with static features that are lowpass filtered and dynamic features that are *not* lowpass filtered.

3. Comparison system $(A, D_{2,L})$ with static features that are lowpass filtered and dynamic features that are *not* lowpass filtered.

5.4.3 Dynamic Features Based on Second Order Polynomials

Fig. 5.10 shows the EER for various combinations of static and dynamic features obtained from the *second* order polynomial filters. Again, the same trends were observed for the MDE. In the figure, from left to right, errors are shown for systems based on two major types of features. Table 5.3 lists the systems and their associated features.

Based on the figures it can be concluded that the first order polynomial filters result in better performance than the second order polynomial filters. As a result, it was decided not to use second order polynomial filters for the remainder of experiments.

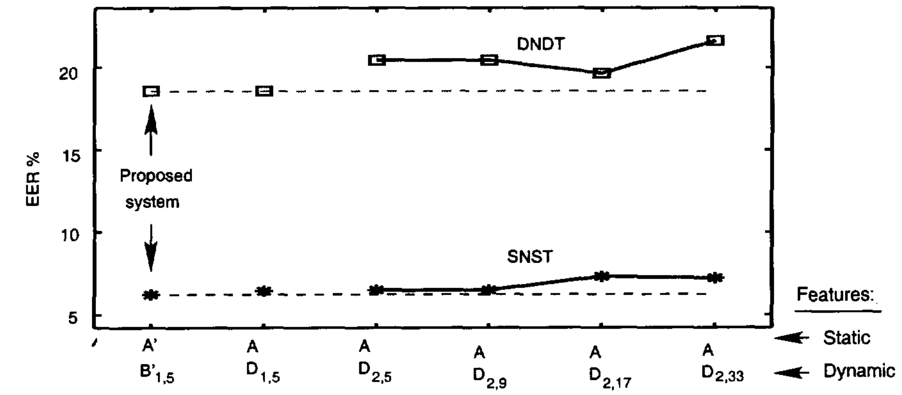


Figure 5.10: EER in percent for various combinations of static and dynamic $f_{2,L}$ features. Errors were averaged for males and females and 3, 10 and 30 second test conditions. 1997 NIST-SRE corpus.

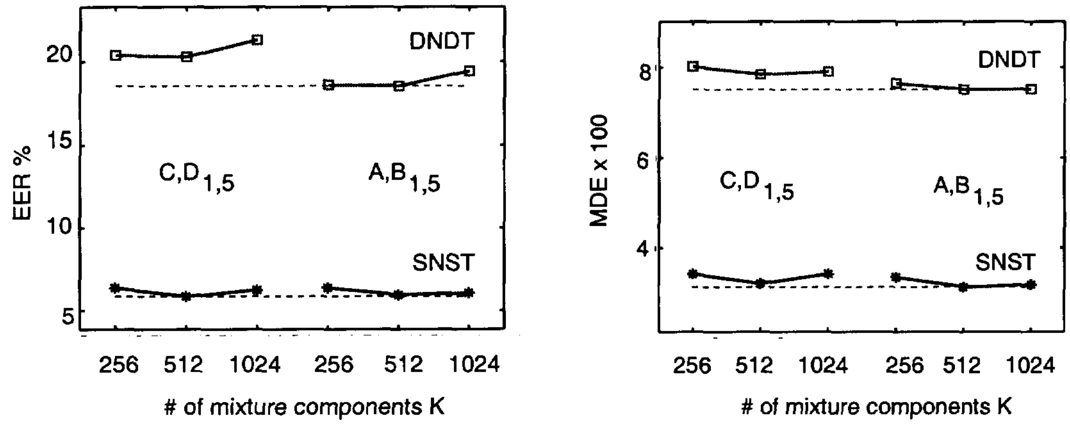


Figure 5.11: EER and MDE as a function of number of components in the GMM. Error rates for the baseline system without lowpass filtering is shown on the left. Error rates for the proposed system with lowpass filtering is shown on the right. 1997 NIST-SRE corpus.

5.4.4 Temporal and Spectral Resolution

It may be argued that the improvements obtained by lowpass filtering are simply the effect of the modeling being biased to a relatively low spectral and temporal resolution. To test this argument the baseline system without lowpass filtering $(C, D_{1,5})$ was compared to the proposed system with lowpass filtering $(A, B_{1,5})$ while increasing the number of components in the GMM from 256 to 1024. Fig 5.11 shows the effect of doing this. Clearly, the results show that the argument is false and that lowpass filtering is beneficial in general, irrespective of the temporal and spectral resolution.

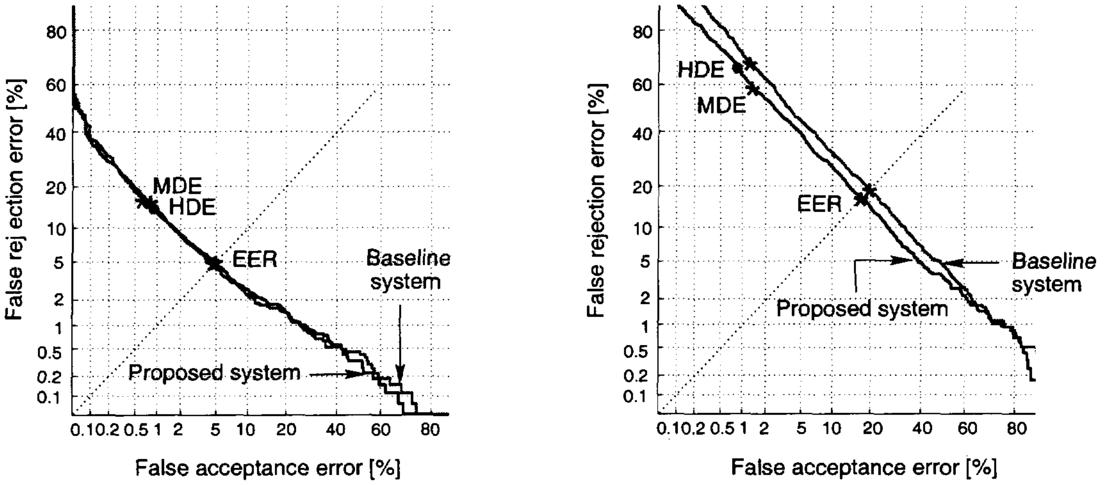
5.5 Test Set Performance

This section presents official results from the 1998 NIST-SRE [59]. This is meant to provide additional support for the results shown previously in this chapter. Recall that these results were obtained on a development data set through a series of optimization experiments using the speech corpus from the 1997 NIST-SRE. It is therefore appropriate to test the main conclusion that the time sequences should be lowpass filtered and down sampled on an independent, unseen data set. For this purpose the speech corpus from the 1998 NIST-SRE was used as a test set and verification results compared for a baseline system and the proposed system on the 30 second test segments for males and females. The system specifications were as follows.

Baseline system The baseline system consisted of static and 5-point dynamic features to which MS filtering was applied. The features were sampled at a rate of 100 Hz. ($(C, D_{1,5})$ in figure 5.7.)

Proposed system The proposed system consisted of static and 5-point dynamic features to which MS and lowpass filtering to 10 Hz was applied. The features were sampled at a rate of 25 Hz. ($(A', B'_{1,5})$ in figure 5.7.)

In the mismatched condition (DNDT), the proposed system performed significantly better than the baseline system, while in the matched condition (SNST), the two systems performed similarly. The difference in performance in the mismatched condition was statistically significant at the level $\alpha = 0.02$. In the mismatched condition the proposed system resulted on average in a relative reduction of the MDE and EER of more than 10%. The results are tabulated in Appendix C together with an indication where the differences in performance are statistically significant. Fig. 5.12 shows DET plots of the system per-



(a) Matched condition (SNST)

(b) Mismatched condition (DNDT)

Figure 5.12: DET plot with EER, MDE and HDE points indicated for the baseline system and the proposed system. (See text for details.)

formances. Results are shown for matched (SNST) and mismatched (DNDT) conditions on 30-second long segments of speech for the 2-session training condition. Scores from males and females were pooled in computing the DET curve. The performance differences are evident.

5.6 Discussion

The highpass filtering, lowpass filtering and down sampling of the features are discussed in this section.

5.6.1 Highpass Filtering

In recent years, RASTA-filtering has become a popular processing technique for dealing with adverse environments. In [37] the 1 Hz highpass cut-off of the RASTA filter was contrasted to the variable cut-off obtained with a MS filtering. In the current work consistently better performance was obtained at a relatively low MS cut-off of around 0.025 Hz – substantially lower than the higher 1 Hz cut-off. This suggests that the 1 Hz highpass cut-off of the RASTA filter should be lowered for speaker verification. Furthermore, in that the performance curves appears to flatten out only as the cut-off is pushed to below about 0.1 Hz (see Figures 5.3 and 5.4) it is surmised that as much as 10 seconds of speech may be needed to obtain a reasonable channel estimate for MS.

5.6.2 Lowpass Filtering

The concept of the modulation spectrum has, to the best of our knowledge, not been used elsewhere to investigate filtering for speaker verification. It should be noted though that benefits of a lowpass type filtering have been reported before. In [26] for example, it was reported that dynamic features obtained from a first order orthogonal polynomial expansion of 90 ms long segments of the time sequence outperformed other choices across a range of conditions and experiments.[3] However, static features were derived directly

[3] The 90 ms duration does not have to be interpreted as relating to lowpass filtering. It can also be interpreted as specifying the context necessary to compute the polynomial expansion.

from the short-term spectrum – without any lowpass filtering. Whether an expansion on 90 ms long segments would improve the static features was therefore unclear at the time. Essentially the same conclusions were reached in [1] where the usefulness of regression features were investigated for ASR. In [92], it was reported that estimating the static and dynamic features within a window of about 75-165 ms (5-11 frames) generally gave better performance.

These results agree with the results reported in Section 5.4.2, where good performance was indicated for dynamic features derived using the filter $f_{1,9}$ which was shown in Section 5.4.1 to have a frequency response with a peak at the relatively low modulation frequency of 7.5 Hz. Given the previous work and the results presented in this dissertation, it seems reasonable to conclude that speaker verification performance improves with the use of lowpass filtered static and dynamic features.

5.6.3 Down Sampling

The benefits of down sampling may be more important than the reduction in error afforded by the lowpass filtering. While down sampling does not increase the error it allows a substantial computational savings. Compared to a baseline system that models and scores the features at a rate of 100 Hz, it can be seen that a system that instead models and scores the features at a rate of 25 Hz, as proposed here, would achieve a 75% computational savings. This savings is immediate, since after down sampling, only one fourth of the frames have to be modeled or scored. In general, the savings can be expected to be linear in the amount of speech to be processed, since the result can always be obtained after evaluating a fixed number of models. If many models have to be evaluated the savings in the scoring stage can be expected to be substantial.

5.7 Summary

Filtering in the time-feature space was used in this chapter to measure the importance of the components of the modulation spectrum for speaker verification. This was done by analyzing the effects that lowpass, highpass, and bandpass filtering of the time sequences of logarithmic energy have on the speaker verification error. It was concluded that modulation frequencies in the 0.1-10 Hz range convey the most useful information for speaker verification.

Dynamic features were derived from a projection onto orthogonal polynomials and the usefulness of different polynomials investigated. The aim was to find a reasonable combination of static and dynamic features. A first order polynomial filter spanning 5 to 9 frames was found to give the best dynamic feature. The frequency responses of these filters exhibited a bandpass characteristic with peaks at modulation frequencies of 3.9 and 7.5 Hz respectively. It was shown that dynamic features appended to static features improve performance in mismatched conditions. Lowpass filtering of the static features specifically, was shown to further improve performance in mismatched conditions and not to affect performance in matched conditions.

The analysis contributed to an understanding of the effects and usefulness of contemporary processing techniques such as CMS and RASTA. Compared to these techniques, it was shown that the time sequences of spectral energy may be lowpass filtered to as low as 10 Hz and down sampled by as much as a factor of four, while preserving, or improving verification accuracy. It was shown that this processing significantly improves verification accuracy in the presence of acoustic mismatch of telephone handsets. While down sampling did not increase the error it allowed a substantial computational savings. Using down sampled features in the modeling and scoring stages translated into a 75% computational savings, since after down sampling only one fourth of the frames had to be modeled and scored. Results from the official 1998 NIST speaker recognition evaluation were provided to further support the claim that components of the modulation spectrum in the 0.1 to 10 Hz range contain the most useful speaker information.

Chapter 6

Conclusion

The goal in this dissertation was to determine the relative importance of components of the modulation spectrum for automatic speaker verification and to use this knowledge to improve the performance of an automatic speaker verification system. This goal was based on the thesis that some components may actually not be useful for speaker verification and that by attenuating them it may be possible to improve speaker verification performance in terms of verification accuracy, and/or computational cost.

An analysis of handset and speech variability at different modulation frequencies in the time sequences of logarithmic energy indicated that handset variability accounts for a substantial portion of the total variability at moderate to high modulation frequencies. This conclusion also held at the very low modulation frequencies, where the variability was ascribed to the effect of a convolutional transmission channel. From an analysis of the speaker verification error for various filters it was concluded that spectral components between 0.1 Hz and 10 Hz contain the most useful speaker information. It was concluded that a RASTA-type processing can be useful for speaker verification, provided that frequency components as low as 0.1 Hz are retained. It was shown that lowpass filtering to 10 Hz preserves salient speaker information while improving robustness. A reduction of the sampling rate of the feature stream from 100 Hz to 25 Hz was proposed. This afforded a 75% computational savings in the scoring stage without affecting accuracy. On the Switchboard corpus, the proposed processing resulted in a 10% relative reduction in error when there was a mismatch of the transmission channel between training and testing. The remainder of this chapter summarizes the work and conclusions described in this dissertation, highlights the original contributions and suggests directions for future research.

6.1 Summary and Results

The first part of the dissertation covered processing in a time-feature space. The main aspect of this processing was a linear filtering of the time sequences of logarithmic energy that when interpreted in the modulation spectral domain, and combined with an analysis-of-variance decomposition, allowed a characterization and comparison of speech and telephone handset variability. This part of the dissertation presented a rather general treatment of telephone handset variability in speech and as such did not specifically deal with speaker variability. It did however serve to motivate and guide the development of a proposed linear filtering of the time sequences of logarithmic energy for attenuating unwanted variability. Whereas the first part was concerned with the effect of telephone handset variability in speech in general, the second and third parts narrowed the focus to the speaker verification task specifically. The second part presented a system for text-independent speaker verification, with detailed motivation and description of the feature extraction, statistical modeling, performance evaluation and parameter choices. The third part examined temporal aspects of speaker verification in the time-feature space. This was done with a systematic investigation of the relative importance of the components of the modulation spectrum for speaker verification.

Chapter 2 covered acoustic feature extraction and processing in a time-feature space. Short-term acoustic features were motivated based on perceptual and physiological considerations. The theory of short-term analysis of the speech signal was reviewed along with common feature representations used in ASR and speaker verification. A medium-term analysis was motivated as necessary for the analysis of the characteristics of the resulting time sequences of spectral features. The modulation spectral domain was defined as a suitable domain for the analysis of the time sequences and for the characterization of various filters that manipulate these time sequences. The effects of the length of the short-term analysis window, analysis sampling rate and transmission channel on the modulation spectrum of speech were examined. The usefulness of the modulation spectrum became apparent when the problem of acoustic mismatch in automatic speaker verification was considered. The effects of convolutional channel distortions and additive noise were studied and various compensation techniques reviewed. An experimental study demonstrated the deleterious effect of acoustic mismatch that is present in adverse environments and suggested that filtering of the time sequences of spectral features may help to alleviate this mismatch. An analysis of the relative importance of the components of the modulation spectrum for speaker verification was suggested to determinate which type of filtering may improve verification performance.

Chapter 3 explored the characteristics of the short-term features in the modulation spectral domain. It was proposed that the modulation spectrum provides information that may be used to understand and reduce the effects of telephone handset variability. The approach that was followed in this dissertation was different from previous approaches in that characteristics of handset types were analyzed along the time dimension of the feature space and not along the frequency dimension. A framework for this analysis was proposed based on a spectral analysis-of-variance model. The model generalized the concept of modulation frequency and allowed the variations to be interpreted in the modulation spectral domain where specifically, handset variation was contrasted to the total and speech variations.

It was shown that telephone handset variability severely contaminates the DC-modulation component as was to be expected from a convolutional model for the transmission channel. It was also shown that the moderate to higher (above about 10 Hz) modulation frequency components can be severely contaminated by sources of extraneous variability. This led to the important conclusion that frequency components in the modulation spectrum of speech that are higher than about 10 Hz convey relatively little speech information. It was suggested that an attenuation of the higher modulation frequency components may make the features less sensitive to adverse environments and likely increase the robustness of an ASR or speaker verification system.

The ratio of speech variation to handset variation (SNR) was observed to be highest in a band of modulation frequencies centered around 3 Hz. It was found that carbon-button transducers exhibited a lower SNR than electret transducers. It was shown that the SNR also depends on the short-term analysis frequency, and is highest for a short-term analysis frequency in a range of 0.3-3 kHz. This observation agreed with a bandwidth for telephone speech in the same range. Irrespective of short-term analysis frequency though, it was observed that very low and moderate to high modulation frequencies consistently have lower SNRs. It was shown that variations in the broad-band energy of the speech signal contribute considerably to the DC variability and that a normalization is appropriate. It was seen that additive noise affects the modulation spectrum across the range of modulation frequencies. By adding noise to the speech that was recorded using the electret transducer, it was concluded that additive noise may be partly responsible for differences in carbon-button and electret handset variations.

Chapter 4 detailed the speaker verification task, describing the feature extraction, modeling and evaluation measures used in the dissertation. The chapter first described the specific processing and parameterization used for feature extraction. It then defined evaluation measures that were used to assess the accuracy of the speaker verification system. The evaluation measures were motivated by the requirements of recent speaker recognition evaluations run by the National Institute of Standards and Technology. Statistical modeling of the features was covered in detail. Past and current approaches to modeling the acoustic features were first reviewed. A test statistic for speaker verification was then motivated from statistical hypothesis testing and a model score defined. It followed that two probability density distribution functions (pdfs) had to be modeled. The one described *speaker independent* (SI) features and the other described *speaker dependent* (SD) features. A Gaussian mixture modeling approach was then motivated and detailed. The well-known Expectation-Maximization algorithm (EM-algorithm) was used to estimate the parameters in the SI model and Bayesian maximum aposteriori (MAP) adaptation of the SI model was used to derive the SD models. The chapter concluded with optimizations experiments for the parameters related to the feature extraction and statistical modeling.

It was indicated that an analysis window with a length from 20-64 ms give reasonable performance. It was concluded that band-limiting the speech signal to the telephone bandwidth improved performance in the mismatched condition. Using a short-term analysis filter bank with between 20 and 26 bands was found to give good performance. It was concluded that mean subtraction and the inclusion of dynamic features in addition to static features improve performance in the mismatched condition but that dynamic features used on their own perform less well.

Verification experiments were modeled on the recent NIST Speaker Recognition Evaluations (NIST-SRE). Matched and mismatched test conditions were introduced that allowed the performance of a speaker verification system to be studied in the presence of acoustic mismatch.

Chapter 5 presented a systematic investigation of the relative importance of the components of the modulation spectrum for speaker verification. An analysis of the error surface was proposed to confirm the observation that higher modulation frequencies are less important for speaker verification. The approach was to measure and analyze the effect on the speaker verification error of various filters designed in the modulation spectral domain and applied in the time-feature space. It was concluded that components with modulation frequencies in the 0.1-10 Hz range convey the most useful information for speaker verification. Highpass filtering of the time sequences at different cut-offs was contrasted to MS and RASTA filtering. Good performance was indicated for MS, while it was shown that the 1 Hz highpass cut-off of the RASTA filter was too high. Based on various results it was subsequently concluded that the time sequences of spectral energy may be lowpass filtered to as low as 10 Hz and down sampled from 100 Hz to as low as 25 Hz, while preserving, or improving verification accuracy. It was pointed out that this processing allows a 75% computational savings in the scoring stage.

Results for different combinations of static and dynamic features were reported. It was shown that dynamic features appended to static features improve performance in mismatched conditions. It was concluded that 5 to 9 frames provide a suitable time span for dynamic features derived from a polynomial filter. Lowpass filtering of the static features specifically, was show to further improve performance in mismatched conditions and not to affect performance in matched conditions.

Based on results from the official 1998 NIST speaker recognition evaluation, when compared to a baseline, the proposed processing resulted in a relative reduction in error of 10% in the mismatched condition. This led to the conclusion that the benefits of the proposed processing carry over to an independent test set.

6.2 Original Contributions

Original contributions of this dissertation to the areas of speech processing, speech recognition and speaker verification are listed in this section.

A Framework for the Characterization of Variability in the Modulation Spectral Domain

This dissertation provided a characterization of the effects of different handset transducers and the presence of additive noise on the time sequences of logarithmic energy. This was done by contrasting the variability in telephone speech that is due specifically to variations among handsets to the overall variability in telephone speech. The different sources of variability were modeled using an analysis-of-variance (ANOVA) that was modified to allow an interpretation at different modulation frequencies. The characterization indicated that noise and handset transducer variability severely contaminate the very low and moderate to high modulation frequency components.

An Analysis and Assessment of the Relative Importance of Components of the Modulation Spectrum for Speaker Verification

This dissertation showed that higher modulation frequencies are in general not particularly important for speaker verification. It was proposed that attenuation of components in the modulation spectrum that are highly affected by the environment or that do not actually convey useful speaker information may improve performance both in terms of verification accuracy and system speed. This was confirmed by careful analysis of the effects that lowpass, highpass, and bandpass filtering of the time sequences of spectral features have on the speaker verification error. It was found that modulation frequencies in the 0.1-10 Hz range convey the most useful information for speaker verification.

The analysis contributed to an understanding of the effects and usefulness of contemporary processing techniques such as CMS and RASTA. Compared to these techniques, it was shown that the time sequences of spectral energy may be lowpass filtered to as low as 10 Hz and down sampled by as much as a factor of four, while preserving, or improving verification accuracy. It was shown that this processing significantly improves verification accuracy in the presence of acoustic mismatch of telephone handsets.

Results from the official 1998 NIST speaker recognition evaluation were provided to further support the claim that components of the modulation spectrum between 0.1 Hz and 10 Hz contain the most useful speaker information.

Design and Implementation of a Speaker Verification Toolkit and a Systematic Optimization of a State-of-the-art Text-Independent Speaker Verification System

A major contribution of this dissertation concerned the design, implementation and optimization of a highly modular, efficient and extendible speaker verification toolkit. This toolkit has been used to build the state-of-the-art text-independent speaker verification system that was used in this dissertation and that was entered in the official 1998 NIST speaker recognition evaluation. The toolkit has contributed substantially to the speaker verification and ASR research effort in our laboratory and elsewhere and have been used by IIT Madras and CSLU among others.

Temporal Processing for ASR

A contribution, which was included in Appendix E as an extension to the dissertation, and that falls in the area of ASR, concerned the use of linear discriminant analysis (LDA) in the automatic derivation of FIR filters that optimize phoneme discriminability. Improved performance over baseline processing techniques was demonstrated on a connected digit recognition task.

6.3 Directions for Future Research

The research discussed in this dissertation is by no means an exhaustive treatment of text-independent speaker verification, feature extraction or the associated applications. This section lists some directions for future research.

6.3.1 Applications

In this dissertation the focus was on speaker verification using telephone handsets with electret or carbon-button transducers. A characterization of other types of handset would be of interest. Cellular and other wire-less telephones in particular have been gaining popularity in recent years. Since there exists a large number of different encoding protocols for these telephones, a study of feature variability among these telephones would be of interest. Similarly, in this dissertation the effects of different types of noise and other degradations were not studied, except for a characterization of the effects of additive white noise. A study using other types of noise, such as pink noise and car noise would be of interest. The effects of reverberation would also be of particular interest. In [7] and elsewhere, it was shown that reverberation exhibits an approximate lowpass filtering effect in the modulation spectrum.

Results in Chapter 3 were obtained independently of whether the task was speaker verification or automatic speech recognition. The main conclusions were corroborated with speaker verification experiments, but it would be of interest to see whether the same conclusions carry over to automatic speech recognition. In the text, experiments were mentioned that suggested that this is the case.

Speaker verification is closely related to speaker identification and speaker detection. Since many of the results in this dissertation pertain to features that describe speaker characteristics, it is expected that these results carry over to the identification and detection tasks. It would be of interest to formally verify this assumption.

6.3.2 Features

The domain that was explored in this thesis was the domain of logarithmic energies. This was motivated in that convolutional degradations are approximately additive in this domain. Noise however, is assumed to be added in the linear domain, which suggests that linear and other domains should also be explored. For example, in [37] the benefits of a log-linear domain was shown for dealing with a combination of convolutional and additive noise.

It was proposed that the time sequences be filtered using an FIR filter. It can be envisaged that the filtering could be made multi-dimensional (i.e. the time sequences in the feature stream could be filtered jointly). Similarly, the linear filtering could be extended to a non-linear filtering to deal with time-varying effects and other shortcomings in the domain of filtering.

Time sequences were *parameterized* using a simple polynomial basis. A more non-parametric temporal approach may be of use. For example, in informal experiments we found that the GMM model described in this dissertation achieved a significantly lower than chance EER when trained using the 1 kHz time sequence only, where feature vectors were obtained as consecutive blocks of 100 samples (i.e. 1 second long segments).

6.3.3 Modeling

Variations among handset transducers were measured using an analysis-of-variance in Chapter 3. Measuring variations using the second order variance statistic may be sub-optimal. For instance, while the variance statistic allows correlations between the input variables to be studied, it may be more appropriate to study statistical dependencies. To this extent, information theoretic measures such as the mutual information (MI) may be useful.

The aim in this dissertation was to attenuate modulation spectral components that were deemed not useful for the speaker verification task. As discussed in Section 1.4, other techniques such as adaptation exist for dealing with adverse environments. Application of these techniques using some of the modulation frequency techniques introduced in

Chapters 2 and 3 would be of interest.

Model scores were not normalized with respect to handset type, speaker and other conditions for the results reported in this dissertation. Based on other work, score normalizations [80] may prove useful in improving the performance of the system described here.

Bibliography

[1] Ted H. Applebaum and Brian A. Hanson. Tradeoffs in the design of regression features for word recognition. In *Proceedings of the European on Speech Communication and Technology (EUROSPEECH'91)*, pages 1203–1206, Genova, Italy, 1991.

[2] T. Arai, M. Pavel, H. Hermansky, and C. Avendano. Intelligibility of speech with filtered time-trajectories of spectral envelopes. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'96)*, pages 2490–2493, Philadelphia, PA, 1996.

[3] B. S. Atal. Automatic speaker recognition based on pitch contours. *Journal of the Acoustical Society of America*, 52(6):1687–1697, December 1972.

[4] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55(6):1304–1312, June 1974.

[5] B. S. Atal. Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, 64(4):460–475, April 1976.

[6] C. Avendano, S. Van Vuuren, and H. Hermansky. Data based filter design for RASTA-like channel normalization in ASR. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'96)*, pages 2087–2090, Philadephia, PA, 1996.

[7] Carlos Avendano. *Temporal Processing of Speech in a Time-Feature Space.* PhD thesis, Oregon Graduate Institute of Science and Technology, Portland, Oregon, April 1997.

[8] Frederic Bimbot, Ivan Magrin-Chagnolleau, and Luc Mathan. Second-order statistical measures for text-independent speaker identification. *Speech Communication*, 17:177–192, 1995.

[9] P. Bojan, O. Anderson, and P. Dalsgaard. On the robust automatic segmentation of spontaneous speech. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'96)*, pages 913–916, Philadelphia, PA, 1996.

[10] R. H. Bolt. Speaker identification by speech spectograms: A scientists view of its reliability for legal purposes. *Journal of the Acoustical Society of America*, 47(2):597–612, February 1970.

[11] L. Boves. Commercial applications of speaker verification: overview and critical success factors. In *Proceedings of Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 150–159, Avignon, France, 1998.

[12] P. Brown. *The Acoustic-Modeling Problem in Automatic Speech Recognition.* PhD thesis, Computer Science Department, Carnegie Mellon University, Pittsburg, PA, May 1987.

[13] A. Buzo, A. H. Gray, R. M. Gray, and J. D. Markel. Speech coding based upon vector quantization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28(5):562–574, October 1980.

[14] Joseph P. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1461, September 1997.

[15] J. R. Cohen. Application of an auditory model to speech recognition. *Journal of the Acoustical Society of America*, 85(6):2623–2629, June 1972.

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

[17] J. De Veth and L. Boves. Comparison of channel normalization techniques for automatic speech recognition over the phone. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'96)*, pages 2332–2335, Philadelphia, PA, 1996.

[18] M. DeGroot. *Optimal Statistical Decisions.* McGraw-Hill, New York, NY, 1970.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 1:1–38, 1977.

[20] G. R. Doddington. Speaker recognition – identifying people by their voices. *Proceedings of the IEEE*, 73:1651–1664, November 1985.

[21] G. R. Doddington. Speaker recognition evaluation methodology - an overview and perspective. In *Proceedings of Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 60–66, Avignon, France, 1998.

[22] R. Drullman, M. Feston, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *Journal of the Acoustical Society of America*, 95(5):2670–2680, May 1994.

[23] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, New York, NY, 1973.

[24] Kevin R. Farrel, R. J. Mammone, and Khaled T. Assaleh. Speaker recognition using neural networks and conventional classifiers. *IEEE Transactions on Speech and Audio Processing*, 2(1):194–204, January 1994.

[25] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, San Diego, CA, 1990.

[26] Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-29(2):254–272, April 1981.

[27] Sadaoki Furui. An overview of speaker recognition technology. In Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, editors, *Automatic Speech and Speaker Recognition - Advanced Topics*, pages 31–56. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[28] L. Gillick and S. J. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'89)*, pages 532–535, Glascow, Scotland, 1989.

[29] Herbert Gish and Michael Schmidt. Text-independent speaker identification. *IEEE Signal Processing Magazine*, 11(4):18–32, October 1994.

[30] Augustine H. Gray, Jr. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-24(5):380–391, October 1976.

[31] G. Green. *Temporal aspects of audition.* PhD thesis, Oxford, 1976.

[32] S. Greenberg. The switchboard transcription project. In *1996 CLSP/JHU Workshop on Innovative Techniques in Continuous Large Vocabulary Speech Recognition*, CLSP, Johns Hopkins University, Baltimore, MD, November 1996.

[33] Brian A. Hanson, Ted H. Applebaum, and Jean-Claude Junqua. Spectral dynamics for speech recognition under adverse conditions. In Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, editors, *Automatic Speech and Speaker Recognition - Advanced Topics*, chapter 14, pages 331–356. Kluwer Academic Publishers, 1996.

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

[35] H. Hermansky. Exploring temporal domain for robustness in speech recognition. In *Proceedings of the 15th International Congress on Acoustics*, pages 61–64, Trondheim, Norway, 1995.

[36] H. Hermansky. Should recognizers have ears? *Speech Communication*, 25:3–27, 1998.

[37] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[38] A. Higgins, L. Bahler, and J. Porter. Voice identification using nonparametric density matching. In Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, editors, *Automatic Speech and Speaker Recognition - Advanced Topics*, pages 211–232. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[39] A. L. Higgins and R. E. Wohlford. A new method of text-independent speaker recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'86)*, pages 869–872, Tokyo, Japan, 1986.

[40] H. G. Hirch, P. Meyer, and H. Ruehl. Improved speech recognition using high-pass filtering of subband envelopes. In *Proceedings of the European on Speech Communication and Technology (EUROSPEECH'91)*, pages 413–416, Genova, Italy, 1991.

[41] T. Houtgast and H. J. M. Steeneken. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *Journal of the Acoustical Society of America*, 77(3):1069–1077, March 1985.

[42] Melvyn J. Hunt. A statistical approach to metrics for word and syllable recognition. *Journal of the Acoustical Society of America*, 66(S1), S35(A), 1979.

[43] S. Itahashi and S. Yokoyama. Automatic formant extraction utilizing mel scale and equal loudness contour. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'76)*, pages 310–313, Philadelphia, PA, 1976.

[44] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel. On the importance of various modulation frequencies for speech recognition. In *Proceedings of the European on Speech Communication and Technology (EUROSPEECH'97)*, pages 1097–1101, Rhodes, Greece, 1997.

[45] D. G. Keehn. A note on learning for Gaussian properties. *IEEE Transactions on Information Theory*, 11:126–132, January 1965.

[46] Peter Ladefoged. *A course in Phonetics.* Harcourt Brace Jovanovich College Publishers, Orlando, FL, third edition, 1993.

[47] C.-H. Lee. A unified statistical hypothesis testing approach to speaker verification and verbal information verification. In *Proceedings of the COST Workshop on Speech Technology in the Public Telephone Network: Where are we today?*, pages 62–73, Rhodes, Greece, 1997.

[48] C.-H. Lee. A tutorial on speaker and speech verification. In *Proceedings of the Nordic Signal Processing Symposium (NORSIG-98)*, pages 9–16, Vigso, Denmark, 1998.

[49] C.-H. Lee and J.-L. Gauvian. Bayesian adaptive learning and MAP estimation of HMM. In Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, editors, *Automatic Speech and Speaker Recognition - Advanced Topics*, pages 83–107. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[50] C. J. Leggetter and P. C. Woodland. Speaker adaptation of hmms using linear regression. Technical Report TR-181, Cambridge University Engineering Department, 1994.

[51] K.-P. Li and E. H. Wrench Jr. An approach to text-independent speaker recognition with short utterances. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'83)*, pages 555–558, Boston, MA, 1983.

[52] Linguistic Data Consortium (LDC). Philadelphia, PA. See LDC's URL: http://www.ldc.upenn.edu.

[53] C.-S. Liu, H.-C. Wang, and C.-H. Lee. Speaker verification using normalized log-likelihood score. *IEEE Transactions on Speech and Audio Processing*, 4(1):56–60, January 1996.

[54] Chi-Shi Liu, Chin-Hui Lee, Wu Chou, Biing-Hwang Juang, and Aaron E. Rosenberg. A study on minimum error discriminative training for speaker recognition. *Journal of the Acoustical Society of America*, 97(1):637–648, January 1995.

[55] Michael A. Lund and C. C. Lee. A robust sequential test for text-independent speaker verification. *Journal of the Acoustical Society of America*, 99(1):609–621, January 1996.

[56] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proceedings of the European on Speech Communication and Technology (EUROSPEECH'97)*, pages 1895–1898, Rhodes, Greece, 1997.

[57] A. Martin et al., editors. *Speaker Recognition Workshop Notebook*, Maritime Institute of Technology, Linthicum Heights, MD, March 1996. Publication of the National Institute of Standards and Technology. See NIST's speech URL: http://www.nist.gov/speech.

[58] A. Martin et al., editors. *Speaker Recognition Workshop Notebook*, Maritime Institute of Technology, Linthicum Heights, MD, June 1997. Publication of the National Institute of Standards and Technology. See NIST's speech URL: http://www.nist.gov/speech.

[59] A. Martin et al., editors. *Speaker Recognition Workshop Notebook*, University of Maryland, College Park, MD, March/April 1998. Publication of the National Institute of Standards and Technology. See NIST's speech URL: http://www.nist.gov/speech.

[60] A. Martin and Mark A. Przybocki. Speaker recognition evaluation plan. Publication of the National Institute of Standards and Technology, 1998. See NIST's speaker recognition URL: http://www.nist.gov/speech/test.htm.

[61] T. Matsui and S. Furui. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'92)*, pages 157–160, San Francisco, CA, 1992.

[62] J. M. Naik, L. P. Netch, and G. R. Doddington. Speaker verification over long distance telephone lines. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'89)*, pages 524–527, Glascow, Scotland, 1989.

[63] J. Oglesby and J. S. Mason. Optimization of neural models for speaker identification. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'90)*, pages 261–264, Albuquerque, NM, 1990.

[64] Joseph P. Olive, Alice Greenwood, and John Coleman. *Acoustics of American English Speech*. Springer-Verlag, New York, NY, 1993.

[65] A. V. Oppenheim and R. W. Schafer. *Discrete-time signal processing.* Prentice-Hall, Englewood Cliffs, New Jersey, second edition, 1989.

[66] D. Ormoneit and V. Tresp. Improved gaussian mixture density estimates using Bayesian penalty terms and network averaging. Technical Report FKI-205-95, Technische Universität München, München, Germany, 1995.

[67] Douglas O'Shaughnessy. *Speech Communication - Human and Machine.* Addison-Wesley, Reading, MA, first edition, 1987.

[68] John K. Ousterhout. *Tcl and the Tk Toolkit.* Addison-Wesley, Reading, MA, 1994.

[69] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, New York, NY, third edition, 1991.

[70] A. B. Poritz. Linear predictive hidden markov models and the speech signal. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'82)*, pages 1291–1294, Paris, France, 1982.

[71] S. Pruzansky. Pattern-matching procedure for automatic talker recognition. *Journal of the Acoustical Society of America*, 35(3):354–358, March 1963.

[72] Mark A. Przybocki and Alvin F. Martin. NIST speaker recognition evaluation -1997. In *Proceedings of Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 120–123, Avignon, France, 1998.

[73] Mark A. Przybocki and Alvin F. Martin. NIST speaker recognition evaluations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 331–335, Grenada, Spain, 1998.

[74] T. F. Quatieri, D. A. Reynolds, and G. C. O'Leary. Magnitude-only estimation of handset nonlinearity with application to speaker recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, pages 745–748, Seattle, WA, 1998.

[75] L. R. Rabiner. *Fundamentals of Speech Recognition.* Prentice-Hall, Englewood Cliffs, NJ, 1993.

[76] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, Englewood Cliffs, NJ, 1978.

[77] Bhiksha Raj, Evandro B. Gouvêa, Pedro J. Moreno, and Richard M. Stern. Cepstral compensation by polynomial approximation for environment-independent speech

recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'96)*, pages 2340–2343, Philadelphia, PA, 1996.

[78] D. A. Reynolds. *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification.* PhD thesis, Georgia Institute of Technology, Atlanta, GA, August 1992.

[79] D. A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17:91–108, 1995.

[80] D. A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proceedings of the European on Speech Communication and Technology (EUROSPEECH'97)*, pages 963–970, Rhodes, Greece, 1997.

[81] D. A. Reynolds. HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, pages 1535–1538, Munich, Germany, 1997.

[82] R. Riesz. Differential intensity sensitivity of the ear for pure tones. *Psychology Review*, 31:867–875, 1928.

[83] B. D. Ripley. *Pattern Recognition and neural networks.* Cambridge University Press, Cambridge, 1996.

[84] A. Rosenberg. Automatic speaker verification: A review. *Proceedings of the IEEE*, 64(4):475–487, April 1976.

[85] A. E. Rosenberg, C.-H. Lee, and F. K. Soong. Sub-word unit talker verification using hidden markov models. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'90)*, pages 269–272, Albuquerque, NM, 1990.

[86] Aaron E. Rosenberg and Frank K. Soong. Recent research in automatic speaker recognition. In Sadoaki Furui and M. Mohan Sondhi, editors, *Advances in Speech Signal Processing*, pages 701–737. Marcel Dekker, New York, NY, 1992.

[87] Johan Schalkwyk, Jacques de Villiers, Sarel Van Vuuren, and Pieter Vermeulen. CSLUsh: an extendible research environment. In *Proceedings of the European on Speech Communication and Technology (EUROSPEECH'97)*, pages 698–701, Rhodes, Greece, 1997.

[88] M. Schmidt, H. Gish, and A. Mielke. Covariance estimation methods for channel robust text-independent speaker identification. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, pages 333–336, Detroit, MI, 1995.

[89] Schmidt-Nielsen and Thomas H. Crystal. Nirdal system. In *Speaker Recognition Workshop Notebook*, University of Maryland, College Park, MD, March/April 1998. Publication of the National Institute of Standards and Technology. See NIST's speech URL: http://www.nist.gov/speech.

[90] M. R. Schroeder. Modulation transfer functions: Definition and measurement. *Acustica*, 49:179–182, 1981.

[91] G. W. Snedecor and W. G. Cochran. *Statistical methods.* Iowa State University Press, Ames, Iowa, sixth edition, 1967.

[92] F. K. Soong and A. E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-36(6):871–879, June 1988.

[93] F. G. Stremler. *Introduction to Communication Systems.* Addison-Wesley, Reading, MA, second edition, 1982.

[94] Don X. Sun and Li Deng. Analysis of acoustic-phonetic variations in fluent speech using TIMIT. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, pages 201–204, Detroit, MI, 1995.

[95] A. C. Surendran, Chin-Hui Lee, and Mazin Rahim. Maximum-likelihood stochastic matching approach to non-linear equalization for robust speech recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'96)*, pages 1836–1839, Philadelphia, PA, 1996.

[96] Matthew Turk and Alex Pentland. Eigen faces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[97] Dirk Van Compernolle. Noise adaptation in a hidden markov model speech recognition system. *Computer Speech and Language*, 3:151–167, 1989.

[98] S. H. J Van Vuuren. Neural network correlates with generalization. Master's thesis, University of Pretoria, Pretoria, South Africa, 1994.

[99] Sarel Van Vuuren. Mx: A package for rapid mathematical prototyping and algorithm development with application to speech and speaker recognition. Technical Report CSE-98-003, Department of Computer Science and Engineering, Oregon Graduate Institute, Portland, OR, March 1998.

[100] Sarel Van Vuuren and Hynek Hermansky. Data-driven design of RASTA-like filters. In *Proceedings of the European on Speech Communication and Technology (EUROSPEECH'97)*, pages 409–412, Rhodes, Greece, 1997.

[101] Sarel Van Vuuren and Hynek Hermansky. !MESS: A modular, efficient speaker verification system. In *Proceedings of Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 189–201, Avignon, France, 1998.

[102] W. D. Voiers. Perceptual bases of speaker identity. *Journal of the Acoustical Society of America*, 36(6):1065–1073, June 1964.

[103] J. Wolf. Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*, 51(6):2044–2056, June 1972.

[104] Howard H. Yang, Sarel Van Vuuren, and Hynek Hermansky. Relevancy of time-frequency features for phonetic classification measured by mutual information. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, Phoenix, AZ, 1999. To appear.

[105] E. Zwicker. Die grenzen der hoerbarkeit der amplitudenmodulation under der frequenzmodulation eines tones. *Acustica*, 2:125–133, 1952.

Appendix A

Experimental Setup

This appendix describes the experimental setup that is used with the speaker verification experiments in this dissertation. The setup is based on the 1996, 1997 and 1998 NIST Speaker Recognition Evaluations (NIST-SRE) [57, 58, 59, 72, 73]. See [60] for the official evaluation plans.

Section A.1 describes the various training and testing conditions of interest. These include matched and mismatched training and testing environments. Section A.2 describes the data set organization. Section A.3 highlights the major similarities between the NIST-SRE and the experimental conditions used in the dissertation.

A.1 Training and Testing Conditions

Segments of conversational telephone speech from the Swb-I and Switchboard-2 phase 1 and 2 corpora are used in the experiments. The experimental conditions of interest are summarized in the following list.

Training condition. The training condition relates to the session and handset variability present in the enrollment speech.

"One-session" (1S). Enrollment speech consists of speech for a single speaker taken from one conversation (session). Approximately two minutes of speech is collected in one-minute segments.

"Two-session" (2S). Enrollment speech consists of equal amounts of speech for the same speaker taken from two different conversations (sessions) collected using the same handset (same telephone number). Approximately one minute of speech is collected from each conversation.

"Two-session-full" (2F). Same as 2S but with more speech available for training.

Gender. Performance is computed separately for males and females.[1] Results are averaged.

Test speech duration. Performance is computed separately for segments of test speech with nominal durations of 3, 10 and 30 seconds.

Test condition. Performance is computed for segments of test speech collected using the training or enrollment handset (matched condition), versus segments collected using a different handset (mismatched condition).

Mismatch occurs when the telephone *number* and/or *handset type* differs between training and testing. Handset types are categorized [80, 58, 59] into *electret* transducer and *carbon-button* transducer types[2]. For the corpus used in the 1996 NIST-SRE the conditions of interest are *same number* (SN) and *different number* (DN). For the corpora used in the 1997 and 1998 NIST-SREs the conditions are *same number* and *same handset type* (SNST) and *different number* and *different handset type* (DNDT).

Statistics related to the corpora, and different training and test conditions as used in this dissertation, are summarized in Table A.1.

Performance is computed separately for six different cases namely, for males and females and each of the three test durations. To simplify the interpretation of results the general approach in this dissertation will be to present either the *averaged* result from these six cases or the averaged result using 30 second segments only. In the cases where differences in the results between cases may lead to a different interpretation, more detailed results will be presented. In general though, it can be shown that conclusions to be

[1] Gender classification is considered to be a separate task from speaker verification in the NIST-SRE. The argument is that gender classification can be performed with high accuracy if needed, by following a different methodology and using, for example, features such as pitch. Gender is therefore considered to be known apriori in this dissertation.

[2] These distinctions were assessed using telephone number and handset type labels distributed by NIST.

Table A.1: Statistics of Switchboard-2 phase 1 and 2 corpora as used for training and testing in this dissertation.

NIST-SRE CORPUS	1996	1997	1998
Training condition	2S	2S	1S, 2S, 2F
Default features (cf. Fig. 5.7)	static (C)	static (C)	static, dynamic (A, B)
Default # of GMM components	128	128	256
# of speakers in SI model	43 (male)	40 (male)	40 (male)
	45 (female)	40 (female)	40 (female)
# of target speakers	21 (male)	67 (male)	250 (male)
	19 (female)	132 (female)	250 (female)
Test segment duration (sec)	3, 10	3, 10, 30	3, 10, 30
# of test segments / segment duration	1183 (male) 1197 (female)	2623 (male) 3266 (female)	2500 (male) 2500 (female)
# of target tests / segment for each gender	21 (male) 19 (female)	10 (male) 10 (female)	10 (male) 10 (female)
Total # of tests / segment duration	47586	58890	50000

drawn for any one case relate to the other cases, thus motivating this strategy. In most experiments then, results will be reported for the matched and mismatched conditions for both the EER and MDE.

A.2 Data Organization

The speech corpora are divided into training, development and test data sets. The 1996 and 1997 NIST-SRE corpora are used for system development and parameter optimization and the 1998 NIST-SRE corpus for formal testing. Each corpus contains a training data set and a test data set that are used respectively to train the speaker dependent (target) models and measure the speaker verification performance. The 1996 NIST-SRE corpus contains additional development data that are used to train a speaker independent model for that corpus. Since the 1997 and 1998 NIST-SRE corpora do not contain dedicated development data, a portion of the 1997 NIST-SRE corpus is set aside and used as development data. The speaker independent model used for the 1997 and 1998 NIST-SRE corpora is trained using this development data.

For the 1996 and 1998 NIST-SRE corpora, the same training and test data sets are used as specified in the actual 1996 and 1998 NIST-SREs. For the 1997 NIST-SRE corpus, training and test data sets were reorganized so as to match the conditions expected in the 1998 NIST-SRE.

In the experiments reported here, the SI models were trained as follows. In experiments using the 1996 NIST-SRE corpus, the SI model was trained using the development portion of that corpus which consists of 43 male and 45 female speakers. In experiments using the 1997 and 1998 NIST-SRE corpora, the SI model was trained using 40 male and 40 female speakers culled from the 1997 corpus. For all the experiments reported here, the SI models were trained on pooled male and female speech from the two-session (2S) data.

A.3 NIST Speaker Recognition Evaluation

This section briefly summarizes the NIST speaker recognition evaluations (NIST-SRE) [60] and motivates the experimental setup that is used in this dissertation.

Evaluation measures. The measures used in this dissertation are those defined and used in the NIST-SRE except that here scores from males and females are not pooled. Instead, the errors are averaged across gender and test speech duration.

Training conditions. The NIST-SREs specify different training conditions. Only the "one-session", "two-session" and "two-session-full" conditions are investigated here.

Test conditions. The test conditions investigated here are the same as in the NIST-SRE.

Design and evaluation rules. The rules specified in the NIST-SRE are followed here. Each decision is based only on the specified segment of test speech and current target speaker. This precludes the use of discriminative training and does not allow normalization across test speech segments. It does however allow extendibility since the system can operate as soon as a single person has enrolled and new user can be added at any time.

A speaker verification system was described in Chapter 4 that used processing that was proposed in Chapter 5. This system and processing were formally evaluated in the NIST-SREs where it performed competitively with systems from other groups.

Appendix B

Estimation of GMM Parameters

This appendix covers the theory and implementation of estimating a Gaussian mixture density for multivariate data. The well-known Expectation Maximization technique which fits the parameters of the density by maximizing the likelihood function is generalized to the case of maximizing the posterior probability. The usefulness of this generalization is examined in terms of a) Bayesian optimal updating of the density estimate given additional data and b) Bayesian regularization of the density estimate.

B.1 Introduction

Consider the problem of estimating the probability density of a continuous random vector $X \in \mathcal{R}^d$ for which a *data set* $\boldsymbol{X} = \{X_1, X_2, ..., X_n, ...X_N\}$ of N independent and identically distributed realizations of X is available. For this purpose assume that the density is well modeled by a weighted mixture of Gaussians

$$p(X|\Theta) = \sum_{k=1}^{K} \omega_k p(X|\mu_k, \Sigma_k), \tag{B.1}$$

where the mixing proportions $\omega_k \geq 0$, $\sum_{k=1}^{K} \omega_k = 1$ and where $\Theta = (\omega_k, \mu_k, \Sigma_k)_{k=1}^{K}$ denotes the parameter vector. To write the likelihood it is convenient to substitute $R_k = \Sigma_k^{-1}$ and use the identity $(X-\mu)^t R(X-\mu)^t = \text{tr } R(X-\mu)(X-\mu)^t$ where tr is the trace operator and t is the transpose operator. Then each Gaussian density can be written as

$$p(X|\mu_k, R_k) = (2\pi)^{-d/2}|R_k|^{1/2}\exp[-\frac{1}{2}\text{tr } (R_k(X-\mu_k)(X-\mu_k)^t)], \tag{B.2}$$

where for the k-th component, μ_k is the mean vector and Σ_k is the covariance matrix.

Table B.1: EM-algorithm.

EM algorithm

E-step: Given a current parameter estimate $\Theta^{(i)}$ compute
$Q(\Theta|\Theta^{(i)}) = E_{\{X,l\}}[\log p(X^*|\Theta) + \log P(\Theta)|\Theta^{(i)}, X]$.

M-step: Choose the new parameter estimate $\Theta^{(i+1)}$ as
the Θ that maximizes $Q(\Theta|\Theta^{(i)})$.

Within the framework of Bayesian learning [23], to find the MAP parameter estimate it is necessary to maximize the posterior probability $P(\Theta|\boldsymbol{X})$. Using Bayes' law, this is equivalent to maximizing $p(\boldsymbol{X}|\Theta)P(\Theta)$, where $p(\boldsymbol{X}|\Theta)$ is the data likelihood and $P(\Theta)$ is a prior distribution for the parameters. Keehn [45] first derived the parameter estimates for a single Gaussian density by comparing the parameters in the prior and posterior distributions. Here we will use the Expectation Maximization (EM) algorithm [19] to derive the parameter estimates for the case of incomplete data, where in the mixture of Gaussians the probability $p(X_n, k)$ of assigning a sample X_n to the k-th Gaussian component is unknown. Assuming a generative model for X, let X^* denote the complete data vector – that is, the incomplete data vector X and the component assignments (X, l), $1 \leq l \leq K$.

The EM algorithm was proposed as an iterative technique to find the mode of the complete-data log-likelihood or complete-data log-posterior [19]. For the latter, given an appropriate prior distribution the complete-data log-posterior expands as $\log p(\Theta, \boldsymbol{X}^*) \propto l(\Theta, \boldsymbol{X}^*) = \log p(\boldsymbol{X}^*|\Theta) + \log P(\Theta)$. The EM algorithm is then formalized in Table B.1.

In the E-step the expectation is estimated over the available data $\{(X, l)\}$. Iterative application of this procedure can be shown [19] to lead to an estimate of the mode of the log-posterior distribution. Closed form solutions for the parameter updates can be derived so that the auxiliary function $Q(\Theta|\Theta^{(i)})$ need not be computed in practice.

To derive these MAP parameter update rules it is necessary to specify the data likelihood and prior distribution. Using Equations (B.1) and (B.2) the likelihood for N independent and identically distributed samples is given by

$$\begin{aligned} p(\boldsymbol{X}|\Theta) &= p(X_1, X_2, ..., X_N|\Theta) \\ &= \prod_{n=1}^{N} \sum_{k=1}^{K} \omega_k (2\pi)^{-d/2} |R_k|^{1/2} \exp[-\frac{1}{2}\text{tr } (R_k(X_n - \mu_k)(X_n - \mu_k)^t)]. \quad \text{(B.3)} \end{aligned}$$

B.2 Prior Distribution

It is convenient to use a *conjugate* family of distributions where the prior distribution $P(\Theta)$ and posterior distribution $P(\Theta|\boldsymbol{X})$ belong to the same family. A thorough treatment of such distributions can be found in [18]. Following [45], a conjugate prior distribution $p(\mu_k, R_k)$ for the multivariate normal distribution in Equation (B.2) that reflects an initial estimate for the mean and covariance parameters ($\hat{\mu}_k$, $\hat{R}_k$) is the product of a normal distribution

$$p(\mu_k|\hat{\mu}_k, \eta R_k) \propto |\eta R_k|^{1/2} \exp[-\frac{1}{2}\text{tr } \eta R_k(\mu_k - \hat{\mu}_k)(\mu_k - \hat{\mu}_k)^t], \quad \eta \geq 0 \qquad \text{(B.4)}$$

and a Wishart distribution[1]

$$p(R_k|\rho, \hat{\Sigma}_k) \propto |R_k|^{(\rho-1)/2} \exp[-\frac{1}{2}\text{tr } \rho\hat{\Sigma}_k R_k], \quad \rho \geq 0. \qquad \text{(B.5)}$$

[1] The Wishart distribution Equation (B.5) is often parameterized as a multivariate generalization of the beta distribution

$$p(R|\alpha, \beta) = c(d, \alpha + 1) \left|\frac{\beta}{2}\right|^{\alpha/2} |R|^{(\alpha-d-1)/2} \exp[-\frac{1}{2} tr\beta R],$$

where $\alpha > d-1$ and $c(d, \alpha) = \left\{\pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\left[\frac{\alpha-i}{2}\right]\right\}^{-1}$. For our purposes it is convenient to substitute $\rho = \alpha - d$ and $\hat{\Sigma} = \frac{\beta}{(\alpha-d)}$.

A conjugate prior distribution for the mixture weights is the Dirichlet distribution [18] with[2]

$$p(\omega|\nu) \propto \prod_{k=1}^{K} \omega_k^{\nu_k}, \text{ with } \omega_k \geq 0, \sum_{k=1}^{K} \omega_k = 1, \text{ and } \nu_k \geq 0. \tag{B.6}$$

The meta-parameters η, ρ and ν reflect the respective confidence in the estimate of μ_k, $\Sigma_k = {R_k}^{-1}$ and ω.

B.3 MAP Parameter Updates

The posterior probability that X_n is assigned to mixture component k given the current parameter estimate $\Theta^{(i)}$ and data $\boldsymbol{X}$ can be written as

$$p^{(i)}(n,k) = p(X_n, k|\Theta^{(i)}, \boldsymbol{X}) = \frac{\omega_k^{(i)} p(X_n|\mu_k^{(i)}, R_k^{(i)})}{\sum_{j=1}^{K} \omega_j^{(i)} p(X_n|\mu_j^{(i)}, R_j^{(i)})}. \tag{B.7}$$

The auxiliary function to be maximized with respect to the parameters $\Theta = (\omega_k, \mu_k, \Sigma_k)_{k=1}^{K}$ can then be written as

$$\begin{aligned} Q(\Theta|\Theta^{(i)}) = & \sum_{n=1}^{N} \sum_{k=1}^{K} p^{(i)}(n,k) \left[\log \omega_k + \log p(X_n|\mu_k, R_k)\right] \\ & + \log p(\omega|\nu) + \sum_{k=1}^{N} \left[\log p(\mu_k|\hat{\mu}_k, \eta R_k) + \log p(R_k|\rho, \hat{\Sigma}_k)\right]. \end{aligned} \tag{B.8}$$

Maximization proceeds by setting the derivatives with respect to each parameter to zero and solving. The following differentiation rules [25], pp. 566-567, will prove useful. For matrices A and B

$$\frac{d}{dA}\text{tr }(AB) = \frac{d}{dA}\text{tr }(BA) = B^t \text{ and } \frac{\partial}{\partial A}\log|A| = (A^{-1})^t.$$

[2] The choice of the Dirichlet distribution stems from the observation that the probability for the assignment of a sequence of samples $\boldsymbol{X}$ among K mixture components follows a multinomial distribution

$$p(X_1, X_2, ...X_N|\omega_1, ..., \omega_K) \propto \prod_{k=1}^{K} \omega_k^{n_k}$$

where n_k is the number of occurrences of the k-th assignment. For our purposes its convenient to substitute $\nu = \gamma - 1$ for the conventional parameterization of the Dirichlet distribution.

B.3.1 Weights

For the weights the constraint that $\sum_{j=1}^{K} \omega_j = 1$ may be enforced using a Lagrange multiplier. Collecting the appropriate terms from $Q(\Theta|\Theta^{(i)})$ this yields,

$$L(\omega_k, \lambda) = \sum_{n=1}^{N} \sum_{k=1}^{K} p^{(i)}(n,k) \log \omega_k + \sum_{k=1}^{K} \nu_k \log \omega_k - \lambda \left[\sum_{j=1}^{K} \omega_j - 1 \right].$$

Taking derivatives and setting to zero,

$$\frac{\partial}{\partial \omega_k} L = \sum_{n=1}^{N} p^{(i)}(n,k) {\omega_k}^{-1} + \nu_k {\omega_k}^{-1} - \lambda = 0$$

$$\frac{\partial}{\partial \lambda} L = - \left[\sum_{j=1}^{K} \omega_j - 1 \right] = 0$$

Solving yields

$$\omega_k^{(i+1)} = \frac{\sum_{n=1}^{N} p^{(i)}(n,k) + \nu_k}{N + \sum_{j=1}^{K} \nu_j}. \tag{B.9}$$

B.3.2 Means

Differentiating $Q(\Theta|\Theta^{(i)})$ with respect to μ_k and setting to zero

$$\frac{1}{2} \sum_{n=1}^{N} p^{(i)}(n,k) R_k (X_n - \mu_k) - \frac{1}{2} \eta R_k (\mu_k - \hat{\mu}_k) = 0.$$

Since $|R_k| \neq 0$ solving yields

$$\mu_k^{(i+1)} = \frac{\sum_{n=1}^{N} p^{(i)}(n,k) X_n + \eta \hat{\mu}_k}{\sum_{n=1}^{N} p^{(i)}(n,k) + \eta}. \tag{B.10}$$

B.3.3 Variances

Differentiating $Q(\Theta|\Theta^{(i)})$ with respect to R_k and setting to zero

$$\frac{1}{2} \sum_{n=1}^{N} p^{(i)}(n,k) \left[{R_k}^{-1} - (X_n - \mu_k)(X_n - \mu_k)^t \right]$$
$$+ \frac{1}{2} \left[{R_k}^{-1} - \eta (\mu_k - \hat{\mu}_k)(\mu_k - \hat{\mu}_k)^t \right] + \frac{1}{2} \left[(\rho - 1) {R_k}^{-1} - \rho \hat{\Sigma}_k \right] = 0.$$

Substituting $\Sigma_k = {R_k}^{-1}$ and solving yields

$$\Sigma_k^{(i+1)} = \frac{\sum_{n=1}^{N} p^{(i)}(n,k)(X_n - \mu_k)(X_n - \mu_k)^t + \eta(\mu_k - \hat{\mu}_k)(\mu_k - \hat{\mu}_k)^t + \rho \hat{\Sigma}_k}{\sum_{n=1}^{N} p^{(i)}(n,k) + \rho}. \tag{B.11}$$

B.3.4 Discussion

To interpret the update equations it is informative to substitute

$$1 - \lambda = \frac{\eta}{\sum_{n=1}^{N} p^{(i)}(n,k) + \eta.}$$

Then Equation (B.10) for example, can be written in the form

$$\mu_k^{(i+1)} = \lambda \mu_k^{\boldsymbol{X}} + (1-\lambda)\hat{\mu}_k, \tag{B.12}$$

where $\mu_k^{\boldsymbol{X}}$ is a sufficient statistic for the data. The adaptation parameter λ can be seen to form the new estimate from the weighted combination of the current and previous estimates. The other update rules can be interpreted similarly.

B.4 Initial Parameter Estimates

Equations (B.9), (B.10) and (B.11) may be used to deduce learning rules for the initial parameters if they are not known. With $p^{(i)}(n,k)$ defined as before, for the weights, setting $\nu_k = 0$ (giving an uninformed prior) yields

$$\hat{\omega}_k^{(i+1)} = \frac{1}{N} \sum_{n=1}^{N} p^{(i)}(n,k). \tag{B.13}$$

For the means, setting $\eta = 0$ yields

$$\hat{\mu}_k^{(i+1)} = \frac{\sum_{n=1}^{N} p^{(i)}(n,k) X_n}{\sum_{n=1}^{N} p^{(i)}(n,k)}. \tag{B.14}$$

For the variances, setting $\eta = 0$ and $\rho = 0$ yields

$$\hat{\Sigma}_k^{(i+1)} = \frac{\sum_{n=1}^{N} p^{(i)}(n,k)(X_n - \hat{\mu}_k)(X_n - \hat{\mu}_k)^t}{\sum_{n=1}^{N} p^{(i)}(n,k)}. \tag{B.15}$$

With the choice of uninformative priors made above these learning rules are the same as given by the EM algorithm when maximizing the complete-data log-likelihood.

B.5 Regularization

For a finite amount of data samples, the auxiliary function $Q(\Theta|\Theta^{(i)})$ typically has many local maxima as well as singularities. For this reason it is usually helpful to introduce a prior distribution for the variances when estimating the initial parameters. A reasonable prior that regularizes smoothness of the estimated density is provided by the Wishart distribution of Equation (B.5). In particular, choosing $\rho\hat{\Sigma}_k = \epsilon I$ in Equations (B.5) and (B.11), where I is the unit matrix, serves to bias the covariance estimate away from becoming singular [66]. The resultant variance estimate is

$$\hat{\Sigma}_k^{(i+1)} = \frac{\sum_{n=1}^{N} p^{(i)}(n,k)(X_n - \mu_k)(X_n - \mu_k)^t + \epsilon I}{\sum_{n=1}^{N} p^{(i)}(n,k) + \epsilon}. \tag{B.16}$$

A reasonable value for ϵ may be obtained from cross-validation, with the value depending on N, d and the scale of the data. For a large amount of data, ϵ may be very small. For example, for speaker verification with normalized data, $d = 19$ and $N = 40000$, cross-validation yielded $\epsilon \approx 1$ while for $N = 10^6$ a value $\epsilon < 0.2$ was found to give good results.

B.6 Numerical Implementation

To implement the EM algorithm it is necessary to have initial (seed) estimates for ω_k, μ_k and Σ_k. The means are initialized using centroids μ_k^{VQ} derived from Vector Quantization [13]. The training algorithm can be summarized as follows.

Given the set $\boldsymbol{X} = \{X_1, X_2, ..., X_n, ...X_N\}$ of N training vectors, in a feature space $\mathcal{S} = \mathcal{R}^d$, find a partitioning of the space $\mathcal{S}$ into $K << N$ non-overlapping regions $\mathcal{S}_1$, $\mathcal{S}_2$, ... $\mathcal{S}_K$. Specifically, $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$, $i \neq j$, where $\emptyset$ denotes the empty set and $\mathcal{S} = \cup_{i=1}^{K} \mathcal{S}_i$. Each partition is represented by a centroid vector μ_k^{VQ}. Training iteratively minimizes the average distortion

$$D = \frac{1}{N} \sum_{n=1}^{N} \min_k d(X_n, \mu_k^{\mathrm{VQ}}), \tag{B.17}$$

where $d(X_n, \mu_k^{\mathrm{VQ}})$ is the distortion between the training vector X_n and the centroid vector μ_k^{VQ} of the k-th partition. In implementation, the algorithm may be initialized with a single centroid $K = 1$, which is optimized for and then split into two centroids $K = 2$, which are jointly optimized for by minimizing the average distortion. This process is iterated until the desired number of centroids is reached and the optimization has converged.

Following vector quantization, the parameters are subsequently initialized to

$$\omega_k^{(0)} = \frac{1}{K} \quad , \quad \mu_k^{(0)} = \mu_k^{\mathrm{VQ}} \quad , \quad \Sigma_k^{(0)} = \xi S_{Tk}, \tag{B.18}$$

where S_{Tk} is a pooled covariance estimate for the data with $\xi = 2$ typical.

To refine the initial parameter estimates using the EM algorithm it is efficient to accumulate for $n = 1...N$ the sufficient statistics. Starting from zero values, for each EM iteration the following statistics are accumulated

$$\omega_k^{\mathrm{acc}} = \sum_n p^{(i)}(n,k) \tag{B.19}$$

$$\mu_k^{\mathrm{acc}} = \sum_n p^{(i)}(n,k) X_n \tag{B.20}$$

$$S_k^{\mathrm{acc}} = \sum_n p^{(i)}(n,k) X_n X_n^t \tag{B.21}$$

$$N^{\mathrm{acc}} = \sum_n 1. \tag{B.22}$$

The parameters are then updated as

$$\mu_k^{(i+1)} = \frac{\mu_k^{\mathrm{acc}}}{\omega_k^{\mathrm{acc}}} \tag{B.23}$$

$$\Sigma_k^{(i+1)} = \frac{{S_k}^{\mathrm{acc}} - \mu_k^{\mathrm{acc}} \left[{\mu_k}^{(i+1)} \right]^t + \epsilon I}{\omega_k^{\mathrm{acc}} + \epsilon} \tag{B.24}$$

$$\omega_k^{(i+1)} = \frac{{\omega_k}^{\mathrm{acc}}}{N^{\mathrm{acc}}}. \tag{B.25}$$

To compute $p^{(i)}(n,k)$ it is convenient to invert $\Sigma_k^{(i)}$ using its Cholesky decomposition LL^t. The determinant can be obtained as the product of the elements on the diagonal of L. The decomposition itself can be computed in place with no need to allocate extra memory.

Convergence for the EM algorithm is assumed when $|\mu_k^{(p+1)}(j) - \mu_k^{(p)}(j)| < \tau$ for $j = 1..d$ or after I iterations of the EM algorithm. Typical values are $\tau = 10^{-8}$ and $5 \leq I \leq 10$.

Instead of computing quantities in $p(X|\mu_k, R_k)$ directly, it is convenient to take logarithms. This serves to simplify the computations by allowing the multiplication to be substituted with a summation

$$p(X|\mu, R) \propto \exp \left[0.5 (\log |R| - \mathrm{tr}\ (R(X-\mu)(X-\mu)^t)) \right]. \tag{B.26}$$

Appendix C

Statistical Significance

This appendix revisits the test results provided at the end of Chapter 5. Statistical significance of the test results are examined in Section C.1 while results from other systems as reported in the 1998 NIST Speaker Recognition Evaluation (NIST-SRE) are considered in Section C.2.

C.1 Statistical Significance

The statistical significance of the test results from Chapter 5 are examined this section.

C.1.1 Exposition

Suppose that systems A and B are to be tested and compared to each other on the *same* data set. Let D_0 be the portion of the data set for which acceptance is the correct decision. Let D_1 be the portion of the data set for which rejection is the correct decision. Let H_0 denote decisions associated with D_0 and let H_1 denote decisions associated with D_1. Typically, D_1 is the larger data set and more decisions will be available for D_1 than for D_0. As in Section 4.2, define the total error or cost associated with system A as

$$E_A = C_{fr} p_A(fr|H_0)P(H_0) + C_{fa} p_A(fa|H_1)P(H_1) \tag{C.1}$$

where C_{fr} and C_{fa} are the costs associated with the false rejection and false acceptance errors. The total error for system B is defined in the same way.

Define n_{ij} as in Table C.1 to be the correct and incorrect counts for systems A and B given D_0. Similarly, define m_{ij} to be the correct and incorrect counts for systems A

Table C.1: Error counts given data set D_0.

		B	
		correct	incorrect
A	correct	n_{00}	n_{01}
	incorrect	n_{10}	n_{11}

and B given D_1. In general, n_{00} and n_{11} will be non zero, so that the error for system A will *not* be independent of the error for system B. This suggests that a test that directly compares the errors for the two systems over the whole data set will be inappropriate.

Let the total number of samples for D_0 and D_1 be

$$\begin{aligned} n &= n_{10} + n_{11} + n_{01} + n_{00}, \\ m &= m_{10} + m_{11} + m_{01} + m_{00}. \end{aligned} \tag{C.2}$$

Then the false rejection and false acceptance error probabilities for system A are

$$p_A(fr|H_0) = \frac{n_{10} + n_{11}}{n} \quad \text{and} \quad p_A(fa|H_1) = \frac{m_{10} + m_{11}}{m}, \tag{C.3}$$

with the false rejection error probability and false acceptance error probability for system B defined in a similar way as $p_B(fr|H_0)$ and $p_B(fa|H_1)$. Substituting into Equation (C.1),

$$\begin{aligned} E_A &= C_{fr}P(H_0)\left(\frac{n_{10} + n_{11}}{n}\right) + C_{fa}P(H_1)\left(\frac{m_{10} + m_{11}}{m}\right), \\ E_B &= C_{fr}P(H_0)\left(\frac{n_{01} + n_{11}}{n}\right) + C_{fa}P(H_1)\left(\frac{m_{01} + m_{11}}{m}\right). \end{aligned} \tag{C.4}$$

C.1.2 McNemar's Test

To determine whether the difference in performance of the two systems is statistically significant, McNemar's test is adopted [28]. Suppose first that we wish to test the null hypothesis $H_0^0 : E_A = E_B$ or alternatively $H_0^0 : d = E_A - E_B = 0$ against the alternative hypothesis $H_1^0 : E_A \neq E_B$. By substituting Equation (C.4) this leads to the test statistic

$$\begin{aligned} d &= \frac{C_{fr}P(H_0)}{n}(n_{10} - n_{01}) + \frac{C_{fa}P(H_1)}{m}(m_{10} - m_{01}), \\ &= \frac{C_{fr}P(H_0)}{n}d_{fr} + \frac{C_{fa}P(H_1)}{m}d_{fa}. \end{aligned} \tag{C.5}$$

Clearly, to determine whether the difference in performance of the two systems is statistically significant, it is sufficient to examine the cases where only one system made an error. Since H_0^0 is composite, i.e. it can be satisfied in more than one way, and since for the purposes here we are mainly interested in demonstrating that the differences in performance between the two systems *are* significant, consider testing two simpler hypotheses instead:

$$\begin{aligned} &H_0^{fr} : d_{fr} = 0 \quad \text{versus} \quad H_1^{fr} : d_{fr} \neq 0, \\ &H_0^{fa} : d_{fa} = 0 \quad \text{versus} \quad H_1^{fa} : d_{fa} \neq 0. \end{aligned} \tag{C.6}$$

Rejection of both H_0^{fr} and H_0^{fa}, together with the condition that both $d_{fr} < 0$ and $d_{fa} < 0$ or that both $d_{fr} > 0$ and $d_{fa} > 0$, would then be grounds for rejection of H_0^0. Conversely, acceptance of both H_0^{fr} and H_0^{fa} would be grounds for acceptance of H_0^0.

Consider for example testing H_0^{fr}. Assuming that both systems are equally likely to make errors, then under H_0^{fr}, it follows that $n_{10} = d_{fr} + n_{01}$ is a random variable drawn from the binomial distribution $\mathcal{B}(k_n, 1/2)$, where $k_n = n_{10} + n_{01}$. The null-hypothesis can thus be tested by applying a two-tailed test to the observation of the random variable n_{10} drawn from the $\mathcal{B}(k_n, \frac{1}{2})$ distribution where

$$\begin{aligned} P &= 2 \sum_{i=n_{10}}^{k_n} \binom{k_n}{i} \left(\frac{1}{2}\right)^{k_n} \quad \text{when } n_{10} > k_n/2, \\ P &= 2 \sum_{i=0}^{n_{10}} \binom{k_n}{i} \left(\frac{1}{2}\right)^{k_n} \quad \text{when } n_{10} < k_n/2. \end{aligned} \tag{C.7}$$

It follows that H_0^{fr} can be rejected if P is less than some significance level α.

C.1.3 Results

A proposed system was compared to a baseline system in Section 5.5.

Baseline system (B)

The baseline system consisted of static and 5-point dynamic features to which MS filtering was applied. The features were sampled at a rate of 100 Hz. ($(C, D_{1,5})$ in figure 5.7.)

Proposed system (A)

The proposed system consisted of static and 5-point dynamic features to which MS and lowpass filtering to 10 Hz was applied. The features were sampled at a rate of 25 Hz. ($(A', B'_{1,5})$ in figure 5.7.)

Table C.2 indicates the conditions for which the differences in optimum[1] performances between the proposed lowpass system (A) and a baseline system (B) are statistically significant at the level $\alpha = 0.02$. The plus symbol (+) in the Table indicates the cases where the proposed system performed significantly better than the baseline system. The minus symbol (-) indicates the cases where the proposed system performed significantly worse than the baseline system. The absence of a symbol indicates no significant difference in performance for the two systems under the hypotheses H_0^{fr} and H_0^{fa}. 30-second long segments of test speech from males and females were used.

C.1.4 Discussion

Based on the results in Table C.2 it can be concluded that the proposed system performs significantly better than the baseline system in the mismatched condition (DNDT), while no significant difference in performance exists in the matched condition (SN). This indicates that the proposed system improves robustness when there is acoustic mismatch.

[1] Performance was measured at the optimum decision threshold as determined on the data set.

Table C.2: Statistical significance at the $\alpha = 0.02$ level for the differences in performances between the proposed system (A) and baseline system (B).

LEGEND: TE = Test condition
TR = Training condition.
RR = Percent relative reduction in error for proposed system over baseline system
E_A = Error for proposed system A
E_B = Error for baseline system B

ERROR	TE	TR	d_{fr}	d_{fa}	% RR	E_A	E_B	k_n	k_m
EER %	DNDT	1S	+	+	13.0	17.1	19.7	93	1618
		2S	+	+	12.9	16.5	19.0	76	1577
		2F	+	+	11.3	16.8	18.9	78	1384
	DNST	1S			0.7	10.0	10.1	29	948
		2S		+	4.0	8.4	8.7	30	844
		2F		+	7.1	7.9	8.5	27	708
	SNST	1S		-	-4.5	7.1	6.8	47	714
		2S			-4.1	4.9	4.7	48	517
		2F			-0.8	4.0	4.0	26	431
MDE x 100	DNDT	1S	+	-	8.8	7.5	8.3	151	253
		2S	+	+	12.2	7.2	8.2	141	288
		2F	+	+	9.5	7.0	7.7	111	176
	DNST	1S		+	4.3	4.2	4.4	78	173
		2S		+	4.1	3.8	4.0	72	163
		2F		+	5.6	3.4	3.6	44	127
	SNST	1S	+	-	-1.5	2.8	2.7	121	120
		2S			1.5	2.1	2.2	112	114
		2F			2.2	1.8	1.8	85	86

C.2 Comparison

This section depicts verification results from the 1998 NIST Speaker Recognition Evaluation (NIST-SRE). The purpose of this depiction is simply to indicate that the proposed system performs competitively compared to other systems. This "state-of-art" performance suggests that the proposed processing improves performance at a realistic level.

Fig. C.1 depicts the minimum decision error (MDE) and the "hard" decision error (HDE) in the matched (SNST) test condition[2] for the two-session (2S) training condition using 30 second long segments of test speech. The MDE was defined in Section 4.2 as the minimum decision error while the HDE was defined as the decision error at a chosen decision threshold. The decision threshold used here for the proposed system was obtained on speech data from the 1997 NIST-SRE. Some of the differences in performance between the proposed system (a) and comparison systems (b,c) are related to the score

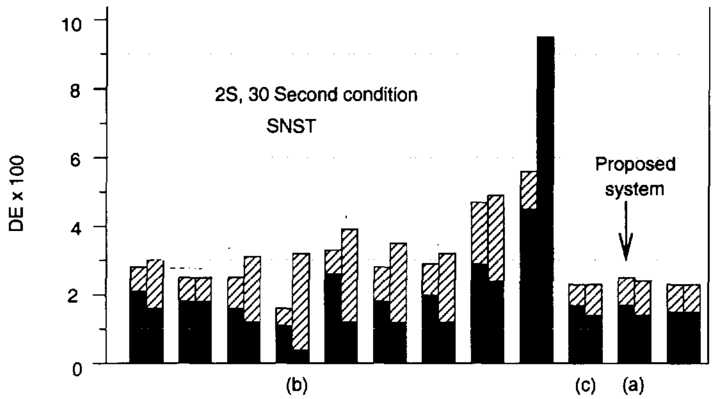


Figure C.1: MDE and HDE performance in the 1998 NIST Speaker Recognition Evaluation for the proposed system and various other systems. Legend: left-side bars show MDE, right-side bars show HDE, solid bars show proportion of DE due to false rejection errors, light bars show proportion of DE due to false acceptance errors. *Reproduced from 1998 NIST Speaker Recognition Evaluation Workshop Notes.*

[2] Results for the matched condition are reported here since that was the condition of interest in the evaluation.

normalization or features that were used. For example, scores for comparison system (b) in the figure were augmented with scores based on a dynamic time warping and subsequent comparison of phoneme-like segments. Scores for comparison system (c) were normalized with respect to handset condition and speaker.

Appendix D

Software Toolkit

Reproduced from: S. van Vuuren and H. Hermansky, "!MESS: A modular, efficient speaker verification system," in *RLA2C*, (Avignon, France), pp. 189–201, April 1998.

This appendix describes features and software aspects of a speaker verification toolkit called !MESS. This toolkit was used in the text-independent speaker verification experiments described in this thesis. !MESS was designed to be a highly modular, efficient speaker recognition system. The system incorporates an extension of the Tcl script language that facilitates rapid prototyping and high overall execution speed. This appendix provides an overview of the system and its implementation, detailing its applicability as a flexible research environment.

Systems for text-independent speaker verification tend to be resource intensive and seldomly run in real time [58]. A system that allows for rapid prototyping and efficient use of resources can greatly aid forays into new feature domains. As a departure point it was decided to design and implement a system that is modular and efficient while giving state-of-the-art performance on a well defined benchmark. This system has since been evaluated formally by NIST in their 1997 and 1998 Speaker Recognition Evaluations (NIST-SRE) with verification performance that was competitive with that of other groups. While the system incorporates existing knowledge on speaker verification, its implementation and applicability is likely to differ from other systems in a number of fundamental ways. While it is difficult to make direct comparisons to other systems – given that many of them are proprietary, the aim here is to highlight some possibly unique characteristics of the !MESS toolkit.

D.1 Modules

A typical modular approach to research systems is a set of precompiled routines, combined in a script and operating on files. Using the extendible scripting language Tcl [68], a somewhat different approach was followed, where the data stays in memory, precompiled routines are dynamically loaded only once and the control flow is scripted in more detail. This facilitates rapid prototyping and algorithmic development while maintaining a small memory footprint and high execution speed.

In a verification task the system has to decide, with some confidence level, to accept or reject the identity of somebody claiming to be the source of the speech signal. The system described here uses spectral and temporal based feature extraction, registration and normalization using a speaker independent mixture model and speaker dependent modeling using a mixture model obtained from Bayesian adaptation of the speaker independent model. Scoring uses a likelihood ratio test. Table D.1 lists the individual modules underlying the speaker verification system. Each module consists of an efficient, and where necessary, pipelined C-code library containing all the basic functionality. These libraries were written to be highly portable [99, 87]. Tcl is used to access functionality in a scripting environment, where each module (a Tcl package) provides specific capabilities such as, for example, mixture modeling. Collectively these modules form an environment in which one can with ease plug and play various components in order to design, debug and execute complex algorithms. The environment is described in [99, 87].

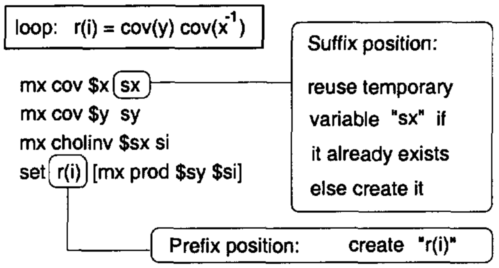
Table D.1: Modules in the speaker verification system.

Category	Module	Description
Math	Mx	General matrix algebra
Processing	Form	Feature analysis
	Seg	Speech-silence segmentation
Modeling	Lda	Data analysis and transformation
	Gvq	Vector quantization
	Gmm	Gaussian mixture modeling
Decoding	Det	Performance evaluation

Generic data objects can be automatically and efficiently saved to and loaded from disk. These objects are conventional C data structures represented at the script level with simple string identifiers.

D.1.1 Mx: Matrix Mathematics

The Mx [99] module provides a generic way to manipulate matrices. Memory management is handled transparently but is *controllable.* This means that while an algorithm is implemented conveniently in the script language, memory usage can still be tailored for maximal efficiency. Fig. D.1 gives an example of this important feature. Here x and y are data matrices and "cov" the covariance. Fast execution speed and a small memory footprint are possible if the temporary variables in the loop are forced to use memory efficiently. The suffix position for sx, sy, and si means that memory should be allocated for them only if they don't exist (the first pass of the loop). On subsequent passes their contents are simply overwritten. The prefix position for the variable $r(i)$ means that memory should not be reused. This example carries over to the other modules.



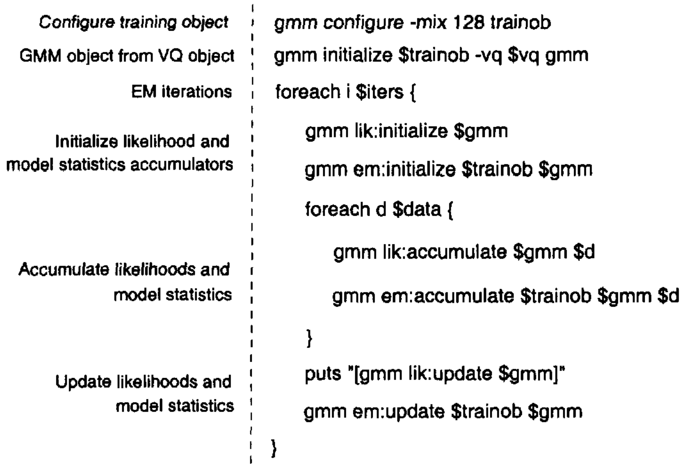
```
mx cov $x sx
mx cov $y sy
mx cholinv $sx si
set r(i) [mx prod $sy $si]
```

loop: r(i) = cov(y) cov(x⁻¹)
Suffix position:
reuse temporary variable "sx" if it already exists else create it
Prefix position: create "r(i)"

Figure D.1: Example of controllable memory usage.

In the script environment the user has full control of when to create, modify, update and destroy objects. Fig. D.2 shows how the Gmm module can be used to train a Gaussian mixture model using accumulators. It can also be used to retrain a Gaussian mixture model using Bayesian MAP training [25]. In the latter case confidence estimates of the model are obtained by cross-validation over the training data for that model[1].

[1] This is done efficiently during training by using separate accumulators for the model statistics for each cross-validation partition. Each permutation of the accumulators (excluding one each time) are combined and the model tested on the relevant training partition.



Configure training object
GMM object from VQ object
EM iterations
Initialize likelihood and model statistics accumulators
Accumulate likelihoods and model statistics
Update likelihoods and model statistics

```
gmm configure -mix 128 trainob
gmm initialize $trainob -vq $vq gmm
foreach i $iters {
    gmm lik:initialize $gmm
    gmm em:initialize $trainob $gmm
    foreach d $data {
        gmm lik:accumulate $gmm $d
        gmm em:accumulate $trainob $gmm $d
    }
    puts "[gmm lik:update $gmm]"
    gmm em:update $trainob $gmm
}
```

Figure D.2: Script for training a Gaussian mixture model.

D.1.2 **Form**: Feature Extraction

The feature extraction module duplicates the feature extraction steps described in Chapter 4. The first stage is a short-term analysis of the speech signal. Popular representations are supported, including filter bank spectrum, logarithmic spectrum, cepstrum [76], LPC [4], Perceptual Linear Prediction coefficients (PLP) [34], and Mel-frequency cepstral coefficients [16]. Bark [34], Mel [16] and linear frequency warpings are supported as are triangular [16] and trapezium shaped [34] frequency integration windows. Band-limiting of the output frequencies are supported. The output of this stage can be cached automatically if required. The second stage entails medium-term feature processing, such as RASTA processing [37], mean subtraction (CMS) [26], and FIR filtering [100]. The medium-term feature processing stage can be duplicated to produce multiple streams of features (such as static and dynamic features.)

D.1.3 **Seg**: Speech-Silence Segmentation

This module allows speech-silence detection and segmentation using an adaptive energy-based detector that is used to discard frames with original energies below the estimated noise floor in the signal.

D.1.4 Lda: Data Analysis and Feature Transformation

This module implements linear discriminant analysis (LDA) [23]. As a special case of LDA, feature vectors can be whitened using the statistics of the training data [25].

D.1.5 Gvq and Gmm: Modeling

The Gvq module implements vector quantization [13]. It is implemented using the LBG algorithm with iterative cluster splitting. The Gmm module implements Gaussian mixture modeling [25]. Training is performed with the Expectation Maximization (EM) algorithm [19] with Bayesian regularization for the covariance parameters [66]. The Gmm module also implements maximum aposteriori (MAP) training [45], allowing a GMM to be updated given new data.

This module allows training of speaker independent (SI) and speaker dependent (SD) models. An SI model is trained on a set of cohort speakers using the EM-algorithm. SD models are trained by MAP estimation of all the parameters in the SI model.

D.1.6 Gmm: Scoring

The SI model can be used to register the subspace of the SD model that pertains to the test utterance. This is done by identifying the N-best scoring components in the SI model for each utterance frame [80]. Target (SD) models are then evaluated using these components. Given the likelihood scores for the SD and SI models, a likelihood ratio test is performed over all the frames of the test utterance.

D.1.7 Det: Results Evaluation

This module allows evaluation measures to be computed for the speaker recognition system. This includes the computation of the identification error, confusion matrix, equal error rate (EER), and general cost functions as described in Chapter 4.

D.2 System Execution Time

The GMM system used in the 1997 NIST-SRE consisted of 256 mixture components. On a Pentium Pro II (266 Mhz) with 128 Megabyte memory running the Solaris operating system, the SD model training ran at 0.38 X R.T. (real time) and verification of a target at 0.11 X R.T. The process size was always under 10 Megabytes. The system needed a fraction of the execution time needed by systems from other groups where execution time was typically more than 1 x R.T. for comparable verification performance. This conforms to the design requirements that the system should be fast enough to facilitate rather exhaustive feature exploration trails.

Appendix E

Automatic Speech Recognition in a Time-feature Space

Reproduced from: S. van Vuuren and H. Hermansky, "Data-driven design of RASTA-like filters," in *Proc. EUROSPEECH'97*, (Rodos, Greece), pp. 409–412, 1997.

This appendix describes use of Linear Discriminant Analysis (LDA) for data-driven automatic design of RASTA-like filters. The LDA applied to rather long segments of time trajectories of critical-band energies yields FIR filters to be applied to these time trajectories in the feature extraction module. Frequency responses of the first three discriminant vectors are in principle consistent with the ad hoc designed RASTA, delta and double-delta filters. On a connected digit task the new features outperform the original RASTA processing.

E.1 Introduction

A typical automatic speech recognition (ASR) system contains a feature extraction module followed by a stochastic classifier. While the classifier is typically trained on training data, the feature extraction module is most often based on knowledge and beliefs. The knowledge applied in the feature extraction module has a critical role in the ASR process. Any information lost during the feature extraction is lost for the recognition process. On the other hand, the knowledge hardwired into the feature extraction module is the knowledge which does not have to be re-acquired from the data every time the recognizer is used for a new task.

In the late seventies, [42] proposed the use of Linear Discriminant Analysis (LDA) for deriving improved features for ASR. The LDA is applied to training data which contain sources of non-linguistic variability and the resulting transformation matrix is then a part of the feature extraction module which thus becomes more robust to the source of the particular non-linguistic variability.

This appendix presents a technique which applies LDA to rather long segments of a single time trajectory of critical band energy. Then, the LDA yields FIR filters to be applied to this time trajectory.

E.1.1 Temporal Domain and RASTA Technique

Acoustic feature vectors typically represent short-term characteristics of the speech signal. Standard HMM-based systems do classification over this short time span under the assumption of independence of the short-term acoustic vectors.

The peripheral human auditory system appears to be able to effectively integrate rather large time-spans (around 200 msec) of the audio signal [35]. Several emergent techniques employ short-term feature vectors from medium-span segments of speech. Among them, the RASTA technique [37] does bandpass filtering of time trajectories of speech features. To alleviate harmful effects of convolutional distortions, frequency components of time trajectories of logarithmic critical-band spectral energies below 1 Hz and above 13 Hz are attenuated. Such processing was found optimal by ASR experiments.

E.1.2 Toward a Data-Driven Design

The initial ad hoc form of the RASTA filters was optimized on a relatively small series of ASR experiments with noisy telephone digits. The optimizations using these ASR experiments are costly and there is no guarantee that the solutions obtained will not be specific to a given ASR problem. Therefore, data-based optimization which would avoid using a specific ASR paradigm is desirable.

The linear discriminant analysis (LDA) is a stochastic technique which optimizes linear discriminability between classes (see e.g. [42] for examples of LDA in ASR). The use of LDA for data-driven design of RASTA-like filters is examined next.

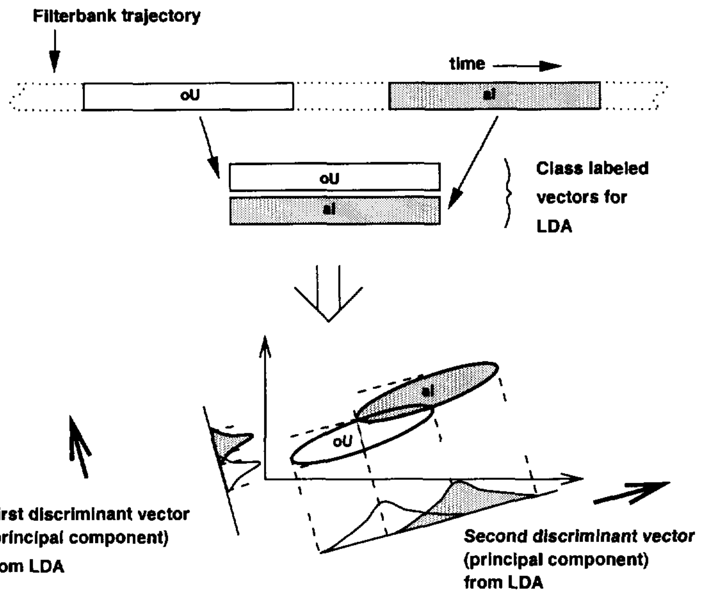

Filterbank trajectory
time
oU
aI
oU
aI
Class labeled vectors for LDA
aI
oU
First discriminant vector (principal component) from LDA
Second discriminant vector (principal component) from LDA

Figure E.1: Linear discriminant analysis on segments of the time trajectory of a single logarithmic critical-band energy.

E.2 Technique

Fig. E.1 shows the Linear Discriminant Analysis technique. The vector space for the LDA is constructed from segments of the time trajectory of a single logarithmic critical-band energy over a relatively long (typically about 1 s) span of time [6]. These segments are overlapped with a one frame spacing.

This approach is different from previous works using LDA [42, 12] since it applies LDA to *rather long time trajectories of features* rather than just to a single feature vector or to a relatively short block of feature vectors. This particular application of LDA results in the principal components (discriminant vectors) forming a set of FIR filters. The LDA on time-shifted segments of the trajectories therefore allows an FIR filtering interpretation of the analysis. This in turn allows us to directly relate the new LDA technique to other processing techniques such as RASTA processing. It should be noted, that up to certain constraints and assumptions[1], the LDA-based FIR filters map most efficiently

[1] It is assumed that the data are homoscedastic so that the within covariances of different classes are the same.

(with respect to the within-class and the across-class variability) the vector space onto several points of the output space.

E.3 Databases

As LDA tries to optimize class separability in the presence of unwanted variability the result depends crucially on the type of nonlinguistic variability present in the data, as well as on the set of classes in the analysis. Three different databases and two sets of classes are examined. First, the LDA is applied to a hand-labeled subset of the Switchboard database. This database is labeled according to standard conventions into a set of 56 American English phonemes. Additionally, this database also contained classes of between-word pauses, and utterance beginning and end silences[32]. Second, the Switchboard database is appended by the identical database but with an added simulated convolutional variability. This is achieved by adding a constant approximately representing 2 standard deviations of the data to each time trajectory. Finally, the English portion of the OGI multi-lingual database is used with a representative set of phoneme classes for the analysis. Essentially this set includes prevalent phonemes in the speech and excludes silence. While the class assignments were obtained from a hand-labeled continuous speech corpus, it should be noted that they may as well be obtained using automatic techniques such as forced alignment. Furthermore, as will be shown, the LDA-based filters need not be designed and used on the same data. It will be shown that even when the filters are designed on a database different from the one on which they are eventually used they can still outperform other processing methods.

E.4 Discriminant Vectors as Filters

In previous work [6] it was showed that the frequency response of the first discriminant vector agrees well with the frequency response of the ad hoc designed RASTA filter that smoothes the feature trajectory. It is important to stress the importance of this result. The discriminant vectors were designed entirely from the data without any intervention whereas the RASTA filter was iteratively optimized for on ASR experiments.

Figures E.2 through E.5 show frequency and impulse responses of the first three discriminant vectors derived on all three above described databases, as well as the frequency and impulse responses of the original RASTA filter and of the RASTA filter combined with the filters approximating the first (delta) and the second (double-delta) derivatives.

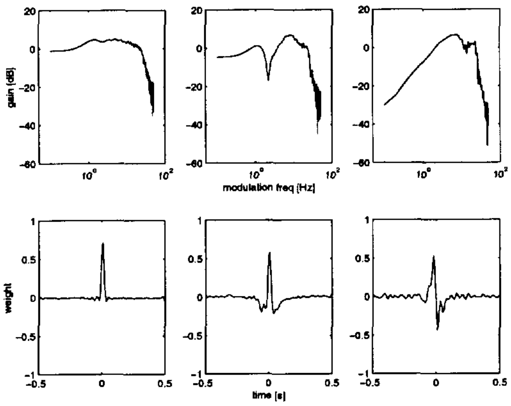


Figure E.2: Frequency and impulse responses of the first three discriminant vectors derived on the clean Switchboard database.

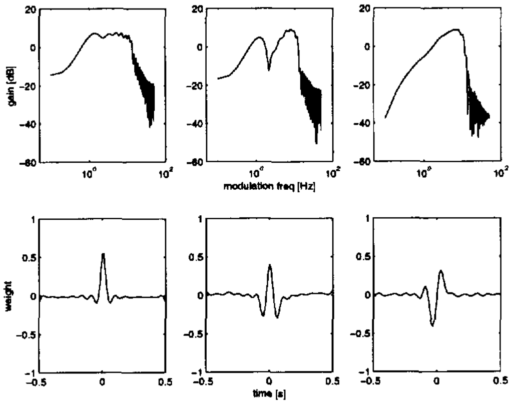


Figure E.3: Frequency and impulse responses of the first three discriminant vectors derived on the Switchboard database with additional steady-state variability.

The first thing to notice is that, as expected, filters designed on the Switchboard data with the additional steady-state variability exhibit stronger suppression of low as well as of high frequencies[2]. The stronger suppression of low frequencies is expected because the additional variability is steady-state.

Filters designed on the OGI multilingual database do contain similar general characteristics as the filters derived on the Switchboard data but differ in details (for example the second and the third filters are interchanged). Note that a reduced class set of only the 20 most common phoneme labels was used with the OGI database.

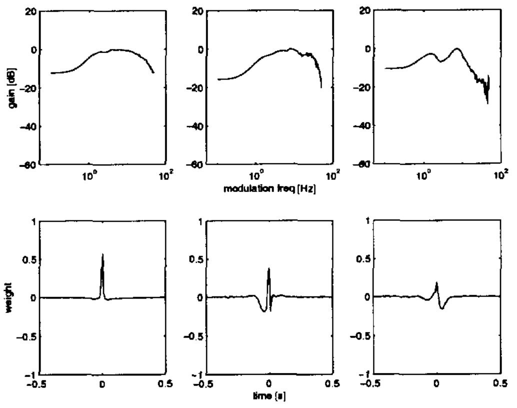


Figure E.4: Frequency and impulse responses of the first three discriminant vectors derived on the English portion of OGI multi-lingual database.

The similarity of the first discriminant vectors from all three databases with the original RASTA filter is noticeable. The impulse responses of the first discriminant vector is approximately symmetric, implying close to zero phase and supporting [17].

The first discriminant vector, while being the most important for discrimination, explains only about 80% of the variability in the data. It was therefore decided to investigate the second and third discriminant vectors as well. The frequency characteristic of the second and third discriminant vectors are somewhat comparable to the second (slope) and third (curvature) orthogonal polynomials approximating the time trajectory of the feature within a 9 frame (90 ms) time interval as proposed in [26]. The second peak at around 1 Hz in the two-peak filters can be simulated by adding a small bias to the double-delta

[2] For the Switchboard experiments, which used only 30 minutes of speech data, to guarantee numerical stability, it was necessary to enforce a condition number of 500 for the with-in covariance matrix. This conditioning caused a slight suppression for the high frequencies. This conditioning was not used for the OGI data.

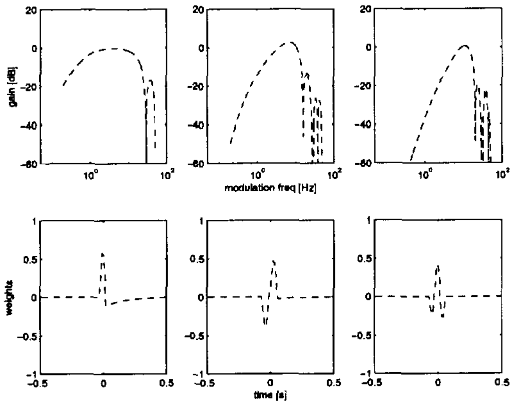


Figure E.5: Frequency and impulse responses of the RASTA filter and the RASTA filter combined with the delta and double-delta filters.

orthogonal polynomial.

As shown in Fig. E.6 which depicts frequency responses of the first discriminant vector at all 15 carrier frequencies (there are 15 critical-band filters covering the telephone-bandwidth), filters at different carrier frequencies are rather similar.

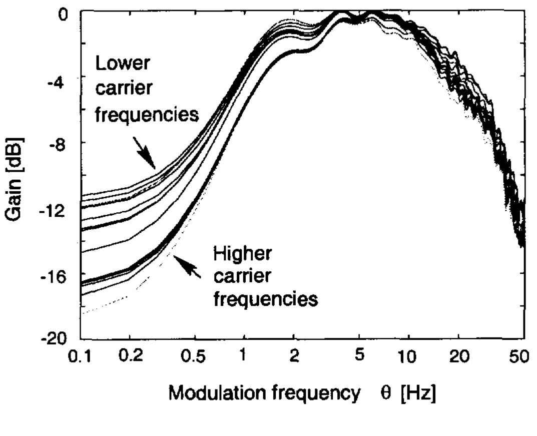


Figure E.6: Frequency response of the first discriminant vector at all 15 carrier frequencies derived on the English portion of OGI multi-lingual database.

To further highlight the modulation frequency selective nature of the LDA-based filters Fig. E.7 shows the frequency response of the resultant first discriminant vector for the case where the log filter bank energies had a disturbance added at modulation frequencies of 5 and 20 Hz with respective amplitudes about 2.5 and 0.5 times the average standard deviation in the log filter bank energies. Such a disturbance can be thought of as a time-varying convolutional disturbance on the speech signal. As expected, the filter attempts to attenuate modulation frequencies at the disturbance.

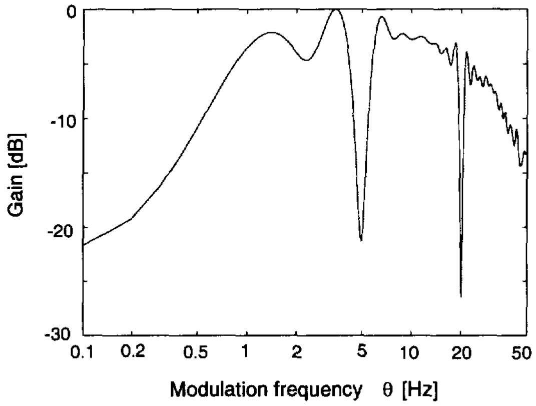


Figure E.7: Frequency response of the first discriminant vector for an artificial non-stationary channel disturbance.

E.5 ASR Results

For the results listed below the filters were derived on the English portion of the OGI multilingual database (OGI-TS). This is a database of almost 3 hours' continuous telephone speech with both speaker and handset variability. Recognition experiments were performed on 500 connected digit utterances from the OGI-Numbers corpus. This database also has speaker and handset variability. The phoneme classes used for deriving the filters were chosen to match the monophone classes expected in the recognition experiment. For the filter design a total of twenty monophones were used. Each monophone was weighted according to its natural frequency of occurrence in the OGI-TS database.

The results are competitive with current filtering schemes such as RASTA. Table E.1 lists the word level accuracy for the connected digit recognition task. A 5 state left-to-right HMM model was used with 3 mixtures per state. Twenty monophone models were trained and a simple single pronunciation grammar used. The baseline features (base) were critical-band log energies from a PLP analysis. The table lists accuracies for the

Table E.1: Percentage word level accuracies for a connected digit recognition task (OGI-Numbers corpus) for the various processing techniques.

base	83.7
rasta	88.0
lda1	91.8
lda2	85.1
lda3	81.7

base	+	delta	91.2
rasta	+	delta	92.4
lda1	+	delta	93.5
lda1	+	lda2	94.3

base	+	delta	+	ddelta	90.0
rasta	+	delta	+	ddelta	92.9
lda1	+	delta	+	ddelta	94.0
lda1	+	lda2	+	lda3	94.6

baseline features processed with RASTA filtering (rasta) and with combinations of the first three LDA-derived filters (lda1, lda2 and lda3). Accuracies for when delta (delta) and double delta (ddelta) features are added are also listed. Features were normalized throughout with a full whitening transform. This normalization was necessary to ensure a fair comparison between the different features for mainly two reasons. a) Decorrelation: The HMM model used diagonal covariances. b) Scale: The HMM model used a numerical floor (1e-4) on the variances parameters.

In practice it was found that the whitening transform gives results similar to the DCT transform. To mitigate effects from the language back-end of the system, and since it is known that different processing techniques exhibit different insertion and deletion trade-off [9], word level accuracies are reported at the optimum cross-word penalty.

The basic LDA derived feature (lda1) is seen to generally outperform the baseline and RASTA processed features. The differences are significant at the level $\alpha = 0.01$ using Mcnemar's test. These results suggest that while RASTA greatly aids performance on this database, other data-derived filters (here from LDA) may yield even better performance. This observation extends to the case where delta and double delta features are added.

Given that the LDA filters were derived from another database and based entirely on the baseline feature and class labels the results are highly encouraging.

E.6 Conclusion

A new temporal filtering technique was proposed to optimize class discriminability. The encouraging performance of the *entirely data-derived* filters in recognition experiments have since led to extensions of the technique forming the subject of ongoing research in our laboratory.

Biographical Note

Sarel van Vuuren was born on April 7, 1970, in Cape Town, South Africa. He grew up in Pretoria, South Africa, where in 1991 he received the Bachelor of Engineering degree with honors from the University of Pretoria, and in 1994 the Master of Engineering degree from the same university. Since 1994 he has been a graduate research assistant at the Oregon Graduate Institute of Science and Technology where he has been a member of the Center for Spoken Language Understanding (CSLU) and the Anthropic Signal Processing Group.

During his studies at OGI, he designed and implemented a complete automatic speaker verification system for telephone speech. This system received top honors in one of the categories of the 1998 official speaker recognition evaluation run by the National Institute for Standards and Technology.

During the summer of 1995, he authored rapid prototyping software for matrix mathematics and speech processing that is being distributed in part with the CSLU speech toolkit. During the summer of 1996 he worked as a student intern at the Cambridge Research Lab operated by Digital Equipment Corporation (now Compaq) in Boston, MA, where he worked on bio-metric technologies under supervision of Dr. William Goldenthal. The work led to an application for a United States patent. In the summer of 1997 he was invited to attend a summer research workshop at the Center for Language and Speech Processing, Johns Hopkins University, Baltimore where he explored new acoustic processing techniques for large vocabulary continuous speech recognition.

During his Ph.D. studies he co-authored scientific papers in several international conferences and prestigious journals – the latter currently in review. His research interests include pattern recognition, neural networks, and signal and image processing, with application in speech recognition, bio-metric authentication, bio-medical signal processing, computational finance and the implementation and optimization of real-time systems and tools.

Selected Publications

1 S. van Vuuren and H. Hermansky, "Relevancy of components of the modulation spectrum for speaker verification," *Speech Communication*, 1999. In review.

2 S. van Vuuren and H. Hermansky, "On the importance of components of the modulation spectrum for speaker verification," in *Proc. ICSLP*, (Sydney, Australia), Nov. 1998.

3 H. H. Yang, S. van Vuuren, and H. Hermansky, "Relevancy of time-frequency features for phonetic classification measured by mutual information," to appear in *ICASSP*, (Phoenix, AZ), 1999.

4 S. van Vuuren, "Mx: A package for rapid mathematical prototyping and algorithm development with application to speech and speaker recognition," Tech. Rep. CSE-98-003, Dept. Comp. Science and Eng., Oregon Graduate Institute, (Portland, OR), March 1998.

5 S. van Vuuren and H. Hermansky, "!MESS: A modular, efficient speaker verification system," in *Proc. Speaker Recognition and its Comme rcial and Forensic Applications*, (Avignon, France), pp. 198-201, April 1998.

6 S. van Vuuren and H. Hermansky, "SR in a Time-Feature Space," in *Proc. of the NIST Speaker Recognition Workshop*, (Baltimore, MD), March/April 1998.

7 W. Wei and S. van Vuuren, "Improved neural network training of inter-word context units for connected digit recognition", in *Proc. ICASSP*, (Seattle, WA), pp. 497-500, May 1998.

8 H. H. Yang, S. van Vuuren, and H. Hermansky, "Relevancy of time-frequency features for phonetic classification," *Speech Communication*, 1998. In review.

9 H. Hermansky, N. Malayath, S. Sharma, and S. van Vuuren, "Towards decreasing error rates in speaker recognition," *Speech Communication*, 1998. Invited paper for special issue on Speaker Recognition. In review.

10 S. van Vuuren and H. Hermansky, "Data-driven design of RASTA-like filters," in *Proc. EUROSPEECH*, (Rodos, Greece), pp. 409-412, 1997.

11 J. Schalkwyk, J. de Villiers, S. van Vuuren, and P. Vermeulen, "CSLUsh: an extendible research environment," in *Proc. EUROSPEECH*, (Rodos, Greece), pp. 698-701, 1997.

12 S. van Vuuren, T. Kamm, J. Luettin, and H. Hermansky, "Data-driven design of speech analysis module," in *1997 CLSP/JHU Summer Workshop on Innovative Techniques for Large Vocabulary Conversational Speech Recognition*, (CLSP, Johns Hopkins University, Baltimore, MD), Aug. 1997.

13 C. Avendano, S. van Vuuren, and H. Hermansky, "Data based filter design for RASTA-like channel normalization in ASR," in *Proc. ICSLP*, (Philadephia, PA), pp. 2087-2090, Oct. 1996.

14 S. van Vuuren, "Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch," in *Proc. ICSLP*, (Philadelphia, PA), pp. 784-1787, Oct. 1996.