Novel Speech Processing Techniques for Robust Speech Recognition

THÈSE N^{o} 3637 (2006)

PRESENTÉE Á LA FACULTÉ INFORMATIQUE ET COMMUNICATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Vivek Tyagi

Bachelor of Technolgy in Electrical Engineering,
Indian Institute of Technology, Kanpur, India

and

Graduate School in Computer Science,
EPFL, Lausanne, Switzerland
et de nationalité Indienne

Acceptée sur proposition du jury:

Prof. Christian Wellekens, Directeur de thèse, Eurecom, France.
Prof. Emre Telatar, President de jury, EPFL, Switzerland.
Prof. Hervé Bourlard, Rapporteur Idiap Research Institute, EPFL, Switzerland.
Dr. Denis Jouvet, Rapporteur, France Telecom Research, France.
Prof. Torbjorn Svendsen, Rapporteur, Norwegian University of Science and Technology, Trondheim, Norway.
Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.
8th September 2006

Abstract

The goal of this thesis is to develop and design new feature representations that can improve the automatic speech recognition (ASR) performance in clean as well noisy conditions. One of the main shortcomings of the fixed scale (typically 20-30 ms long analysis windows) envelope based feature such as MFCC, is their poor handling of the non-stationarity of the underlying signal. In this thesis, a novel stationarity-synchronous speech spectral analysis technique has been proposed that sequentially detects the largest quasi-stationary segments in the speech signal (typically of variable lengths varying from 20-60 ms), followed by their spectral analysis. In contrast to a fixed scale analysis technique, the proposed technique provides better time and frequency resolution, thus leading to improved ASR performance. Moving a step forward, this thesis then outlines the development of theoretically consistent amplitude modulation and frequency modulation (AM-FM) techniques for a broad band signal such as speech. AM-FM signals have been well defined and studied in the context of communications systems. Borrowing upon these ideas, several researchers have applied AM-FM modeling for speech signals with mixed results. These techniques have varied in their definition and consequently the demodulation methods used therein. In this thesis, we carefully define AM and FM signals in the context of ASR. We show that for a theoretically meaningful estimation of the AM signals, it is important to constrain the companion FM signal to be narrow-band. Due to the Hilbert relationships, the AM signal induces a component in the FM signal which is fully determinable from the AM signal and hence forms the redundant information. We present a novel homomorphic filtering technique to extract the leftover FM signal after suppressing the redundant part of the FM signal. The estimated AM message signals are then down-sampled and their lower DCT coefficients are retained as speech features. We show that this representation is, in fact, the exact dual of the real cepstrum and hence, is referred to as fepstrum. While Fepstrum provides amplitude modulations (AM) occurring within a single frame size of 100ms, the MFCC feature provides static energy in the Mel-bands of each frame and its variation across several frames (the deltas). Together these two features complement each other and the ASR experiments (hidden Markov model and Gaussian mixture model (HMM-GMM) based) indicate that Fepstrum feature in conjunction with MFCC feature achieve significant ASR improvement when evaluated over several speech databases.

The second half of this thesis deals with the noise robust feature extraction techniques. We have designed an adaptive least squares filter (LeSF) that enhances a speech signal corrupted by broad band noise that can be non-stationary. This technique exploits the fact that the autocorrelation coefficients of a broad-band noise decay much more rapidly with increasing time lag as compared to those of the speech signal. This is especially true for voiced speech as it consists of several sinusoids at the multiples of the fundamental frequency. Hence the autocorrelation coefficients of the voiced speech are themselves periodic with period equal to the pitch period. On the other hand, the autocorrelation coefficients of a broad band noise are rapidly decaying with increasing time lag. Therefore, a high order (typically 100 tap) least square filter that has been designed to predict a noisy speech signal (speech + additive broad band noise) will predict more of the clean speech components than the broad band noise. This has been analytically proved in this thesis and

we have derived analytic expressions for the noise rejection achieved by such a least squares filter. This enhancement technique has led to significant ASR accuracy in the presence of real life noises such as factory noise and aircraft cockpit noise.

Finally, the last two chapters of this thesis deal with feature level noise robustness technique. Unlike the least squares filtering that enhances the speech signal itself (in the time domain), the feature level noise robustness techniques as such do not enhance the speech signal but rather boosts the noise-robustness of the speech features that usually are non-linear functions of the speech signal's power spectrum.

The techniques investigated in this thesis provided a significant improvement in the ASR performance for the clean as well noisy acoustic conditions.

Keywords: Robust speech recognition, speech enhancement, speech processing, feature extraction, stationary analysis, amplitude modulation, frequency modulation, fepstrum, least squares filtering, adaptive filtering.

Version abrégée

Le but de cette thèse est de développer et concevoir de nouvelles représentations caractéristiques qui peuvent améliorer la performance de la reconnaissance automatique de la parole (ASR) avec ou sans conditions de bruit. Un des principaux défauts d'une enveloppe de taille fixe (typiquement une fenêtre d'analyse de 20 à 30 ms) comme MFCC, est le traitement insuffisant du signal implicite non stationnaire. Dans cette thèse, une nouvelle technique d'analyse spectrale stationnaire et synchrone de la parole est proposée, elle détecte séquentiellement les plus larges segments quasi-stationnaires dans le signal de la parole (typiquement avec des longueurs variant entre 20 et 60 ms), suivie de leur analyse spectrale. Contrairement à une technique d'analyse avec une échelle fixe, celle proposée apporte une meilleure résolution de temps et de fréquence permettant d'améliorer les performances de l'ASR. Cette thèse souligne le développement des techniques logiques et théoriques de modulations d'amplitude et de fréquence (AM-FM) pour un signal à bandes larges comme la parole. Les signaux AM et FM ont été définis et étudiés dans le contexte des systèmes de communications. S'appuyant sur ces idées, de nombreux chercheurs ont appliqué les modèles AM et FM pour les signaux de la parole avec des résultats mitigés. Ces techniques ont evolué ainsi que les méthodes de démodulation. Dans cette thèse, nous définissons soigneusement les signaux AM et FM dans le contexte ASR. Nous montrons que pour une estimation théorique significative des signaux AM, il est important de contraindre les signaux FM à d'étroites bandes. Avec les relations d'Hilbert, le signal AM crée un composant dans le signal FM qui est déterminable à partir du signal AM et qui forme une information redondante. Nous présentons une nouvelle technique filtrante homomorphique pour extraire le surplus du signal FM après avoir supprimé la partie redondante du signal FM. Les signaux AM estimés sont de basses amplitudes et leurs bas coefficients DCT sont retenus comme ces caractéristiques de la parole. Nous montrons que cette représentation est le dual exact du "cepstrum" réel et est noté "fepstrum". Tandis que Festrum fournit des modulations d'amplitude (AM) se produisant dans une seule fenêtre de 100 ms, la caractéristique MFCC fournit une énergie statique dans les "Mel-bands" de chaque fenêtre et des variations à travers plusieurs fenêtres. Ces deux caractéristiques se complémentent et les expériences ASR indiquent que la caractéristique Fesptrum en conjonction avec celle MFCC accomplit des améliorations ASR significatives en évaluant d'autres bases de données de la parole.

Le second point de cette thèse traite des techniques d'extraction des caractéristiques robustes au bruit. Nous avons élaboré un filtre adaptatif de moindres carrés qui améliore le signal de la parole corrompu par une large bande de bruit pouvant être non stationnaire. Cette technique exploite le fait que le coefficient d'autocorrélation d'une large bande de bruit baisse rapidement avec un temps de retard croissant comparé à celui du signal de la parole. Ceci est vrai pour la voix car elle consiste en plusieurs sinusoïdes de fréquences fondamentales. Les coefficients d'autocorrélation de la voix sont eux-mêmes périodiques. Ceux d'une large bande de bruit décroient rapidement avec l'augmentation du temps de retard. Donc un filtre carré qui est concu pour predire un signal de la parole bruité predira plus des composants de la parole non bruités qu'un bruit de large bande. Ceci est analytiquement prouvé dans cette thèse et nous avons dérivé les expressions analytiques pour

le rejet du bruit par de tels filtres carrés. Cette amélioration technique a conduit a une précision significante de l'ASR en présence de bruit naturel comme le bruit d'usine ou celui d'un cockpit d'avion.

Finalement, les deux derniers chapitres de cette thèse traitent des techniques robustes des caractéristiques de niveau de bruit.

Les techniques abordées dans cette thèse apportent une amélioration significative dans la performance de ASR dans des conditions acoustiques bruitées ou non.

Mots clés : Reconnaissance robuste de la parole, débruitage de la parole, extraction de traits, analyse de la stationarité, modulation d'amplitude, modulation de fréquence, fepstrum, filtrage au sens des moindres carrés, filtrage adaptatif

Contents

List of Figures

List of Tables

Acknowledgments

I would first like to thank my thesis advisor Prof. Christian J. Wellekens who has been highly supportive during my stay at Institute Eurecom. His sincerity, integrity, intelligence, and latitude were extremely valuable. The other members of my committee included Prof. Hervé Bourlard, Dr. Denis Jouvet, Prof. Torbjorn Svendsen and Prof. Emre Telatar. I would like to thank all of them for being on my committee. I would also like to especially thank Hervé who introduced me to speech research and mentored me as a young researcher.

This thesis is dedicated to my mother, my sister and my father who have constantly been the sources of inspitation and immense support to me. I will also like to extend my gratitude and respect for the Indian Institute of Technology, Kanpur which as an institution has left an indelible mark on my academic as well as personal life.

I will also like to thank Institute Eurecom and IDIAP Research Institute for hosting me as a doctoral student at the various stages of my doctoral research. Finally, this work was supported by the European Commission's 6th Framework Program project DIVINES under the contract number FP6-002034. I duly acknowledge their institutional and financial support.

Chapter 1

Thesis Overview

1.1 Introduction

Automatic Speech Recognition (ASR) systems have come a long way since the first systems appeared in early 1970s. Much of the advances in speech signal processing were already made in early 1970s when the linear prediction technique was invented by Atal (Atal and Hanauer, 1971) and several other researchers. With the advent of statistical techniques in the ASR systems in the form of the hidden Markov Models (Jelinek et al., 1975; Bahl et al., 1983; Rabiner and Juang, 1986) and artificial neural networks (ANNs) (Bourlard and Wellekens, 1990; Bourlard and Morgan, 1994) in the 1990s, the performance of the ASR systems improved dramatically from the limited vocabulary isolated word ASR systems to large vocabulary continuous speech ASR systems. State-of-the-art ASR systems continue to improve, yet there remain many tasks for which the technology is inadequate. Under good conditions human phone error rate for nonsense syllables has been estimated to be as low as 1.5% (Allen, 1994), as compared with rates that are an order of magnitude higher for the best machine phone recognizers (Deng and Sun, 1994). This is partly due to the fact that apart from the message content, the speech signal also carries variabilities such as speaker characteristics, emotions and background noise. Moreover, speech is a highly non-stationary signal and the speech signal processing techniques that were developed over the past few decades have made simplifying assumptions regarding the signal stationarity over small time periods (20ms). This is a simple engineering approximation and it has served the ASR technology well over the past few decades when the computational power was limited. However, this assumption does not take into account the highly non-stationary nature of certain speech segments (eg. stops, bursts and plosives). The core acoustic operation has essentially remained the same for decades: a single *feature vector* (derived from the power spectral envelope over a 20-30 ms window, stepped forward by 10 ms per frame) is compared to a set of *probability distributions* derived from a training data for an inventory of subword units (usually monophones or triphones). While many systems also incorporate time derivatives and/or projections from five or more frames to a lower dimension, the fundamental character of the acoustic features has remained quite similar. To better extract the information from the speech signal one needs to develop speech processing techniques that can better describe the inherent non-stationary structure in the speech signal.

The above mentioned limitations coupled with the problem of noisy environments makes the ASR systems particularly vulnerable to variabilities such as acoustic environment mismatch (noise) and more inherent variabilities such as speaking rate, speaker accent and the speaker's emotional state. To develop generally applicable and useful recognition techniques, we must overcome the limitations of the current acoustic processing techniques. Interestingly, even human phonetic cate-

gorization is poor for extremely short segments (e.g., $<$ 100ms), suggesting that analysis of longer time is somehow important for phonemic recognition.

1.2 Objective of the thesis

One of the principal objectives of this thesis is to study the shortcomings of the current fixed scale (20-30ms), spectral envelope based features such as MFCC (Davis and Mermelstein, 1980) and PLP (Hermansky, 1990) and to propose solutions to overcome these deficiencies. In particular, we will revisit the following assumptions in this thesis:

- The speech signal can be assumed to be stationary for the time durations of 20-30ms.
- A discrete Fourier transform (DFT) of these fixed sized (20-30ms) speech segments, followed by Mel-filtering and its cepstrum estimation (MFCC feature) (Davis and Mermelstein, 1980) captures all the pertinent information for the ASR applications.

In the course of this thesis, we have shown that the speech feature extraction techniques that can describe the non-stationarity inherent in the speech signal can significantly improve the ASR performance even in the clean conditions over the state-of-the art features such as MFCCs[1]. The second aim of this thesis is to improve the robustness of the ASR systems by developing speech processing techniques that lead to features that are noise robust and hence ensure satisfactory ASR performance in noisy conditions as well. As will be shown later on in the course of this thesis, feature extraction for clean and noisy acoustic conditions are two practically different endeavors. While in the noise robust feature design, one can use the a-priori information about the noise signal such as different bandwidths of the noise and the clean speech signals (Bourlard et al., 1997)[2], perform Wiener filtering if the noise reference signal is available (Kay, 1998), or perform modulation spectrum filtering (RASTA filtering) if the noise and the clean speech's modulation spectra do not overlap significantly (Hermansky and Morgan, 1994a). Whereas, in the case of designing features for the clean speech which can improve the ASR performance beyond that of the MFCCs, one plausible scheme could be to bring in complementary sources of information that is lacking in the MFCC features themselves.

1.3 Motivation for the present work

1.3.1 Stationarity-synchronous spectral analysis

In the current state-of-the art ASR systems, MFCCs or PLPs are acoustic features of choice as they yield the best performance in clean conditions as opposed to the other features such as RASTA-PLP (Hermansky and Morgan, 1994a), or the spectral centroid features (Paliwal, 1998) that are designed to be noise robust and hence have better performance in noisy conditions than the MFCC or PLP but are worse in clean conditions. However, the phone recognition accuracies obtained through the use of the MFCC features, even in the clean conditions, are an order of magnitude higher than that achieved by humans (Deng and Sun, 1994). Therefore, there is a significant interest in studying the deficiencies of the spectral envelope based features, such as MFCC, to further improve the ASR accuracies even in the clean conditions. Perhaps one of the most commonly known deficiency of MFCC features is that they are not invariant to the frequency warping induced due to the varying vocal

[1] We note that despite the simple assumptions involved in the computations of the MFCCs, it still remains one of the best features for ASR in clean acoustic conditions (roughly SNR of 18dB and above)

[2] This is exploited in the multi-stream paradigm where the evidence from the noisy bands are either rejected or weighted down as compared to those from the uncorrupted speech bands.

tract lengths of the speakers. This problem has been addressed by maximum likelihood (ML) warping of the power spectrum that is, in turn, used to compute the MFCC features. It has been shown that vocal tract length normalization (VTLN) provides roughly 5%-10% relative improvement over the unnormalized MFCCs.

However, one of the major limitations of the MFCC feature vectors is that they handle the inherent non-stationarity of the underlying speech signal in a very ad-hoc way that usually results in the smearing of the time and frequency resolutions of the underlying speech signal and hence may be detrimental to the ASR performance. Mel-Frequency Cepstral Coefficient (MFCC) (Davis and Mermelstein, 1980) or Perceptual Linear Prediction (PLP) (Hermansky, 1990), are based on some sort of representation of the smoothed spectral envelope, usually estimated over fixed analysis windows of typically 20ms to 30ms of the speech signal (Davis and Mermelstein, 1980; Hermansky, 1990). Such analysis is based on the assumption that the speech signal can be assumed to be quasi-stationary over these segment durations. Typically, the sustained-stationary segments in a vowel can typically last from 30 to 80ms, while stops are time-limited by less than 20ms (Rabiner and Juang, 1993). Therefore, it implies that the spectral analysis based on a fixed size window of 20ms or 30ms has some limitations, including:

- The frequency resolution obtained for quasi-stationary segments (QSS) longer than 20 or 30ms is quite low compared to what could be obtained using larger analysis window. Although, most of the frequency resolution is lost due to averaging by 24 Mel filters. However, power spectrum estimation (DFT) over quasi-stationary segments will still lead to low-variance Mel-filter bank energies as compared to those obtained with a fixed scale spectral analysis that does not take quasi-stationarity into account.

- In certain cases, the analysis window can span the transition between two QSSs, thus blurring the spectral properties of the QSSs, as well as of the transitions. Indeed, in theory, Power Spectral Density (PSD) cannot even be defined for such non stationary segments (Haykin, 1993). Furthermore, on a more practical note, the feature vectors extracted from such transition segments do not belong to a single unique (stationary) class and may lead to poor discrimination in a pattern recognition problem.

To overcome these limitations there is a need to first detect the quasi-stationary segments in the speech signal which typically are of variable lengths. This is followed by the spectral analysis of these variable sized segments and the MFCC computation thereof. Such an approach can provide better time and frequency resolution as compared to the fixed scale spectral analysis, thereby, improving the ASR performance.

1.3.2 Modulation Spectral Analysis

A central result from the study of the human speech perception is the importance of slow changes in speech spectrum for speech intelligibility (Dudley, 1939). A second key to human speech recognition is the integration of phonetic information over relatively long intervals of time. Speech is a dynamic acoustic signal with many sources of variation. As noted by Furui (Furui, 1986, 1990), spectral changes are a major cue in phonetic discrimination. Therefore, it is desirable to incorporate the amplitude modulations (that account for the spectral dynamics) in the speech signal, as a feature vector for ASR. In past several years, significant efforts have been made to develop new speech signal representations which can better describe the non-stationarity (spectral dynamics) inherent in the speech signal. Some representative examples are temporal patterns (TRAPS) features (Hermansky, 2003; Athineos et al., 2004) and the several modulation spectrum related techniques (Tyagi et al., 2003; Athineos et al., 2004; Zhu and Alwan, 2000; Kingsbury et al., 1998a; Kanedera et al., 1998). In the past (Tyagi et al., 2003; Athineos et al., 2004; Zhu and Alwan, 2000; Kingsbury et al.,

1998a; Kanedera et al., 1998), the modulation spectrum that has been used as a feature vector for ASR has been defined and extracted in a slightly ad-hoc manner. For instance, several researchers have extracted the speech modulation spectrum by computing a discrete Fourier transform (DFT) of the Mel or critical band spectral energy trajectories, where each sample of the trajectory has been obtained through a power spectrum (followed by Mel filtering) over 20-30ms long windows. An illustration of this is provided in Fig.1.1. The major limitations of these techniques are that:

- They implicitly assumes that within each Mel or critical band, the amplitude modulation (AM) signal remains constant within the duration of the window length that is typically 20-30ms long.
- Instead of modeling the constantly and slowly changing amplitude modulation signal in each band, they mostly models the spurious and abrupt modulation frequency changes that occur due to the frame shifting of $10ms$.

Therefore, to be able to realize the true potential of the modulation spectrum as a feature vector for the ASR applications, it is important to develop and design theoretically consistent AM demodulation technique that is suitable for the speech signals.

1.4 Contributions of the thesis

This thesis focuses on addressing certain deficiencies of the existing spectral envelope based feature extraction techniques to improve the ASR performance in clean conditions. Moreover, it is also important to design new adaptive feature extraction techniques to achieve satisfactory noise robust ASR performance. The contributions of this thesis in the form of new feature extraction algorithms for clean as well noisy acoustic conditions are listed below.

1.4.1 Features for clean speech

1. **Variable scale piece-wise quasi-stationary analysis of speech signals:** It is often acknowledged that speech signals contain short-term and long-term temporal properties (Rabiner and Juang, 1993) that are difficult to capture and model by using the usual fixed scale (typically 20ms) short time spectral analysis used in hidden Markov models (HMMs), based on piecewise stationarity and state conditional independence assumptions of acoustic vectors. For example, vowels are typically quasi-stationary over 40-80ms segments, while plosive typically require analysis below 20ms segments. Thus, fixed scale analysis is clearly sub-optimal for "optimal" time-frequency resolution and modeling of different stationary phones found in the speech signal. In this work, we have studied the potential advantages of using variable size analysis windows towards improving state-of-the-art speech recognition systems in clean acoustic conditions. Based on the usual assumption that the speech signal can be modeled by a time-varying autoregressive (AR) Gaussian process, we estimate the largest piecewise quasi-stationary speech segments, based on the likelihood that a segment was generated by the same AR process. This likelihood is estimated from the Linear Prediction (LP) residual error. Each of these quasi-stationary segments is then used as an analysis window from which spectral features are extracted. Such an approach thus results in a variable scale time spectral analysis, adaptively estimating the largest possible analysis window size such that the signal remains quasi-stationary, thus the best temporal/frequency resolution tradeoff.
2. **Fepstrum representation of speech:** In this work, a theoretically consistent amplitude modulation (AM) signal analysis technique has been developed as compared to the previous

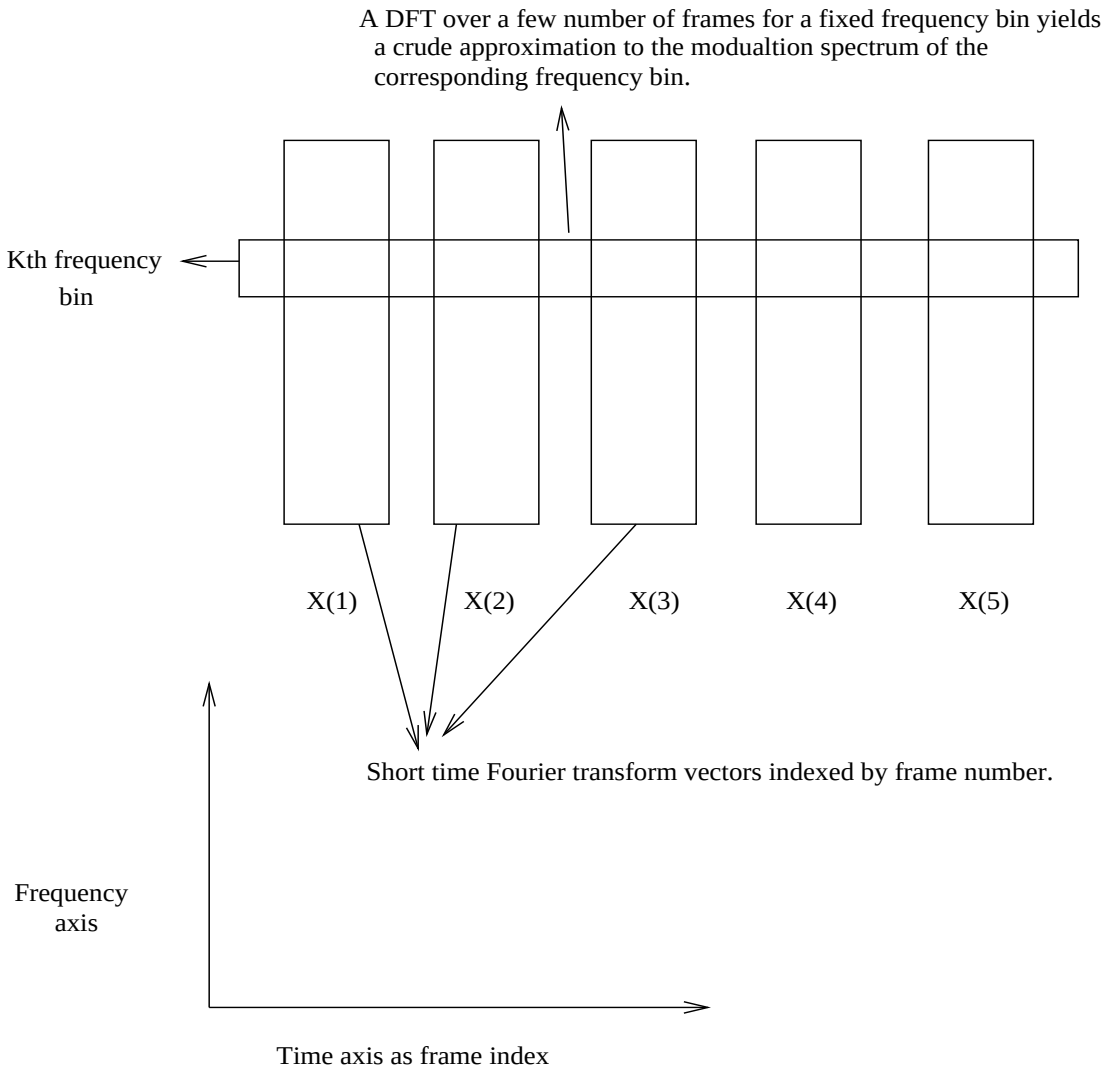

Figure 1.1. An approximate AM demodulation technique.

ones (Tyagi et al., 2003; Athineos et al., 2004; Zhu and Alwan, 2000; Kingsbury et al., 1998a; Kanedera et al., 1998). We have shown that a "meaningful" AM signal estimation is possible only if we decompose the speech analytic signal using several narrow-band filters which results in narrow-band carrier signals. Secondly, we use the lower modulation frequency spectrum of the downsampled AM signal, as a feature vector (termed FEPSTRUM as it is shown to be an exact dual of the well known quantity, cepstrum). While Fepstrum provides amplitude modulations (AM) occurring within a single speech frame of size 80-100 ms, the MFCC provides a description of static energy in each of the Mel-bands of each frame and its variation across several frames (through the use of the delta and double delta feature). The Fepstrum provides complementary information to the MFCC features and we show that a combination of the two features provides a significant ASR accuracy improvement in clean conditions over several speech databases.

1.4.2 Speech signal enhancement and noise robust features

1. **Least Square filtering of speech signal:** We have developed an adaptive filtering techniques that enhances a speech signal that is corrupted by additive broad-band noise. We have analyzed the behavior of the least squares filter (LeSF) operating on noisy speech signal. Speech analysis is performed on a block by block basis and a LeSF filter is designed for each block of signal, using a computationally efficient algorithm. Unlike the classical spectral subtraction and Wiener filtering techniques that require the noise to be stationary, the proposed LeSF technique makes no such assumption as this technique works on a block by block basis. Moreover, the proposed techniques does not require a reference noise signal as is required in Wiener filtering. This renders this technique as a highly practical signal enhancement technique as usually in the realistic scenarios, the reference noise channel is not available.

   The key contribution in the approach proposed in this work is that we relax the assumption of the input signal being stationary. The method of least squares may be viewed as an alternative to Wiener filter theory (Haykin, 1993). Wiener filters are derived from *ensemble averages* and they require good estimates of the clean signal power spectral density (PSD) as well as the noise PSD. Consequently, one filter (optimum in a probabilistic sense) is obtained for all realizations of the operational environment, assumed to be wide-sense stationary. On the other hand, the method of least squares is *deterministic* in approach. Specifically, it involves the use of time averages over a block of data, with the result that the filter depends on the number of samples used in the computation. Moreover, the method of least squares does not require the noise PSD estimate. Therefore the input signal is blocked into frames and we analyze a L-weight least squares filter (LeSF), estimated on each frame which consists of N samples of the input signal. In this work, analytic expressions for the weights and the output of the LeSF are derived as a function of the block length and the signal SNR computed over the corresponding block.

   Unlike other feature level noise robustness technique, the LeSF filter enhances the signal waveform itself and a MFCC feature computed over this enhanced signal leads to a significant improvement in the noisy speech recognition accuracies as compared to the other competing feature level noise robustness techniques such as RASTA-PLP and spectral subtraction. In distributed speech recognition (DSR) in the context of mobile telephony and voice-over IP systems, it may be desirable not only to have noise robust feature extraction algorithm but also to enhance the noisy speech signal for the human listener. Therefore, a signal enhancement technique that also leads to noise robust ASR is desirable. The proposed LeSF filtering technique falls into this category as it not only enhances the signal, a simple MFCC feature computed over this enhanced signal leads to significant ASR accuracy improvements in several realistic noisy conditions.

2. **Cepstral desensitization to noise and its modulations as a noise robust feature** As is well known, in the presence of commonly encountered additive noise levels, the formants are less affected as compared to the spectral "valleys" which exhibit spurious ripples. The DCT of a log Mel-filter bank spectrum (logMelFBS) which is commonly known as MFCC (Davis and Mermelstein, 1980) feature vector, is sensitive to ripples in the spectral valleys which, otherwise, do not characterize the speech sounds. This is one of the reasons for the poor performance of MFCC features in additive noisy conditions. Observing that the higher amplitude portions (such as formants) of a spectrum are relatively less affected by noise, Paliwal proposed spectral subband centroids (SSC) as features (Paliwal, 1998; Chen et al., 2004). In this work, we analytically show that exponentiating the logMelFBS can decrease the sensitivity of the cepstra to the spurious perturbations in the logMelFBS valleys as compared to the peaks.

   Lim has proposed the use of spectral root homomorphic deconvolution system (SRDS) (Lim,

1979) as an approximately more general case of logarithmic homomorphic deconvolution system (LHDS) (Oppenheim and Schafer, 1989). SRDS uses a root compression $(.)^{\gamma}$, $\gamma < 1$ of the mel-filter bank energies instead of the logarithmic compression used by LHDS. Although, Lockwood et. al (Alexandre and Lockwood, 1993) and Tokuda et. al (Tokuda et al., 1994) have proposed a unified approach to root Mel-cepstral coefficients (RMFCC), many researchers have used RMFCC with a motivation based on auditory and perceptual data. However, in this work, we use LHDS based MFCC features (Davis and Mermelstein, 1980). We provide a signal processing reason for the high sensitivity of the MFCC features towards additive noise and propose a solution to alleviate this problem by exponentiating the logMelFBS by a suitable positive power greater than unity. In (Tyagi et al., 2003), we proposed the use of Mel-cepstrum modulation spectrum (MCMS) features for robust ASR. MCMS features (Tyagi et al., 2003) are obtained by filtering cepstral trajectories using a bank of band-pass filters in the range $[2, 20]Hz$. In this work we derive MCMS features from the cepstra of the exponentiated logMelFBS. The experimental results show that these two sequential processing techniques improve the recognition rate in presence of additive non-stationary noise as compared to the MFCC and RASTA-PLP feature vectors.

1.5 Organization of the thesis

This thesis has been organized into eight chapters. The next chapter (Chapter 2) provides an overview of the current state-of-the-art and describes various state-of-the-art feature extraction techniques that are widely used for speech recognition in the clean and the noisy acoustic conditions. In this chapter we explain different components of the widely used HMM-GMM (hidden Markov model and Gaussian mixture model) ASR system. Subsequently, we provide description of the databases that have been used to evaluate the performance of the feature extraction techniques that have been proposed in this thesis.

In Chapter 3, we address the problem of the variable scale piece-wise quasi-stationary spectral analysis of speech signals. We show that we can overcome the time-frequency resolution limitations of a fixed scale (typically 20-30ms) spectral analysis of the speech signals that usually consist of piece-wise quasi-stationary segments of variable durations (3-80 ms).

Chapter 4 describes the modulation spectral analysis of the speech signal. Amplitude Modulation(AM) and frequency modulation(FM) have been well defined and studied in the context of communications systems. Borrowing upon these ideas, several researchers have applied AM-FM modeling for speech signals with mixed results. These techniques have varied in their definition and consequently the demodulation methods used therein. In Chapter 4, we carefully define AM and FM signals in the context of ASR. We show that for a theoretically meaningful estimation of the AM signals, it is important to constrain the companion FM signal to be narrow-band. The estimated AM message signals are down-sampled and their lower DCT coefficients are retained as speech features. We show that this representation is, in fact, the exact dual of the real cepstrum and hence, is noted as fepstrum. These features carry information that is complementary to the MFCCs. A combination of these two features is shown to significantly improve the recognition accuracies over two databases, (TIMIT phoneme recognition and University of Oldenburg non-sense syllable (Logatome) database).

Chapters 5, 6, 7 deal with noise robustness. In Chapter 5, a novel least squares filtering algorithm has been proposed that enhances the speech signal (in time-domain) that has been corrupted by the additive broad-band noise. A usual MFCC feature computed from the enhanced signal achieves significant ASR accuracies in noisy conditions as compared to the state-of-the art feature level noise robustness techniques such as Constant JRASTA-PLP and soft spectral subtraction.

Chapter 6 and 7 deal with feature level noise robustness technique. Unlike the least squares filtering that enhances the speech signal itself (in the time domain), the feature level noise robustness techniques as such do not enhance the speech signal but rather boosts the noise-robustness of the speech features that usually are non-linear functions of the speech signal's power spectrum. In the last chapter, we summarize the techniques studied in this thesis and draw conclusions. We also suggest future directions that can be pursued to improve the performance further.

Chapter 2

Review of the noise robustness in ASR systems

This chapter highlights some of the problems inherent in modern ASR systems, especially the acoustic mismatch. This is followed by a review of the prominent noise robust techniques that have been developed in the past. While many of them work quite well for specific situations, in general, they do not generalize to all the conditions. For the sake of completeness, this chapter starts with a brief introduction to the state-of-the-art ASR systems and help to familiarize the reader with the notation used in this thesis.

2.1 State-of-the-art ASR systems

The most successful approaches for ASR are based on pattern matching of the statistical representations of the speech signals. ASR typically involves a sequence of operations: 1) feature extraction, and 2) statistical modeling. Feature extraction computes a sequence of vectors representing linguistic information in the speech signal. Statistical modeling estimates the likelihood of match between that vector sequence and a set of reference probability density functions, to facilitate message decoding. The reference density functions are learned from a set of speech data called training data. It could be argued that the existence of the feature extraction as a separate block is not required, as the statistical modeling can also be performed directly on the speech signal. However, the existing statistical modeling techniques are unable to cope with the all kinds of variabilities present in the raw speech signal. Feature extraction using some sort of a-priori knowledge in form of parametric feature extraction can help to discard some of these undesirable variabilities by transforming the signal to another domain, with the help of some external knowledge. It also helps to reduce the dimensionality of the signal vectors, thereby saving the statistical modeling step from the curse of dimensionality.

Moreover, because of the infinitely large number of possible word sequences, there are infinitely large number of possible distinct representations of the whole speech signal. This makes it highly impossible to perform a statistical modeling of the whole vector sequence. A divide-and-conquer strategy is followed to simplify this problem, where the word sequences are divided into smaller segments, with the total number of distinct segments being restricted to a finite number. Typically, such a segmentation is done at the phonetic level. This is followed by the use of powerful dynamic programming algorithms to recognize the whole sequence.

2.1.1 Feature extraction

The ideal aim of the feature extraction in ASR systems is to extract representations from the speech signal that carries only the linguistic information. However this is hard to achieve as till date there exist no exact mathematical models of the linguistic information in the speech signal. Therefore, it is not possible to design feature extraction modules that can extract only and only the linguistic information from the speech signal. Various steps involved in a typical feature extraction module are explained in detail below.

Digitization of the signals: The speech signals generated by the humans are continuous-time signals. For processing these signal by machines, which can only perform digital processing, the signal are first digitized by an analog-to-digital converter (A/D) converter. A/D converter outputs the digital version of the continuous time signal by sampling it at equidistant points in time and then by quantizing those amplitudes. Telephone quality speech is the most common speech signal used in the ASR systems, whose bandwidth is typically from 200Hz to 3400Hz. Following Nyquist's sampling theorem, these signal are sampled at 8000Hz sampling frequency.

Signal pre-emphasis Signal pre-emphasis is originally motivated by the production model for voiced speech, according to which there is a spectral roll-off of -6dB/octave due to glottal closure and radiation from the lips. This is typically compensated by a pre-emphasis filter of form $(1 - az^{-1})$ which flattens the spectrum of the voiced speech. Typical values for "a" in the filter equation are in the range 0.95 to 1.00.

Short-term analysis of the signals(framing): Most of the state-of-the-art features used for speech recognition are based on Fourier analysis of the signals. Fourier analysis requires the characteristics of the signal taken for the analysis to be stationary over the analysis window. It is assumed that the signal is roughly stationary over 20ms to 30ms of signal durations. Hence, for further processing, the speech signal is divided into a sequence of short segments called frames, by performing a sequence of shifting and windowing operations on the original signal. The typical length of the window used is 20-30ms and the typical window shift size is 10ms.

Windowing: Windowing in time domain results in a convolution in the frequency domain of the signal spectrum and the window spectrum. Usually a hamming (Rabiner and Juang, 1993) window is used which has the following form.

$$w[n] = 0.54 - 0.46\cos(\tfrac{2\pi n}{N-1}),\ 0 \leq n \leq N-1 \tag{2.1}$$

Spectral Analysis Typical features used for speech recognition are based on the power spectral representation of the speech signal. The power spectrum is estimated as the absolute magnitude of the complex valued DFT(Discrete Fourier Transform) coefficients. If N represents the number of samples in a windowed segment $\mathbf{s} = s[0], s[1], s[2], \cdots, s[N-1]$, then the DFT coefficients $S[k]$ are given by, (Oppenheim and Schafer, 1989)

$$S[k] = \textstyle\sum_{n=0}^{N-1} s[n]\exp(-j2\pi kn/N) \tag{2.2}$$

Feature Transformations The knowledge that has been gained over the past few decades about the human speech perception mechanism and the human speech production apparatus, is utilized to transform the power spectrum to feature vectors that may be useful for the ASR. The goal of all of these transformations is to emphasize the linguistic information in the signal and to suppress the undesired variabilities present in the signal that do not carry any information about the linguistic message conveyed by the signal. This is one of the toughest problems in ASR as there is no precise analytic (mathematical) definition available of what is the undesired variability. As a result, we cannot design precise mathematical transformations that can rid the signal of the undesired variabilities. However, few examples such Mel-frequency cepstral coefficients(MFCC) (Davis

and Mermelstein, 1980), linear predictive cepstral coefficients(LPCC) (Atal and Hanauer, 1971) and perceptual linear prediction(PLP) (Hermansky, 1990) have been shown to be quite successful for the clean speech ASR.

Instead of incorporating the external knowledge during the feature extraction, the algorithms can also be made to learn transformations that would discard undesired variabilities in the power spectrum. Such feature extraction schemes are called data-driven feature extraction. Examples of such data-driven feature extraction techniques are principal component analysis (PCA) (Bourlard and Kamp, 1988), and its non-linear equivalent using artificial neural networks (Ikbal et al., 1999), linear discriminant analysis(LDA), its non-linear equivalent called TANDEM approach (Ellis et al., 2001a). The MFCC features, which are quite successful in the state-of-the-art speech recognition systems (and will also be used throughout this thesis as a reference feature), are explained in more detail below. The development of MFCC has its roots in the outcome of the studies (Davis and Mermelstein, 1980) on human auditory perception. It tried to mimic the human auditory periphery by utilizing the knowledge about the human perception system in the feature transformation stage. The computation of the MFCC proceeds as follows: The power spectral values are integrated within overlapping Mel-scaled triangular filters to obtain what is called Mel-scale filter bank energies. The Mel-filter bank energies are then compressed by a logarithmic function. The resultant values are then transformed through a discrete cosine transform (DCT) to obtain the MFCC coefficients. The higher order cepstral coefficients are usually dropped to as it is believed that they mostly correspond to the non-linguistic information. A consequence of this is the smoothing in the spectral domain. State-of-the-art speech recognition systems typically incorporate the temporal dynamics of the speech signal in the feature representation by including the first and the second derivatives of the static feature vectors. the first derivative of the static feature vector is referred to as the delta feature and the second derivative is referred to as the double-delta (acceleration) feature. The delta is computed over a short time window using equation:

$$\Delta x_t = \frac{\sum_{d=1}^{D} d(x_{t+d} - x_{t-d})}{2\sum_{d=1}^{D} d^2} \tag{2.3}$$

where D denoted the time window length over which the delta is computed. The same equation is used to compute the double-delta features, by replacing the static features with delta features. This sequence of operations in the feature extraction block finally outputs feature vectors every 10 ms. These feature vectors then go as input to the next stage, the statistical modeling stage.

2.1.2 Statistical modeling

If $\mathbf{X} = \{\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_{T-1}}\}$ represents the sequence of the feature vectors extracted from the speech signal , statistical modeling technique formulates the speech recognition problem as a maximum a posteriori (MAP) problem as follows.

$$W^* = \arg\max_{W \in \mathcal{L}} P(W|X) \tag{2.4}$$

The most likely word sequence W^* from the set of all possible word sequences $\mathcal{L}$, given $\mathbf{X}$, is chosen as the recognized string of the words. The MAP formulation of the speech recognition is hard

to deal with directly. It is usually reformulated into a problem based on the likelihood estimation using the Bayes rule[1].

$$W^* = \arg\max_{W \in \mathcal{L}} \frac{P(X,W)}{P(X)} \sim \arg\max_{W \in L} P(X|W)P(W) \tag{2.5}$$

$P(X)$ has been dropped from the above equation as it serves just as a scaling factor. In the above equation, $P(X|W)$, the conditional probability of $\mathbf{X}$ given W, is usually referred to as the *acoustic model*, and $P(W)$, the prior probability of the word sequence W, is referred to as the *language model*. In practice, both the acoustic and language models are assumed to fit in some parametric form, say $P_\Theta(.)$ and $P_\Gamma(.)$, with parameter sets Θ and Γ respectively. Then $P_\Theta(\mathbf{X}|W,\Theta)$ and $P_\Gamma(W|\Gamma)$ are used in (2.5) as estimates for $P(\mathbf{X}|W)$ and $P(W)$. The values of model parameters Θ and Γ are estimated from the training database containing a large collection of utterances with known transcriptions. If ξ represents the set of all the training utterances along with the corresponding transcriptions, ideally the parameters can be estimated according to:

$$\Theta^*, \Gamma^* = \arg\max \left[\prod_{W \in \xi} P_\Theta(\mathbf{X}|W,\Theta) P_\Theta(W|\Gamma) \right] \tag{2.6}$$

But practical constraints do not allow the joint estimation of Θ and Γ. They are usually estimated independently of each other from different training sets, say ξ_a and ξ_l, respectively, yielding:

$$\Theta^* = \arg\max_\Theta \left[\prod_{\mathbf{X} \in \xi_a} P_\Theta(\mathbf{X}|W,\Theta) \right] \tag{2.7}$$

$$\Gamma^* = \arg\max_\Gamma \left[\prod_{W \in \xi_l} P_\Gamma(W|\Gamma) \right] \tag{2.8}$$

Equation (2.7) is referred to as maximum likelihood (ML) training. A popular ML training algorithm is expectation-maximization (EM) algorithm (Dempster et al., 1977) where a few hidden variables are postulated in addition to the existing parameter set in order to make the otherwise intractable training problem to be tractable. EM is an iterative procedure where, in each iteration the new values for the parameter set, Θ^{new}, are found from the old values, Θ^{old}, so that the overall likelihood of the training data is increased:

$$\prod_{\mathbf{X} \in \xi_a} P_\Theta(\mathbf{X}|W,\Theta^{new}) \geq \prod_{\mathbf{X} \in \xi_a} P_\Theta(\mathbf{X}|W,\Theta^{old}) \tag{2.9}$$

Every iteration of EM has two steps: E and M steps. In E step, we find the expected value of the complete data log likelihood with respect to the probability distribution of the hidden variables given the observed variables $\mathbf{X}$ and the current estimates of the parameters. In the M step, we maximize this expected complete data log-likelihood. These two steps are repeated as necessary. Each iteration is guaranteed to increase the likelihood of the observed variables $\mathbf{X}$.

State-of-the-art ASR system use hidden Markov models (HMMs) (Rabiner and Juang, 1993; Bourlard and Morgan, 1994) for acoustic modeling and bigram/trigram probabilities for the language modeling. As language modeling does not fall within the scope of this thesis, it will not be discussed further in this thesis. HMM used for acoustic modeling is explained in more detail below.

[1] $P(A|B) = \frac{P(A,B)}{P(B)}$

Hidden Markov Models (HMM): The most successful approach developed so far for the acoustic modeling task of the ASR is the hidden Markov model (HMM). HMM is basically a stochastic finite state automaton, i.e, a finite state automaton with stochastic output process associated with each state. HMM models the speech by assuming that the feature vector sequence $\mathbf{X} = \{\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_{T-1}}\}$ to be a piece-wise stationary stochastic process that has been generated by a sequence of HMM states, denoted by $Q = \{q_0, q_1, \ldots, q_{T-1}\}$, that transit from one to another over time. The stochastic output process associated with each state is assumed to govern the generation of feature vectors by the states. If C represents the set of all possible state sequences, the acoustic model in (2.5) can be rewritten as:

$$P(X|W) = \sum_{Q \in C} P(X, Q|W) \tag{2.10}$$

In the above equation, Θ as in (2.7) is dropped for the reasons of simplicity. To make the model simple and computationally tractable a few simplifying assumptions are made while applying HMMs to the acoustic model problem. They are:

1. First order hidden Markov model assumption, i.e.,

$$P(q_t|q_0, q_1, \ldots, q_{t-1}) = P(q_t|q_{t-1}) \tag{2.11}$$

   where $P(q_t|q_{t-1})$ is referred to as the state-transition probability.

2. Feature independence (i.i.d) assumption, i.e.,

$$P(x_t|x_0, x_1, \ldots, x_{t-1}, Q) = p(x_t|q_t) \tag{2.12}$$

   where $p(x_t|q_t)$ is referred to as the emission probability, i.e., the probability of the state q_t emitting the feature vector x_t.

With these assumptions 2.10 becomes,

$$P(X|W) = \sum_{Q \in C} P(q_0)p(x_0|q_0) \prod_{t=1}^{T-1} P(q_t|q_{t-1})p(x_t|q_t) \tag{2.13}$$

The above equation gives an exact formula for the computation of the likelihood. However, some times it is also approximated as the likelihood of the best state sequence as follows:

$$P(X|W) = \max_{Q \in C} p(q_0)p(x_0|q_0) \prod_{t=1}^{T-1} P(q_t|q_{t-1})p(x_t|q_t) \tag{2.14}$$

This is called as Viterbi approximation.

An illustration of the use of the HMM as an acoustic model with the above assumption is given in Fig.2.1. If the number of the states in the HMM is M, the complete set of parameters that describe the complete HMM are the following:

1. Transition probabilities, $P(q_t = \text{state } j|q_{t-1} = \text{state } i)$,denoted by $a_{ij},\ 0 \leq i, j, \leq M-1$ and satisfying the constraints $\sum_{j=0}^{M-1} a_{ij} = 1$, and
2. State emission density functions, $p(x_t|q_t = \text{state i})$, denoted by $p_i(.), 0 \leq i \leq M-1$.

A detailed tutorial on EM algorithm and its use to estimate the HMM parameters for ASR application is provided in (Bilmes, 1998). Most commonly used techniques for modeling the emission density are Gaussian Mixture models and multi layer perceptrons (MLP) and they are briefly explained below:

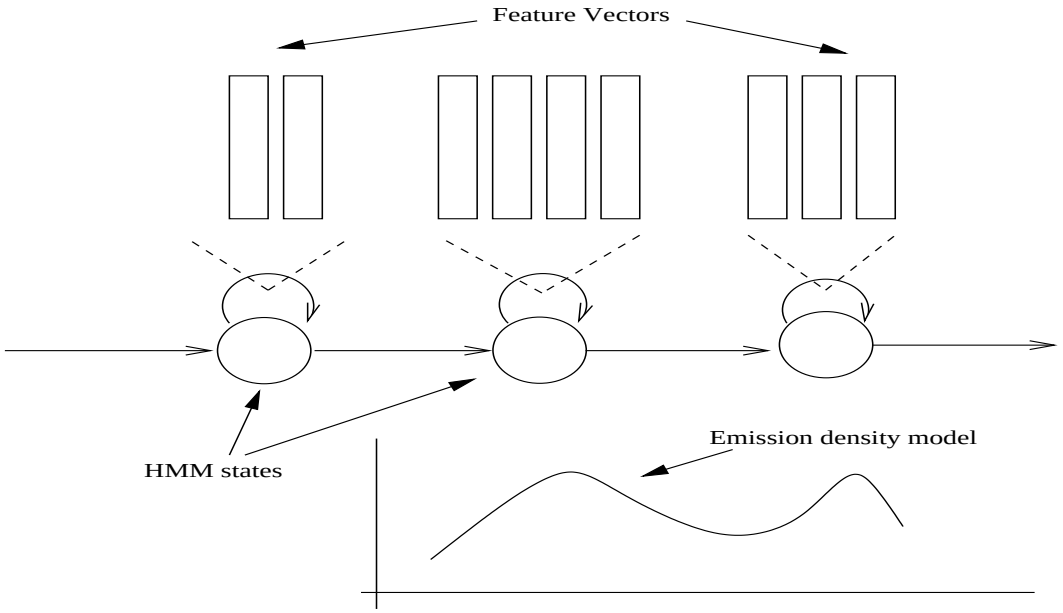

Figure 2.1. Illustration of Hidden Markov Models (HMM).

1. Gaussian Mixture Model (GMM) is a weighted mixture of several Gaussian densities. It is fully characterized by weighting factors, mean vectors and covariance matrices of all the constituent Gaussians. The expression for density function $p(\mathbf{x})$ for GMM is given by,

$$p(\mathbf{x}) = \sum_{k=0}^{K-1} c_k G_k(\mathbf{x}) \tag{2.15}$$

where K denotes the number of Gaussians in the GMM, and c_k denoted the weighting factor for the k^{th} Gaussian, $G_k(.)$. If μ_k and Σ_k denote respectively the mean vector and covariance matrix of the k^{th} Gaussian, and if D denote the feature vector dimension, the expression for $G_k(\mathbf{x})$ is given by,

$$G_k(\mathbf{x}) = \frac{1}{2\pi^{D/2}|\Sigma_k|^{D/2}} \exp(-\tfrac{1}{2}(\mathbf{x}-\mu_k)^T \Sigma_k^{-1}(\mathbf{x}-\mu_k)) \tag{2.16}$$

2. Multi layer perceptrons (MLP) (Bourlard and Wellekens, 1990; Bourlard and Morgan, 1994), a special case of artificial neural networks (ANNs) has a series of layers of artificial neurons. Neurons in each layer are fully connected with the neurons of the following layer. The first layer is called the input layer, the last layer is the output layer, and all the in between layers are called hidden layers. Every neuron of all the layers, except the input layer, perform non-linear operation; a weighted sum of the outputs of all the neurons from which it receives the input, followed by a sigmoid or a softmax operation. The connection weighting factors between the neurons are called the weights. The vector applied at the input layer propagates through the hidden layers one by one until it finally reaches the output layer. MLPs can be used for classification or function mapping. Arbitrary complex decision hyper-surfaces can be formed by the MLPs while using them in the classification mode. Any continuous mapping between input and output space can be represented by the MLPs while using them in the function mapping mode. The connection weights of the MLP can be trained using error back propagation

method. For emission probability density modeling, MLPs are used in the classification mode, where the number of the output classes, i.e., the number of output neurons, corresponds to the number of HMM states. In pattern classification mode, the outputs of MLP estimate the posterior probability of the input vector $\mathbf{x}$, $p(q_k|\mathbf{x})$. Using the Bayes rule, likelihood as used in (2.14) can be recalculated as,

$$p(\mathbf{x}|q_k) = \frac{p(q_k|\mathbf{x})p(\mathbf{x})}{p(q_k)} \tag{2.17}$$

where $p(q_k)$ denote the apriori class probability and can be estimated from the training set. However, since $p(\mathbf{x})$ is a constant for all the classes and thus appearing just as a scale factor, a scaled likelihood, as calculated by the equation below, is used in (2.14) instead of the likelihoods $p(x|q_k)$.

$$\frac{p(\mathbf{x}|q_k)}{p(\mathbf{x})} = \frac{p(q_k|\mathbf{x})}{p(q_k)} \tag{2.18}$$

2.2 Noise robust speech recognition

Solution to the problem of the sensitivity of the speech recognition systems to the external noise can be approached in two different ways. Accordingly, the noise robust techniques are grouped into classes:

1. Model based approaches
2. Feature based approaches

Model based approaches assume the feature vector to be sensitive to the external noise and attempt to handle this sensitivity at the statistical modeling level. Whereas, the feature based approaches try to make the feature vectors insensitive to the external noise. Several successful techniques developed under both the approaches have been reported in the literature. A few prominent techniques are explained in the next two sections. However, before going into the details of various noise robust methods, it will be useful to have a look at the various types of noise and understand the manner in which they affect the speech signal, which is discussed in the next subsection.

Effect of noise on speech signal: The external noise mainly affects the speech signal as it is propagated from the speaker to the receiver. The noise could be correlated with the speech signal or could be uncorrelated. Correlated noise results from the reflections and reverberations. Externally generated noises are usually uncorrelated. The uncorrelated noise could be stationary or non-stationary, wide-band or narrow-band (colored), and could last for only a short time (impulsive) or for long time periods. A noise type is said to be stationary if its statistical characteristics do not change over time and is said to be non-stationary if its statistical characteristics do change over time. Wide band noise has a spectral energy distribution that is widely distributed over the frequencies of interest. A few examples of the non-stationary short time noises are door slamming and the noise generated by a passing car. Noises from a factory floor and competing speakers are examples of long-duration non-stationary noises. Fan and air-conditioning noise are examples of stationary noise while siren is an example of non-stationary and colored noise.

The noise types that are considered in this thesis are only the externally generated noises that are uncorrelated with the speech signal. The resultant signal of two sound sources is approximately an addition of the individual sources. Suppose $s[n]$ denotes the speech signal generated by the speaker, and $r[n]$ denoted the resultant signal of all the noise sources in the surrounding environment. Then the resultant signal that reaches the receiver is $s[n] + r[n]$. Suppose the power

spectral density of the clean speech is $S[k]$ and that of the additive noise is $R[k]$. If the noise is uncorrelated with the speech signal, then the power spectral representation of the noisy speech, $\tilde{S}[k]$ is given by,

$$\tilde{S}[k] = S[k] + R[k] \tag{2.19}$$

At this point, it is important to define a term called signal-to-noise ratio (SNR) that gives a measure of the extent to which the speech signal is affected by the noise. SNR is basically a ratio between the powers of the signal component and the noise component, usually specified in decibel (dB), which is 20 times the logarithm of the ratio given below,

$$SNR = 10 \log \left(\frac{\text{signal power}}{\text{noise power}} \right) \tag{2.20}$$

2.3 Model based approaches

The variability due to the external noise is accounted for in the model based approaches either by adapting the statistical model to match the new acoustic environment(through the estimation of the noise distribution or through the estimation of the perturbations in the speech distributions caused by the noise) or by helping the statistical model to discard the unreliable part of the feature vector. The model based approaches, especially adaptation based techniques, are computationally expensive. Some specific techniques need an impractical requirement of larger amount of transcribed speech data for the adaptation during the recognition. A few model based approaches are explained briefly in the following subsections.

2.3.1 Multi-condition training

A simple and direct model based method for achieving noise robustness is the inclusion of all possible testing noise conditions in the training set (Furui, 1992). By this way the statistical modeling will be able to model the all possible variabilities observed in the feature vectors due to external noise. This in fact has been shown experimentally to improve the noisy speech recognition performance. However, this method is unrealistic in the sense that it is impossible to include all possible noise types in the training set. A slight variant of this approach is to include a set of representative noise conditions in the training set and to make the statistical models generalize to the unseen noise. This has been observed to result in improved robustness, though the relative degradations are more severe than the case when statistical models are directly trained on the appropriate noise conditions.

2.3.2 Signal decomposition

The idea in signal decomposition (Varga and Moore, 1990) is to recognize the concurrent signals simultaneously using a set of HMMs, one each for the components into which the signal is to be decomposed. Recognition is carried out by searching through the combined state space of the constituent models. For example, if the signal considered is speech added with noise from a single source, the search will be through a three-dimensional space. If $\mathbf{r}_t$ represents the noise component added to the speech component $\mathbf{x}_t$ to obtain the resultant representation $\widetilde{\mathbf{x}}$, and if q_t and p_t represents the states of speech and noise HMMs respectively, then the likelihood to be used in the three dimensional search is given by:

$$P(\widetilde{\mathbf{x}}_t|q_t,p_t) = \int_{S_{\mathbf{x_t},\,\mathbf{r_t}}} P(\mathbf{x_t},\mathbf{r_t}|q_t,p_t) \tag{2.21}$$

where $S_{\mathbf{x_t},\,\mathbf{r_t}}$ represents the set of all possible pairs $\{\mathbf{x_t},\mathbf{r_t}\}$ such that $\widetilde{\mathbf{x}}_t = \mathbf{x_t} + \mathbf{r_t}$.
As in multi-condition training, the signal decomposition also ideally relies on the need to include all possible noise conditions in the training set, which is clearly practically infeasible. Moreover, the computation cost also increases exponentially with the increase in the number of the number of noise components to be recognized in the signal. Even the presence of a single noise component requires a search through the three dimensional space, which is computationally quite expensive. Besides this, it is not known apriori as to how many signal components are present in the signal.

2.3.3 Parallel model combination

A variation of signal decomposition is parallel model composition (PMC) (Gales and Young, 1996) where, instead of using independent speech and noise models to perform recognition of observations sequence in a three dimensional space, the speech and noise models are first combined to form a noisy speech model and then the recognition is performed using this model. In this case, in order to devise an appropriate model combination algorithm, it is important to understand quite well as to how the component signals are combined. Additionally, the domain in which the model combination should be performed is also important. For example, if the speech is independent of noise, they are additive in the power spectral domain. However, if the feature vectors used are in the cepstral domain, then first the model parameters should be transformed from the cepstral domain to the spectral domain and then after combining the models,the combined model should be brought back to the cepstral domain.

2.3.4 Maximum likelihood linear regression (MLLR)

MLLR (Legetter and Woodland, 1995), originally developed to handle the speaker related variabilities in the speech signal can also be used to handle noise related variabilities. In MLLR based method, the GMM parameters, such as the mean vectors and possible the covariance matrices, are adapted to new environmental condition using data from the new environment. The linear transformation to adapt the mean vectors, μ, of one of the constituent Gaussians of a state is given by the following equation:

$$\widetilde{\mu} = [\mathbf{b}\ \mathbf{A}]\left[1\ \mu^T\right]^T \tag{2.22}$$

Where **A** represents the regression matrix and the **b** is the bias vector. The matrix **A** and the bias vector b are estimated using speech data from the new acoustic environment in a maximum likelihood fashion. This requirement of the adaptation data to adapt the model parameters stands in as a disadvantage as usually the estimation of parameters require a large amount of data during recognition. The effective number of parameters in the regression matrix can be reduced by tying the parameters, which in turn requires lesser amount of data from the new environmental condition. However, the requirement of even a reasonable amount of data during recognition is impractical, except for a few specific applications. Additionally, the transformation required to adapt the model parameters to new acoustic conditions may not be linear. In such as case, non-linear transformations can be realized by a mixture of linear regression classes, which in turn require a large amount of adaptation data.

2.3.5 Multi-band and multi-stream processing

In multi-band processing (Bourlard et al., 1997), the spectral representation of the speech signal is split into several frequency bands and each band is processed separately to first extract sub-band feature vectors. These features are then processed with different HMMs for every sub-band to extract sub-band likelihoods or probabilities. These probabilities are then combined to yield an effective likelihood that can be used for recognition. The contribution from different sub-bands for computing the combined probability is varied based upon the reliability of the sub-band. If a particular sub-band is corrupted its contribution to the combined probability estimate is lowered assuming that thus will help the recognition. This method is effective for the cases where speech is affected by the colored noise. However, independent processing assumes each spectral band to be independent, which in turn causes the performance to degrade for clean speech. This can be partly avoided by the recently proposed full combination method, where all possible combinations of sub-bands including the full-band, which does not treat different bands independently, are considered while computing the final likelihood to use in decoding.

Multi-stream processing combines evidences from several streams of feature vectors extracted from the same speech signal. Different processing techniques used for extracting different features may emphasize different aspects of the signal which may be complementary in nature. Thus an adaptive combination of the stream may yield a recognition performance better than that of the individual streams. Multi-stream feature combination in ASR systems has its roots in the psychoacoustic studies on human beings, which shows evidence for multiple representation of the speech signal at different stages of human auditory processing and an integration of them in order to get a robust final representation. Different processing steps done in different feature extraction schemes for ASR system may provide different kinds of evidence under different environmental conditions. Hence an appropriate combination of such features, making use of their complementary information, is expected to improve the overall recognition performance. The combination can be performed at various stages of the ASR system as follows:

1. Feature combination: The feature vectors are combined before the statistical modeling. One simple method to combine the features is to concatenate the features to get a single feature vector of larger dimension, as done in the case of static and dynamic features.

2. Posterior Combination: The probability outputs of the different acoustic models, processing different feature streams, are combined in order to obtain the combined probability.

2.3.6 Missing data approach

This method relied on the fact that some of the spectro-temporal regions in the spectrogram will be dominated by the external noise. Thus during recognition, by treating these regions as missing or unreliable the overall robustness can be improved. The recognition is only based on the regions that are tagged as reliable (Raj et al., 2004; Seltzer et al., 2004; Cooke et al., 2001). If the components in $\mathbf{x}$ belonging to the reliable part are denoted by $\mathbf{x}_r$ and those belonging to the unreliable part are denoted by $\mathbf{x}_u$, then, during recognition, the missing data is dealt with in one of the following ways.

1. Marginalization, i.e, the local emission probability is estimated as just the emission probability of the reliable part, $\mathbf{x}_r$, as follows:

$$p(\mathbf{x}_r|q_t) = \int_{\mathbf{x}_u} p(\mathbf{x}_r, \mathbf{x}_u|q_t) d\mathbf{x}_u \tag{2.23}$$

2. Data imputation, where values corresponding to the unreliable regions are estimated to produce an estimate of the complete observation vectors $\hat{\mathbf{x}}$, which is further used for computing the local emission probability as $p(\mathbf{x}|q_t) = p(\hat{\mathbf{x}}|q_t)$.

The practical implementation of the missing data approach requires a robust algorithm to identify the reliable regions in the spectrogram. In the related work, reported in the literature, simple noise estimation techniques are used as a basis for the identification task.

2.3.7 Tandem Modeling

In the Tandem modeling (Ellis et al., 2001b; Hermansky, 2003; Zhu et al., 2004), a multi layer perceptron (MLP) that is used to perform a data-driven transformation of the input features, learns the transformation by getting trained in a supervised, discriminative mode, with phoneme labels as the output classes. Such a training make the MLP to perform a non-linear discriminative analysis in the input feature space and thus makes it to learn a transformation that projects the input features onto a subspace of maximum class discriminatory information. This transformation is able to suppress the variabilities which as such do not characterize the speech sound (phonemes). Tandem modeling is also useful in integrating several feature streams for their final use in an HMM-GMM system.

2.4 Feature based approaches

Feature based methods avoid the computationally intensive model based methods by generating feature representations that are invariant to the noise. A review of prominent feature techniques can be found in (Stern et al). These methods often involve the use of external knowledge about the effect of the noise on the features, in order to devise an appropriate algorithm. Such knowledge is basically used to design transformations that would supposedly remove the noise prone aspects of the features.

2.4.1 The use of psychoacoustic and neuro-physical knowledge

Early feature based methods involve incorporation of various psychoacoustic and neuro-physical knowledge, obtained form the human auditory system, into the feature extraction algorithm. As human auditory system is the best speech processing system till date, imitating a few functionalities of it in the feature extraction algorithm is expected to improve the noise robustness of the ASR systems. Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) features are widely used features falling in this category. The MFCC uses the Mel-warped frequency axis and approximates the power law of hearing by taking the logarithm of the critical band power spectrum. With these operations the features vectors have been shown to improve the recognition performance. One of the widely used feature vectors during the early stages of the speech recognition was the linear prediction (LP) cepstrum (Makhoul, 1975). LP cepstrum can be computed recursively from the linear prediction coefficients. An improvement over the simple linear prediction analysis (LP) utilizing the auditory peripheral knowledge, is the perceptual linear prediction (PLP) (Hermansky, 1990). Before doing the LP analysis an estimate of the auditory spectrum is obtained from the power spectrum by applying several transformations that are assumed to happen at the human auditory periphery front-end. The series of transformations include critical band integration (on bark scale), equal loudness pre-emphasis, cubic root compression (to account for power law of hearing). The auditory spectrum obtained is then used to predict the LP coefficients, and then the equivalent PLP spectrum, and finally the PLP cepstrum. Alike MFCC, PLP cepstrum is also expected to have reduced undesirable variabilities as a results of the incorporation of various auditory like transformations.

2.4.2 Speech enhancement

A different class of feature based noise robust methods try to enhance the speech specific aspects of the spectrum by suppressing the noise-specific aspects. An early method falling in this category is the spectral subtraction. It gets an estimate of the enhanced spectrum $\hat{S}[k]$ from the original spectrum $S[k]$ using an estimate of the noise spectrum $R[k]$ as follows:

$$\hat{S}[k] = S[k] - R[k] \tag{2.24}$$

The success of this method relies on the reliable estimation of the noise power spectrum. The noise power spectrum is usually estimated from the non-speech intervals of the signal. Thus a reliable speech versus non-speech detector is required. Especially, in the low SNR conditions, due to the spectral similarities between the unvoiced speech and the noise, the noise estimation becomes a difficult task. Furthermore this technique is suitable only for the case of stationary noise. In case of non-stationary noise, even if we have perfect speech/non-speech detector, it is not possible to accurately follow the noise spectral statistics as they may change quite rapidly. Therefore, it usually results in the removal of the significant speech information. The subtraction of the noise power can result in negative values if the noise estimate exceeds the actual noise magnitude. This can be partially taken care of by setting a threshold for the power values, which introduces a residual noise(also called musical noise) in the signal domain. An improvement over the spectral subtraction is the nonlinear spectral subtraction (NSS), which combines the spectral subtraction with noise masking. NSS has been demonstrated to improve the speech recognition performance in car noisy conditions (Lockwood and Boudy, 1992). A relatively new technique that has been shown to be quite successful for recognition of speech corrupted by slow varying noise is relative spectra (RASTA) processing (Hermansky and Morgan, 1994a). It tries to suppress these noise components whose temporal properties are quite different from that of the speech, in the spectral domain. The temporal properties of different frequency bands in the spectrum are modeled by the modulation spectrum. The lower bound of the modulation spectral band-width of the clean speech gives a measure of the lowest possible rate at which the signal components of speech can be generated, while higher bound gives the highest possible rate. Thus the modulation spectral components beyond the bandwidth of the clean speech can be assumed to be from the noise source. A band-pass filter, whose band-width is equal to the modulation spectral bandwidth of the clean speech, is applied to each frequency band of the spectrum, to filter out the noise components. The transfer function of the filter is:

$$H(z) = 0.1z^4 \frac{2 + z^{-1} + z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{2.25}$$

where $z^{-1} = e^{-j\omega T}$ and sampling frequency of the spectral trajectories is 100 Hz. The above filter is best suited for channel effects in the logarithmic spectral domain. To handle both noise and channel effects simultaneously, RASTA filtering is more effective when applied to an equivalent spectrum, $\hat{S[k]}$, computed according to the following equation:

$$\hat{S[k]} = log(1 + JS[k]) \tag{2.26}$$

where J is the scaling factor to be found empirically. This procedure is called constant J-RASTA (CJ-RASTA) processing. PLP cepstrum obtained from CJ-RASTA filtered spectrum is called CJ-RASTA-PLP.

2.4.3 Wiener Filtering

Wiener filtering (Haykin, 1993) is an optimal technique for the signal enhancement in the minimum mean square sense when the noise and the clean speech satisfy the following conditions:

- The noise and speech are statistically independent of each other and the noise is assumed to be stationary.
- Besides the speech + noise audio channel, one has an access to the 'noise only' channel as well. This condition is required in order to estimate the noise power spectral density.

Let us denote the clean speech and the noise power spectral densities by $\tilde{X}[k]$ and $\tilde{N}[k]$ respectively. Then the enhanced signal's spectrum ($\hat{S}[k]$), given the noisy speech spectrum ($S[k]$) is given as:

$$\hat{S}[k] = S[k]H[k] = S[k]\frac{\tilde{X}[k]}{\tilde{X}[k] + \tilde{N}[k]} \tag{2.27}$$

where $H[k]$ is the Wiener filter. The major limitation of the Wiener filter is that, in the most of the practical scenarios, one does not have an access to the 'noise only' channel. The leads to the problem of the estimation of the noise power spectral density which is, in turn, required to estimate the Wiener filter. Nevertheless, Wiener filter is one of the most celebrated results in the adaptive filter theory.

2.4.4 Noise Masking

Noise masking is a psychological phenomenon observed in humans where the perceptibility of the signal is reduced in the presence of noise, to decrease the effect of the noise. As a result of the masking, acoustic stimuli lower than certain threshold, fixed adaptively based on the noise level, cannot be perceived. Based on our knowledge of perception, this involves reduction of contribution of the lower energy regions of the spectrum during the recognition process. Employing this idea in the ASR system, a simple noise flooring and its extension in the HMM framework were shown to provide improved noise robustness. Noise masking in the logarithmic spectral domain and the cepstral domain have also been tried.

Spectral root homomorphic deconvolution scheme introduced in (Lim, 1979) perform a root operation instead of logarithmic operation on the spectral values, before transforming them to the cepstral domain. An appropriate root value for the root operation, in effect, relatively emphasizes and deemphasizes the peaks and the valleys respectively.

2.5 Databases and experimental setup

The speech databases used for the experimental evaluation of novel front-end techniques developed in this thesis is primarily OGI Numbers95 database (Cole et al., 1994). Chapters 3 and 4 deal with the design of novel features that have been specially designed to improve the ASR performance beyond that of the MFCC features in the clean acoustic conditions. These novel features have also been tested on the TIMIT and the University of Oldenburg non-sense syllable OLLO (Wesker et al., 2005) database. Chapters 5, **??**, 6 describe a noisy speech signal enhancement technique and several noise robust feature extraction techniques that have been developed in the course of this thesis. To evaluate the performance of these novel noise robust ASR techniques, the ASR experiments were performed on the OGI Numbers95 database (Cole et al., 1994). To simulate the noisy conditions, different types of non-stationary noises are added to the OGI Numbers95 clean test-set as explained in the following subsection.

2.5.1 OGI Numbers95 database

The OGI numbers95 database consists of naturally spoken connected numbers pronounced by American English speakers. The utterances were recorded over the telephone lines and are hand-

labeled with phonetic transcriptions by trained phoneticians. It has a lexicon size of 30 words[2] and 27 different phonemes. The database is divided into two independent subsets: the training set (including a cross-validation set) and the test-set. The training set consists of 3233 utterances comprising roughly of 3 hours of speech. 20% of the training utterances are used as the cross-validation data. The test-set consists of 1206 utterances.

2.5.2 Noise data

For the noisy speech recognition, different noises are added to the test-set's clean speech utterances from the OGI Numbers95 database. The noise types considered are Factory and F-16 aircraft cockpit noise from the Noisex92 (Varga et al., 1992) database. These noises are added to the clean utterances of only the test set at varying SNRs of 12dB, 6dB and 0dB. We emphasize that unlike the multicondition training, throughout this thesis, the acoustic model parameters are trained using only the clean utterances, while testing is done on noisy as well as clean speech utterances.

2.5.3 University of Oldenburg non-sense syllable (logatome) database OLLO

OLLO database (Wesker et al., 2005) has been specially designed to study the effects of the intrinsic variabilities (speaking rate and speaking style) present in the usual speech signal. This database is rich in various speech variabilities such as different speaking styles (slow, fast, statement, questioning, loud and soft) and with almost equal sampling of the male and female speakers. The lexicon consists of 150 logatomes[3] which are either CVCs or VCVs such as uttu, acsha, atta etc. The train set and test-set consists of roughly 13,500 and 13,800 utterances respectively and they correspond to the no-accent part of the database[4]. The experiments reported in this thesis are for the entire logatome recognition on the No-accent test part of the OLLO database that consists of roughly 13,800 utterances. Each of these utterances correspond to an instance of a logatome.

2.5.4 The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus

The TIMIT acoustic-phonetic speech corpus was designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. The train set consists of roughly 3850 utterances. The test data has a core portion containing 24 speakers, 2 male and 1 female from each dialect region. The core-test set consists of 192 utterances.

2.5.5 Experimental Setup

The main speech recognition system used for the experiments is HMM-GMM based system. For the OGI NUmbers95 task, it consists of 80 tri-phones, 3 left-to-right states per triphone, and 12 mixture GMM to estimate the emission probability within each state. Training is performed using HMM tool kit (HTK). In some of the experiments Tandem acoustic modeling (Ellis et al., 2001a) is also used to further validate the efficacy of the proposed features. Tandem modeling consists of using the a-posterior probabilities at the output of a trained MLP (that has been trained using novel features) as a feature vector to the HMM-GMM system.

[2] with confusable triplets of words such as nine, nineteen and ninety.
[3] A logatome is CVC or VCV that consists of nonsense syllables
[4] The no-accent portion of the database correspond to the normal German accent

2.6 Conclusion

In this chapter, we have given a brief introduction to the state-of-the-art ASR systems and gave a comprehensive coverage of the prominent noise robustness techniques developed in the past. The contribution of this thesis is explained starting from the next chapter. The first two chapters are devoted toward improving the ASR performance in the clean speech environment and the remaining three chapters concern noise robustness techniques developed in the course of this thesis.

Chapter 3

Variable-Scale Piecewise Stationary Spectral Analysis of Speech Signals for ASR.

It is often acknowledged that speech signals contain short-term and long-term temporal properties (Rabiner and Juang, 1993) that are difficult to capture and model by using the usual fixed scale (typically 20ms) short time spectral analysis used in hidden Markov models (HMMs), based on piecewise stationarity and state conditional independence assumptions of acoustic vectors. For example, vowels are typically quasi-stationary over 40-80ms segments, while plosive typically require analysis below 20ms segments. Thus, fixed scale analysis is clearly sub-optimal for "optimal" time-frequency resolution and modeling of different stationary phones found in the speech signal. In this chapter, we investigate the potential advantages of using variable size analysis windows toward improving state-of-the-art speech recognition systems in clean acoustic conditions. Based on the usual assumption that the speech signal can be modeled by a time-varying autoregressive (AR) Gaussian process, we estimate the largest piecewise quasi-stationary speech segments, based on the likelihood that a segment was generated by the same AR process. This likelihood is estimated from the Linear Prediction (LP) residual error. Each of these quasi-stationary segments is then used as an analysis window from which spectral features are extracted. Such an approach thus results in a variable scale time spectral analysis, adaptively estimating the largest possible analysis window size such that the signal remains quasi-stationary, thus the best temporal/frequency resolution trade-off.

3.1 Introduction

Most of the Automatic Speech Recognition (ASR) acoustic features, such as Mel-Frequency Cepstral Coefficient (MFCC) (Davis and Mermelstein, 1980) or Perceptual Linear Prediction (PLP) (Hermansky, 1990), are based on some sort of representation of the smoothed spectral envelope, usually estimated over fixed analysis windows of typically 20ms to 30ms of the speech signal (Davis and Mermelstein, 1980; Hermansky, 1990). Such analysis is based on the assumption that the speech signal can be assumed to be quasi-stationary over these segment durations. Typically, the sustained-stationary segments in a vowel can typically last from 30 to 80ms, while stops are time-limited by less than 20ms (Rabiner and Juang, 1993). Therefore, it implies that the spectral analysis based on a fixed size window of 20ms or 30ms has some limitations, including:

- The frequency resolution obtained for quasi-stationary segments (QSS) longer than 20 or 30ms is quite low compared to what could be obtained using larger analysis window. Although, most of the frequency resolution is lost due to averaging by 24 Mel filters. However, power spectrum estimation (DFT) over quasi-stationary segments will still lead to low-variance Mel-filter bank energies as compared to those obtained with a fixed scale spectral analysis that does not take quasi-stationarity into account.

- In certain cases, the analysis window can span the transition between two QSSs, thus blurring the spectral properties of the QSSs, as well as of the transitions. Indeed, in theory, Power Spectral Density (PSD) cannot even be defined for such non stationary segments (Haykin, 1993). Furthermore, on a more practical note, the feature vectors extracted from such transition segments do not belong to a single unique (stationary) class and may lead to poor discrimination in a pattern recognition problem.

In this work, we make the usual assumption that the piecewise quasi-stationary segments (QSS) of the speech signal can be modeled by a Gaussian AR process of a fixed order p as in (Itakura, 1975; Svendsen et al., 1989; Brandt, 1983; Obrecht, 1988). We then formulate the problem of detecting QSSs as a Maximum Likelihood (ML) detection problem, defining a QSSs as the longest segment that has most probably been generated by the same AR process.[1] As is well known, given a p^{th} order AR Gaussian QSS, the Minimum Mean Square Error (MMSE) linear prediction (LP) filter parameters $[a(1),\ a(2), ... a(p)]$ are the most "compact" representation of that QSS among-st all the p^{th} order all pole filters (Haykin, 1993). In other words, the normalized "coding error"[2] is minimum among-st all the p^{th} order LP filters. When erroneously analyzing two distinct p^{th} order AR Gaussian QSSs in the same non-stationary analysis window, it can be shown that the "coding error" will then always be greater than the ones resulting of QSSs analyzed individually in stationary windows (Kay, 1998). This is intuitively satisfying since, in the former case, we are trying to encode $'2p'$ free parameters (the LP filter coefficients of each of the QSS) using only p parameters (as the two distinct QSS are now analyzed within the same window). Therefore, higher coding error is expected in the former case as compared to the optimal case when each QSS is analyzed in a stationary window. As further explained in the next sections, this forms the basis of our criteria to detect piecewise quasi-stationary segments. Once the "start" and the "end" points of a QSS are known, all the speech samples coming from this QSS are analyzed within that window, resulting in (variable-scale) acoustic vectors.

Working under the similar framework, Brandt (Brandt, 1983) had proposed a maximum likelihood algorithm for speech segmentation. However, there are certain subtle theoretical as well practical differences in the proposed approach and the Brandt's algorithm which are described in Section 3.3. Brandt's approach was again followed in (Obrecht, 1988), where the authors proposed several speech segmentation algorithms. However, none of these papers (Obrecht, 1988; Brandt, 1983) attempted to perform stationary spectral analysis as has been done in this paper. Using a parametric model that the speech signal is generated by a time-varying auto-regressive process, we have shown the relationship between ML segmentation and piece-wise stationary spectral analysis in Section 3.4. Although, there has been plenty of research on speech signal segmentation (including speaker change detection), quite limited work has been done to interlink signal segmentation and quasi-stationary spectral analysis as has been done in this work.

In (Obrecht, 1988), the author has illustrated certain speech waveforms with segmentation boundaries overlaid. The validity of their algorithm is shown by a segmentation experiment, which on an average, segments phonemes into 2.2 segments. This result is quite useful as a pre-processor for the manual transcription of speech signals. However, the author in (Obrecht, 1988) did not dis-

[1] Equivalent to the detection of the transition point between the two adjoining QSSs.

[2] The power of the residual signal normalized by the number of samples in the window

cuss or extend the ML segmentation algorithm as a variable-scale quasi-stationary spectral analysis technique suitable for ASR, as done in the present work.

In (Atal, 1983), Atal has described a temporal decomposition technique, with applications in speech coding, to represent the continuous variation of the LPC parameters as a linearly weighted sum of a number of discrete elementary components. These elementary components are designed such that they have the minimum temporal spread (highly localized in time) resulting in superior coding efficiency. However, the relationship between the optimization criterion of "the minimum temporal spread" and the quasi-stationarity is not obvious. Therefore, the discrete elementary components are not necessarily quasi-stationary and vice-versa.

Coifman et al (Coifman and Wickerhauser, 1992) have described a minimum entropy basis selection algorithm to achieve the minimum information cost of a signal relative to the designed orthonormal basis. In (Srinivasan and Kleijn, 2004), Srinivasan et. al. have proposed a multi-scale QSS technique for noisy speech enhancement which is based on Coifman's technique (Coifman and Wickerhauser, 1992). In (Svendsen et al., 1989), Svendsen et al have proposed a ML segmentation algorithm using a single fixed window size for speech analysis, followed by a clustering of the frames which were spectrally similar for sub-word unit design. We emphasize here that this is different from the approach proposed here where we use variable size windows to achieve the objective of piecewise quasi-stationary spectral analysis. More recently, Achan et al (Achan et al., 2004) have proposed a segmental HMM for speech waveforms which identifies waveform samples at the boundaries between glottal pulse periods with applications in pitch estimation and time-scale modifications.

Our emphasis in this work is on better spectral modeling of the speech signal rather than achieving better coding efficiency or reduced information cost. Nevertheless, we believe that these two objectives are somewhat fundamentally related. The main contribution of this chapter is to demonstrate that the variable-scale QSS spectral analysis technique can possibly improve the ASR performance as compared to the fixed scale spectrum analysis. Moreover, we show the relationship between the maximum likelihood QSS detection algorithm and the well known spectral matching property of the LP error measure (Makhoul, 1975). This chapter is organized as follows. Section 3.2 formulates the ML detection problem for identifying the transition points between QSS. Section 3.3 compares the proposed approach with Brandt's (Brandt, 1983) approach. In Section 3.4, we illustrate an analogy of the proposed technique with spectral matching property of the LP error measure. Finally, the experimental setup and results are described in Section 3.5.

3.2 ML Detection of the change-point in an AR Gaussian random process

Consider an instance of a p^{th} order AR Gaussian process, $\mathbf{x}[n], n \in [1, N]$ whose generative LP filter parameters can either be $\mathbf{A}_0 = [1, a_0(1), a_0(2)....a_0(p)]$ or can change from $\mathbf{A}_1 = [1, a_1(1), a_1(2)....a_1(p)]$ to $\mathbf{A}_2 = [1, a_2(1), a_2(2)....a_2(p)]$ at time n_1 where $n_1 \in [1, N]$. As usual, the excitation signal is assumed to be drawn from a white Gaussian process and its power can change from $\sigma^2 = \sigma_1^2$ to $\sigma^2 = \sigma_2^2$. The general form of the Power Spectral Density (PSD) of this signal is then known to be

$$P_{xx}(f) = \frac{\sigma^2}{|1 - \sum_{i=1}^{p} a(i) \exp(-j2\pi i f)|^2} \tag{3.1}$$

where $a(i)$s are the LPC parameters. The hypothesis test consists of:

- $\mathbf{H}_0$: No change in the PSD of the signal $x(n)$ over all $n \in [1, N]$, LP filter parameters are $\mathbf{A}_0$ and the excitation (residual) signal power is σ_0^2.

$$x(n) = \sum_{k=1}^{k=p} a_0(k)x(n-k) + e_0(n),\ n \in [1, N] \tag{3.2}$$

where $e_0(n)$ is drawn from a white Gaussian noise process with power(variance) σ_0^2.

- $\mathbf{H}_1$: Change in the PSD of the signal $x(n)$ at n_1, where $n_1 \in [1, N]$, LP filter parameters change from $\mathbf{A}_1$ to $\mathbf{A}_2$ and the excitation(residual) signal power changes from σ_1^2 to σ_2^2.

$$\begin{array}{r} x(n) = \sum_{k=1}^{k=p} a_1(k)x(n-k) + e_1(n),\ n \in [1, n_1] \\ x(n) = \sum_{k=1}^{k=p} a_2(k)x(n-k) + e_2(n),\ n \in [n_1+1, N] \end{array} \tag{3.3}$$

where $e_1(n)$ and $e_2(n)$ are drawn from independent white Gaussian noise processes of powers σ_1^2 and σ_2^2 respectively.

Let, $\hat{\mathbf{A}}_0$ denote the maximum likelihood estimate (MLE) of the LP filter parameters and $\hat{\sigma}_0^2$ denote the MLE of the residual signal power under the hypothesis $\mathbf{H}_0$. The MLE estimate of the filter parameters is equal to their MMSE estimate due to the Gaussian distribution assumption (Itakura, 1975) and, hence, can be computed using the Levinson Durbin algorithm (Haykin, 1993) without significant computational cost.

Let $\mathbf{x}_1$ denote $[x(1), x(2), ...x(n_1)]$ and $\mathbf{x}_2$ denote $[x(n_1+1), ...x(N)]$. Under the hypothesis $\mathbf{H}_1$, ($\hat{\mathbf{A}}_1$, $\hat{\sigma}_1^2$) are the MLE of ($\mathbf{A}_1$, σ_1^2) estimated on $\mathbf{x}_1$, and ($\hat{\mathbf{A}}_2$, $\hat{\sigma}_2^2$) are the MLE of ($\mathbf{A}_2$, σ_2^2) estimated on $\mathbf{x}_2$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ have been assumed to be independent of each other. A Generalized Likelihood Ratio Test (GLRT) (Kay, 1998) would then pick hypothesis $\mathbf{H}_1$ if

$$\log L(\mathbf{x}) = \log(\frac{p(\mathbf{x}_1|\hat{\mathbf{A}}_1, \hat{\sigma}_1)p(\mathbf{x}_2|\hat{\mathbf{A}}_2, \hat{\sigma}_2)}{p(\mathbf{x}|\hat{\mathbf{A}}_0, \hat{\sigma}_0)}) > \gamma \tag{3.4}$$

where γ is a decision threshold that will have to be tuned on some development set. Given that the total number of samples in $\mathbf{x}_1$ and $\mathbf{x}_2$ is the same as in $\mathbf{x}_0$, their likelihoods can be compared directly in (3.4). Under the hypothesis $\mathbf{H}_0$ the entire segment $\mathbf{x} = [x(1)...x(N)]$ is considered stationary and the MLE $\hat{\mathbf{A}}_0$ is computed via the Levinson-Durbin algorithm using all the samples in segment $\mathbf{x}$. It can be shown that the MLE $\hat{\sigma}_0^2$ is the power of the residual signal (Kay, 1998; Itakura, 1975). Under $\mathbf{H}_1$, we assume that there are two distinct QSS, namely $\mathbf{x}_1$ and $\mathbf{x}_2$. The MLE $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{A}}_2$ are computed via the Levinson-Durbin algorithm using samples from their corresponding QSS. MLE $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are computed as the power of the corresponding residual signals. In fact, $p(\mathbf{x}|\hat{\mathbf{A}}_0, \hat{\sigma}_0)$ is equal to the probability of residual signal obtained using the filter parameters $\hat{\mathbf{A}}_0$, yielding:

$$p(\mathbf{x}|\hat{\mathbf{A}}_0, \hat{\sigma}_0) \quad = \quad \frac{1}{(2\pi\hat{\sigma}_0^2)^{N/2}} \exp\left[\frac{-1}{2\hat{\sigma}_0^2}\sum_{n=1}^{N}(e_0^2(n))\right] \tag{3.5}$$

where $e_0(n)$ is the residual error and

$$e_0(n) = x(n) - \sum_{i=1}^{p} a_0(i)x(n-i),\ \ n \in [1, N]$$

and

$$\hat{\sigma}_0^2 = \frac{1}{N}\sum_{n=1}^{N} e_0^2(n)$$

Similarly, $p(\mathbf{x}_1|\hat{\mathbf{A}}_1, \hat{\sigma}_1)$ and $p(\mathbf{x}_2|\hat{\mathbf{A}}_2, \hat{\sigma}_2)$ are the likelihoods of the residual signal vectors of the AR models $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively, and have the same functional forms as above. Substituting these expressions into (3.4) and simplifying yields

$$\log L(\mathbf{x}) = \frac{1}{2}\log\left[\frac{\hat{\sigma}_0^N}{\hat{\sigma}_1^{n_1}\hat{\sigma}_2^{(N-n_1)}}\right] \tag{3.6}$$

In the present form, the likelihood ratio (LR) $\log L(\mathbf{x})$ has now a natural interpretation. Indeed, if there is a transition point in the segment $\mathbf{x}$ then it has, in effect, $2p$ degrees of freedom. Under hypothesis $\mathbf{H}_0$, we encode $\mathbf{x}$ using only p degrees of freedom (LP parameters $\hat{\mathbf{A}}_0$) and, therefore, the coding (residual) error $\hat{\sigma}_0^2$ will be high. However, under hypothesis $\mathbf{H}_1$, we use $2p$ degrees of freedom (LP parameters $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{A}}_2$) to encode $\mathbf{x}$. Therefore, the coding (residual) errors $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ can be minimized to reach the lowest possible value.[3] This will result in $L(\mathbf{x}) > 1$. On the other hand, if there is no AR switching point in the segment $\mathbf{x}$ then it can be shown that, for large n_1 and N, the coding errors are all equal ($\hat{\sigma}_0^2 = \hat{\sigma}_1^2 = \hat{\sigma}_2^2$). This will result in $L(\mathbf{x}) \simeq 1$.

Distinction with BIC: Bayesian Information Criterion (BIC) (Ajmera et al., 2004) in association with a penalty term is widely used while comparing the likelihoods from two different probability density functions *(pdf)* which may have different number of parameters, for instance, the number of Gaussians in the Gaussian mixture model (GMM) may be different in the two *pdfs* that we may be comparing. However, at this stage, we will like to emphasize the point that the prediction order 'p' in the autoregressive (AR) processes, that we are discussing in this chapter, is not a parameter in the same sense as that of the number of Gaussians in the GMM. In fact, unlike being the number of parameters of a *pdf*, the autoregressive order 'p' is the generative filter's parameter as in (3.2) and (3.3). Therefore there is no direct comparison of the proposed GLRT with the BIC.

An interesting and useful property of the likelihood ratio (LR) in (3.6) is that it is invariant to any multiplicative scale factor of signal $\mathbf{x}$. For example let us consider a scaled segment $\mathbf{y} = c \times \mathbf{x}$, where c is a constant. As the LP filter and the inverse LP filter are both linear filters, a scaled input signal will result in an output signal with the same multiplicative scale factor. Therefore, the residual signals obtained after analyzing $\mathbf{y}$ will be $e_0^y(n) = c \times e_0(n)$, $e_1^y(n) = c \times e_1(n)$, $e_2^y(n) = c \times e_2(n)\ \forall n \in [1, N]$. Therefore the LR in (3.6) will become,

$$\begin{aligned} \log L(\mathbf{y}) &= \frac{1}{2} \log \left[\frac{c^N \hat{\sigma}_0^N}{c^{n_1} \hat{\sigma}_1^{n_1} c^{N-n_1} \hat{\sigma}_2^{(N-n_1)}} \right] \\ &= \log L(\mathbf{x}) \end{aligned} \tag{3.7}$$

This ensures that even if the speech signal might have varying power levels (different scale factors) LR in (3.7) can still be compared to a fixed threshold γ. However, if there are abrupt energy changes within a segment, then the LR will most likely classify them as different QSSs.

An example is illustrated in Fig. 3.1. The top pane shows a segment of a voiced speech signal. In the bottom figure, we plot the LR as the function of the hypothesized change over point n. Whenever, the right window i.e the segment $\mathbf{x}_2$ spans the glottal pulse in the beginning of the window, the LR exhibits strong downward spikes (negative values of the LR), which is due to the fact that the LP filter cannot predict large samples occurring in the beginning of the window. However, these negative spikes of the LR do not affect our decision as we are comparing LR to a large positive threshold (typically 4.5). Therefore, in a way, we are comparing only the positive envelope of the LR to a pre-selected positive threshold. Consequently, the sharp negative spikes in LR caused due to the occurrence of a glottal pulse in right window i.e $\mathbf{x}_2$ will bring down the LR value, thus preventing the LR from exceeding a positive threshold. This has a desirable effect that as long as the pitch periods are nearly similar(stationary voiced segment), they will never be segmented into different QSSs. Instead, they will be glued together to form one QSS. The minimum sizes of the left and the right windows are 160 and 100 samples respectively and the reasons for this choice are explained in Section 3.5. This explains the zero value of the LR at the beginning and the end of the whole test segment. The LR peaks around sample 500 which marks a strong AR model switching point. In Fig. 3.2, we plot the LR of an unvoiced speech segment that consists of two QSSs. As can be seen

[3] When $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{A}}_2$ are estimated, strictly based on the samples from the corresponding quasi-stationary segments.

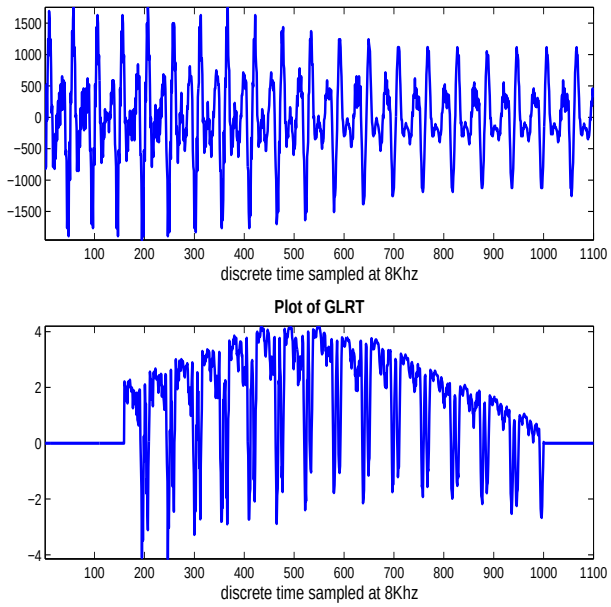

Figure 3.1. Typical plot of the log likelihood ratio for a voiced speech segment. The sharp downward spikes in the LR are due to the presence of a glottal pulse at the beginning of the right analysis window ($\mathbf{x}_2$). The LR peaks around the sample 500 which marks as a strong AR model switching point

from Fig. 3.2, in the case of the unvoiced speech, the LR has a rather smooth envelope due to the absence of the glottal pulses. The LR peaks around sample 200 that marks an AR model switching point. The algorithm presented in this chapter does not make any distinction between voiced and unvoiced speech segments. LR of all the segments in an utterance are compared to a fixed threshold that has been tuned on a development set. This results in a sequence of speech segments that are usually of variable lengths. We note that the segments returned by the algorithm are quasi-stationary only in a probabilistic sense that the event that two adjacent segments are instances of the same stationary process is $e^{-Threshold}$ times as likely as the event that they are instances of two different stationary processes. Therefore the choice of the threshold decides the trade-off between false acceptance and false rejection of QSSs. However, as we are primarily interested in improved recognition accuracies, we have tuned the threshold on a development set based on the recognition accuracies.

3.3 Comparison with Brandt's algorithm

The likelihood ratio (LR) in both Brandt's approach and the proposed approach are the same as in (3.6). This is not surprising as in both the approaches LR is a maximum likelihood solution under the assumption that the speech signal is a realization of a Gaussian AR process where the AR parameters can change over time. However, the differences lie in the methods employed to estimate the residual powers $\sigma_0, \sigma_1, \sigma_2$. In our approach, the AR parameters $\mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2$ and the residual powers $\sigma_0, \sigma_1, \sigma_2$. are estimated by solving the least squares equations (Haykin, 1993) over their corresponding segments, $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$ respectively. To solve these least squares equations, we use the so-called autocorrelation method (Haykin, 1993)(page:486) that leads to an autocorrelation matrix that is Toeplitz. Toeplitz matrices can be inverted quite efficiently using the Levinson Durbin algorithm (Haykin, 1993)(pages:254). Therefore the computational overload of our approach is quite

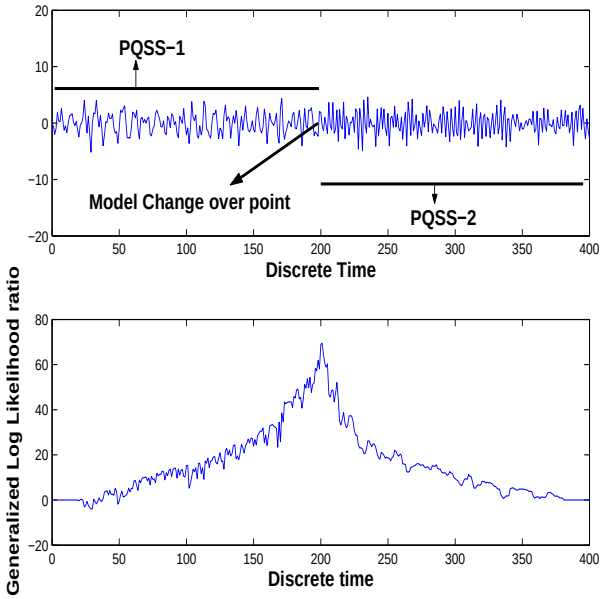

Figure 3.2. Typical plot of the log likelihood ratio for an unvoiced speech segment that consists of two piece-wise quasi-stationary segments (PQSS). The LR peaks around the sample 200 which is indeed an AR model switching point.

low.

Whereas, Brandt used the so-called covariance method (Haykin, 1993)(page:486) to solve for the AR parameters and the residual powers, $\sigma_0, \sigma_1, \sigma_2$.. This method leads to an autocorrelation matrix that is non-Toeplitz, thus excluding the use of fast Levinson-Durbin algorithm. Brandt has used the lattice-filters (Haykin, 1993)(pages:280) to estimate the AR parameters and the residual powers. Lattice filters are also quite efficient as compared to the Gram-Schmidt orthogonalization, but less so as compared to Levinson-Durbin algorithm. Therefore, the proposed approach is faster than Brandt's approach. Use of the autocorrelation method (Haykin, 1993) guarantees a minimum phase all-pole filter (Haykin, 1993). However, the covariance method (Haykin, 1993) does not necessarily lead to a minimum phase all-pole filter. Therefore, the proposed approach ensures a stable all-pole filter as opposed to Brandt's approach.

In Brandt's algorithm, the left window i.e $\mathbf{x}_1$ uses a growing memory covariance ladder algorithm and the right window $\mathbf{x}_2$ uses a sliding memory covariance ladder algorithm. Initialization of the growing memory covariance ladder algorithm requires certain intermediate quantities that are provided by the sliding memory covariance ladder algorithm which operates on $\mathbf{x}_2$. Hence, the AR parameters $\mathbf{A}_1$ and the residual power σ_1 are indirectly influenced by the samples in the right window $\mathbf{x}_2$. To compensate for this, Brandt's algorithm uses a second search called "jump time optimization process" to estimate the stationarity change-over point.

Whereas, in our approach the AR parameters $\mathbf{A}_1, \mathbf{A}_2$ and the residuals σ_1, σ_2 are estimated using samples strictly from their corresponding segments i.e. $\mathbf{x}_1, \mathbf{x}_2$ respectively. Therefore, in the proposed approach there is just one step stationarity change point detection. Whenever the LR in (3.6) exceeds the threshold γ, a stationarity change point is recorded. This is in contrast to Brandt's method where this step is followed by a "jump time optimization process" which, finally estimates the stationarity change point.

3.4 Relation of the generalized likelihood ratio test (GLRT) to Spectral Matching

In the section will show the relationship between the maximum likelihood segmentation and the spectral matching which is one of the main contributions of this work. As is well known the LP error measure possesses the spectral matching property (Makhoul, 1975). Specifically, given a speech segment $\mathbf{x}$, let its power spectrum (periodogram) be denoted by $\mathbf{S}(e^{j\omega})$. Let the all pole model spectrum of the segment $\mathbf{x}$ be denoted as $\hat{\mathbf{S}}_0(e^{j\omega})$. Then it can be shown that the MMSE error σ_0^2 of the LP filter estimated over the entire segment $\mathbf{x}$ is given by (Makhoul, 1975).

$$\sigma_0^2 = \int_{-\pi}^{\pi} \frac{\mathbf{S}(e^{j\omega})}{\hat{\mathbf{S}}_0(e^{j\omega})} d\omega \text{ where,} \tag{3.8}$$

$$\hat{\mathbf{S}}_0(e^{j\omega}) = \frac{1}{|1 - \sum_{i=1}^{p} a_0(i) \exp(-j2\pi i f)|^2} \tag{3.9}$$

Therefore minimizing the residual error σ_0^2 is equivalent to the minimization of the integrated ratio of the signal power spectrum $\mathbf{S}(e^{j\omega})$ to its approximation $\hat{\mathbf{S}}_0(e^{j\omega})$ (Makhoul, 1975). Substituting (3.8) in (3.6) we obtain,

$$\log L(\mathbf{x}) = \frac{1}{2} \log \frac{\left(\int_{-\pi}^{\pi} \frac{\mathbf{S}(e^{j\omega})}{\hat{\mathbf{S}}_0(e^{j\omega})} d\omega \right)^N}{\left(\int_{-\pi}^{\pi} \frac{\mathbf{S}_1(e^{j\omega})}{\hat{\mathbf{S}}_1(e^{j\omega})} d\omega \right)^{n_1} \left(\int_{-\pi}^{\pi} \frac{\mathbf{S}_2(e^{j\omega})}{\hat{\mathbf{S}}_2(e^{j\omega})} d\omega \right)^{N-n_1}} \tag{3.10}$$

where, $\mathbf{S}(e^{j\omega})$, $\mathbf{S}_1(e^{j\omega})$ and $\mathbf{S}_2(e^{j\omega})$ are the power spectra of the segments $\mathbf{x}$, $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. Similarly $\hat{\mathbf{S}}_0(e^{j\omega})$, $\hat{\mathbf{S}}_1(e^{j\omega})$ and $\hat{\mathbf{S}}_2(e^{j\omega})$ are the MMSE p^{th} order all-pole model spectra estimated over the segments $\mathbf{x}$, $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. Therefore, $\hat{\mathbf{S}}_0(e^{j\omega})$, $\hat{\mathbf{S}}_1(e^{j\omega})$ and $\hat{\mathbf{S}}_2(e^{j\omega})$ are the best LP spectral matches to their corresponding power spectra. One way of interpreting (3.10) is that it is a measure of the relative goodness between the best LP spectral match achieved by modeling $\mathbf{x}$ as a single QSS and the best LP spectral matches obtained by assuming $\mathbf{x}$ to consist of two distinct QSS, namely $\mathbf{x}_1$ and $\mathbf{x}_2$. This is further explained as follows. If $\mathbf{x}_1$ and $\mathbf{x}_2$ are indeed two distinct QSS, then $\mathbf{S}_1(e^{j\omega})$ and $\mathbf{S}_2(e^{j\omega})$ will be quite different and $\mathbf{S}(e^{j\omega})$ will be a gross average of these two spectra. In other words, the frequency support of $\mathbf{S}(e^{j\omega})$ will be a union of those of the $\mathbf{S}_1(e^{j\omega})$ and $\mathbf{S}_2(e^{j\omega})$. $\hat{\mathbf{S}}_1(e^{j\omega})$ and $\hat{\mathbf{S}}_2(e^{j\omega})$, having p poles each, will match their corresponding power spectra reasonably well, resulting in a lower value of the denominator in (3.10). However, $\hat{\mathbf{S}}_0(e^{j\omega})$ will be a relatively poorer spectral match to $\mathbf{S}(e^{j\omega})$ as it has only p poles to account for the wider frequency support. Therefore we incur a higher spectral mismatch by assuming $\mathbf{x}$ to be a single QSS when in fact it is composed of two distinct QSS $\mathbf{x}_1$ and $\mathbf{x}_2$. This results in the LR $\log L(\mathbf{x})$ taking up a high value. Whereas if $\mathbf{x}_1$ and $\mathbf{x}_2$ are the instances of the same quasi-stationary process, then so is $\mathbf{x}$. Therefore $\mathbf{S}_1(e^{j\omega})$, $\mathbf{S}_2(e^{j\omega})$ and $\mathbf{S}(e^{j\omega})$ are nearly the same with similar all-pole models, resulting in a value of the LR close to zero. The above discussion points to the fact that the QSS analysis based on the proposed LR is constantly striving to achieve a better time varying spectral modeling of the underlying signal as compared to single fixed scale spectral analysis. However, the above discussion is true only if the speech signal can be assumed to have been generated from a Gaussian AR process of a fixed order p, where the AR parameters can change over time. This limitation arises from the fact that one needs to assume a signal generative model based on which one can develop a criterion of stationarity.

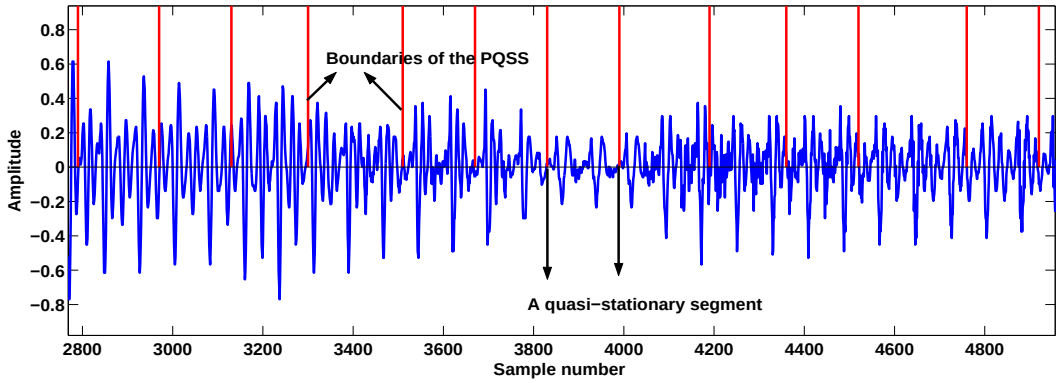

Figure 3.3. Quasi stationary segments (QSS) of a speech signal as detected by the algorithm with $\gamma = 4.5$ and LP order $p = 10$.

3.5 Experiments and Results

We have used the LR L($\mathbf{x}$) in (3.6) to perform QSS spectral analysis of speech signals for ASR applications. We initialize the algorithm with a left window size W_L = 20ms and a right window size W_R = 12.5ms. We compute their corresponding MMSE residuals and the MMSE residual of the union of the two windows using the Levinson-Durbin algorithm. Then, the LR is computed using (3.6) and is compared to the threshold. The choice of the threshold $\gamma = 4.5$ was decided on the basis of ASR results on a development set. In figure (3.3), we illustrate the boundaries of the QSS as detected by the algorithm with $\gamma = 4.5$. In general, the ASR results are slightly sensitive to the threshold, although not in a huge way. If the LR is greater than the threshold γ, W_L is considered the largest possible QSS and we obtain a spectral estimate using all the samples in W_L. Otherwise,W_L is incremented by INCR=1.25ms and the whole process is repeated until LR exceeds γ or W_L becomes equal to the maximum window size WMAX=$60ms$. The computation of a MFCC feature vector from a very small segment (such as 10ms) is inherently very noisy.[4] Therefore, the minimum duration of a QSS as detected by the algorithm was constrained to be $20ms$. Ideally the right window size W_R should be as small as possible so that we can instantaneously detect a stationarity change point. However, a reliable estimate of the AR parameters and the corresponding residual signal requires sufficiently large number of samples in the analysis window. Therefore as a compromise between these two opposing factors, we have chosen the W_R $= 12.5ms$. Throughout the experiments, a fixed LP order $p = 10$ was used. The likelihood ratio test is quite widely used for speaker segmentation (Ajmera et al., 2004) where the average length of a single speaker segment may last from 1sec to several seconds. This provides a relatively large amount of samples to estimate the parameters of the probability density functions as compared to the present problem where we have to first estimate the generative AR parameters and the corresponding residuals to detect stationarity change over point within 20ms to 60ms. Moreover, in speaker change detection one can use the apriori-information that a speaker will at least speak for a second or so. Therefore most of the time it can be safely assumed that there will not be more than two speakers within one second long speech segment. Hence, a local maxima of the LR within a time-span of one second can be used as a speaker-change point. This approach has been successfully used in (Ajmera et al., 2004). However, in our case the QSSs can have much more variable durations ranging from 3ms to 80ms[5]. Therefore, there is no

[4] Due to very few DFT samples falling under the the Mel-filter bins resulting in high variance of the mel-filter bank energies

[5] However, as we require sufficiently large number of samples to reasonably estimate the AR parameters and the residuals to compute the LR, the proposed algorithm can only detect QSSs larger than and equal to 20ms

minimum duration in which we can assume that only two QSSs will be present, thus excluding the use of local maxima of the LR as an estimate of the stationarity change-over point.

Before proceeding further, however, we feel necessary to briefly discuss certain inconsistencies between variable-scale spectral analysis and state-of-the-art Hidden Markov models ASR using Gaussian mixture models (HMM-GMM). HMM-GMM systems typically use spectral features based on a constant window size (typically $20ms$) and a constant shift size (typically $10ms$). The shift size determines the Nyquist frequency of the cepstral modulation spectrum (Tyagi et al., 2003), which is typically measured by the delta features of the static MFCC or PLP features. In a variable-scale piecewise quasi-stationary analysis, the shift size should preferably be equal to the size of the detected QSS. Otherwise, if the shift size is x% of the duration of the QSS, then the next detected QSS will be the same but of duration $(100-x)$% and the following one will be of duration $(100-2x)$% and so on until we have shifted past the entire duration of the QSS. This results in the undesirable effect that the same QSS gets analyzed by successively smaller windows, hence increasing the variance of the feature vector of this QSS. On the other hand, the use of a shift size equal to the variable window size will change the Nyquist frequency of the cepstral modulation spectrum (Tyagi et al., 2003). Therefore, the modulation frequency pass-band of the delta filters (Tyagi et al., 2003) will vary from frame to frame and may suffer from aliasing for shift sizes in excess of $20ms$.

To avoid fluctuating Nyquist frequency of the cepstral modulation spectrum (Tyagi et al., 2003), a fixed shift size of $12.5ms$ was used in the algorithm. As explained above, this sometimes resulted in the undesirable effect that the same QSS gets analyzed by progressively smaller windows. To alleviate this problem, the zeroth cepstral coefficient $c(0)$, which is a non-linear function of the windowed signal energy and, hence, of the window size, was normalized such that its dependence on the window size is minimized.

We believe that in order to realize the true potential of variable scale piece-wise quasi-stationary analysis, we will have to research new statistical modeling techniques that can handle the spectral vectors derived from the variable sized segments in a suitable way. One example could be Dynamic Bayesian Networks(DBNs) (Stephenson et al., 2004) where the length of the QSS can be an auxiliary (Stephenson et al., 2004) variable that conditions the emitted MFCC vector. However, this discussion is beyond the scope of our present work.

3.5.1 OGI Numbers95 database

In order to assess the effectiveness of the proposed algorithm, speech recognition experiments were conducted on the OGI Numbers corpus (Cole et al., 1994). This database contains spontaneously spoken free-format connected numbers over a telephone channel. The lexicon consists of 31 words. Figure (3.4) illustrates the distribution of the QSSs as detected by the proposed algorithm. Nearly 47% segments were analyzed with the smallest window size of $20ms$ and they mostly corresponded to short-time limited segments. However, voiced segments and long silences were mostly analyzed by using longer windows in the range $30ms - 60ms$. The short peak at 60ms is due to the accumulated value over all the segments that should have been longer than 60ms but were constrained by our choice of the largest window size.

Throughout the experiments, MFCC coefficients and their temporal derivatives were used as speech features. However, three feature sets were compared and they are listed below

1. [39 dim. MFCC:] computed over a fixed window of length 20ms.
2. [39 dim. MFCC:] computed over a fixed window of length 50ms.
3. [Variable-scale QSS MFCC+Deltas:] For a given frame, the window size in the range (20,60ms), with increments of 1.25 ms, is dynamically chosen using the proposed algorithm ensuring that the windowed segment is the largest quasi-stationary segment.

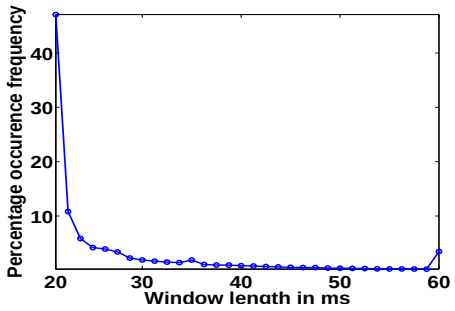

Figure 3.4. Distribution of the QSS window sizes detected and then used in the training set

Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition systems were trained using public domain software HTK (Young et al., 1995) on the clean training set from the original Numbers corpus. The speech recognition results in clean conditions for various spectral analysis techniques are given in table 3.1. The fixed scale MFCC features using 20ms and 50ms long analysis windows have 5.7% and 5.9% word error rate (WER) respectively. The proposed variable-scale system which adaptively chooses a window size in the range [20ms, 60ms], followed by the usual MFCC computation, has a 5.0% WER. This corresponds to a relative improvement of nearly 10% over the fixed scale features.

Table 3.1. Word error rate in clean conditions

MFCC 20ms	5.7
MFCC 50ms	5.9
Proposed Variable-scale QSS MFCC	5.0

3.5.2 University of Oldenburg non-sense syllable database

OLLO (Wesker et al., 2005) database has been specially designed to study the effects of the intrinsic variabilities (speaking rate and speaking style) present in the usual speech signal. This database is rich in various speech variabilities such as different speaking styles (slow, fast, statement, questioning, loud and soft) and with almost equal sampling of the male and female speakers. The lexicon consists of 150 logatomes[6] which are either CVCs or VCVs such as uttu, acsha, atta etc. The train set and test-set consists of roughly 13,500 and 13,800 utterances respectively and they correspond to the no-accent part of the database[7]. The experiments reported in this thesis are for the entire logatome recognition on the No-accent train/test parts of the OLLO database. Each of these utterances correspond to an instance of a logatome.

Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition systems were trained using public domain software HTK on the NO-accent part of the OLLO (Wesker et al., 2005) training set. Three state left to right HMM models were trained for each of the 26 phonemes in the OLLO database including silence as well. The lexicon consists of 150 logatomes. The ASR performance of the baseline system trained on the MFCCs saturated around 50-65 Gaussians per mixture and hence Gaussian mixture models(GMMS) with 65 Gaussians per state and diagonal covariance matrices were used to model the emission probability densities of the feature

[6]A logatome is CVC or VCV that consists of nonsense syllables

[7]The no-accent portion of the database correspond to the normal German accent

vectors. The number of parameters used in the HMM-GMM system is the same for all the features reported. The logatome recognition results for various features are given in Table 3.2.

As can be noted from the Table 3.2, the proposed Variable-scale QSS MFCCs have significantly improved recognition accuracies for the Slow, Questioning and Normal speech. Whereas for the Fast , Loud and Soft speech, the performance gains are insignificant. When averaged over all the variabilities, the proposed techniques achieves an absolute improvement of 1.2% . We believe that in order to realize the true potential of variable scale piece-wise quasi-stationary analysis, we will have to research new statistical modeling techniques that can handle the spectral vectors derived from the variable sized segments in a suitable way. One potential candidate could be Dynamic Bayesian Networks(DBNs) citepToddDBN where the length of the QSS can be an auxiliary (Stephenson et al., 2004) variable that conditions the emitted MFCC vector. However, this discussion is beyond the scope of our present work.

Table 3.2. Logatome recognition rates reported over each variabilities of the OLLO database.

Feature	Fast	Slow	Loud	Soft	Questioning	Normal	Overall
MFCC 20ms	71.3	76.0	76.9	64.6	80.1	79.6	74.7
Proposed Variable-scale QSS MFCC	71.7	77.4	77.0	66.4	81.3	81.6	75.9
Absolute Improvement	0.4	1.4	0.1	-0.2	1.2	2.0	1.2

3.6 Summary

In this chapter, we have demonstrated that the variable-scale piecewise quasi-stationary spectral analysis of speech signal can possibly improve the recognition accuracies of the state-of-the-art ASR systems in clean acoustic conditions. Such a technique can partially overcome the time-frequency resolution limitations of the fixed scale spectral analysis techniques. However, it can be argued that most of the frequency resolution is anyway lost due to the Mel-filter binning of the DFT samples. Nevertheless, a spectrum(DFT) estimated over a quasi-stationary segment will help to reduce the variance of the estimated Mel-filter bank energies and consequently those of the MFCC feature vectors. However, as we need certain minimum number of samples to estimate the AR parameters and the residuals, our algorithm cannot detect QSSs below 20ms. We believe that in order to realize the true potential of variable scale piece-wise quasi-stationary analysis, we will have to research new statistical modeling techniques that can handle the spectral vectors derived from the variable sized segments in a suitable way. One example could be Dynamic Bayesian Networks(DBNs) (Stephenson et al., 2004) where the length of the QSS can be an auxiliary (Stephenson et al., 2004) variable that conditions the emitted MFCC vector. However, this discussion is beyond the scope of our present work.

As the linear prediction analysis is very sensitive to noise, we expect the proposed technique to be sensitive to noisy speech. In particular if the acoustic model (HMM-GMM) have been trained using only the clean speech and the test utterances are noisy, then we can expect a mismatch between the LPC analysis of the train set and test set. Consequently, this will affect the values of the likelihood ratio in (3.4). As the threshold γ in (3.4) has been determined over a clean development set, this will lead to a mismatch whenever the test conditions are noisy. However, if the train and the development set include the noisy utterances as well, then this mismatch can be reduced. At this point, we will like to stress the point that the proposed technique is aimed toward improving the ASR performances in the clean acoustic conditions ($SNR \geq 18db$). To address the problem of noise robust ASR, this thesis proposed other features which are elaborated in chapters 5, 6 and 7.

As, the performance gains obtained through the variable-scale QSS analysis were a bit modest,

this led us to design new features that can inherently describe the non-stationary signals such as speech. Amplitude modulations and Frequency modulations (AM-FM) (Haykin, 1994) of any given signal, can reasonably well model the non-stationarity inherent in that signal. In the next chapter we will study the AM-FM demodulation techniques that has been specially developed and designed for the analysis of the speech signals and in particular for its use as a feature vector in ASR. We will identify the shortcomings of the previously proposed modulation spectrum related techniques (Tyagi et al., 2003; Athineos et al., 2004; Zhu and Alwan, 2000; Kingsbury et al., 1998a; Kanedera et al., 1998). These previously published modulation spectrum related techniques have primarily improved the word recognition accuracies in noisy conditions but have had a much poorer performance than the MFCC features in the clean acoustic conditions.

We will outline the development and design of a novel feature (Fepstrum) that has been derived using a theoretically consistent AM-FM demodulation technique and has led to significant improvements in word recognition accuracies in the clean acoustic conditions.

Chapter 4

Fepstrum representation of speech signal

4.1 Introduction

In past several years, significant efforts have been made to develop new speech signal representations which can better describe the non-stationarity (spectral dynamics) inherent in the speech signal. Some representative examples are temporal patterns (TRAPS) features(Hermansky, 2003; Athineos et al., 2004) and the several modulation spectrum related techniques(Tyagi et al., 2003; Athineos et al., 2004; Zhu and Alwan, 2000; Kingsbury et al., 1998a; Kanedera et al., 1998). In TRAPS technique, temporal trajectories of spectral energies in individual critical bands over windows as long as one second are used as features for pattern classification.

The notion of the amplitude modulation (AM) and the frequency modulation (FM) were initially developed for the communication signals (Haykin, 1994). In theory, the AM signal modulates a narrow-band carrier signal (specifically, a monochromatic sinusoidal signal). Therefore to be able to extract the AM signals of a wide-band signal such as speech (typically 4KHz), it is necessary to decompose the speech signal into several narrow spectral bands where each band's output signal can be modeled as an AM signal modulating a single carrier (FM) signal. In the past (Tyagi et al., 2003; Athineos et al., 2004; Zhu and Alwan, 2000; Kingsbury et al., 1998a; Kanedera et al., 1998), the modulation spectrum that has been used as a feature vector for ASR has been defined and extracted in a slightly ad-hoc manner. For instance, several researchers have extracted the speech modulation spectrum by computing a discrete Fourier transform (DFT) of the Mel or critical band spectral energy trajectories, where each sample of the trajectory has been obtained through a power spectrum (followed by Mel filtering) over 20-30ms long windows. An illustration of this is provided in Fig.4.1. The major limitation of such a technique is that,

- It implicitly assumes that within each Mel or critical band, the amplitude modulation (AM) signal remains constant within the duration of the window length that is typically 20-30ms long.
- Instead of modeling the constantly and slowly changing amplitude modulation signal in each band, it mostly models the spurious and abrupt modulation frequency changes that occur due to the frame shifting of $10ms$.

In this chapter, we have proposed an algorithm to perform AM-FM demodulation of the speech

signal in the time domain. As the AM-FM signal model is defined in time domain[1], a demodulation in time domain leads to robust estimation of the continuously though slowly changing AM signals. It also leads to a better understanding of the relationships between various signal sub-components. Through examples, we will show that for a theoretically meaningful estimation of the AM signals, it is important to constrain the companion FM signal to be narrow-band. Similar arguments from the modulation filtering point of view as applied to speech coding, were presented by Schimmel and Atlas (Schimmel and Atlas, 2005). In their experiment, they consider a wide-band filtered speech signal $x(t) = a(t)c(t)$, where $a(t)$ is the AM signal and $c(t)$ is the broad-band carrier signal. Then, they perform a low-pass modulation filtering of the AM signal $a(t)$ to obtain $a_{LP}(t)$. The low-pass filtered AM signal $a_{LP}(t)$ is then multiplied with the original carrier $c(t)$ to obtain a new signal $\tilde{x}(t)$. They show that the acoustic bandwidth of the reconstructed signal $\tilde{x}(t)$, is not necessarily less than that of the original signal $x(t)$. This unexpected result is a consequence of the signal decomposition into wide spectral bands that results in a broad-band carrier (Schimmel and Atlas, 2005) and is explained below. Let us consider the original signal $x(t)$ and its Fourier transform $X(f)$ which is obtained as the convolution of the spectra of the AM and the carrier signals, namely $A(f)$ and $C(f)$.

$$x(t) = a(t)c(t),\ X(f) = A(f) * C(f) \tag{4.1}$$

The Fourier transform of the reconstructed signal $\tilde{x}(t)$ can be expressed as follows,

$$\tilde{x}(t) = a_{LP}(t)c(t),\ \tilde{X}(f) = A_{LP}(f) * C(f) \tag{4.2}$$

where '*' denotes convolution and $A(f)$, $A_{LP}(f)$ and $C(f)$ and are the Fourier transforms of the AM signal $a(t)$, its low-pass filtered version $a_{LP}(t)$ and the carrier signal $c(t)$. Now if $c(t)$ is a sinusoidal carrier i.e. $c(t) = sin(2\pi f_0 t)$ then it can be seen that the acoustic bandwidth of $\tilde{x}(t)$ is less than that of the original signal $x(t)$ which is the desired result due to the low-pass filtering of the AM signal $a(t)$. However, this is not necessarily the case if the carrier $c(t)$ is not sinusoidal and is broad-band. This can be seen as follows. As $x(t)$ has finite bandwidth (lets say $X(f)$ is non-zero only over the interval $[100 - 200]\ Hz$), therefore, the broad-band carrier spectrum $C(f)$ and the AM spectrum $A(f)$ have a special structure such that their convolution in (4.1) is zero outside the interval $[100 - 200]\ Hz$. The low-pass filtering operation [2] will not preserve this "special structure" between $A_{LP}(f)$ and $C(f)$. Therefore, the convolution in (4.2) is not guaranteed to be zero outside the interval $[100 - 200]Hz$, thus increasing the bandwidth of $x_{LP}(t)$ and defying the intent of the low-pass filter. Therefore, it is important to ensure that the carrier signal is narrow-band (ideally monochromatic). We realize that is not only a serious problem for modulation filtering as applied to speech coding (Schimmel and Atlas, 2005), but also for modulation spectrum analysis (which is used as feature vector for ASR and is the topic of this chapter). As a solution, we propose using narrow-band filters to decompose speech signal, followed by the AM signal estimation in each band, using analytic signals in time domain. The usefulness of this modification is further explained, later on in this chapter.

Over the past few decades, pole-zero transfer functions that are used for modeling the frequency response of a signal, have been well studied and understood (Atal and Hanauer, 1971; Makhoul, 1975; Haykin, 1993). In this work we will denote them by "F-PZ". Lately, Kumaresan and his colleagues (Kumaresan and Rao, 1999; Kumaresan, 1998) have proposed to model analytic signals (Haykin, 1994) using pole-zero models in the temporal domain (denoted by T-PZ to distinguish them from the F-PZ). Along similar lines, Athineos et al. (Athineos and Ellis, 2003; Athineos et al., 2004) have used the dual of the linear prediction in the frequency domain to improve upon the TRAP features.

[1] $x(t) = a(t)cos(\int_0^t 2\pi f(t)dt)$, here $x(t)$ is a narrow band-pass filtered speech signal where, $a(t)$ is the corresponding AM signal and $f(t)$ is the corresponding FM signal

[2] In fact, any filtering operation except for the identity one

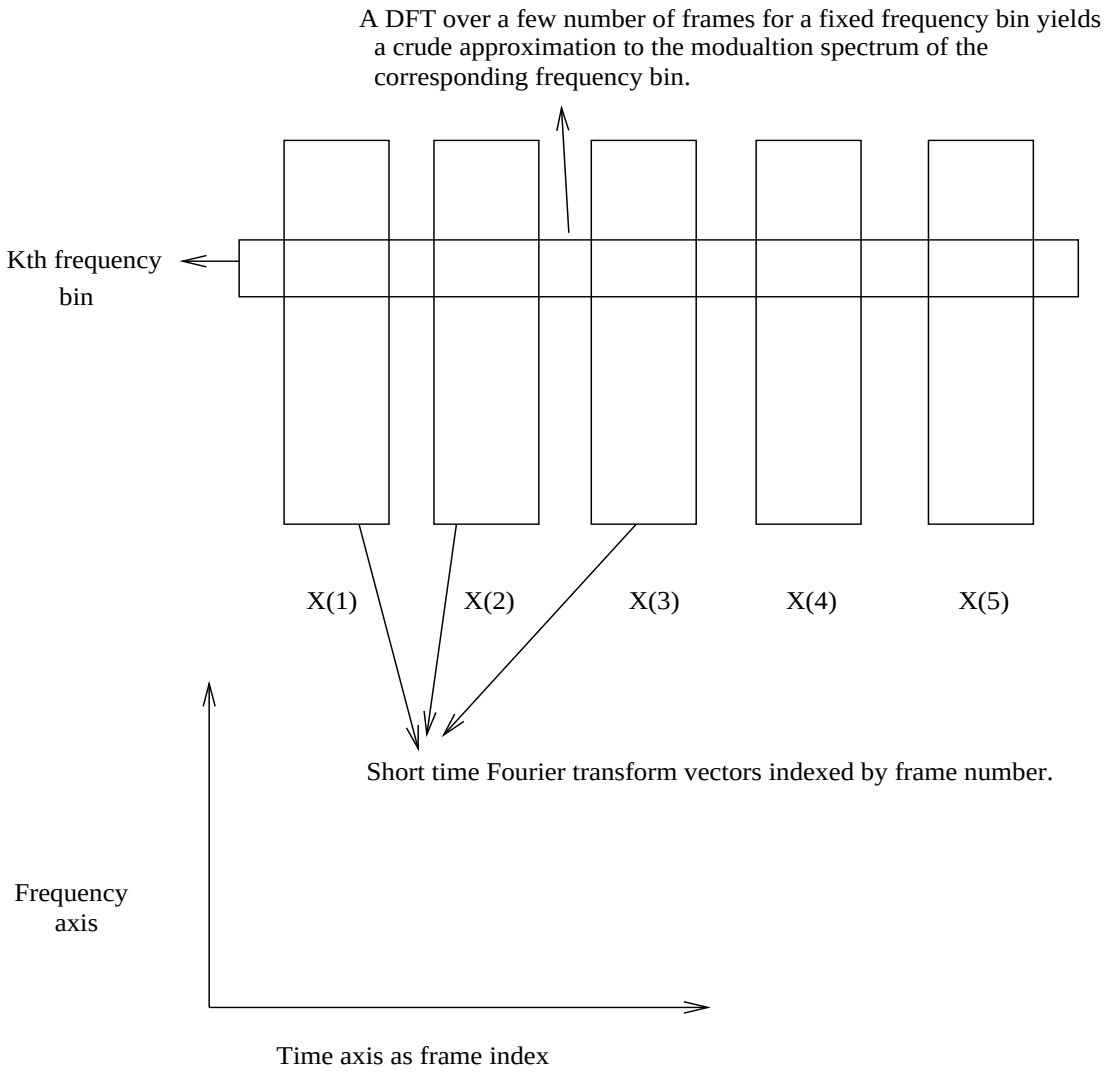

Figure 4.1. An approximate AM demodulation technique.

An inherent advantage of working with the analytic signal is that it elegantly allows the decomposition of an arbitrary signal (possibly non-stationary) into its amplitude modulation (AM) and frequency modulation (FM) signals. We make extensive use of T-PZ representation in this paper. For the sake of completeness and clarity, we state and prove several interesting time-frequency dualities for the analytic signals. These properties are then used to develop "meaningful" AM-FM decomposition of the speech signal.

4.2 Pole-zero models (elementary signals) in the temporal domain

Traditionally, the pole-zero transfer functions have been used to approximate a discrete time frequency response which is inherently periodic with a period of 2π. Voelcker and Kumaresan have

used the T-PZ to approximate analytic signals in the temporal domain. We recall that given a real periodic[3] signal $x(t)$ with period T seconds, its analytic version $s(t)$ is given by,

$$s(t) = x(t) + j\hat{x}(t) \tag{4.3}$$

where $\hat{x}(t)$ denotes the Hilbert transform of $x(t)$. If $x(t)$ is band-limited, then so is $s(t)$. Moreover $s(t)$ has non-zero spectrum for only positive frequencies. Therefore $s(t)$ can be expressed in terms of a finite number of Fourier series coefficients at positive frequencies.

$$s(t) = e^{j\omega_t t} \sum_{k=0}^{M} a_k e^{jk\Omega t} \tag{4.4}$$

where ω_t is an arbitrary frequency translation, $\Omega = 2\pi/T$ and M is sufficiently large. Noting that $s(t)$ is a polynomial, it can be factored in terms of T-PZ as follows,

$$s(t) = a_0 e^{j\omega_t t} \prod_{i=1}^{P}(1 - p_i e^{j\Omega t}) \prod_{l=1}^{Q}(1 - q_l e^{j\Omega t}) \tag{4.5}$$

where $P+Q = M$ and p_i and q_l are the complex roots, inside and outside the unit circle respectively. We note that this is a unit circle in the time domain, $e^{j\Omega t},\ t \in [0, T],\ \Omega = \frac{2\pi}{T}$. More generally, if s(t) is not band-limited, it can be represented using poles and zeros.

$$s(t) = a_0 e^{j\omega_t t} \frac{\prod_{i=1}^{P}(1 - p_i e^{j\Omega t})}{\prod_{i=1}^{U}(1 - u_i e^{j\Omega t})} \prod_{l=1}^{Q}(1 - q_l e^{j\Omega t}) \tag{4.6}$$

where, p_i and q_i are the zeros inside and outside the unit circle respectively. The poles u_i are guaranteed to be inside the unit circle as proved in the following lemma.

Lemma 1 *The T-PZ factorization of an analytic signal $s(t)$ has all the poles u_i inside the unit circle.*

Proof: Lets assume that there is a pole $r = |r|e^{j\phi}$ outside the unit circle, with $|r| > 1$. The expansion of $s(t)$ will then have a term,

$$\begin{aligned} \frac{A}{(1 - re^{j\Omega t})} &= \frac{-A}{re^{j\Omega t}} \frac{1}{1 - r^{-1}e^{-j\Omega t}} \\ &= \frac{-A}{re^{j\Omega t}} \sum_{k=0}^{\infty} r^{-k} e^{-jk\Omega t} \end{aligned} \tag{4.7}$$

where, A is a constant. (4.7) implies that $s(t)$ has non-zero spectrum for negative frequencies. This is in contradiction to the fact that $s(t)$ being an analytic signal has zero spectral energy for negative frequencies. Hence $|r| < 1$.

The importance of lemma 1 will become apparent later on. Let us now specify the dual analogues of three well known properties which are,

- Minimum-phase: Traditionally, minimum phase is a frequency domain phenomenon. A frequency response (F-PZ) is termed minimum-phase (F-MinP) if all its poles and zeros are inside the unit circle. Similarly, a T-PZ is called T-MinP if all its poles and zeros are inside the unit circle.

[3]This is not a limitation as in short-time Fourier analysis, we implicitly make the signal periodic with the base period equal to the T second long windowed segment.

- All-pass: Traditionally, all-pass is a frequency domain phenomenon. A frequency response, (F-PZ), is said to be all-pass (F-AllP) if its magnitude is unity at all frequencies. Similarly, a T-PZ is called T-AllP if it has unity magnitude for $t \in (-\infty, \infty)$.

- Causality: Traditionally, causality is a time-domain phenomenon. A signal $x(t)$ is said to be causal (T-causal) if it is non-zero only for the $t \geq 0$. Similarly, we define a frequency response to be F-causal if it is non-zero only for the $f \geq 0$. Therefore, an analytic signal is F-causal.

With these definitions in place, we are ready to describe the decomposition of an analytic signal $s(t)$ into its T-MinP and T-AllP part which will lead to its AM and FM parts. Therefore, reflecting the zeros q_i inside the unit circle, we get,

$$s(t) = a_0 e^{j\omega_t t} \underbrace{\frac{\prod_{i=1}^{P}(1-p_i e^{j\Omega t})}{\prod_{i=1}^{U}(1-u_i e^{j\Omega t})} \prod_{i=1}^{Q}(1-1/q_i^* e^{j\Omega t})}_{\textbf{T-MinP}} \times \prod_{i=1}^{Q}(-q_i^*) \underbrace{\prod_{i=1}^{Q} \frac{(e^{-j\Omega t}-q_i)}{(1-q_i^* e^{-j\Omega t})}}_{\textbf{T-AllP}} \tag{4.8}$$

We recall the following two well-known lemmas,

Lemma 2 *Given a frequency response (F-PZ)* $X(f)$ $= |X(f)|e^{j\phi(f)}$*, its phase response* $\phi(f)$ *is the Hilbert transform of its log-envelope* $\log|X(f)|$*, if and only if the frequency response is minimum phase (i.e a F-PZ with all the poles and zeros inside the unit circle).*

Lemma 3 *Given a frequency response (F-PZ)* $X(f)$ $= |X(f)|e^{j\phi(f)}$*, it is minimum phase, if and only if, its complex cepstrum (CC)* $x_{cc}(n)$ *is causal (i.e* $x_{cc}(n) = 0, n \in [-\infty, -1]$*)*

The proof of above two lemmas can be found in the pages 782-783 of (Oppenheim and Schafer, 1989). Using the *time-frequency* duality, we will state and prove a dual of the lemmas (2), (3).

Lemma 4 *Given an analytic T-PZ signal* $s(t)$ $= \frac{\prod_{i=1}^{P}(1-p_i e^{j\Omega t})}{\prod_{i=1}^{U}(1-u_i e^{j\Omega t})} = |s(t)|e^{j\Psi(t)}$*, all of its poles and zeros are within the unit-circle (i.e s(t) is T-MinP) if and only if its phase* $\Psi(t)$ *is the Hilbert transform of its log envelope* $\log|s(t)|$.

Proof: Let $\tilde{S}(f)$ be the Fourier transform (FT) of $\log s(t) = \log|s(t)| + j\Psi(t)$. We note that $\tilde{S}(f)$ consists of spectral lines at integral multiple of Ω[4] and hence is a discrete sequence. Let us assume that the phase $\Psi(t)$ is the Hilbert transform of the log envelope $\log|s(t)|$. This implies that $\log s(t)$ is an analytic signal and hence its FT $\tilde{S}(f)$ is zero for negative frequencies (i.e. $\tilde{S}(f)$ is a discrete and F-causal sequence). Using the duality principle we note that $\log s(-f)$ is the FT of $\tilde{S}(t)$. In fact, $\tilde{S}(t)$ is the complex cepstrum(CC) of a signal whose FT is $s(-f)$. As $\tilde{S}(t)$ has the same functional form as $\tilde{S}(f)$, this implies that $\tilde{S}(t)$ is a discrete and causal CC sequence. Therefore in light of lemma

[4]This can be seen by series expansion of $\log(1-pe^{j\Omega t}) = \sum_{k=1}^{\infty} -p^k e^{jk\Omega t}/k$

(3), it follows that $s(-f)$ is minimum-phase F-PZ with all the zeros and poles inside the unit circle. Therefore we get,

$$s(-f) = \frac{\prod_{i=1}^{P}(1 - p_i e^{j\Omega(-f)})}{\prod_{i=1}^{U}(1 - u_i e^{j\Omega(-f)})}$$

substituting t for '-f' we get,

$$s(t) = \frac{\prod_{i=1}^{P}(1 - p_i e^{j\Omega t})}{\prod_{i=1}^{U}(1 - u_i e^{j\Omega t})} \tag{4.9}$$

This proves that the T-PZ $s(t)$ that is T-MinP results in its phase being the HT of its log-envelope.

Therefore, using Lemma (4), $s(t)$ can be expressed as follows,

$$s(t) = \underbrace{a_0 \prod_{i=1}^{Q}(-q_i^*)}_{A_c} \underbrace{e^{\alpha(t)+j\hat{\alpha}(t)}}_{\textbf{T-MinP}} \underbrace{e^{j\gamma(t)}}_{\textbf{T-AllP}} \tag{4.10}$$

where A_c is a constant, $\alpha(t)$ is the logarithm of the AM signal, $\hat{\alpha}(t)$ its HT and $\hat{\alpha}(t) + \gamma(t)$ is the phase signal and its derivative is the FM signal. As $\hat{\alpha}(t)$ can be determined from the log AM signal $\alpha(t)$ [5], it forms the redundant information and hence is excluded from the FM signal. Therefore, $\gamma'(t)$ is the FM (instantaneous frequency) signal of interest, where $'$ denotes derivative with respect to time.

The next step is to develop algorithms that can automatically achieve the decomposition as in (4.10). Noting that the all-pole F-PZ as estimated using classical linear prediction technique is guaranteed to be minimum phase, Kumaresan et. al. used the dual of linear prediction in the spectral domain (LPSD) (Kumaresan, 1998), with sufficiently high prediction order 'M', to derive the T-MinP signal. The T-AllP signal was obtained as the residual signal of the LPSD.

However, it is well known that the LP technique overestimates the peaks and poorly models the valley. Moreover, the results are highly susceptible to the model order 'M' whose actual value is not known. Therefore, in this work, we use a non-parametric technique to estimate the AM signals. From (4.10), we note that $\log|s(t)| = \alpha(t) + \log(A_c)$, where $\log(A_c)$ is a constant over the frame. Therefore the logarithm of the absolute magnitude of the analytic signal in each band is an estimate of the corresponding AM signal + a constant term.

4.2.1 Carrier Signal (FM) Extraction

Fig. 4.2 illustrates the FM signal extraction through homomorphic filtering. Consider a narrow band analytic signal $s_b(t)$, $t \in [0, N-1]$. Our objective is to represent $s_b(t)$ as a product of a T-MinP signal and a T-AllP signal as done in (4.10). Specifically,

$$s_b(t) = \underbrace{e^{\alpha(t)+j\hat{\alpha}(t)}}_{\textbf{T-MinP}} \underbrace{e^{j\gamma(t)}}_{\textbf{T-AllP}} \tag{4.11}$$

The phase of the T-AllP signal, $\gamma(t)$ is the FM signal of interest. Let $realFepstrum(k) = FFT\{log|s_b(t)|\} = FFT\{\alpha(t)\}$. In fact $realFepstrum(k)$ is the dual of the well known quantity i.e. real cepstrum. The

[5] Due to the HT relationship between the two

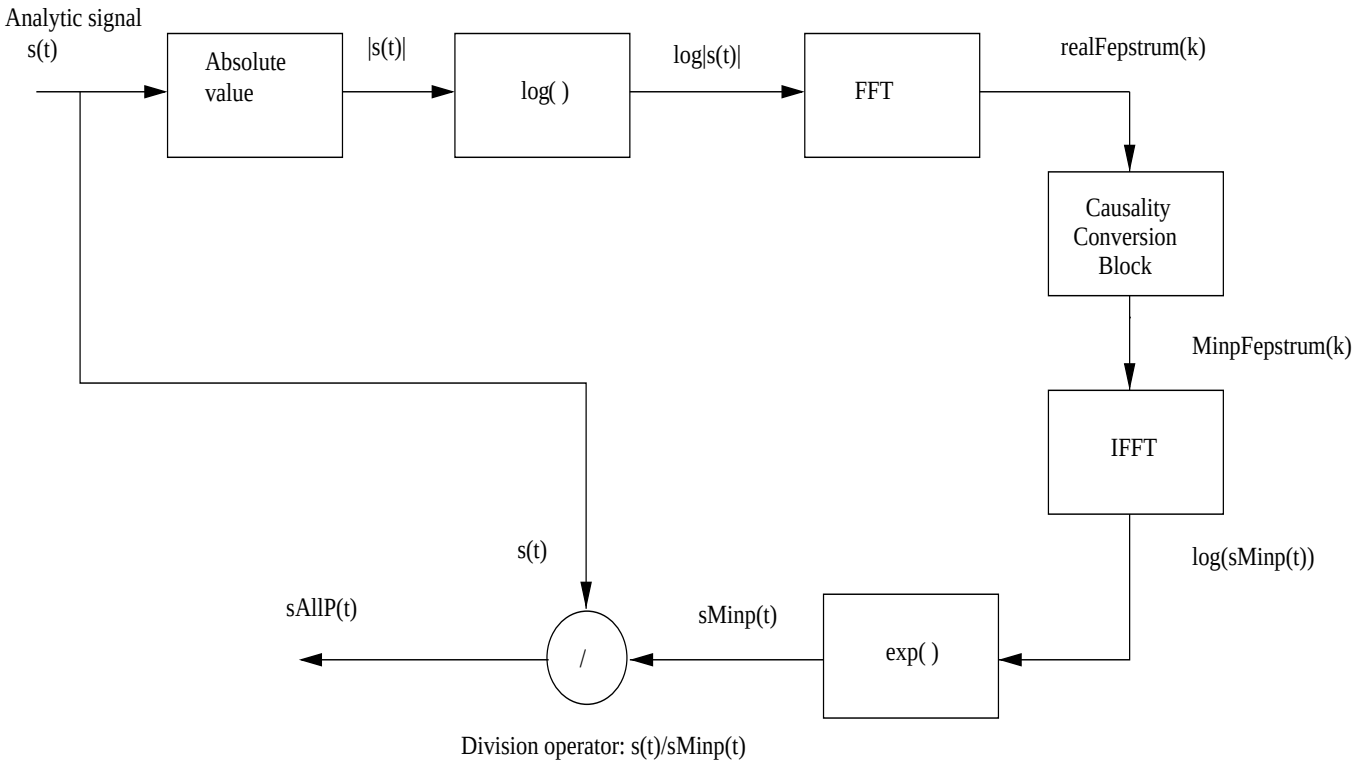

Figure 4.2. Carrier signal decomposition via homomorphic filtering.

causality conversion block in Fig.4.2 performs the following operation.

$$\begin{aligned} MinpFepstrum[k] &= 2 \times realFepstrum[k] \\ k &\in [1, N/2-1] \\ MinpFepstrum[k] &= realFepstrum[k] \\ k &= 0, N/2 \\ MinpFepstrum[k] &= 0 \\ k &\in [N/2+1, N-1] \end{aligned} \tag{4.12}$$

As $MinpFepstrum[k]$ is a causal sequence (by construction above), its IFFT, $log(sMinp(t)) = \alpha(t) + j\hat{\alpha}(t)$ is an analytic signal. In light of lemma (4), it can be seen that, $sMinp(t) = e^{\alpha(t)+j\hat{\alpha}(t)}$ is the T-MinP signal in (4.10). Moreover, according to (4.11), T-AllP signal $sAllP(t)$ that corresponds to the original signal $s_b(t)$ can be obtained as,

$$sAllP(t) = \frac{s_b(t)}{sMinp(t)} \tag{4.13}$$

We note that the T-ALLP signal $sAllp(t) = e^{j\gamma(t)}$ has a unity magnitude for all time $'t'$ and it accounts for only the FM signal. The unique FM signal can be obtained from $sAllP(t)$ as the derivative of its unwrapped phase.

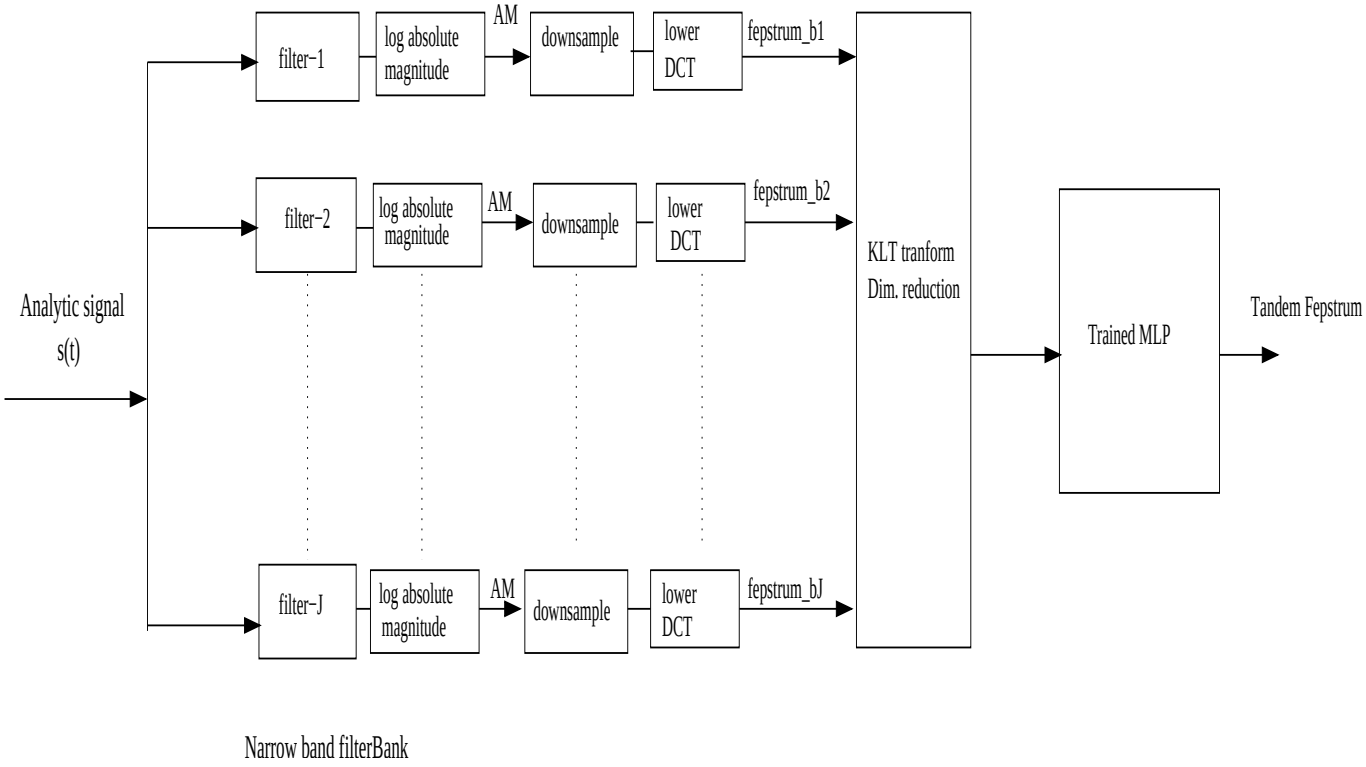

Figure 4.3. The FEPSTRUM feature extraction

4.3 FEPSTRUM feature extraction

Fig.4.3 illustrates our feature extraction scheme. A full-band analytic speech signal s(t) is decomposed into J linearly spaced, non-overlapping narrow bands. We have used narrow-bandwidth filters to achieve our objective of a more "meaningful" modulation analysis by keeping the carrier signal narrow-band (ideally, a sinusoid) (Schimmel and Atlas, 2005). We take the log magnitude of the output of each filter to obtain its corresponding AM signal $\alpha(t)$. The AM signal is then downsampled and its lower DCT coefficients are retained as the feature vector. To distinguish this representation from the previous use of the word "modulation spectrum" (Tyagi et al., 2003; Kingsbury et al., 1998a; Kanedera et al., 1998), which has been weakly specified/defined in the ASR literature (Schimmel and Atlas, 2005), we have termed this representation as FEPSTRUM. In fact, fepstrum is an exact dual of the well-known quantity called real cepstrum.[6] The fepstrum features from each band are concatenated together, then uncorrelated using a KL transform (Principal Component Analysis) followed by the dimensionality reduction. This representation can then be used by itself as a feature vector or can be fed to a Tandem (Ellis et al., 2001b; Hermansky, 2003; Zhu et al., 2004) system to finally derive a Fepstrum-Tandem feature. In the Tandem modeling, phoneme aposteriors obtained at the output of the multi-layer perceptron (MLP), which has been trained on MFCC or Fepstrum features to classify phonemes, are used as the final features in a usual HMM-GMM system.

Fig. 4.4 illustrates the case when we use narrow-band filters to decompose the speech analytic signal, followed by the AM signal estimation in each band. Second and third pane shows the nar-

[6] the name FEPSTRUM denotes this dual nature

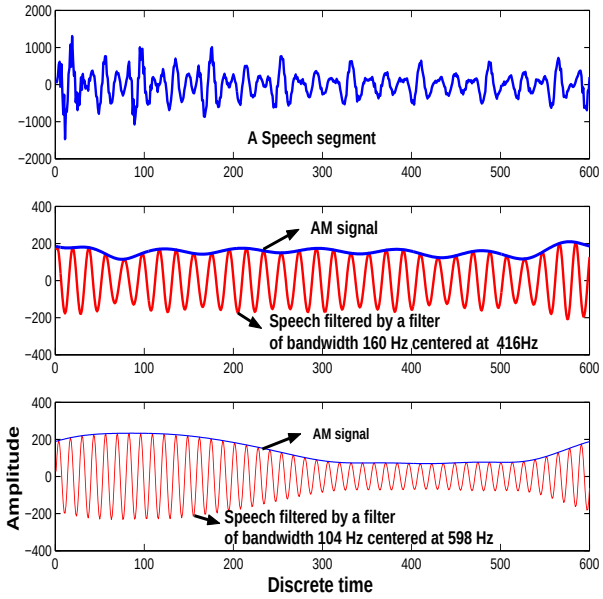

Figure 4.4. The AM signal derived using narrow-band filters

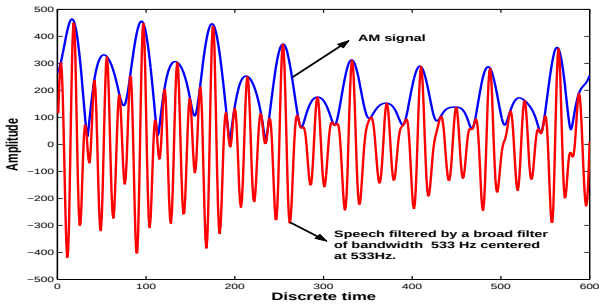

Figure 4.5. The AM signal derived using broad-band filters

row band-pass filtered speech signals and their corresponding AM signals. We note that these AM signals are low modulation frequency signals. The narrow band-pass filters used have band-widths 160 Hz and 104 Hz respectively. Fig.4.5 illustrates the case where a broad-band filter (bandwidth 533 Hz) has been used. For voiced speech, each pitch harmonic can be roughly seen as a monochromatic sinusoidal carrier signal. The spectrum of the signal at the output of a broad-band filter will have several pitch harmonics in it and therefore will violate the condition of a single narrow-band carrier signal. As can be noted in the Fig. 4.5, the pitch component manifests itself as sharp spikes in the AM signal. Therefore a modulation spectrum of this AM signal will reflect the pitch frequency as well, which is undesirable in the context of a speaker independent ASR system. We present these arguments to emphasize the fact that while demodulating an AM component from a speech signal, it is important to make sure that the companion carrier signal (FM) is narrowband.

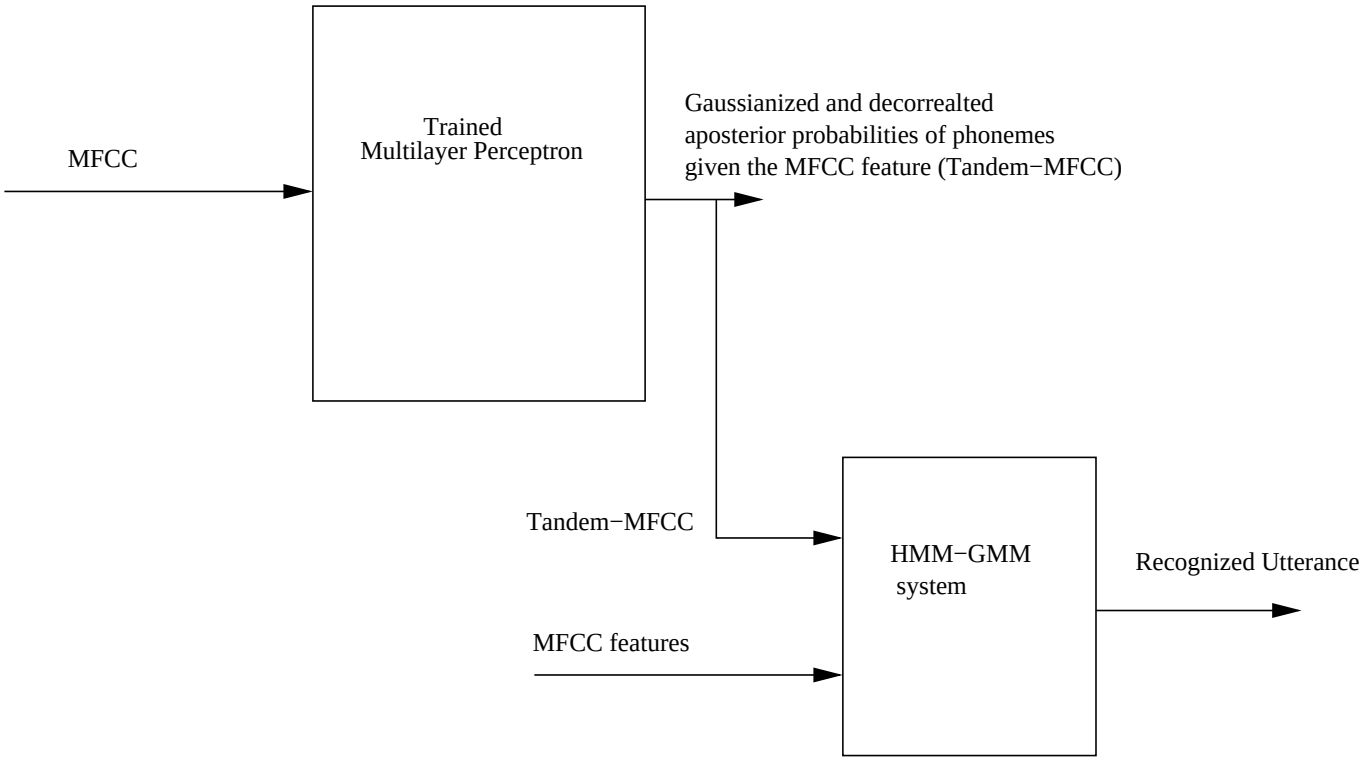

Figure 4.6. Illustration of the Tandem-MFCC system. A concatenation of the Tandem-MFCC and the MFCC features is used to train the HMM-GMM system.

4.4 Experiments and Results

4.4.1 OGI Numbers95

In order to assess the effectiveness of the fepstrum features, speech recognition experiments were conducted on the OGI Numbers corpus (Cole et al., 1994). The OGI Numbers95 database consists of spontaneously spoken free-format connected numbers over a telephone channel. The lexicon consists of 31 words.[7] The Fepstrum features were extracted as per the scheme outlined in Fig. 4.3. We have used 20 linearly spaced, non-overlapping rectangular filters to decompose the speech analytic signal into narrow-band signals of bandwidth 200Hz each. The AM signal is obtained as the logarithm of the absolute magnitude of the narrow-band filter output. At this stage, the AM signal has the same sampling frequency as the original speech signal ($8KHz$). As can be noted in the Fig. 4.4, the AM signals are low modulation frequency signals. Therefore, we filter the AM signals through a low-pass filter of cutoff-frequency 100 Hz and then down-sample them by a factor of 40. Long rectangular windows of size 85 ms were used to frame the narrow band-pass filtered analytic signals. This was done to ensure that we have sufficient number of AM signal samples after down-sampling the AM signal in each band. We chose a rectangular shape of the window to avoid any artificial tilt in the lower DCT coefficients. We then retain its first 5 DCT coefficients (Fepstrum) that roughly correspond to $[0, 50]$ Hz. Fepstrum sub-vector from each band are concatenated together to form a vector of dimensionality 100 (5×20). We then perform a KL transform on this vector, followed by a dimensionality reduction to obtain a 60 dim. feature vector. At this stage, this 60 dim. feature

[7] with confusable words like nine, ninety and nineteen.

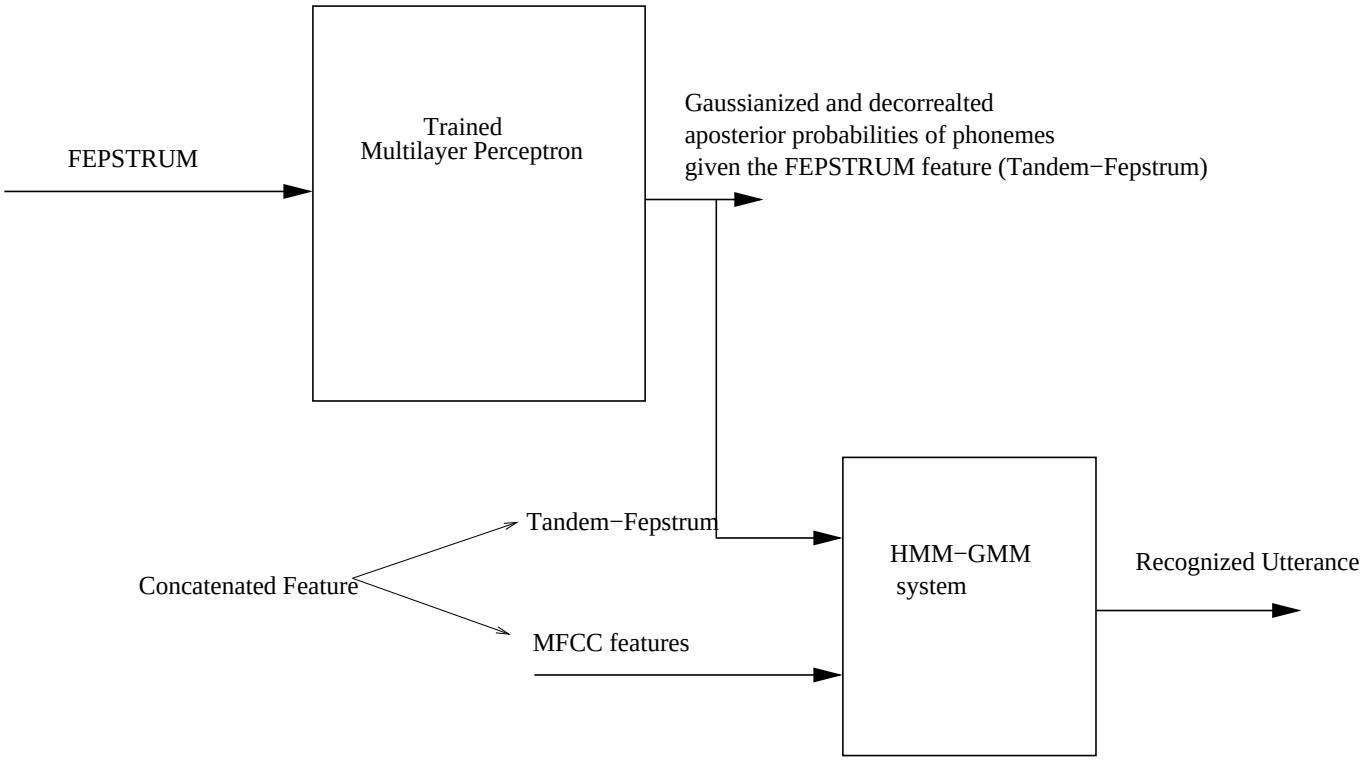

Figure 4.7. Illustration of the Tandem-FEPSTRUM system. A concatenation of the Tandem-Fepstrum and the MFCC features is used to train the HMM-GMM system.

vector can be concatenated with the MFCC feature and can be used as a composite feature vector for a HMM-GMM system. Otherwise, the 60 dim. Fepstrum vector can also be used as an input to a multi-layer perceptron to obtain the Tandem-Fepstrum features. The fepstrum features are fed to a multi-layer perceptron (MLP) to obtain phoneme aposteriors which are again KL transformed to obtain 27 dimensional Tandem-Fepstrum features[8]. Tandem (Ellis et al., 2001b; Hermansky, 2003; Zhu et al., 2004) has been shown to be an effective technique for combining different kind of features. Fepstrum features measure the slow amplitude modulations within a window of $80 - 100ms$ long duration and hence provide information that is complementary to the usual spectral envelope based MFCC features. While Fepstrum provides amplitude modulations (AM) occurring within a single frame of approximately $100ms$ long duration, the MFCC feature provides static energy in the Mel-bands of each frame and its variation across several frames (the deltas). Therefore we have concatenated MFCC feature with the Fepstrum-Tandem features. Figs. 4.6, 4.7 illustrate the (Concatenated MFCC +Tandem-MFCC) and the (Concatenated MFCC+ Tandem-FEPSTRUM) features. As illustrated in these figures the augmented features are used as an observation by the Gaussian mixture hidden Markov model (HMM-GMM) system in the tandem framework.

Mel-frequency cepstral coefficients (MFCC) and their temporal derivatives along with cepstral mean subtraction have been used as additional features. For comparison, seven feature sets were generated:

1. [MFCC:] 39 dimensional MFCC + delta features.
2. [Fepstrum:] 60 dimensional Fepstrum.

[8] each dimension corresponds to a monophone which are 27 in number

3. [Concat. Fepstrum +MFCC:] 60 dimensional Fepstrum feature concatenated with the MFCCs.
4. [T-MFCC:] 27 dim. Tandem representation of MFCC + delta features.
5. [T-Fepstrum:] 27 dim. Tandem representation of Fepstrum features
6. [Concat. MFCC+ (T-MFCC):] (39+27) dim. feature vector which is a concatenation of the MFCC and Tandem-MFCC
7. [Concat. MFCC+ (T-FEPSTRUM):] (39+27) dim. feature vector which is a concatenation of the MFCC and Tandem-Fepstrum features.

All the above features were then used in a Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition system that was trained using public domain software HTK (Young et al., 1995) on the clean training set from the original Numbers corpus. The system consisted of 80 tied-state triphone HMMs with 3 emitting states per triphone and 12 mixtures per state.[9] Table 4.1 indicates the performance of these feature sets. Fepstrum and T-Fepstrum features have only the modulation frequency information and hence they perform slightly worse than the MFCC and the T-MFCC features respectively. However, as they carry complementary information, their concatenation (MFCC+T-FEPSTRUM) results in the lowest (4.1%) WER. For a fair comparison, we compare this to a concatenation of the MFCC+ T-MFCC features which has a WER of 4.6%.

Table 4.1. Word error rate (WER) in clean conditions

MFCC	5.7
Fepstrum	6.8
Concat. Fepstrum + MFCC	4.6
T-MFCC	5.2
T-FEPSTRUM	5.5
Concat. (T-MFCC) + MFCC	4.6
Concat. (T-FEPSTRUM) + MFCC	4.1

In this work, we have only used the AM modulation spectrum (fepstrum) as a feature as the inclusion of the FM signal as a feature did not provide any further performance gains over the use of the combination of the fepstrum and the MFCC features. The FM signal estimation requires the phase unwrapping which apart from being computationally expensive, is also a numerically error-prone process. In fact, there is no algorithm in the literature that performs error-free phase-unwrapping for any arbitrary signal. Therefore, in this thesis, we have solely worked with the Fepstrum feature which is a representation of the AM signal that is suitable for ASR applications.

4.4.2 TIMIT Phoneme recognition Task

In the TIMIT phoneme recognition task, monophone HMMs with 3 state per monophone were trained for each of the features that were compared. Each state emission density consisted of 11 Gaussians with diagonal covariance matrices.[10] Unlike the telephony speech, the TIMIT database

[9]The performance of the baseline system using the MFCC features saturated around 9-12 Gaussians per mixture and hence throughout the experiments, the emission density of each HMM state, is modeled by 12 Gaussians per GMM. We believe that such an approach provides as many parameters to the baseline system as it requires to achieve the best ASR performance. As is well known, increasing the number of parameters beyond a certain optimal number, leads to the over-fitting and consequently performance degradation.

[10]The ASR performance of the baseline system trained on the MFCCs saturated around 10-11 Gaussians per mixture and hence the number of Gaussians per state was set to 11.

has a bandwidth of 8kHz. Therefore in order to use a parsimonious representation we used the conventional 24 Mel filters to decompose the speech analytic signal into narrow band-pass filtered signals as in Fig.4.3. As most of the speech signal's energy is contained within 0-2 KHz bandwidth and as the lower Mel-filters have narrow-bandwidth, we still were able to ensure that the carrier signals in the lower (10-15) Mel filters, were narrowband. Rest of the steps to compute the Fepstrum feature were the same as in the previous example of the OGI Numbers95.

Five feature sets were computed for comparing the ASR accuracy on the TIMIT core-test set and the ASR results are provided in Table 4.2. As can be noted form this table, Fepstrum features in conjunction with the MFCC features, significantly reduced the phoneme recognition error rates

1. [MFCC:] 39 dim. MFCC + delta features.
2. [Fepstrum:] 60 dimensional Fepstrum.
3. [Concat Fepstrum + MFCC:] 60 dimensonal Fepstrum feature concatenated with 39 dim. MFCC feature
4. [Concat. MFCC+ (T-MFCC):] (48+39) dim. feature vector which is a concatenation of the MFCC and Tandem-MFCC
5. [Concat. MFCC+ (T-FEPSTRUM):] (48+39) dim. feature vector which is a concatenation of the MFCC and Tandem-Fepstrum features.

Table 4.2. Phoneme recognition error rate on the TIMIT core-test set.

MFCC	33.5
Fepstrum	35.0
Concat Fepstrum + MFCC	31.7
Concat. MFCC+ (T-MFCC)	29.1
Concat. MFCC+ (T-FEPSTRUM)	28.1

4.4.3 OLLO non sense syllable recognition task

To further validate the results, we conducted similar experiments on the OLLO (Wesker et al., 2005) database. In Table 4.3, we provide results on a University of Oldenburg nonsense syllable (OLLO) (Wesker et al., 2005) recognition task. OLLO database is rich in various speech variabilities such as different speaking styles (slow, fast, statement, questioning, loud and soft) and with almost equal sampling of the male and female speakers.Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition systems were trained using public domain software HTK on the NO-accent part of the OLLO training set that roughly consist of 13,500 utterances. Three state left to right HMM models were trained for each of the 26 phonemes in the OLLO (Wesker et al., 2005) database including silence as well. The lexicon consists of 150 logatomes[11] which are either CVCs or VCVs such as uttu, acsha, atta etc. The experiments reported in this paper are for the entire logatome recognition on the No-accent test part of the OLLO database that consists of roughly 13,800 utterances. Each of these utterances correspond to an instance of a logatome. The ASR performance of the baseline system trained on the MFCCs saturated around 60-65 Gaussians per mixture and hence Gaussian mixture models(GMMS) with 65 Gaussian per state and diagonal covariance matrices were used to model the emission probability densities of the feature vectors. The logatome recognition results for various features are given in Table 4.3.

[11]A logatome is CVC or VCV that consists of nonsense syllables

As the OLLO database too has a bandwidth of $8kHz$, the Fepstrum feature extraction module was the same as the one employed for the TIMIT-database with the initial analysis done by 24 Mel-filters. Unlike the TIMIT and the OGI Numbers95 database, the OLLO database is not time-labeled. Therefore, we did not train any Tandem system for this database.[12] We note from the Table 4.3 that the fepstrum features have a performance similar to the MFCCs. However, as can be noted from the Table 4.3, even a simple concatenation of the Fepstrum features with MFCCs provide a significant improvement over the MFCC features alone in clean acoustic conditions (18db SNR). The absolute improvement averaged over all the variabilities is 3.3%. , we note that improvement gains in case of fast and normal speech are quite substantial and stand in excess of 4.0% absolute.

Table 4.3. Logatome recognition rates reported over each variabilities of the OLLO database.

Feature	Fast	Slow	Loud	Soft	Questioning	Normal	Overall
MFCC	71.3	76.0	76.9	64.6	80.1	79.6	74.7
Fepstrum	68.4	77.9	74.5	66.2	80.8	80.4	74.6
Concat Fepstrum + MFCC	75.4	78.5	79.5	67.1	83.4	84.3	78.0
Absolute Improvement	4.1	2.5	2.6	2.5	3.3	4.7	3.3

4.5 Summary

In this chapter we have developed a theoretically sound AM-FM decomposition technique using analytic signals in the time domain. In the past, several researchers (Tyagi et al., 2003; Athineos et al., 2004; Zhu and Alwan, 2000; Kingsbury et al., 1998a; Kanedera et al., 1998) have extracted the speech modulation spectrum by computing a discrete Fourier transform (DFT) of the Mel or critical band spectral energy trajectories, where each sample of the trajectory has been obtained through a power spectrum (followed by Mel filtering) over 20-30ms long windows. We point out the deficiencies in such techniques namely,

- It implicitly assumes that within each Mel or critical band, the amplitude modulation (AM) signal remains constant within the duration of the window length that is typically 20-30ms long.
- Instead of modeling the continuously and slowly changing amplitude modulation signal in each band, it mostly models the spurious and abrupt modulation frequency changes that occur due to the frame shifting of $10ms$.

Finally we present a suitable representation of the AM signal in form of the lower modulation frequencies of the downsampled AM signals (fepstrum) in each band. While Fepstrum provides amplitude modulations (AM) occurring within a single frame of size 100ms, the MFCC feature provides static energy in the Mel-bands of each frame and its variation across several frames (the deltas). Together these two features complement each other and the ASR experiments indicate that Fepstrum feature in conjunction with MFCC feature achieves significant ASR improvement over several speech databases.

Until now, we have described two feature extraction techniques that are geared for improving the ASR performance in the clean (SNR conditions of roughly 18dB and above) acoustic conditions by better describing the inherent non-stationary structure in the speech signals.

[12]We could have performed a forced Viterbi alignment of the training utterances to get time labeled train set that is required for training a multi-layer perceptron. However, our goal was primarily to show the efficacy of the Fepstrum feature irrespective of the modeling technique. Our experiments over OGI numbers95, TIMIT and OLLO indicate that fepstrum features in conjunction with the MFCCs provide significant improvement in ASR accuracies.

The next three chapters cover study and design of noise robust features. It may be argued that why do we need to design different kinds of features for clean and noisy speech instead of a single feature that may work, both in the clean and the noisy conditions. However, as the noise robust features are usually designed to combat the noise, they usually introduce artifacts in the features whenever the noise is not present in the speech signal. Therefore, they have poorer performance as compared to the MFCC features in the clean acoustic conditions.

In the next chapter we will describe a least-squares filtering technique that enhances the speech signal that is corrupted by additive broad-band noise.

Chapter 5

Least Squares filtering of the speech signals for robust ASR

In this chapter we introduce an adaptive filtering technique that enhances a speech signal which has been corrupted by additive broad-band noise. One of the key advantage of the proposed technique is that unlike the Wiener filtering, it does not require a reference noise signal. This renders the least-squares filtering technique as a highly practical enhancement technique that can work with only a single noisy speech channel. Furthermore, unlike the classical spectral subtraction and Wiener filtering techniques that require the noise to be stationary, the proposed LeSF technique makes no such assumption as this technique works on a block by block basis.

In this work we have developed an analytic solution of the least-squares filter that helps us to understand input-output gain relationship and the filter's bandwidth versus input signal's SNR relationship. We analyze the behavior of the least squares filter(LeSF) operating on noisy speech signal. Speech analysis is performed on a block by block basis and a LeSF filter is designed for each block of signal, using a computationally efficient algorithm. Unlike other feature level noise robustness technique, the LeSF filter enhances the signal waveform itself and a MFCC feature computed over this enhanced signal leads to a significant improvement in speech recognition accuracies as compared to the other competing feature level noise robustness techniques such as RASTA-PLP and spectral subtraction. In distributed speech recognition (DSR) in the context of mobile telephony and voice-over IP systems, it may be desirable not only to have noise robust feature extraction algorithm but also to enhance the noisy speech signal for the human listener. Therefore, a signal enhancement technique that also leads to noise robust ASR is desirable. The proposed LeSF filtering technique falls into this category as it not only enhances the signal, a simple MFCC feature computed over this enhanced signal leads to significant ASR accuracy improvements in several realistic noisy conditions.

5.1 Introduction

Speech enhancement, amongst other signal de-noising techniques, has been a topic of great interest for past several decades. The importance of such techniques in speech coding and automatic speech recognition systems can only be understated. Towards this end, adaptive filtering techniques have been shown to be quite effective in various signal de-noising applications. Some representative examples are echo cancellation (Sondhi and Berkley, 1980), data equalization (Gersho, 1969; Satorius and Alexander, 1979; Satorius and Pack, 1981) narrow-band signal enhancement (Widrow, 1975;

Bershad et al., 1980) beamforming (Griffiths, 1969; Frost, 1972; Compton, 1980), radar clutter rejection (Gibson and Haykin, 1980), system identification (Rabiner et al., 1978; Marple, 1981) and speech processing (Widrow, 1975).

Most of the above mentioned representative examples require an explicit external noise reference to remove additive noise from the desired signal as discussed in (Widrow, 1975). In situations where an external noise reference for the additive noise is not available, the interfering noise may be suppressed using a Wiener linear prediction filter (for stationary input signal and stationary noise) if there is a significant difference in the bandwidth of the signal and the additive noise (Widrow, 1975; Zeidler et al., 1978; Anderson et al., 1983). One of the earliest use of the least mean square (LMS) filtering for speech enhancement is due to Sambur (Sambur, 1978). In his work, the step size of the LMS filter was chosen to be one percent of the reciprocal of the largest eigenvalue of the correlation matrix of the first voiced frame. However, speech being a non-stationary signal, the estimation of the step size based on the correlation matrix of just single frame of the speech signal, may lead to divergence of the LMS filter output. Nevertheless, the exposition in (Sambur, 1978) helped to illustrate the efficacy of the LMS algorithm for enhancing naturally occurring signals such as speech. In (Zeidler et al., 1978), Zeidler et. al. have analyzed the steady state behavior of the adaptive line enhancer (ALE), an implementation of least mean square algorithm that has applications in detecting and tracking narrow-band signals in broad-band noise. Specifically, they have shown that for a stationary input consisting of multiple ($'N'$) sinusoids in white noise, the L-weight ALE, can be modeled by the $L \times L$ Wiener-Hopf matrix equation and that this matrix can be transformed into a set of $2N$ coupled linear equations. They have derived the analytical expression for the steady-state L-weight ALE filter as function of input SNR and the interference between the input sinusoids. It has been shown that the coupling terms between the input sinusoid pairs approach zero as the ALE filter length increases.

In (Anderson et al., 1983), Anderson et al extended the above mentioned analysis for a stationary input consisting of finite band-width signals in white noise. These signals consist of white Gaussian noise (WGN) passed through a filter whose band-width α is quite small relative to the Nyquist frequency, but generally comparable to the bin width $1/L$. They have derived analytic expressions for the weights and the output of the LMS adaptive filter as function of input signal band-width and SNR, as well as the LMS filter length and bulk delay 'z^{-P}' (please refer to Fig. 5.1).

In this work, we extend the previous work in (Anderson et al., 1983; Zeidler et al., 1978) for enhancing a class of non-stationary signals that are composed of either (a) multiple sinusoids (voiced speech) whose frequencies and the amplitudes may vary from block to block or (b) are the output of an all-pole filter excited by white noise input (unvoiced speech segments) and which are embedded in white noise. The number of sinusoids may also vary from block to block. The key difference in the approach proposed in this work is that we relax the assumption of the input signal being stationary. The method of least squares may be viewed as an alternative to Wiener filter theory pg.483 (Haykin, 1993). Wiener filters are derived from *ensemble averages* and they require good estimates of the clean signal power spectral density (PSD) as well as the noise PSD. Consequently, one filter (optimum in a probabilistic sense) is obtained for all realizations of the operational environment, assumed to be wide-sense stationary. On the other hand, the method of least squares is *deterministic* in approach. Specifically, it involves the use of time averages over a block of data, with the result that the filter depends on the number of samples used in the computation. Moreover, the method of least squares does not require the noise PSD estimate. Therefore the input signal is blocked into frames and we analyze a L-weight least squares filter (LeSF), estimated on each frame which consists of N samples of the input signal.

Working under the assumptions that the clean signal spectral vector and noise spectral vector are Gaussian distributed with k^{th} spectral value independent of j^{th} spectral value, Ephraim and Malah derived the optimum minimum mean square error (MMSE) estimator of the clean speech's spectral amplitude (MMSE-STA) (Ephraim and Malah, 1984) and its log spectral ampli-

tude (MMSE-LSA) (Ephraim and Malah, 1985). This assumption is valid only if the clean signal and the noise are both stationary processes and the spectrum is estimated over an infinitely long window. Clearly the speech signal is neither a stationary process nor does it have a Gaussian distributed spectrum. Moreover, in most of the situations, the noise is not a stationary process. Besides this, MMSE-LSA, MMSE-STA, spectral subtraction (SS) and Wiener filter (WF) based techniques need a good estimate of noise spectrum. It is often claimed that the estimate of the noise PSD can be obtained from "non-speech" frames which can be detected using a pre-tuned threshold (Ephraim and Malah, 1985, 1984). However, if the noise power changes (varying SNR conditions), there is no single threshold which can detect the non-speech frames. Moreover if the noise is non-stationary, the noise PSD estimate obtained through "non-speech" frames may not be able to track the noise statistics quite well as it is dependent on the availability of non-speech frames which are unevenly distributed in an utterance. Martin (Martin, 2001) has proposed a noise PSD estimator based on the minimum statistics. However even this techniques relies on certain parameters which need to be tuned depending on the degree of non-stationarity of the noise. Several researchers have tried to use a multitude of well-tailored tuning-parameters dependent on the a-prior knowledge of non-speech frames[1], highest and lowest SNR range, and several other ad-hoc weighting factors[2] pg. 438 (Kim and Rose, 2003) to achieve noise robustness in ASR.

Therefore it is desirable to develop a new enhancement technique that does not require a noise PSD estimate. The least squares filter (LeSF) based techniques fall in this category as they do not require a noise PSD estimate. We have derived the analytical expressions for the impulse response of the L-weight least squares filter (LesF) as a function of the input SNR (computed over the current frame), effective band-width of the signal (due to finite frame length), filter length 'L' and frame length 'N'. We have applied the block estimated LeSF filter for de-noising speech signals embedded in broad-band noise.

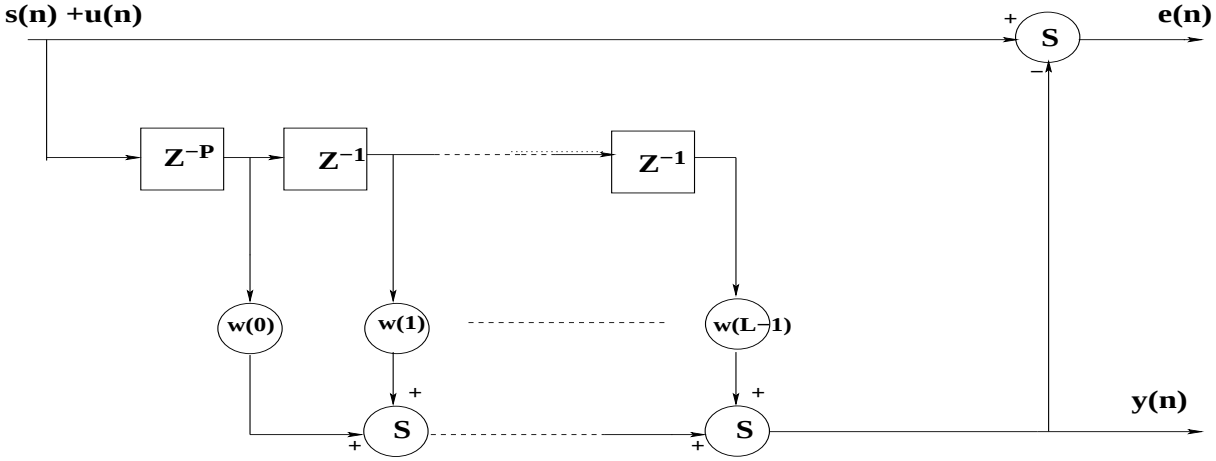

Figure 5.1. The basic operation of the LeSF. The input to the filter is noisy speech, ($x(n) = s(n) + u(n)$), delayed by bulk delay =P. The filter weights w_k are estimated using the least squares algorithm based on the samples in the current frame. The output of the filter $y(n)$ is the enhanced signal.

[1]For example, just by the design of the speech databases, the initial few frames always correspond to the silence and hence can be used for noise PSD estimation. However in a realistic ASR task such assumptions cannot be made.

[2]such as raising the Wiener filter or spectral subtraction gain function to a certain power which is empirically tuned, dependent on the SNR conditions.

5.2 Least Squares filter (LeSF) for signal enhancement

The basic operation of the LeSF is illustrated in figure (5.1) and it can be understood intuitively as follows. The autocorrelation sequence of the additive noise $u(n)$ that is broad-band decays much faster for higher lags than that of the speech signal. Therefore the use of a large filter length ('L') and the bulk delay P causes de-correlation between the noise components of the input signal, namely ($u(n-L-P+1),\ u(n-L-P+2), ..., u(n-P)$) and the noise component of the reference signal, namely ($u(n)$). The LeSF filter responds by adaptively forming a frequency response which has pass-bands centered at the frequencies of the formants of the speech signal while rejecting as much of broad-band noise (whose spectrum lies away from the formant positions). Denoting the clean and the additive noise signals by $s(n)$ and $u(n)$ respectively, we obtain the noisy signal $x(n)$.

$$x(n) = s(n) + u(n) \tag{5.1}$$

The LeSF filter consists of L weights and the filter coefficients w_k for k $\in [0, 1, 2..L-1]$ are estimated by minimizing the energy of the error signal $e(n)$ over the current frame, $n \in [0, N-1]$.

$$e(n) = x(n) - y(n) \tag{5.2}$$

$$\text{where } y(n) = \sum_{i=0}^{L-1} w(i)x(n-P-i) \tag{5.3}$$

Let $\mathbf{A}$ denote the $(N+L)\times L$ data matrix (Haykin, 1993) of the input frame $\mathbf{x} = [x(0), x(1), ...x(N-1)]$ and $\mathbf{d}$ denote the $(N+L)\times 1$ desired signal vector which in this case is signal $\mathbf{x}$ appended by L zeros. The LeSF weight vector $\mathbf{w}$ is then given by

$$\mathbf{w} = \left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{d} \tag{5.4}$$

As is well known, $\mathbf{A}^H\mathbf{A}$ is a symmetric $L \times L$ Toeplitz matrix whose (i, j) element is the temporal autocorrelation of the signal vector $\mathbf{x}$ estimated over the frame length (Haykin, 1993).

$$\left[\mathbf{A}^H\mathbf{A}\right]_{i,j} = r(|i-j|) \tag{5.5}$$

$$= \sum_{n=0}^{N-|i-j|} x(n)x(n+|i-j|) \tag{5.6}$$

In practice, $\mathbf{A}^H\mathbf{A}$ can always be assumed to be non-singular due to presence of additive noise (Haykin, 1993) for filter length $L < N$. The weight vector $\mathbf{w}$ in (5.4) can be obtained using Levinson Durbin algorithm (Haykin, 1993) without incurring a significant computational cost.

5.3 LeSF applied to Speech

In this section, we will analytically solve (5.4) to obtain the LeSF $\mathbf{w}$.To this aim, we model voiced speech using sinusoidal model (McAulay and Quatieri, 1986), while unvoiced speech is modeled by a source-filter model. However, we show that the functional form of the equations remain the same except for a change in the parameter values.

5.3.1 Voiced Speech

As proposed in (McAulay and Quatieri, 1986), voiced speech signals can be modeled as a sum of multiple sinusoids whose amplitudes, phases and frequencies can vary from frame to frame. Let us assume that a given frame of speech signal $s(n)$ can be approximated as a sum of M sinusoids. The

number of sinusoids M may vary from block to block. Then the noisy signal $x(n)$ can be expressed as

$$x(n) = \sum_{i=1}^{M} A_i \cos(\omega_i n + \phi_i) + u(n) \tag{5.7}$$

where $n \in [0, N-1]$ and $u(n)$ is a realization of white noise. Then the k^{th} lag autocorrelation can be shown to be

$$\begin{array}{c} r(k) = \sum_{n=0}^{N-k-1} x(n)x(n+k) \\ \simeq \sum_{i=1}^{M}(N-k)A_i^2 cos(2\pi f_i k) + N\sigma^2\delta(k) \end{array} \tag{5.8}$$

where it is assumed that the noise $u(n)$ is white and uncorrelated with the signal s(n) and $N \gg 1/(f_i - f_j)$ for all frequency pairs (i, j). The latter condition ensures that all the interference terms between all the sinusoids pairs (i, j) sum up to zero. The LeSF weight vector $w(k)$ is then obtained as the solution of the Normal equations,

$$\begin{array}{c} \sum_{k=0}^{L-1} r(l-k)w(k) = r(l+P) \\ l \in [0, 1, 2..L-1] \end{array} \tag{5.9}$$

The set of L linear equations described in (5.9) can be solved by elementary methods if the z-transform ($S_{xx}(z)$) of the symmetric autocorrelation sequence ($r(k)$) is a rational function of 'z' (Satorius et al., 1978). $S_{xx}(z)$ is given by,

$$S_{xx}(z) = \sum_{k=-\infty}^{\infty} r(k)z^{-k} \tag{5.10}$$

Consider then, a real symmetric rational z transform with M pairs of zeros and M pairs of poles.

$$S_{xx}(z) = G\frac{\prod_{m=1}^{M}(z - e^{-\beta_m + j\Psi_m})(z^{-1} - e^{-\beta_m - j\Psi_m})}{\prod_{m=1}^{M}(z - e^{-\alpha_m + j\omega_m})(z^{-1} - e^{-\alpha_m - j\omega_m})} \tag{5.11}$$

If the signal $x(n)$ is real, then so is its autocorrelation sequence, $r(k)$. In this case the power spectrum, $S_{xx}(z)$, has quadruplet sets of poles and zeros because of the presence of conjugate pairs at $z = exp(\pm\alpha_m \pm j\omega_m)$ and $z = exp(\pm\beta_m \pm j\Psi_m)$. Anderson et. al. (Anderson et al., 1983) have derived the general form of the solution to (5.9) for input signal with rational power spectra such as that described by (5.11). In this case, the LeSF weights are given by,

$$w(k) = \sum_{m=1}^{M} \left(B_m e^{-\beta_m k + j\Psi_m k} + C_m e^{+\beta_m k + j\Psi_m k}\right) \tag{5.12}$$

As can be seen, LeSF consists of an exponentially decaying term and an exponentially growing term attributed to reflection (Widrow, 1975), that occurs due to the finite filter length L. The value of the coefficients B_m and C_m can be determined by solving the set of coupled equations obtained by substituting the expression for $w(k)$ given in (5.12) into (5.9).

To be able to use the general form of the solution of the LeSF filter as in (5.12), we need a pole-zero model of the input autocorrelation in the form as described in (5.11). For sufficiently large frame length N, such that filter length $L \ll N$, we can make the following approximation.

$$\begin{aligned} (N-k) &\simeq Ne^{-k/N} \\ k \in [0,1,2,\ldots,L] \text{ and } L &\ll N \end{aligned} \tag{5.13}$$

The above can be verified by using the Taylor series expansion of $Ne^{-\alpha k}$ and using only the linear term as $k \ll N$. We call ($\alpha = 1/N$) as α^{voiced}. Using this approximation in (5.8), we get,

$$r(k) = Ne^{-\alpha^{voiced}k}\sum_{i=1}^{M} A_i^2 cos(\omega_i k) + N\sigma^2\delta(k) \tag{5.14}$$

In this form, $r(k)$ corresponds to a sum of multiple decaying exponential sequences and its z transform takes up the form,

$$\begin{aligned} S_{xx}(z) &= \sum_{m=1}^{M} \frac{NA_i^2(1-e^{-2\alpha})}{2} \times \\ &\quad \Big(\frac{1}{(z-e^{-\alpha_m+j\omega_m})(z^{-1}-e^{-\alpha_m-j\omega_m})} \\ &\quad + \frac{1}{(z-e^{-\alpha_m-j\omega_m})(z^{-1}-e^{-\alpha_m+j\omega_m})}\Big) + N\sigma^2 \\ \text{where } \alpha_m &= \alpha^{voiced} = 1/N \ \forall\, m \in [1.M] \end{aligned} \tag{5.15}$$

5.3.2 Unvoiced Speech

We model unvoiced speech $s(n)$ as the output of an all pole transfer function (whose poles are at $z = e^{-\alpha_i^{unvoiced} \pm \omega_i}$) excited by a white noise signal $e(n)$. Specifically,

$$S(z) = \frac{E(z)}{\prod_{i=1}^{Q}(z-e^{-\alpha_i^{unvoiced}+j\omega_i})(z-e^{-\alpha_i^{unvoiced}-j\omega_i})} \tag{5.16}$$

where $S(z), E(z)$ are the z-transforms of unvoiced speech signal $s(n)$ and white noise excitation signal $e(n)$ respectively. Then it can be shown that the autocorrelation coefficients of the unvoiced speech are also decaying exponentials (pg. 118, (Haykin, 1993)) i.e

$$r_{unvoided}(k) = \sum_{i=1}^{Q} e^{-\alpha_i^{unvoided}k} cos(\omega_i k), \tag{5.17}$$

where the decaying factor $\alpha_i^{unvoiced} > \alpha^{voiced} = 1/N$ (where N is the block length). This is due to the fact that voiced speech has sharper spectral peaks than the unvoiced speech. Consequently the autocorrelation coefficients of the unvoiced speech decay much faster than those of the voiced speech. However, the functional form for the autocorrelation coefficients of the voiced and unvoiced speech is the same, except that $\alpha^{voiced} < \alpha^{unvoiced}$. In presence of white noise, the power spectral density of the noisy unvoiced speech segment is given by,

$$
\begin{aligned}
S_{xx}(z) &= \sum_{i=1}^{Q} \frac{NA_i^2(1-e^{-2\alpha_i})}{2} \times \\
&\left(\frac{1}{(z-e^{-\alpha_i+j\omega_i})(z^{-1}-e^{-\alpha_i-j\omega_i})}\right. \\
&\left.+\frac{1}{(z-e^{-\alpha_i-j\omega_i})(z^{-1}-e^{-\alpha_i+j\omega_i})}\right)+N\sigma^2
\end{aligned}
\tag{5.18}
$$

where α_i is a decay factor of the i^{th} pole pair. We note that the functional form of the power spectral densities in (5.18) and (5.15) are the same except that α_i in (5.18) will in general be greater than α^{voiced} in (5.15). Therefore the functional form of the LeSF filter $\mathbf{w}$ in (5.12) remains the same for both voiced and unvoiced speech. It is just that for the unvoiced speech the bandwidth of the pass-bands of the LeSF will be wider than that of voiced LeSF. In Fig. 5.2, we show a transfer function with two complex-pole pairs (at conjugate symmetric positions) that is used to synthesize unvoiced speech by exciting it with white noise. First pane shows the pole-zero plot. Second pane shows the frequency response of this all-pole model. In the third pane, blue, red and green curves are the FFT magnitudes of the clean speech, noisy speech corrupted by white noise at SNR -3dB and the LeSF enhanced speech respectively. The fact that the green curve matches the blue curve closely, shows that the LeSF has been able to filter out the noise component.

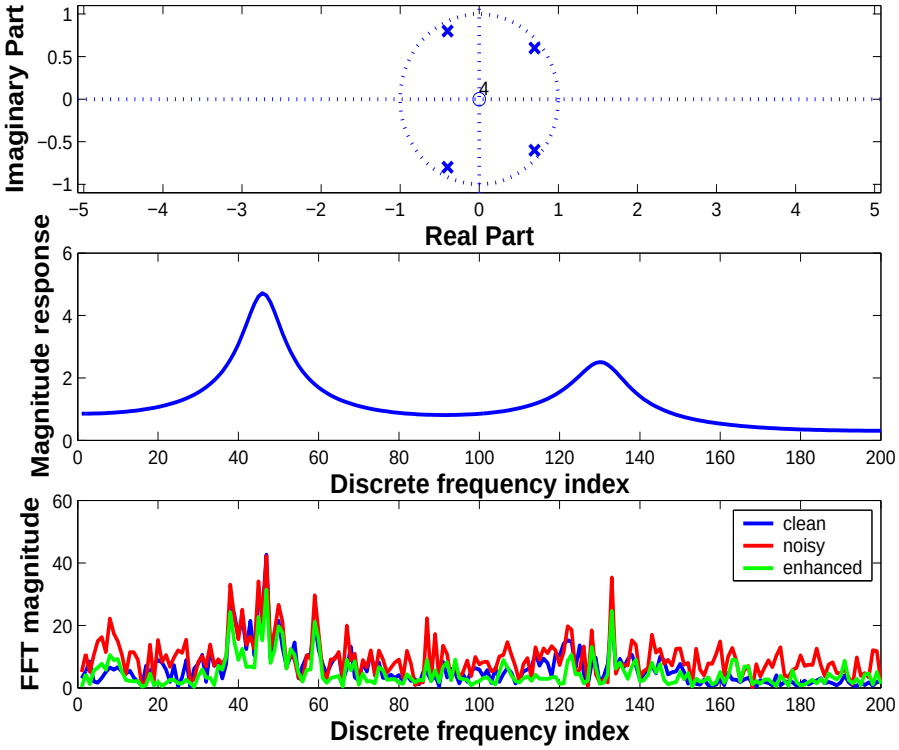

Figure 5.2. A example of a two-formant vocal-tract frequency response which is excited by white noise to synthesize unvoiced speech.

5.3.3 Analytic form of LeSF

From now onward we will not make any distinction between the exponential decay factors α^{voiced} and $\alpha^{unvoiced}$ as the functional form of the equations remain the same. Therefore the following discussion is valid for both voiced speech and unvoiced speech.

To be able to use the general form of the solution of the LeSF filter as in (5.12), we need a pole-zero model of the input autocorrelation in the form as described in (5.11). Under the approximation that the decaying exponentials are widely spaced along the unit circle, the power spectrum $S_{xx}(z)$ in (5.15) that consists of sum of certain terms can be approximated by a ratio of the product of terms (of the form $(z - e^{\rho+j\theta})$), leading to a rational 'z' transform. Specifically, as explained in (Anderson et al., 1983; Satorius et al., 1978) and making the following assumptions,

- The pole pairs in (5.15) lie sufficiently close to the unit circle (easily satisfied as $\alpha \simeq 0$.)
- All the frequency pairs (ω_i, ω_j) in (5.15) are sufficiently separated from each other such that their contribution to the total power spectrum do not overlap significantly.

the z transform of the total input can be expressed as,

$$\begin{aligned} S_{xx}(z) &= \sigma^2 \frac{\prod_{m=1}^{M}(z - e^{-\beta_m + j\omega_m})(z - e^{+\beta_m + j\omega_m})}{\prod_{m=1}^{M}(z - e^{-\alpha_m + j\omega_m})(z - e^{+\alpha_m + j\omega_m})} \\ &\quad \times \frac{(z - e^{+\beta_m - j\omega_m})(z - e^{-\beta_m - j\omega_m})}{(z - e^{+\alpha_m - j\omega_m})(z - e^{-\alpha_m - j\omega_m})} \\ \text{where } \alpha_m &= 1/N \end{aligned} \tag{5.19}$$

Corresponding to each of the sinusoidal component in the input signal there are four poles at locations $z = e^{\pm\alpha \pm \omega_m}$ and there are four zeros on the same radial lines as the signal poles but at different distances away from the unit circle. Using the general solution described in (5.12), which has been derived at length in (Anderson et al., 1983), the solution of the LeSF weight vector to the present problem is,

$$w(n) = \sum_{m=1}^{M} \left(B_m e^{-\beta_m n} + C_m e^{+\beta_m n}\right) \cos \omega_m (n + P) \tag{5.20}$$

The values of β_m, B_m and C_m can be determined by substituting (5.20) and (5.14) in (5.9). The detailed solution of this equation is provided in the Appendix A. The l^{th} equation in the linear-system described in (5.9) has terms with coefficients $exp(-\beta_m l)$, $exp(+\beta_m l)$, $exp(-\alpha l)\cos(\omega_m(l + P))$ and $exp(\alpha l)\cos(\omega_m(l + P))$. Besides these, there are two other kind of terms that can be neglected.

- "Non-stationary" terms that are modulated by a sinusoid at frequency $2\omega_m$ where $m \in [1, M]$. For $\omega_m \neq 0$, $\omega_m \neq \pi$, their total contribution is approximately zero.[3]
- Interference terms that are modulated by a sinusoid at frequency $\Delta\omega = (\omega_i - \omega_j)$ where $(i, j) \in [1, \ldots, M]$. If filter length $L \gg 2\pi/\Delta\omega$, these interference terms approximately sum up to zero and hence can be neglected.

The coefficients of the terms $exp(-\beta_m l)$, $exp(+\beta_m l)$ are the same for each of the L equations and setting them to zero leads to just one equation which relates β_m to α and the SNR. Let ρ_i denote

[3] due to self cancelling positive and negative half periods of a sinusoid.

the "partial" SNR of the sinusoid at frequency ω_i i.e $\rho_i = A_i^2/\sigma^2$ and the complementary signal SNR be denoted as $\gamma_i = (\sum_{m=1,m\neq i}^{M} A_i^2)/\sigma^2$. Then we have the following relation,

$$\cosh\beta_i = \cosh\alpha + \frac{\rho_i}{2\gamma_i + \rho_i + 2}\sinh\alpha \tag{5.21}$$

There are two interesting cases. First case is when the sinusoid at frequency ω_i is significantly stronger than other sinusoids such that γ_i is quite low. This is illustrated in figure (5.3), where we plot the bandwidth β_i of the LeSF's pass-band that is centered around ω_i as a function of the partial SNR of the i^{th} sinusoid, ρ_i. The complementary signal's SNR is quite low at $\gamma_i = -6.99db$. We plot curves for different "effective" input sinusoid's bandwidth α. From (5.15), we note that α is reciprocal of frame length N. The vertical line in figure (5.3) corresponds to the case when $\rho_i = \gamma_i$. We note that for a given partial SNR ρ_i, the LeSF bandwidth becomes narrower as the frame length N increases, indicating a better selectivity of the LeSF filter.

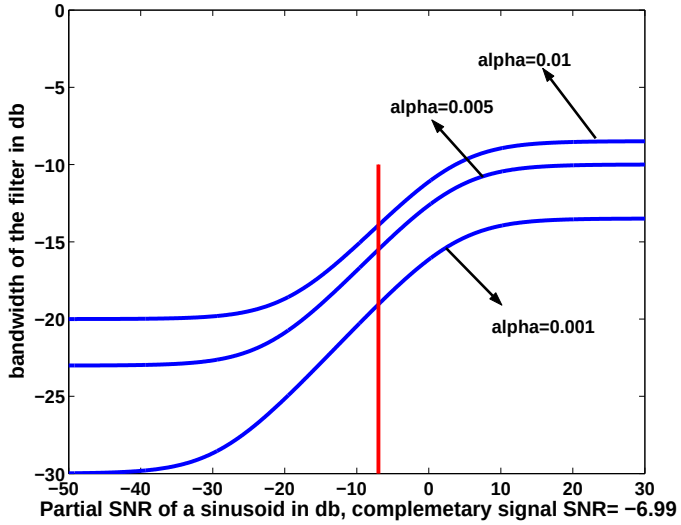

Figure 5.3. Plot of the filter bandwidth β_i centered around frequency ω_i as a function of partial sinusoid SNR ρ_i for a given complementary signal SNR $\gamma_i = -6.99db$ and "effective" input bandwidth $\alpha(alpha) = 0.01, 0.005, 0.001$ respectively. The vertical line meets the three curves when $\rho_i = \gamma_i$.

In figure (5.4), we plot the bandwidth β_i as a function of ρ_i for the cases when complementary signal SNR is high at $\gamma_i = 10db$ and is low at $\gamma_i = -6.99db$. The two dots correspond to the case when $\rho_i = \gamma_i$. We note that $\gamma_i = 10db$ corresponds to a signal with high overall SNR[4]. Therefore the cross-over point ($\gamma_i = \rho_i$) for low γ_i occurs at narrower bandwidth as compared to high γ_i case. This is so because in the former case the overall signal SNR is low and thus the LeSF filter has to have narrower pass-bands to reject as much of noise as possible.

B_i and C_i in (5.20) are determined by equating their respective coefficients. The "non-stationary" interference terms between all of the pairs of the frequency (ω_i, ω_j), can be neglected if $(\omega_i - \omega_j) >> 2\pi/L$. This requires that LeSF's frequency resolution $(2\pi/L)$ should be able to resolve the constituent sinusoids.

$$\begin{aligned} B_i &= \frac{2e^{-\beta_i}e^{-\alpha P}(\alpha+\beta_i)^2(\beta_i-\alpha)}{((\alpha+\beta_i)^2 - e^{-2\beta_i L}(\beta_i-\alpha)^2)} \\ C_i &= \frac{2e^{-\beta_i(2L+1)+1}e^{-\alpha P}(\alpha+\beta_i)(\beta_i-\alpha)^2}{((\alpha+\beta_i)^2 - e^{-2\beta_i L}(\beta_i-\alpha)^2)} \end{aligned} \tag{5.22}$$

[4]As overall SNR of the signal = $10\log_{10}(10^{10\gamma_i} + 10^{10\rho_i})$

We note from (5.21) that the various sinusoids are coupled with each other through the dependence of their bandwidth β_i on the complementary signal SNR γ_i. As a consequence of that B_i, C_i are also indirectly dependent on the powers of the other sinusoids through β_i.

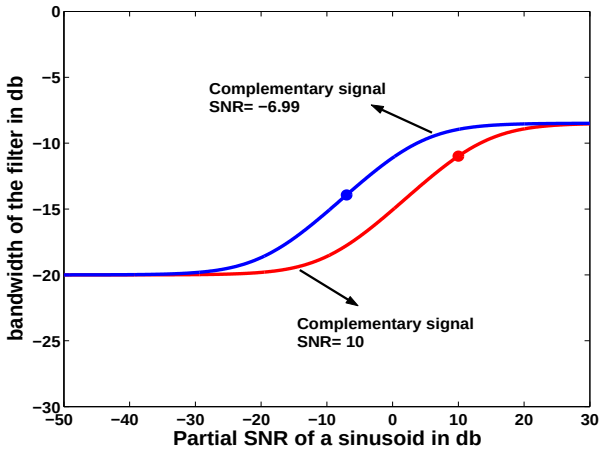

Figure 5.4. Plot of the filter bandwidth β_i centered around frequency ω_i as a function of partial sinusoid SNR ρ_i for given complementary signal SNRs $\gamma_i = -6.99db, 10db$ respectively. The "effective" input bandwidth $\alpha(alpha) = 0.01$ for both the curves. The two dots correspond to the cases when the partial SNR ρ_i is equal to complementary signal SNR γ_i.

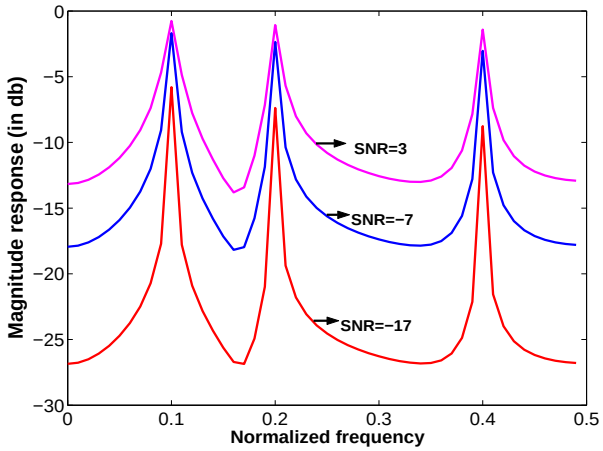

Figure 5.5. Plot of the magnitude response of the LeSF filter as a function of the input SNR. The input consists of three sinusoids at normalized frequencies (0.1, 0.2, 0.4) with relative strength $(1 : 0.6 : 0.4)$ respectively.

In Fig.5.5, the magnitude response of the LeSF filter is plotted for various SNR. The input in this case consist of three sinusoids at normalized frequencies (0.1, 0.2, 0.4). The frame length is $N = 500$ and filter length is $(L = 100)$. As the signal SNR decreases, the bandwidth of the LeSF filter starts to decrease in order to reject as much of noise as possible. The LESF filter's gain decreases with decreasing SNR. Similar results were reported in (Anderson et al., 1983; Zeidler et al., 1978) for the case of stationary inputs.

In Fig. 5.6, we plot the spectrograms of a clean speech utterance. Fig. 5.7 and Fig. 5.8 display the same utterance embedded in F16-cockpit noise at SNR 6dB and its LeSF enhanced version

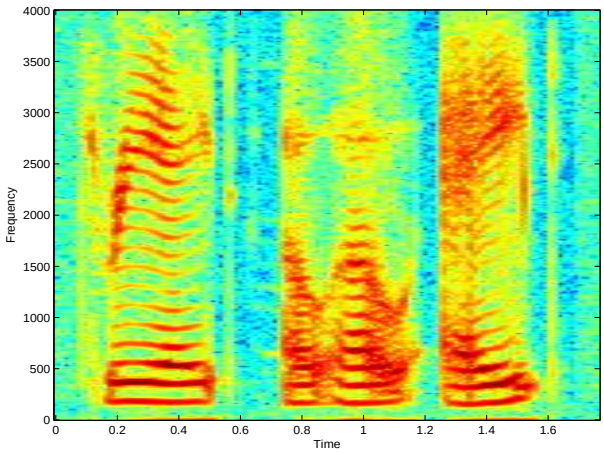

Figure 5.6. Clean spectrogram of an utterance from the OGI Numbers95 database

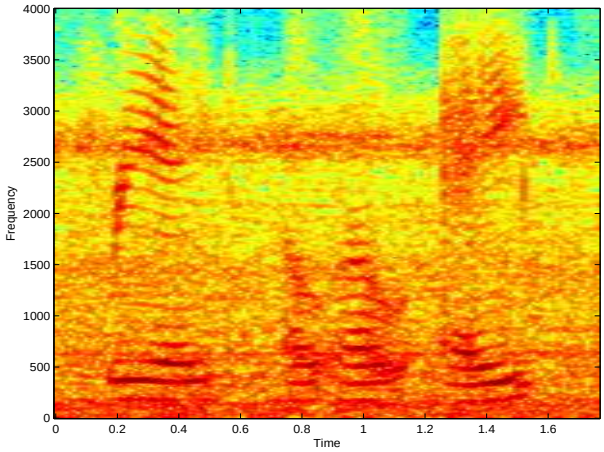

Figure 5.7. Spectrogram of the utterance corrupted by F16-cockpit noise at 6dB SNR.

respectively. As can be seen from the spectrograms, the LeSF filter has been able to reject significant amount of additive F-16 cockpit noise (Varga et al., 1992) from the speech signal.

5.4 Gain of the LeSF filter

The LeSF filter output consist of filtered sinusoids and the filtered noise signal. For the input signal described by (5.7) that is filtered by a LeSF filter with coefficients as in (5.20), the output filtered signal power P_{signal} and the output filtered noise power P_{noise} are approximately[5] given by,

[5]assuming $N \gg L$ such that initial L samples can be used to initialize the filter

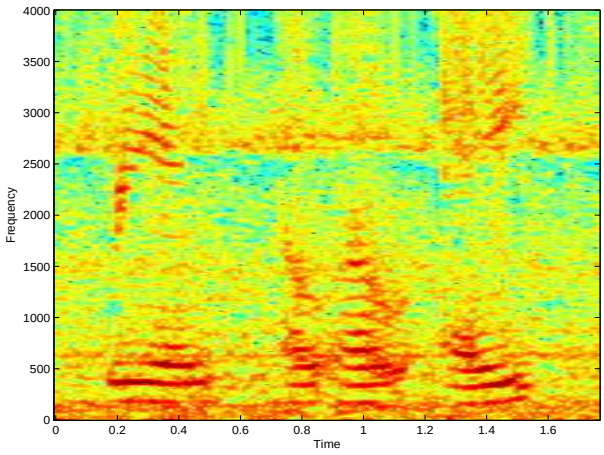

Figure 5.8. Spectrogram of the noisy utterance enhanced by a $(L = 100)$ tap LeSF filter that has been estimated over blocks of length $(N = 500)$.

$$P_{signal} = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} w(i)w(j)r(|i-j|) \tag{5.23}$$

$$= \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} w(i)w(j) \times \frac{(N-|i-j|)}{N}\sum_{m=1}^{M} A_m^2 cos(2\pi f_m(i-j)) \tag{5.24}$$

$$P_{noise} = \sum_{n=0}^{L-1} \sigma^2 w^2(n)$$

$$= \sum_{i=1}^{M}\left[B_i^2 e^{-\beta_i L} + C_i^2 e^{\beta_i L}\right]\frac{\sigma^2 \sinh(\beta_i L)}{2\beta_i} + \sigma^2 BCL \tag{5.25}$$

The output SNR of the LeSF filter is given by (5.23) divided by (5.25). The LeSF gain is given by the ratio of the output SNR to the input $SNR = \sum_{i=1}^{M} A_i^2/(2\sigma^2)$.

In Fig. 5.9, we plot the LeSF broadband gain as a function of input SNR, for a fixed block length of $N = 500$ and varying filter lengths ($L = 100, 80, 60$). As can be noted from the Fig. 5.9, the LeSF broadband gain approaches a horizontal asymptote for decreasing input SNR. This is in agreement with the fact that the bandwidth β_i of the LeSF filter approaches a horizontal asymptote (Fig. 5.3) for decreasing SNR. We note from Fig. 5.9 that the LeSF broadband gain increases as the filter length L increases for a fixed block length $N = 500$. However, since the L tap LeSF filter is estimated using the N samples from a given block, the filter length cannot be increased arbitrarily and is limited from above by the block length 'N'.

In Fig. 5.10, we plot the LeSF broadband gain as a function of input SNR, for a fixed filter length $L = 100$ and varying block lengths ($N = 300, 400, 500$). We note that the LeSF broadband gain increases as the block length N increases. However, for a non-stationary signal such as speech, as the block length increases, the corresponding power spectrum will become more broad-band.

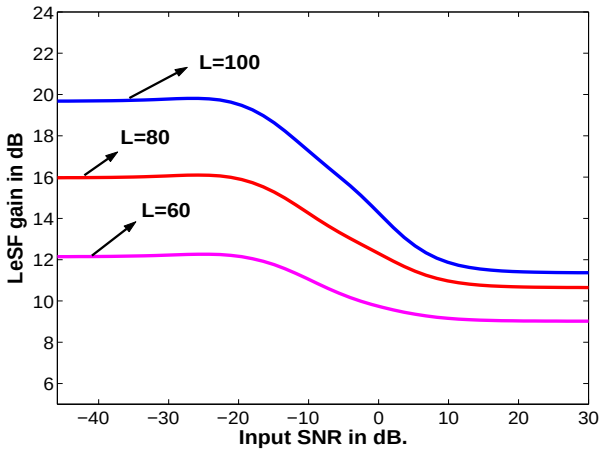

Figure 5.9. LeSF gain plotted as a function of input SNR for fixed block length $N = 500$ and various filter lengths $L = 100, 80, 60$.

Therefore we will not be able to model the corresponding block as a sum of a small number of sinusoids M as done in (5.7). As a result the number of sinusoids M will be large and possibly closely spaced to each other, leading to significant interference terms between the constituent sinusoids in (5.8) and (5.14).

5.5 Experiments and Results: OGI Numbers95

In order to assess the effectiveness of the proposed algorithm, speech recognition experiments were conducted on the OGI Numbers95 (Cole et al., 1994) corpus. This database contains spontaneously spoken free-format connected numbers over a telephone channel. The lexicon consists of 31 words. Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition systems were trained using public domain software HTK (Young et al., 1995) on the clean training set from the original Numbers corpus. The system consisted of 80 tied-state triphone HMM's with 3 emitting states per triphone and 12 mixtures per state. To verify the robustness of the features to noise, the clean test utterances were corrupted using Factory and F-16 cockpit noise from the Noisex92 (Varga et al., 1992) database.

5.5.1 Bulk Delay P

Noting that the autocorrelation coefficients of a periodic signal are themselves periodic with the same period (hence they do not decay with the increasing lag), Sambur (Sambur, 1978) has used a bulk delay equal to the pitch period of the voiced speech for its enhancement. However, for the un-voiced speech a high bulk delay will result in a significant distortion by the LeSF filter as its autocorrelation coefficients decay much more rapidly than those of the voiced speech. Therefore, we kept the bulk delay at '$P = 1$' as a good choice for enhancing both the voiced and un-voiced speech frames.

5.5.2 Block length N and filter length L

Speech signals were blocked into frames of (N=500) samples (62.5ms) each and a (L=100) tap LeSF filter was derived using (5.4) for each frame that could be either voiced or unvoiced. The relatively

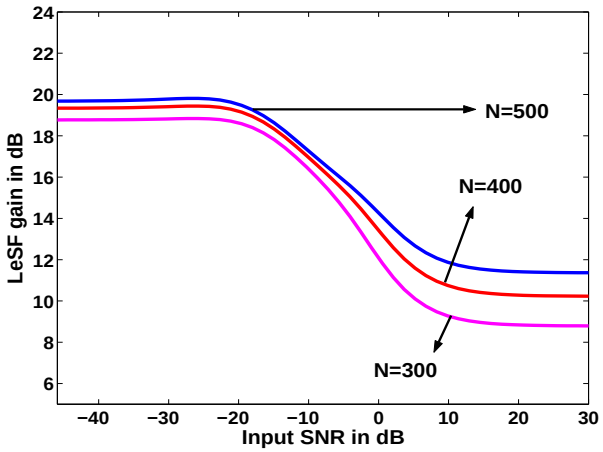

Figure 5.10. LeSF gain plotted as a function of input SNR for fixed filter length $L = 100$ and various block lengths $N = 300, 400, 500$.

wer

high order $(L = 100)$ of the LeSF filter is required to be able to have sufficiently high frequency resolution $(2\pi/L)$ to resolve the constituent sinusoids in case of voiced speech. Each speech frame was then filtered through its corresponding LeSF filter to derive an enhanced speech frame. Finally MFCC feature vector was computed from the enhanced speech frame. These enhanced LeSF-MFCC were compared to the baseline MFCC features and noise robust CJ-RASTA-PLP (Hermansky and Morgan, 1994a) features. The MFCC feature vector computation is the same for the baseline and the LeSF-MFCC features. The only difference is that the MFCC baseline features are computed directly from the noisy speech while the LeSF-MFCC features are computed from LeSF enhanced speech signal. We also compared our technique with the soft-decision spectral subtraction based technique. In (Lathoud et al., 2005), authors have used a speech presence probability in conjunction with spectral subtraction to achieve noise robustness. This can be seen as a soft-decision spectral subtraction which has been shown to be superior than hard-decision spectral subtraction by Mcaulay et al. (McAulay and Malpass, 1980). As the train set, test set and the factory noise environment in our experiments and those of (Lathoud et al., 2005) are the same, we quote the ASR results for the factory noise directly from (Lathoud et al., 2005). The authors in (Lathoud et al., 2005) propose three features based on the soft-spectral subtraction, which vary in their preprocessing steps and are termed POST-FILT, POWER-FILT and PSIL. In table 5.1, we quote their ASR word error rates in the factory noise environment, directly from the results reported in (Lathoud et al., 2005). We note that the proposed technique outperforms all three soft-decision spectral subtraction variants.

The speech recognition results for the baseline MFCC, C-JRASTA-PLP (Hermansky and Morgan, 1994a) and the proposed LeSF-MFCC, in various levels of noise are given in Tables 5.2 and 5.3. We have used the default value of constant $J = 10^{-6}$ in the C-JRASTA-PLP feature. All the reported features in this paper have cepstral mean subtraction (CMS). The proposed LeSF processed MFCC performs significantly better than others in all noise conditions. The slight performance degradation of the LeSF-MFCC in the clean is due to the fact that the LeSF filter being an all-pole filter does not model the valleys of the clean speech spectrum well. As a result, the LeSF filter sometimes amplifies the low spectral energy regions of the clean spectrum.

Table 5.4 shows the word error rate of the LeSF enhanced MFCC features for a fixed block length of 500 samples $(62.5ms)$ and varying LeSF filter length 'L'. We note that the word error rate decreases as the filter length increases. This is so because a higher filter length results in a sharper

Table 5.1. Word error rate results for factory noise using soft-decision spectral subtraction. All features have cepstral mean subtraction.

SNR	LeSF MFCC	POST-FILT	POWER-FILT	PSIL
Clean	6.6	8.1	8.3	7.1
12 dB	11.3	16.2	17.0	15.7
6 dB	20.0	30.7	31.2	28.7
0 db	41.3	63.1	61.9	58.2

Table 5.2. Word error rate results for factory noise. Parameters of the LeSF filter, L=100 and N=500. C-JRASTA-PLP used with the constant $J = 10^{-6}$ which is the default value. All features have cepstral mean subtraction.

SNR	MFCC	C-JRASTA-PLP	LeSF MFCC
Clean	5.7	7.8	6.6
12 dB	14.0	12.2	11.3
6 dB	31.5	23.8	20.0
0 db	75.7	59.8	41.3

frequency response of the LeSF filter(narrower band-width of the passbands), thereby enabling it to reject as much of the broad-band noise as possible that lies away from the frequencies of the constituent sinusoids of the clean signal.

5.6 Conclusion

We consider a class of non-stationary signals as input that are composed of either (a) multiple sinusoids (voiced speech) whose frequencies and the amplitudes may vary from block to block or, (b) output of an all-pole filter excited by white noise input (unvoiced speech segments) and which are embedded in white noise. We have derived the analytical expressions for the impulse response of the L-weight least squares filter (LesF) as a function of the input SNR (computed over the current frame), effective band-width of the signal (due to finite frame length), filter length 'L' and frame length 'N'. Recognizing that such a time-varying sinusoidal model (McAulay and Quatieri, 1986) and the source-filter model are a reasonable approximation to the voiced speech and unvoiced speech respectively, we have applied the block estimated LeSF filter for de-noising speech signals embedded in the realistic (Varga et al., 1992) broad-band noise as commonly encountered on a factory floor and an aircraft cockpit. The proposed technique leads to a significant improvement in ASR performance as compared to noise robust CJ-RASTA-PLP (Hermansky and Morgan, 1994a), speech presence probability based spectral subtraction (Lathoud et al., 2005) and the MFCC features computed from the unprocessed noisy signal.

Least squares filtering is a speech signal enhancement technique that "cleans" the noisy speech signal to make it as similar to the clean speech signal as possible. Therefore, once the noisy speech signal has been cleaned, the mismatch between the clean acoustic model and the noisy test utterance (that is enhanced by the least squares filtering) is minimized leading to noise robust ASR. Therefore, in a way the least squares filtering plays a dual role of not only enhancing the speech signal but also of providing noise robustness to the features. This is particularly suitable for the mobile telephony applications where the same signal processing algorithm, typically running on a handheld device, enhances the noisy speech signal and also provides noise robust features for the distributed ASR. In the next two chapters, we will present "feature-level" noise robustness techniques. Unlike the least squares filtering technique, these techniques do not work at the sampled speech signal level but at the power spectrum or the Mel-filter bank energy level. Therefore, these

Table 5.3. Word error rate results for F16-cockpit noise. Parameters of the LeSF filter, L=100 and N=500. C-JRASTA-PLP used with the constant $J = 10^{-6}$ which is the default value. All features have cepstral mean subtraction.

SNR	MFCC	C-JRASTA-PLP	LeSF MFCC
Clean	5.7	7.8	6.6
12 dB	15.8	14.2	12.5
6 dB	32.8	25.3	21.0
0 db	75.1	59.2	41.0

Table 5.4. Word error rate results for factory noise for varying length, $L = 100, 50, 20$ of the LeSF filter. The block length, N is 500 (62.5ms).

SNR	LeSF L=20	LeSF L=50	LeSF L=100
Clean	9.3	7.3	6.6
12 dB	14.3	12.3	11.3
6 dB	24.4	22.0	20.0
0 db	46.5	43.0	41.3

techniques do not enhance the speech signal but rather make the final feature representation quite immune to the noise through non-linearly transforming the Mel-filter bank energies.

(5.25)where,

$x'(n)$ denotes "discrete" derivative of x. Therefore the sensitivity of the DCT of the logMelFBS can be "approximately" measured in terms of the sensitivity of derivatives of the logMelFBS. We define the sensitivity index $\rho(a, b)$ as the ratio of derivatives of the function $log(x)$ at a Mel-formant energy $x = a$ and a low Mel-filter bank energy value $x = b$. Given (**??**), we expect $\rho(a, b)$ to measure the relative contributions of a peak of the logMelFBS and the low energy Mel-filter bank energies in a DCT coefficient which is a cepstral coefficient.

$$\begin{aligned}\rho(a,b) &= \frac{log'(x)|_{x=a}}{log'(x)|_{x=b}} = \frac{1/a}{1/b} \\ &= b/a \; where \; a \gg b \\ &\Rightarrow \rho(a,b) \ll 1.00\end{aligned} \tag{5.26}$$

Similarly we define the sensitivity index $\sigma(a, b)$ as the ratio of the derivatives of the function $sign(log(x))[log(x)]^P$ at a Mel-formant energy $x = a$ and a low Mel-filter bank energy value $x = b$.

$$\begin{aligned}\sigma(a,b) &= \frac{P[sign(log(a))][log(a)]^{P-1}/a}{P[sign(log(b))][log(b)]^{P-1}/b} \\ &= \frac{[sign(log(a))][log(a)]^{P-1}}{[sign(log(b))][log(b)]^{P-1}}(b/a) \\ &= \frac{[sign(log(a))][log(a)]^{P-1}}{[sign(log(b))][log(b)]^{P-1}}\rho(a,b) \; where \; a \gg b \\ &\Rightarrow \sigma(a,b) > \rho(a,b) \; where \; a \gg b, P > 1.0\end{aligned} \tag{5.27}$$

The value of $\rho \ll 1.0$ in (5.26) implies that a unit change in the low Mel-filter bank energy value, namely "b" will have a far greater influence on the computation of the DCT of logMelFBS as compared to a unit change in the Mel-formant energy, namely "a". Therefore, it can be seen in the light of (5.26) that the DCT of the logMelFBS is quite sensitive to the perturbations in the low-energy

regions as compared to those around the formants. However, for the domain $1.0 \leq b \ll a < \infty$ and $P > 1$, $\sigma(a,b)$ is always greater than $\rho(a,b)$. This can be achieved by using $(log(x+1))^P$ as x being power spectral energy never takes negative values. The fact that the $\sigma(a,b)$ is always greater than the $\rho(a,b)$ implies that we have been able to decrease the sensitivity of cepstral coefficients to spurious low energy perturbations. An important parameter in the above mentioned processing scheme is the exponent P. As can be seen from (5.27), the sensitivity ratio $\sigma(a,b)$ increases exponentially as the exponent P increases. However, a large value of P will result in the case where the spectral modulations of the largest formant takes very high numerical values which render the spectral modulations of the other formants numerically insignificant relative to those of the largest formant. Therefore an intermediate value of P is the most suitable for such a processing scheme.[6]

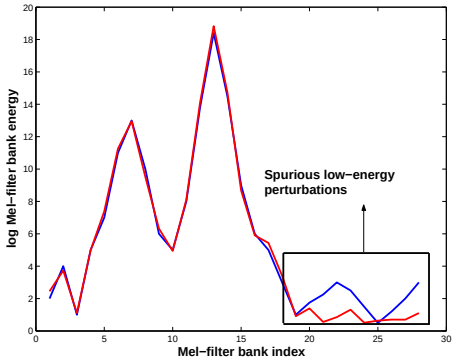

Figure 5.11. Log Mel-filter bank energies of clean and noisy(perturbed) speech.

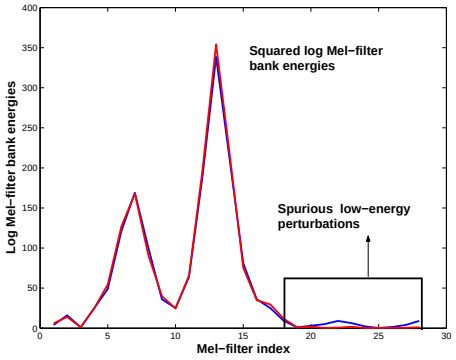

Figure 5.12. Square of the log Mel-filter bank energies of clean and noisy(perturbed) speech.

5.7 Relationship between cepstrum and exponentiated cepstrum

Let $c_{log}(n) = IDFT \log|X(k)|$ be the log cepstrum, $c_{root}(n) = IDFT|X(k)|^{\gamma}$ be the root cepstrum, $c_{expoLog}(n) = IDFT(\log|X(k)|)^P$ be the proposed exponentiated log cepstrum. Then Lim (Lim,

[6]The experiments results with different values of P reconfirmed these observations.

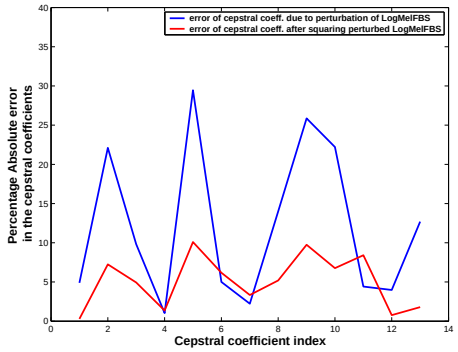

Figure 5.13. Absolute percentage error between the cepstral coefficients due to perturbations. Blue curve corresponds to the DCT of the log Mel-filter bank spectrum while red curve corresponds to the DCT of the squared log Mel-filter bank spectrum.

1979) has shown that for root γ close to zero,

$$c_{root}(n) = \delta(n) + c_{log}(n) \tag{5.28}$$

In general, the proposed exponentiated log cepstrum $c_{expoLog}(n)$ is,

$$c_{expoLog}(n) = \textbf{IDFT}(\log(|X(e^{j\omega})|^2))^P \tag{5.29}$$

In particular for $P = 2$, we have,

$$c_{expoLog}(n) = c_{log}(n) * c_{log}(n) \tag{5.30}$$

$$= R_{temporal}(n) \tag{5.31}$$

where '*' denotes convolution and $R_{temporal}(n)$ is the temporal autocorrelation of the log cepstrum. This leads to the interpretation that sharp "spikes" in the log cepstrum are smoothed out by the proposed technique(the implicit convolution of the $c_{log}(n)$ with itself), leading to a more smoother cepstrum.

5.8 Experiments and Results: OGI Numbers95

In order to assess the effectiveness of the proposed scheme for reducing the effect of spurious perturbations in the low Mel-filter bank energies, speech recognition experiments were conducted on the OGI Numbers95 corpus (Varga et al., 1992) using the proposed processing scheme for the log-MelFBS. The lexicon size for this connected digits recognition task is 30 words with 27 different phonemes. To verify the robustness of the features to noise, the clean test utterances were corrupted using additive non-stationary *factory* noise and *f16 cockpit* noise from the Noisex92 (Varga et al., 1992) database. Throughout the experiments, Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980) and their temporal derivatives have been used as speech features. Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition systems were trained using public domain software HTK (Young et al., 1995) on the clean training set from the original Numbers95 corpus. The system consisted of 80 tied-state triphone HMM's with 3 emitting states per triphone and 12 mixtures per state. Three kinds of feature sets were generated:

- [MFCC+Deltas:] 13 MFCCs with deltas.

- [RMFCC+Deltas: generated by root Mel-filter bank spectrum with $R = 0.10$] 13 root Mel-cepstral coefficients with deltas.

- [ExpoMFCC+Deltas: generated by exponentiated logMelFBS with $P = 2.7$] 13 exponentialted log-Mel-cepstral coefficients with deltas.

Per utterance cepstral mean subtraction was applied to each of the above feature vectors. The speech recognition results using the above mentioned feature sets in clean and noisy conditions are reported in table 5.5. The root $R = 0.10$ and the exponent $P = 2.7$ gave the best recognition results for the RMFCC and ExpoMFCC features respectively. The exponentiated logMelFBS MFCC system performs better than the usual MFCC features in the noisy conditions. We note that the performance of the proposed features is similar to that of RMFCC features using the optimal value of the root $R = 0.10$. Figure 5.14 illustrates the fact that the proposed technique can reduce the mismatch between clean and noisy MFCC features.

In the next chapter, we will incorporate cepstral modulation features that have in-turn been derived from the exponentiated MFCC. The resulting representation benefits from the noise robustess properties of the Expo-MFCC and those of the cepstral modulation spectrum.

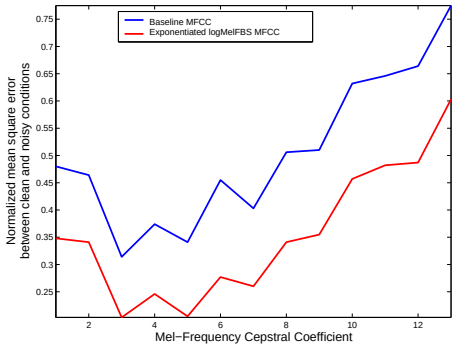

Figure 5.14. Mean square error of MFCC vectors in clean and noisy conditions, normalized by the average power of the corresponding MFCC feature vector in clean condition. Blue curve corresponds to baseline MFCC while red curve corresponds to MFCC derived by squaring the log Mel-filter bank spectrum. These mean estimates were computed using nearly 160000 speech frames.

Table 5.5. Word error rate results for factory and f16 noise. The best results for RMFCC (R=0.10) and Exponentiated MFCC (P=2.7) are reported.

SNR	MFCC	RMFCC	**ExpoMFCC**
Clean	5.7	6.1	6.2
Fact SNR 12	14.0	12.0	11.6
Fact SNR 6	31.5	20.6	20.3
Fact SNR 0	75.7	45.7	44.3
F16 SNR 12	15.8	12.3	12.1
F16 SNR 6	32.8	20.8	20.9
F16 SNR 0	75.1	44.2	43.4

5.9 Conclusion

We identify a numerical sensitivity problem with the MFCC (Davis and Mermelstein, 1980) features. It is analytically shown that by exponentiating the logMelFBS one can desensitize the MFCC coefficients to spurious low-energy spectral perturbations.We show the relationship between the conventional MFCC features and the Expo-MFCC features. In particular for the power $P = 2$, the expo-MFCC features can be thought to be obtained by convolving the usual MFCC features with itself. This leads to the interpretation that that sharp "spikes" in the MFCCs are smoothed out by the proposed technique, leading to a more smoother cepstrum. Experimental results show that useful noise robustness can be achieved by the use of the proposed features in all conditions as compared to the MFCC features.

In the next chapter, we will incorporate cepstral modulation features that have in-turn been derived from the exponentiated MFCC. As will be shown in the next chapter, the resulting representation benefits from the noise robustess properties of the Expo-MFCC and those of the cepstral modulation spectrum leading to a composite feature that is highly noise robust.

Chapter 6

Mel-Cepstrum Modulation Spectrum (MCMS) features for Robust ASR

6.1 Introduction

A central result from the study of the human speech perception is the importance of slow changes in speech spectrum for speech intelligibility (Dudley, 1939). A second key to human speech recognition is the integration of phonetic information over relatively long intervals of time. Speech is a dynamic acoustic signal with many sources of variation. As noted by Furui (Furui, 1986, 1990), spectral changes are a major cue in phonetic discrimination. Moreover, in the presence of acoustic interference, the temporal characteristics of speech appear to be less variable than the static characteristics (Kingsbury et al., 1998b). Therefore, representations and recognition algorithms that better use the information based on the specific temporal properties of speech should be more noise robust (Hermansky and Morgan, 1994b). Temporal derivative features (Furui, 1986, 1990) of static spectral features like filter-bank, Linear Prediction (LP) (Markel and Gray, 1976) , or mel-frequency cepstrum (Davis and Mermelstein, 1980) have yielded improvements in noise robust ASR performances. Similarly, the RASTA processing (Hermansky and Morgan, 1994b) and cepstral mean normalization (CMN) techniques, which perform cepstral filtering, have provided a remarkable amount of noise robustness.

Using these temporal processing ideas, we have developed a speech representation for the noise robust ASR applications in particular which, factors the cepstral changes over time into slow and fast moving orthogonal components. Any DFT coefficient of a speech frame, considered as a function of frame index with the discrete frequency fixed, can be interpreted as the output of a linear time-invariant filter with a narrow-bandpass frequency response. Therefore, taking a second DFT of a given spectral band, across frame index, with discrete frequency fixed, will capture the spectral changes in that band with different rates. This effectively extracts an approximate modulation frequency response of the spectral band. However, the final feature representation used in an ASR system are typically the cepstral features as they are known to be highly uncorrelated and are modeled very well by the diagonal covariance matrices. Therefore instead of using a DFT of the spectral trajectories as a feature, we have used a DFT of the cepstral trajectories as a noise robust feature. We note this representation as the Mel cepstral modulation spectrum (MCMS).

We note that although MCMS is a form of a modulation spectrum in the cepstral domain, it is

very different from the fepstrum feature that was derived in the Chapter 4. As described in Chapter 4, demodulating an AM signal through the use of its analytic signal in the time domain, is an exact demodulation technique. Whereas, computing a DFT over cepstral trajectories- where each of the cepstral samples in the trajectory have been computed through the Mel-smoothed DFT of 20ms of speech signal stepped forward by 10ms- is an inexact demodulation technique. Nevertheless, as the MCMS features are intricately related to the cepstral changes over time, they can be seen as a generalization of the delta features (Furui, 1986, 1990) and our goal here, is to use them as a substitute for the delta features.

In Section 6.2, we first give an overview and visualisation of the modulation frequency response. The visual representation is shown to be very stable in presence of additive noise. The proposed MCMS dynamic features are then derived in Section 6.3. Finally, Section 6.4 compares the performance of the MCMS features with standard temporal derivative features in recognition experiments on the Numbers database for non-stationary noisy environments. In Section 6.5, we have computed the MCMS features from the cepstral trajectories of the exponentiated MFCC features that were derived in the Chapter **??**. The experimental result indicate that the performance gains are additive from both the techniques and it results in a highly noise robust feature.

6.2 Approximate Modulation Frequency Response of Speech

Let $X[n,\ k]$ be the DFT of a speech signal $x[m]$, windowed by a sequence $w[m]$. Then, by rearrangement of terms, the DFT operation could be expressed as:

$$X[n,\ k] = x[n] * h_k[n] \tag{6.1}$$

where $'*'$ denotes convolution and:

$$h_k[n] = w[-n]e^{\frac{j2\pi kn}{M}}. \tag{6.2}$$

From (6.1) and (6.2), we can make the well-known observation that the k^{th} DFT coefficient $X[n,\ k]$, as a function of frame index n, and with discrete frequency k fixed, can be interpreted as the output of a linear time invariant filter with impulse response $h_k[n]$. Taking a second DFT, of the time sequence of the k^{th} DFT coefficient, will factorize the spectral dynamics of the k^{th} DFT coefficient into slow and fast moving modulation frequencies. We call the resulting second DFT the "Modulation Frequency Response" of the k^{th} DFT coefficient. Let us define a sequence $y_k[n] = X[n,\ k]$. Then taking a second DFT of this sequence over P points gives:

$$Y_k(q) = \sum_{p=0}^{P-1} y_k(n+p)e^{\frac{-j2\pi qp}{P}}\ ,\ \ q \in [0,\ P-1] \tag{6.3}$$

$$Y_k(q) = \sum_{p=0}^{P-1} X[n+p,\ k]e^{\frac{-j2\pi qp}{P}}$$

where $Y_k(q)$ is termed the q^{th} modulation frequency coefficient of k^{th} primary DFT coefficient. Lower $q's$ correspond to slower spectral changes and higher $q's$ correspond to faster spectral changes. For example, if the spectrum $X[n,\ k]$ varies a lot around the frequency k, then $Y_k(q)$ will be large for higher values of modulation frequency, q. This representation should be noise robust, as the temporal characteristics of speech appear to be less variable than the static characteristics.

To illustrate the modulation frequency response, in the following we derive a modulation spectrum based on (6.3), and plot it as a series of modulation spectrograms. This representation emphasizes the temporal structure of the speech and displays the fast and slow modulations of the spectrum. This modulation spectrum is a function of three quantities with time n (6.1), linear frequency k (6.1) and modulation frequency q (4.5) being the three variables.

Let $C[n,\ l]$ be the real cepstrum of the DFT $X[n,\ k]$.

$$C[n,\ l] = \frac{1}{K-1}\sum_{k=0}^{K}\log(\mid X[n,\ k]\ \mid)e^{\frac{+j2\pi kl}{K}}\ ,\ \ l \in [0,\ K-1] \tag{6.4}$$

Using a rectangular low quefrency lifter which retains only the first 12 cepstral coefficients, we obtain a smoothed estimate of the spectrum, noted $S[n,\ k]$.

$$\log S[n,\ k] = C[n,\ 0] + \sum_{l=1}^{L-1} 2C[n,\ l]\cos(\frac{2\pi lk}{K}) \tag{6.5}$$

where we have used the fact that $C[n,\ l]$ is a real symmetric sequence. The resulting smoothed spectrum $S[n,\ k]$ is also real and symmetric. $S[n,\ k]$ is divided into B linearly spaced frequency bands and the average energy, $E[n,\ b]$, in each band is computed.

$$E[n,\ b] = \frac{1}{K/B}\sum_{i=0}^{K/B-1} S[n,\ b\frac{K}{B}+i]\ ,\ \ b \in [0,\ K/B-1] \tag{6.6}$$

Let $M[n,\ b,\ q]$ be the magnitude modulation spectrum of band b computed over P points.

$$\begin{aligned} M[n,\ b,\ q] =& \mid \textstyle\sum_{p=0}^{P-1} E[n+p,\ b]e^{\frac{-j2\pi pq}{P}}\ \mid, \\ & with\ q \in [0, P],\ b \in [0,\ K/B-1] \end{aligned} \tag{6.7}$$

The modulation spectrum $M[n,\ b,\ q]$ is a 4-dimensional quantity being a function of time n, frequency-band b and modulation frequency q. Keeping the frequency band number b fixed, it can be plotted as a conventional spectrogram. Figures 6.1 and 6.2 show conventional spectrograms of a clean speech utterance and its noisy version at SNR 6. Whereas, the Figures 6.3 and 6.4 show modulation spectrograms of the same clean and noisy utterance as above. The stability of modulation spectrogram towards additive noise can be easily noticed in these figures. The figures consists of 16 modulation spectrograms, corresponding to each of 16 frequency bands in (6.6), stacked on top of each other. In our implementation, we have used a frame shift of 3 ms and the primary DFT window of length 32 ms. The secondary DFT window has a length $P = 41$ which is equal to 3 ms*40=120 ms. This size was chosen, assuming that this would capture phone specific modulations rather than average speech like modulations. We divided $[0,\ 4kHz]$ into 16 bands for the computation of modulation spectrum in (6.7). For the second DFT the Nyquist frequency is 333.33 Hz. We have only retained the modulation frequency response up to 50 Hz as there was negligible energy present in the band [50 Hz, 166 Hz]. For every band, we have shown the modulation spectrum with $q \in [1,\ 20]$, which corresponds to the modulation frequency range [0 Hz, 160 Hz].

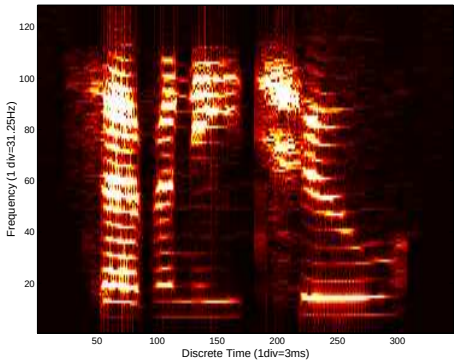

Figure 6.1. Conventional Spectrogram of a clean speech utterance.

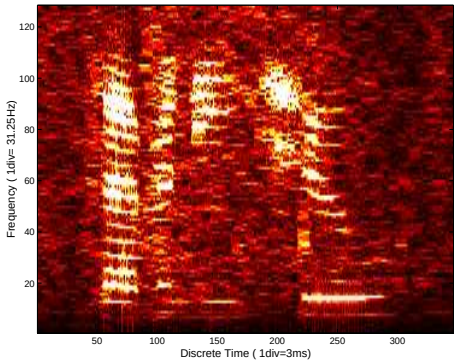

Figure 6.2. Conventional Spectrogram of a noisy speech utterance at SNR6.

6.3 Mel-Cepstrum Modulation Spectrum Features

As the spectral energies $E[n,\ b]$ in adjacent bands in (6.6) are highly correlated, the use of the magnitude modulation spectrum $M[n,\ b,\ q]$ as features for ASR would not be expected to work well (this has been verified experimentally). Instead, we here compute the modulation spectrum in the cepstral domain, which is known to be highly uncorrelated. The resulting features are referred to here as Mel-Cepstrum Modulation Spectrum (MCMS) features.

Consider the modulation spectrum of the cepstrally smoothed power spectrum $\log(S[n,\ k])$ in (6.5). Taking the DFT of $\log(S[n,\ k])$ over P points and considering the q^{th} coefficient $M^{'}[n,\ k,\ q]$, we obtain:

$$M^{'}[n,\ k,\ q] = \sum_{p=0}^{P-1} \log(S[n+p,\ k])e^{\frac{-j2\pi pq}{P}} \tag{6.8}$$

Using (6.5), (6.8) can be expressed as:

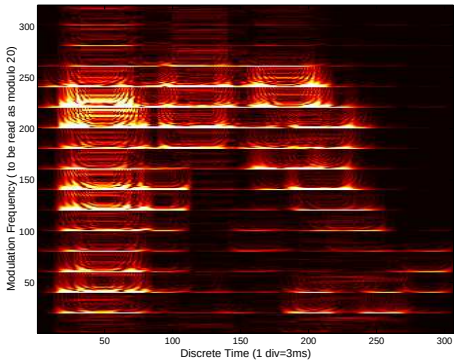

Figure 6.3. Modulation Spectrum across 16 bands for a clean speech utterance. The above figure is equivalent to 16 modulation spectrums corresponding to each of 16 bands. To see q^{th} modulation frequency sample of b^{th} band, go to number $(b-1)*6+q$ on the modulation frequency axis.

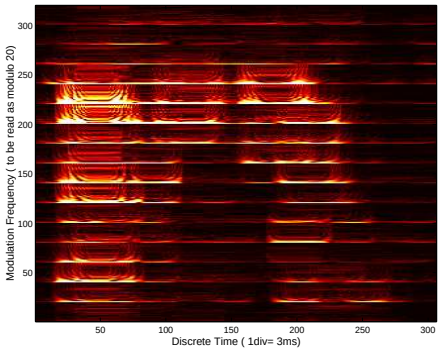

Figure 6.4. Modulation Spectrum across 16 bands for a noisy speech utterance at SNR6. The above figure is equivalent to 16 modulation spectrums corresponding to each of 16 bands. To see q^{th} modulation frequency sample of b^{th} band, go to number $(b-1)*6+q$ on the modulation frequency axis.

$$M^{'}[n,\ k,\ q] = \sum_{p=0}^{P-1} C[n+p,\ 0]e^{\frac{-j2\pi pq}{P}}$$
$$+ \sum_{l=1}^{L-1} \cos(\frac{2\pi kl}{K}) \underbrace{\sum_{p=0}^{P-1} 2C[n+p,\ l]e^{\frac{-j2\pi pq}{P}}} \tag{6.9}$$

In (6.9) we identify that the under-braced term is the cepstrum modulation spectrum. Therefore, $M^{'}[n,\ k,\ q]$ is a linear transformation of the cepstrum modulation spectrum. As cepstral coefficients are mutually uncorrelated, we expect the cepstrum modulation spectrum to perform better than the power spectrum modulation spectrum $M^{'}[n,\ k,\ q]$. Therefore, we define:

$$MCMS_{DFT}[n,\ k,\ q] = \sum_{p=0}^{P-1} C[n+p,\ l]e^{\frac{-j2\pi pq}{P}} \tag{6.10}$$

An alternative interpretation of the MCMS features, is as filtering of the cepstral trajectory in the cepstral modulation frequency domain. Temporal derivatives of the cepstral trajectory can also

be viewed as performing filtering operation. Figure 6.5 shows the cepstral modulation frequency response of the filters corresponding to first and second order derivatives of the MFCC features, while Figure 6.6 shows few of the filters employed in the computation of the MCMS features. On direct comparison, we notice that both of the temporal derivative filters emphasize the same cepstral modulation frequency components around 15 Hz. This is in contrast to the MCMS features, which emphasize different cepstral modulation frequency components. This further illustrates the fact that the different MCMS features carry complementary information.

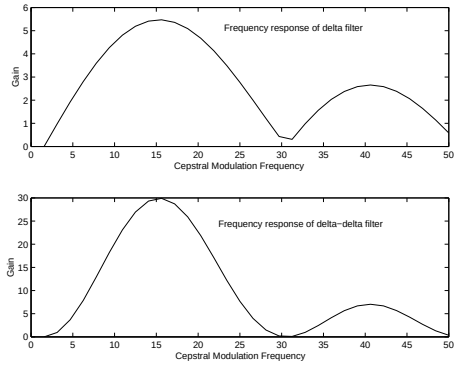

Figure 6.5. Cepstral Modulation Frequency responses of the filters used in computation of derivative and acceleration of MFCC features

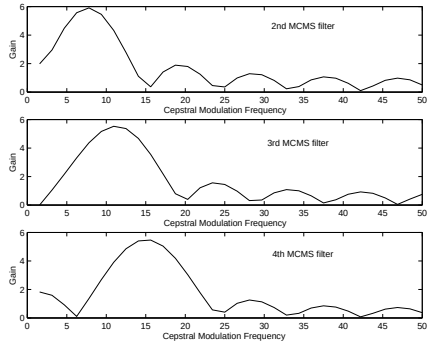

Figure 6.6. Cepstral Modulation Frequency responses of the filters used in computation of MCMS features

Let $MCMS_{DCT}[n,\ k,\ q]$ be the q^{th} DCT coefficient of the k^{th} cepstral trajectory taken across P frames.

$$MCMS_{DCT}[n,\ k,\ q] = \sum_{p=0}^{P-1} C[n+p-P/2,\ k] cos \frac{j\pi(q)(p+0.5)}{P} \tag{6.11}$$

$$with\ q \in [0, P-1],\ k \in [0,\ L-1]$$

6.4 Experiments: OGI Numbers95

In order to assess the effectiveness of the proposed MCMS features for speech recognition, experiments were conducted on the OGI Numbers (Cole et al., 1994) corpus. Four feature sets were

generated :

MFCC+Deltas: 39 element feature vector consisting of 13 MFCCs (including 0^{th} cepstral coefficient) with cepstral mean subtraction and variance normalization and their standard delta and acceleration features.

PLP + Deltas C-JRASTA Processed 39 element feature vector consisting of 13 PLPs which have been filtered by constant J-RASTA filter and their standard delta and acceleration features. All the coefficients have cepstral mean subtraction and have variance normalization.

$MCMS_{DFT}$**:** 78 element feature vector consisting of first three real and imaginary $MCMS_{DFT}$ coefficients derived from a basis of sines and cosines as in (6.10) with variance normalization

MFCC+MCMS: 78 element feature vector consisting of 13 MFCCs (including 0^{th} cepstral coefficient) with their first five $MCMS_{DCT}$ dynamic features as in (6.11) with variance normalization.

In this work we investigate the use of the range of MCMS filters covering (3-25) Hz[1] of the cepstral modulation frequency (6 MCMS filters).

The speech recognition systems were trained using HTK on the clean training set from the original Numbers corpus. The system consisted of 80 tied-state triphone HMM's with 3 emitting states per triphone and 12 mixtures per state. The context length for the MCMS features has been kept at 11 frames which corresponds to 120 ms. The MCMS features used in the experiment cover a cepstral modulation frequency range from 3Hz to 25 Hz. This range was selected as it yielded the best recognition performance. The MCMS features used in these experiments are computed over a 120 ms long time window. To verify the robustness of the features to noise, the clean test utterances were corrupted using Factory and F-16 aircraft cockpit noises from the Noisex92 database (Varga et al., 1992). The results for the baseline and MCMS systems in various levels of noise are given in Tables 6.1 and 6.2. We note that the MFCC baseline and the C-JRASTA PLP baseline features reported in these tables have been rendered highly noise robust as they have been cepstral mean and variance normalized. Therefore, even the slight gains obtained by MCMS features (which have also been cepstral mean and variance normalized) are useful.

From the results in Tables 6.1 and 6.2 we see a number of interesting points. First, unlike other noise robust features, the MCMS features achieve almost the same ASR accuracy as the MFCC features in the clean conditions. This is an important result, as most noise robust features generally lead to some degradation in clean conditions (such as RASTA-PLP, for example). The performance gains of MCMS features over the C-JRASTA PLP are in fact modest. This led us to combining the noise robustness properties of the exponentiated MFCC features which were introduced in Chapter **??** with the MCMS features. These features are discussed in the next section and they further improved the ASR performance.

Table 6.1. Word error rate results for F-16 cockpit noise. All the features have been cepstral mean and variance normalized. C-JRASTA-PLP is used with a constant $J = 10^{-6}$.

SNR	MFCC+Deltas	C-JRASTA PLP	$MCMS_{DFT}$	MFCC+$MCMS_{DCT}$
Clean	5.1	6.5	5.6	4.9
12 dB	10.5	11.2	10.6	10.4
6 dB	19.2	17.9	17.6	17.3
0 dB	36.3	34.8	32.8	33.2

[1] while excluding the zeroth frequency component

Table 6.2. Word error rate results for factory noise. All the features have been cepstral mean and variance normalized. C-JRASTA-PLP is used with a constant $J = 10^{-6}$.

SNR	MFCC+Deltas	C-JRASTA PLP	$MCMS_{DFT}$	MFCC+$MCMS_{DCT}$
Clean	5.1	6.5	5.6	4.9
12 dB	11.2	10.6	10.3	9.6
6 dB	20.2	18.4	18.2	19.3
0 dB	41.4	37.9	37.8	39.0

6.5 Combining MCMS with Expo-MFCC

In previous chapter, we had described the expo-MFCC features which were derived from exponentiated log Mel-filter-bank energies and were analytically shown to be less sensitive to random perturbation in the log Mel-filter-bank energies. As a results the expo-MFCC features leads to a noise robust feature representation. As MCMS features bring in complementary information by extracting the information from the useful range of the cepstral modulation spectrum, we thought that it would be ideal to combine these two complementary feature extraction techniques. Therefore, instead of computing MCMS features from the trajectories of the usual MFCC, we have computed MCMS from the trajectories of the Expo-MFCC features and have termed it as (ExpoMFCC+MCMS) features. The speech recognition performance of these features is listed in Table 6.3. As can be noted from this table this new feature synergistically combines the noise robustness properties of the expoMFCC feature and the MCMS features, resulting in the best ASR performance across all the noise conditions. The superior performance of ExpoMFCC+MCMS features can be noticed in the last column of the table 6.3. The average word error rate (WER) for the ExpoMFCC+MCMS features in clean and all the noisy conditions in 15.8%. This corresponds to a relative improvement of 24.0% over RASTA-PLP features and 11.4% over the optimal RMFCC features. Interestingly enough, and unlike the RASTA-PLP and Root MFCC features, the ExpoMFCC+MCMS features also preserves the good ASR performance in clean conditions.

Table 6.3. Word error rate results for factory and f16 noise. All the features in this case have cepstral mean and variance normalization. C-JRASTA-PLP is used with a constant $J = 10^{-6}$.

SNR	MFCC +Deltas	C-JRASTA PLP	Root MFCC	**ExpoMFCC+MCMS**
Clean	5.1	6.5	6.1	5.0
Fact SNR 12	11.2	10.6	10.4	9.2
Fact SNR 6	20.2	18.4	16.7	15.2
Fact SNR 0	41.4	37.9	35.3	31.6
F16 SNR 12	10.5	11.2	10.2	9.5
F16 SNR 6	19.2	17.9	15.7	14.4
F16 SNR 0	36.3	34.8	28.9	26.0

6.6 Summary

In this chapter we have proposed a new feature representation that exploits the temporal structure of speech, which we referred to here as the Mel-Cepstrum Modulation Spectrum (MCMS). These features can be seen as the outputs of an array of band-pass filters applied in the cepstral modulation frequency domain, and as such factor the spectral dynamics into orthogonal components moving at different rates. In our experiments, we found that a context length of 120 ms for the

computation of MCMS features, performs the best. In these experiments we have used 6 MCMS coefficients which cover the cepstral modulation frequency in the range (3, 25) Hz. In experiments, the proposed MCMS dynamic features are compared to standard delta and acceleration temporal derivative features and constant J-RASTA features. Recognition results demonstrate that the MCMS features lead to performance improvement in non-stationary noise, while importantly also achieving the performance similar to the MFCCs in the clean acoustic conditions. By computing the MCMS features through the use of the cepstral trajectories of the exponentiated MFCC features, we synergistically combine the noise robustness properties of the expoMFCC feature and the MCMS features, resulting in the best ASR performance across all the noise conditions.

Chapter 7

Conclusion

The goal of this thesis was to develop and design novel speech processing techniques that can overcome some of the deficiencies of the existing fixed scale (20-30ms)spectral envelope based front-ends. In an information bearing signal such as speech the information is propagated through the slow evolution of one quasi-stationary segment into another. For instance vowels slowly evolve to consonants and vice versa. However, the current ASR systems make a simplified assumption that all the stationary events are of uniform duration and the duration is typically assumed to be $20ms$. This poses a major limitation as certain sounds (events) such as vowels last for typically $(40, 80)ms$ while certain short-time-limited sounds such as plosive and stop last for $10ms$. The specific instants in a signal waveform when this stationarity switching happens, the rate at which this switching occurs and the spectra of sustained stationary segments are all very important quantities which need to be detected and estimated to extract all the useful information from the speech signal.

In this thesis we have investigated the potential benefits of using variable sized stationarity synchronous analysis windows for the MFCC feature computation. Although, this provided little improvement over the fixed scale analysis, we believe that the true potential of the variable scale analysis is limited due to the use of the HMM-GMM back-end which, in a way, is incompatible with the variable scale analyzed features. In particular,

- Firstly, although the algorithm proposed in this thesis uses variable sized analysis windows, it still uses a fixed shift size (for advancing the time) of $10\ ms$. This leads to an undesirable effect which is explained as follows. Lets us consider a long quasi-stationary segment (QSS) of length 100ms, followed by another QSS of arbitrary length and so forth. As we are using a constant shift size of $10\ ms$, we will end up analyzing the first QSS with progressively smaller windows of sizes, $[100, 90, 80, ...10]ms$ until we have entirely shifted past the first QSS. This will increase the variance of the MFCC features (as they have been computed form variable sized windows) that belong to the same QSS (the first one). The increase in the variance will adversely affect the discriminative ability of the first QSS.
- Ideally, we can use a shift size equal to the size of the detected QSS. For example if the detected QSS has a size of $x\ ms$, then we can use of shift size of $x\ ms$. Although this will solve the first problem, it introduces problems of its own:
  - First of all, the variable shift size will fluctuate the Nyquist frequency of the modulation spectrum which is roughly estimated through the delta features (Tyagi et al., 2003).
  - Secondly, the Viterbi decoding algorithm (Rabiner and Juang, 1993) needs the feature vectors computed from equal sized segments for comparing the scores of different hypothesis. With the features computed from the variable length segments, thereby affecting

the cost function per unit time, the optimality of the Viterbi decoding can no longer be ensured.

This thesis has also proposed fepstrum feature which is an improvement over the previous uses of the modulation spectrum as a feature in ASR. It has been shown in this thesis that fepstrum is an exact dual of the well known quantity cepstrum. While fepstrum provides amplitude modulations (AM) occurring within a single speech frame of size 80-100 ms, the MFCC provides a description of static energy in each of the Mel-bands of each frame and its variation across several frames (through the use of the delta and double delta feature). The Fepstrum provides complementary information to the MFCC features and we show that a combination of the two features, in form of simple concatenation or tandem modeling provides a significant ASR accuracy improvement in clean conditions over several speech databases.

In the second half of this thesis, we have focused on the noise robust feature design. Toward this aim, this thesis proposes a least squares filtering technique for speech signal enhancement in presence of additive broad band noise. One of the key advantage of the proposed technique is that unlike the Wiener filtering, it does not require a reference noise signal. This renders the least-squares filtering technique as a highly practical enhancement technique that can work with only a single noisy speech channel. Furthermore, unlike the classical spectral subtraction and Wiener filtering techniques that require the noise to be stationary, the proposed LeSF technique makes no such assumption as this technique works on a block by block basis. Unlike other feature level noise robustness technique, the LeSF filter enhances the signal waveform itself and a MFCC feature computed over this enhanced signal leads to a significant improvement in speech recognition accuracies as compared to the other competing feature level noise robustness techniques such as RASTA-PLP and spectral subtraction. In distributed speech recognition (DSR) in the context of mobile telephony and voice-over IP systems, it may be desirable not only to have noise robust feature extraction algorithm but also to enhance the noisy speech signal for the human listener. Therefore, a signal enhancement technique that also leads to noise robust ASR is desirable. The proposed LeSF filtering technique falls into this category as it not only enhances the signal, a simple MFCC feature computed over this enhanced signal leads to significant ASR accuracy improvements in several realistic noisy conditions.

Finally, this thesis proposes two feature level noise robustness technique. We have identified a numerical sensitivity problem with the MFCC (Davis and Mermelstein, 1980) features. It is analytically shown that by exponentiating the log Mel filter bank spectrum, one can desensitize the MFCC coefficients (derived from the exponentiated log Mel filter bank spectrum) to spurious low-energy spectral perturbations. We show the relationship between the conventional MFCC features and the proposed expo-MFCC features. In particular for the power $P = 2$, the expo-MFCC features can be thought to be obtained by convolving the usual MFCC features with itself. This leads to the interpretation that that sharp "spikes" in the MFCCs are smoothed out by the proposed technique, leading to a more smoother cepstrum.

This is supplemented by yet another representation that exploits the temporal structure of speech, which we refer to here as the Mel-Cepstrum Modulation Spectrum (MCMS). These features can be seen as the outputs of an array of band-pass filters applied in the cepstral modulation frequency domain, and as such factor the spectral dynamics into orthogonal components moving at different rates. In our experiments, we found that a context length of 120 ms for the computation of MCMS features, performs the best. In these experiments we have used 6 MCMS coefficients which cover the cepstral modulation frequency in the range (3, 25) Hz. By computing the MCMS features through the use of the cepstral trajectories of the expo-MFCC features, we synergistically combine the noise robustness properties of the expo-MFCC feature and the MCMS features, resulting in the best ASR performance across all the noise conditions.

7.1 Future Directions

We believe that in order to realize the true potential of variable scale piece-wise quasi-stationary analysis, we will have to research new statistical modeling techniques that can handle the spectral vectors derived from the variable sized segments in a suitable way. One example could be Dynamic Bayesian Networks(DBNs) (Stephenson et al., 2004) where the length of the QSS can be an auxiliary (Stephenson et al., 2004) variable that conditions the emitted MFCC vector.

In the Chapter 4, we have developed a theoretically sound AM-FM decomposition technique using analytic signals in the time domain. Future work in this direction will entail a further study of the statistical properties of the demodulated AM-FM signals and their possible relationships with quantities such as prosody, speaking rate and stress. To the best of our knowledge there has been very limited work which provides a signal processing measure of these quantities (stress, prosody and speaking rate). We believe that developing a consistent AM-FM demodulation technique for a wideband signal such as speech may lead to signal processing measures for stress, prosody and speaking rate that will eventually benefit ASR.

It will also be interesting to see if the fepstrum features and the MFCC features can be modeled together in the DBN framework. As the fepstrum features correspond to the slow moving amplitude modulations, they can potentially be used to condition the distribution of the MFCC features. This appears to be a more plausible scheme as compared to concatenating the two feature streams and modeling them in a HMM-GMM framework.

In Chapter 5, we have developed an adaptive speech signal enhancement technique based on the least squares filtering (LeSF) of the speech signal . We have shown that unlike Wiener filtering, LeSF does not require a reference noise signal channel and is capable of filtering out non-stationary noises as well. Moreover, since LeSF is a signal enhancement technique, it not only provides noise robust features, it provides an enhanced speech signal too. This dual utility of LeSF is particularly interesting in the context of the distributed ASR, where the same signal processing algorithm can perform the dual functions of noise robust feature extraction and speech signal enhancement for the human listener. We believe that this work can be further extended in the context of microphone arrays and audio-visual speech recognition with applications in robust hands free ASR.

Appendix A

Solution to LeSF Equation

A.1 Solution to the equation

A.1.1 Autocorrelation over a block of samples

We consider the signal $x(n)$ as

$$x(n) = \sum_{i=1}^{M} A_i \cos(\omega_i n + \phi_i) + u(n) \tag{A.1}$$

where $n \in [0, N-1]$ and $u(n)$ is a realization of white noise. Then the k^{th} lag autocorrelation is given by,

$$
\begin{aligned}
r(k) &= \textstyle\sum_{n=0}^{N-k-1} x(n)x(n+k) \\
&= \textstyle\sum_{n=0}^{N-k-1}(\sum_{i=1}^{M} A_i \cos(\omega_i n + \phi_i) + u(n))(\sum_{j=1}^{M} A_j \cos(\omega_j (n+k) + \phi_j) + u(n+k)) \\
&= \textstyle\sum_{n=0}^{N-k-1}(\sum_{i=1}^{M} A_i^2 \cos(\omega_i n + \phi_i)\cos(\omega_i (n+k) + \phi_i) \\
&\quad + \textstyle\sum_{n=0}^{N-k-1}\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M} A_i A_j \cos(\omega_i n + \phi_i)\cos(\omega_j (n+k) + \phi_j) \\
&\quad + \textstyle\sum_{n=0}^{N-k-1}\sum_{j=1}^{M} A_j u(n)\cos(\omega_j (n+k) + \phi_j) + \sum_{n=0}^{N-k-1}\sum_{i=1}^{M} A_i u(n+k)\cos(\omega_i n + \phi_i) \\
&\quad + \textstyle\sum_{n=0}^{N-k-1} u(n)u(n+k) \\
&= \textstyle\sum_{n=0}^{N-k-1}(\sum_{i=1}^{M} \frac{A_i^2}{2}(\cos(\omega_i k) + \cos(2\omega_i n + \omega_i k + 2\phi_i))) \\
&\quad + \textstyle\sum_{n=0}^{N-k-1}\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M} \frac{A_i A_j}{2}(\cos((\omega_i - \omega_j)n - \omega_j k + \phi_i - \phi_j) + \cos((\omega_i + \omega_j)n + \omega_j(k) + \phi_i + \phi_j)) \\
&\quad + \textstyle\sum_{n=0}^{N-k-1}\sum_{j=1}^{M} A_j u(n)\cos(\omega_j (n+k) + \phi_j) + \sum_{n=0}^{N-k-1}\sum_{i=1}^{M} A_i u(n+k)\cos(\omega_i n + \phi_i) \\
&\quad + \textstyle\sum_{n=0}^{N-k-1} u(n)u(n+k) \\
&= \underbrace{(N-k)\sum_{i=1}^{M}\frac{A_i^2}{2}\cos(\omega_i k)}_{SIG} + \underbrace{\textstyle\sum_{i=1}^{M}\frac{A_i^2}{2}\sum_{n=0}^{N-k-1}\cos(2\omega_i n + \omega_i k + 2\phi_i)}_{A} \\
&\quad + \textstyle\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M} \underbrace{\frac{A_i A_j}{2}\sum_{n=0}^{N-k-1}\cos((\omega_i - \omega_j)n - \omega_j k + \phi_i - \phi_j)}_{B} \\
&\quad + \textstyle\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M} \underbrace{\frac{A_i A_j}{2}\sum_{n=0}^{N-k-1}\cos((\omega_i + \omega_j)n + \omega_j(k) + \phi_i + \phi_j)}_{C} \\
&\quad + \underbrace{\textstyle\sum_{j=1}^{M}\sum_{n=0}^{N-k-1} A_j u(n)\cos(\omega_j (n+k) + \phi_j)}_{D} + \underbrace{\textstyle\sum_{i=1}^{M}\sum_{n=0}^{N-k-1} A_i u(n+k)\cos(\omega_i n + \phi_i)}_{E} + \underbrace{\textstyle\sum_{n=0}^{N-k-1} u(n)u(n+k)}_{F}
\end{aligned}
\tag{A.2}
$$

Let us consider the under-braced terms A, B, C. They are the sums of the samples of a cosine wave at frequencies, $2\omega_i$, $\omega_i - \omega_j$, $\omega_i + \omega_j$ respectively. If (N-k) is much greater than the $\frac{2\pi}{\omega_i}$ and $\frac{2\pi}{\omega_i - \omega_j}$ for all frequency pairs (i, j), then the sums A, B, C will contain several periods of their corresponding cosine waves. Sum over 'Q' periods of the samples of a cosine wave at any nonzero frequency is zero. This can be seen by the following integral which approximates the sum of the samples in A, B, C,

$$\int_{t=0}^{t=Q2\pi/\omega} A\cos(\omega t + \phi)dt = 0 \tag{A.3}$$

This happens as the negative and positive swings of the cosine cancel each other. Therefore A, B, C are approximately zero. Let $(N-k) = Q \times T + \Delta$, where Q is an integer and T is the period of a certain sinusoid at frequency ω and Δ is the left over part as $N-k$ is not an exact multiple of T. Hence $\Delta < T$. Then we have,

$$
\begin{aligned}
\textstyle\sum_{n=0}^{n=N-k} A\cos(\omega n + \phi) &= \textstyle\sum_{n=0}^{n=QT} A\cos(\omega n + \phi) + \sum_{n=QT+1}^{n=N-k} A\cos(\omega n + \phi) \\
&= \textstyle\sum_{n=QT+1}^{n=N-k} A\cos(\omega n + \phi) \\
&< A \times \Delta \ll A \times (N-k)
\end{aligned}
\tag{A.4}
$$

This proves the A, B, C can safely be ignored in comparison to SIG term which is proportional to (N-k). Moreover D, E are also approximately zero as the noise $u(n)$ is assumed uncorrelated with signal $s(n)$. The term $F = N\sigma^2\delta(k)$ as the noise is assumed to be white. Hence ignoring A, B, C, D, E, we get

$$\begin{aligned} r(k) &= \sum_{n=0}^{N-k-1} x(n)x(n+k) \\ &\simeq \sum_{i=1}^{M}(N-k)A_i^2 cos(2\pi f_i k) + N\sigma^2\delta(k) \\ &\simeq \sum_{i=1}^{M}(N\exp(-\alpha k))A_i^2 cos(2\pi f_i k) + N\sigma^2\delta(k) \end{aligned} \tag{A.5}$$

where $\alpha = 1/N$ and hence $\alpha \ll 1$.

A.1.2 Solving the least squares matrix equation

In the section, we will analytically solve the LeSF equation for the form of the autocorrelation coefficients given above in (A.5). The $(L \times L)$ matrix LeSF equation is reproduced below

$$\begin{pmatrix} r(0) & r(1) & r(2) & \cdots & r(L-1) \\ r(1) & r(0) & r(1) & \cdots & r(L-2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r(L-1) & r(L-2) & \cdots & \cdots & r(0) \end{pmatrix} \times \begin{pmatrix} w(0) \\ w(1) \\ \cdots \\ w(L-1) \end{pmatrix} = \begin{pmatrix} r(P) \\ r(P+1) \\ \cdots \\ r(P+L-1) \end{pmatrix} \tag{A.6}$$

where the $w(0), w(1), \cdots, w(L-1)$ are the LeSF filter tap weights and the autocorrelation coefficients $r(k)$ are given by (A.5). In (Anderson et al., 1983), it has been shown that the functional form of filter tap weights in (A.6) for the form of the $r(k)$ as in (A.5), is given by,

$$w(k) = \sum_{i=1}^{M-1}\left(B_i\exp(-\beta_i k) + C_i\exp(\beta_i k)\right)\cos(\omega_i(k+P)) \tag{A.7}$$

where P is the bulk delay. In (A.7), the quantities C_i, B_i, β_i are unknown. Our objective is to solve (A.6) for the unknown quantities in the filter tap weights w in closed form. Towards this end, let's consider the $(p+1)^{th}$ equation in the system of the equations (A.6) which is reproduced below,

$$\sum_{k=0}^{k=p-1} r(p-k)w(k) + r(0)w(p) + \sum_{k=p+1}^{L-1} r(k-p)w(k) = r(P+p) \tag{A.8}$$

Next, we substitute the functional forms of $r(k)$ and $w(k)$ in (A.8). We collect the terms that correspond to the i^{th} sinusoid together and call them as "self-terms" while the terms that have contribution from the i^{th} and the j^{th} sinusoid are called "cross terms". As, we will show that some of these cross terms can be ignored in comparison to the "self-terms", thus facilitating analytical solution. The "self-terms" due to the i^{th} sinusoid, on the right hand side of (A.8) are,

$$
\begin{aligned}
&\textstyle\sum_{k=0}^{p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\cos(\omega_i(k+P))\left[A_i^2\exp(-\alpha(p-k))\cos(\omega_i(p-k))\right]\\
&\qquad+\left[B_i\exp(-\beta_i p)+C_i\exp(\beta_i p)\right]\cos(\omega_i(p+P))\left[\textstyle\sum_{i=1}^{M}A_i^2+\sigma^2\right]\\
&\textstyle\sum_{k=p+1}^{L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\cos(\omega_i(k+P))\left[A_i^2\exp(-\alpha(k-p))\cos(\omega_i(k-p))\right]\\
&\quad=\underbrace{\frac{1}{2}A_i^2\cos(\omega_i(p+P))\sum_{k=0}^{k=p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(p-k))}_{stationary}\\
&\quad+\underbrace{\frac{1}{2}A_i^2\sum_{k=0}^{k=p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]exp(-\alpha(p-k))\cos(2\omega_i k-\omega_i(p-P))}_{non-stationary}\\
&\qquad+\underbrace{\left[B_i\exp(-\beta_i p)+C_i\exp(\beta_i p)\right]\cos(\omega_i(p+P))(\sum_{i=1}^{M}A_i^2+\sigma^2)}_{stationary}\\
&\qquad\underbrace{\frac{1}{2}A_i^2\cos(\omega_i(p+P))\sum_{k=p+1}^{k=L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(k-p))}_{stationary}\\
&\qquad+\underbrace{\frac{1}{2}A_i^2\sum_{k=p+1}^{k=L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]exp(-\alpha(k-p))\cos(2\omega_i k-\omega_i(p-P))}_{non-stationary}
\end{aligned}
\tag{A.9}
$$

The "non-stationary" terms approximately sum up to zeros due to the self-canceling positive and negative swings of the sinusoid at frequency $(2\omega_i)$ and hence can be ignored in comparison to the "stationary terms". Similarly in the $(p+1)^{th}$ equation, there are "cross-terms" that get contribution from the i^{th} and the j^{th} sinusoid. These terms are given below.

$$
\begin{aligned}
&\textstyle\sum_{k=0}^{p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\cos(\omega_i(k+P))\left[A_j^2\exp(-\alpha(p-k))\cos(\omega_j(p-k))\right]\\
&+\textstyle\sum_{k=p+1}^{L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\cos(\omega_i(k+P))\left[A_j^2\exp(-\alpha(k-p))\cos(\omega_j(k-p))\right]\\
&=\underbrace{\frac{1}{2}A_j^2\sum_{k=0}^{k=p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(p-k))\left[\cos((\omega_i-\omega_j)(k)+\omega_i P+\omega_j p)\right]}_{non-stationary}\\
&+\underbrace{\frac{1}{2}A_j^2\sum_{k=0}^{k=p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(p-k))\left[\cos((\omega_i+\omega_j)(k)+\omega_i P-\omega_j p)\right]}_{non-stationary}\\
&+\underbrace{\frac{1}{2}A_j^2\sum_{k=p+1}^{k=L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(k-p))\left[\cos((\omega_i-\omega_j)k+\omega_i P+\omega_j p)\right]}_{non-stationary}\\
&+\underbrace{\frac{1}{2}A_j^2\sum_{k=p+1}^{k=L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(k-p))\left[\cos((\omega_i+\omega_j)(k)+\omega_i P-\omega_j p)\right]}_{non-stationary}
\end{aligned}
\tag{A.10}
$$

If $(\omega_i - \omega_j) \gg \frac{2\pi}{L}$ and $(\omega_i + \omega_j) \neq 2\pi$, then these "non-stationary" terms approximately sum up to zero too. Therefore, all the cross-terms noted above can be safely neglected on the right hand side of the equation (A.8). Therefore neglecting all the non-stationary terms in (A.9), (A.10), we get,

$$\begin{aligned}
&\underbrace{\frac{1}{2}A_i^2\cos(\omega_i(p+P))\sum_{k=0}^{k=p-1}[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)]\exp(-\alpha(p-k))}_{stationary} \\
&\quad+\underbrace{[B_i\exp(-\beta_i p)+C_i\exp(\beta_i p)]\cos(\omega_i(p+P))(\sum_{i=1}^{M}A_i^2+\sigma^2)}_{stationary} \\
&+\underbrace{\frac{1}{2}A_i^2\cos(\omega_i(p+P))\sum_{k=p+1}^{k=L-1}[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)]\exp(-\alpha(k-p))}_{stationary} \\
&\quad=\textstyle\sum_{i=1}^{M}\exp(-\alpha(p+P))A_i^2 cos(\omega_i(p+P))
\end{aligned} \tag{A.11}$$

Next, we collect all the terms in (A.11) with the coefficients $\exp(-\beta_i p)$,$\exp(+\beta_i p)$ for each of the i^{th} sinusoid and set them to zero as there are no terms on the right hand side of (A.11) with these coefficients. Consider the terms with the coefficient $\exp(-\beta_i p)$, which are given below, and is set to zero as explained above.

$$\begin{aligned}
&\textstyle\sum_{i=1}^{M}\left[-\frac{A_i^2}{2}\frac{\cos(\omega_i(p+P))B_i\exp(-\beta_i p)}{(1-\exp(-(\beta_i-\alpha)))}+B_i\exp(-\beta_i p)\cos(\omega_i(p+P))\left(\textstyle\sum_{i=1}^{M}A_i^2+\sigma^2\right)\right] \\
&\quad+\textstyle\sum_{i=1}^{M}\left[\frac{A_i^2}{2}\frac{B_i\exp(-\beta_i p)\exp(-(\alpha+\beta_i))\cos(\omega_i(p+P))}{(1-\exp(-(\alpha+\beta_i)))}\right] \\
&\quad=0
\end{aligned} \tag{A.12}$$

Therefore for each "i", we get the relationship,

$$\cosh\beta_i=\cosh\alpha+\frac{\rho_i}{2\gamma_i+\rho_i+2}\sinh\alpha \tag{A.13}$$

where, ρ_i denotes the "partial" SNR of the sinusoid at frequency ω_i i.e $\rho_i = A_i^2/\sigma^2$ and the complementary signal SNR is denoted as $\gamma_i = (\sum_{m=1,m\neq i}^{M}A_i^2)/\sigma^2$. The coefficients of the terms $exp(-\beta_i p)$, $exp(+\beta_i p)$ are the same for each of the L equations and setting them to zero leads to just one equation which relates β_i to α, ρ_i and γ_i. Next step is to solve for B_i and C_i. Towards this end, we equate the coefficient of $\exp(-\alpha p)$ on both the left and right hand sides of (A.11). This leads to,

$$\begin{aligned}
&\frac{A_i^2\cos(\omega_i(p+P))B_i\exp(-\alpha p)}{1-\exp(-(\alpha-\beta_i))}+\frac{A_i^2\cos(\omega_i(p+P))C_i\exp(-\alpha p)}{1-\exp(-(\alpha+\beta_i))} \\
&\quad=A_i^2\exp(-\alpha p)\exp(-\alpha P)\cos(\omega_i(p+P))
\end{aligned} \tag{A.14}$$

Similarly we set the coefficient of $\exp(+\alpha p)$ to zero as there is no term with this coefficient on the right hand side of (A.11). This leads to,

$$\begin{aligned}
&\frac{A_i^2\cos(\omega_i(p+P))B_i\exp(-(\alpha+\beta_i))\exp(-\alpha L+\alpha p+\alpha-\beta_i L+\beta_i)}{1-\exp(-(\alpha+\beta_i))} \\
&+\frac{A_i^2\cos(\omega_i(p+P))C_i\exp(-\alpha+\beta_i)\exp(-\alpha L+\alpha p+\alpha+\beta_i L-\beta_i)}{1-\exp(-\alpha+\beta_i)}=0
\end{aligned} \tag{A.15}$$

There are two unknowns B_i and C_i in (A.14) and (A.15). Solving them simultaneously gives,

$$\begin{aligned} B_i &= \frac{2e^{-\beta_i}e^{-\alpha P}(\alpha+\beta_i)^2(\beta_i-\alpha)}{((\alpha+\beta_i)^2-e^{-2\beta_i L}(\beta_i-\alpha)^2)} \\ C_i &= \frac{2e^{-\beta_i(2L+1)+1}e^{-\alpha P}(\alpha+\beta_i)(\beta_i-\alpha)^2}{((\alpha+\beta_i)^2-e^{-2\beta_i L}(\beta_i-\alpha)^2)} \end{aligned} \tag{A.16}$$

This concludes the analytic solution of the LeSF equation.

Bibliography

Achan, K., Roweis, S., Hertzmann, A., and Frey, B. (2004). A segmental HMM for speech waveforms. Technical Report UTML Techical Report 2004-001, University of Toronto.

Ajmera, J., McCowan, I., and Bourlard, H. (2004). Robust speaker change detection. *IEEE Signal Processing Letters*.

Alexandre, P. and Lockwood, P. (1993). Root cepstral analysis: A unified view. application to speech processing in car noise environments. *Speech Communication*.

Allen, J. (1994). How do humans process and recognize speech. *IEEE Trans on Speech and Audio Processing*, 2.

Anderson, C. M., Satorius, E. H., and Zeidler, J. R. (1983). Adaptive enhancement of finite bandwidth signals in white gaussian noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*.

Atal, B. (1983). Efficient coding of lpc parameters by temporal decomposition. In *Proc. of IEEE ICASSP, 1983*.

Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. of America*.

Athineos, M. and Ellis, D. (2003). Frequency domain linear prediction for temporal features. In *In the Proc. of IEEE ASRU 2003, St. Thomas, Virgin Islands, USA*.

Athineos, M., Hermansky, H., and Ellis, D. (2004). Lp-trap: Linear predictive temporal patterns. In *Proc. of SAPA, Jeju, S. Korea, April 2004*.

Bahl, L. R., Jelinek, F., and Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5:179–190.

Bershad, N., Feintuch, P., Reed, F., and Fisher, B. (1980). Tracking characteristics of the LMS adaptive line-enhancer -response to a linear chirp signal in noise. *IEEE Transactions on Audio, Speech and Signal Processing*.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its applications to parameter estimation for gaussian mixture and hidden markov models. Technical Report 97-021, ICSI, Berkeley, CA, USA.

Bourlard, H., Dupont, S., Hermansky, H., and Morgan, N. (1997). Towards sub-band based speech recognition. In *Proc. of EUROSPEECH*, Rhodes, Greece.

Bourlard, H. and Kamp, Y. (1988). Autoassociation by multi-layer perceptrons and singular value decomposition. *Biological Cybernetics*.

Bourlard, H. and Morgan, N. (1994). *Connectionist Speech Recognition*. Kluwer Academic Publishers, Boston.

Bourlard, H. and Wellekens, C. (1990). Links between markov models and multilayer perceptrons. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 12(12).

Brandt, A. V. (1983). Detecting and estimating the parameters jumps using ladder algorithms and likelihood ratio test. In *Proc. of ICASSP 1983*.

Chen, J., Huang, Y., Li, Q., and Paliwal, K. K. (2004). Recognition of noisy speech using dynamic spectral subband centroids. *IEEE Signal processing Letters*.

Coifman, R. R. and Wickerhauser, M. V. (1992). Entropy based algorithms for best basis selection. *IEEE Trans. on Information Theory*, 38(2).

Cole, R., Fanty, M., and Lander, T. (1994). Telephone speech corpus at CSLU. In *Proc. of ICSLP, Yokohama, Japan, 1994*.

Compton, R. (1980). Pointing accuracy and dynamic range in a steered beam antenna array. *IEEE. Trans. on Aerosp. Electron. Syst.*

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable data. *Speech Communication*, 34:267–285.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. ASSP*, pages 357–366.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum-likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society of Britain*, 39:1–38.

Deng, L. and Sun, D. (1994). Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of english sounds. In *Proc. of ICASSP*, pages 45–48.

Dudley, H. (1939). Remaking speech. *JASA*, pages 169–177.

Ellis, D., Singh, R., and Sivadas, S. (2001a). Tandem acoustic modeling in large-vocabulary recognition. In *In the Proc. of ICASSP-2001*.

Ellis, D., Singh, R., and Sivadas, S. (2001b). Tandem acoustic modeling in large-vocabulary recognition. In *In the Proc. of ICASSP-2001, Salt Lake City, USA*.

Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum mean square error short-time spectral magnitude estimator. *IEEE Trans. on Acoustics, Speech and Signal Processing*.

Ephraim, Y. and Malah, D. (1985). Speech enhancement using a minimum mean square error log spectral amplitude estimator. *IEEE Trans. on Acoustics, Speech and Signal Processing*.

Frost, O. L. (1972). An algorithm for linearly constrained adaptive array processing. *Proceedings for IEEE*.

Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. ASSP*, 34:52–59.

Furui, S. (1990). On the use of hierarchial spectral dynamics in speech recognition. In *Proc. ICASSP*, pages 789–792.

Furui, S. (1992). Towards robust speech recognition under adverse conditions. In *Proc. of ESCA Workshop on Speech Processing in Adverse Conditions*, Cannes, France.

Gales, M. J. F. and Young, S. J. (1996). Robust continuous speech recognition using parallel model compensation (pmc). *IEEE Trans on Speech and Audio Prcoessing*, 4.

Gersho, A. (1969). Adaptive equalization of highly dispersive channels for data transmission. *Bell Syst. Tech. Journal*.

Gibson, C. and Haykin, S. (1980). Learning characteristics of adaptive lattice filtering algorithms. *IEEE Trans. on Audio, Speech and Signal Processing*.

Griffiths, L. J. (1969). A simple adaptive algorithm for real time processsing in antenna arrays. *Proceedings of IEEE*.

Haykin, S. (1993). *Adaptive Filter Theory*. Prentice-Hall Publishers, N.J., USA.

Haykin, S. (1994). *Communication Systems*. John Wiley Sons, New York, 1994.

Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech. *JASA*, 87(4):1738–1752.

Hermansky, H. (2003). Trap-tandem: Data driven extraction of the features from speech. In *In the Proc. of IEEE ASRU 2003, St. Thomas, Virgin Islands, USA*.

Hermansky, H. and Morgan, N. (1994a). Rasta processing of speech. *IEEE Trans. on Speech and audio processing*.

Hermansky, H. and Morgan, N. (1994b). Rasta processing of speech. *IEEE Trans. on Speech and Audio Processing*, pages 578–589.

Ikbal, S., Misra, H., and Yegnanarayana, B. (1999). Analysis of autoassociative mapping neural network. In *In the Proc. of IJCNN-99, Washington, USA*.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Trans. on Audio and Speech signal processing*.

Jelinek, F., Bahl, L. R., and Mercer, R. L. (1975). Design of linguistic statistical decoder for teh recognition of continuous speech. *IEEE Transactions on Information Theory*, IT-21:250–256.

Kanedera, N., Hermansky, H., and Arai, T. (1998). Desired characteristics of modulation spectrum for robust automatic speech recognition. In *In the Proc. of IEEE-ICASSP, 1998.*

Kay, S. M. (1998). *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice-Hall Publishers, N.J., USA.

Kim, H. K. and Rose, R. C. (2003). Cepstrum domain acoustic feature compensation based on decomposition of speech and noise for asr in noisy environments. *IEEE Tran. on Speech and Audio Processing*.

Kingsbury, B., Morgan, N., and Greenberg, S. (1998a). Robust speech recognition using the modulation spectrogram. *Speech Communication*.

Kingsbury, B., Morgan, N., and Greenberg, S. (1998b). Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1-3).

Kumaresan, R. (1998). An inverse signal approach to computing the envelope of a real valued signal. *IEEE Signal Processing Letters*.

Kumaresan, R. and Rao, A. (1999). Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications,. *J. Acoust. Soc. Am.*

Lathoud, G., Doss, M. M., and Mesot, B. (2005). A spectrogram model for enhanced source localization and noise robust asr. In *Proc. of Eurospeech 2005, Lisbon, Portugal*.

Legetter, C. J. and Woodland, P. C. (1995). Flexible speaker adaptation using maximum likelihood linear regression. In *Proc. of ARPA Workshop on Spoken Language Systems Technology*.

Lim, J. (1979). Spectral root homomorphic deconvolution system. *IEEE Trans. on Acoustics, Speech and Signal Processing*.

Lockwood, P. and Boudy, J. (1992). Experiments with a non-linear spectral subtractor NSS, hidden markov model and the projection, for robust speech recognition in cars. *Speech Communication*, 11(2-3):215–228.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of IEEE*, 63(4).

Markel, J. and Gray, A. (1976). *Linear Prediction of Speech*. Springer Verlag.

Marple, L. (1981). Efficient least squares FIR system identification. *IEEE Trans. on Audio, Speech and Signal Processing*.

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Tran. on Speech and Audio Processing*.

McAulay, R. J. and Malpass, M. L. (1980). Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech and Signal Processing*.

McAulay, R. J. and Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*.

Obrecht, R. A. (1988). A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Trans. ASSP*.

Oppenheim, A. V. and Schafer, R. W. (1989). *Discrete-time signal Processing*. Prentice-Hall, Inc., 1989, Englewood Cliffs, New Jersey.

Paliwal, K. K. (1998). Spectral subband centriod features for speech recognition. In *Proc. of ICASSP, 1998*.

Rabiner, L., Crochiere, R., and Allen, J. (1978). FIR system modeling and identification in the presence of noise and band-limited inputs. *IEEE Trans. on Audio, Speech and Signal Processing*.

Rabiner, L. and Juang, B. H. (1993). *Fundamentals of speech recognition*, chapter 2, pages 20–37. Prentice Hall PTR, Englewoood Cliffs, NJ, USA.

Rabiner, L. R. and Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE ASSP magazine*, 3(1).

Raj, B., Seltzer, M., and Stern, R. (2004). Reconstruction of missing features for robust speech recognition. *Speech Communication*, 43:275–296.

Sambur, M. R. (1978). Adaptive noise canceling for speech signals. *IEEE Trans. on Acoustics, Speech and Signal Processing*.

Satorius, E. and Alexander, S. T. (1979). Channel equalization using adaptive lattice algorithms. *IEEE Trans. Commun.*

Satorius, E. and Pack, J. (1981). Application of least squares lattice algorithms for adaptive equalization. *IEEE Trans. Commun.*

Satorius, E., Zeidler, J., and Alexander, S. (1978). Linear predictive digital filtering of narrowband processes in additive broad-band noise. Technical report, Naval Ocean Systems Center, San Diego, CA.

Schimmel, S. and Atlas, L. (2005). Coherent envelope detection for modulation filtering of speech. In *Proceedings of IEEE ICASSP 2005, Philadephia, USA*.

Seltzer, M., Raj, B., and Stern, R. (2004). A bayesain classifier for a spectrographic mask estimation for missing feature speech recognition. *Speech Communication*, 43:379–393.

Sondhi, M. and Berkley, D. (1980). Silencing echoes on the telephone network. *Proceedings of IEEE*.

Srinivasan, S. and Kleijn, W. B. (2004). Speech enhancement using adaptive time-domain segmentation. In *Proc. of ICSLP 2004*.

Stephenson, T., Magimai-Doss, M., and Bourlard, H. (2004). Speech recognition with auxiliary information. *IEEE Trans. on Speech and Audio Processing*, 12(3):189–203.

Svendsen, T., Paliwal, K. K., Harborg, E., and Husoy, P. O. (1989). An improved sub-word based speech recognizer. In *Proc. of IEEE ICASSP, 1989.*

Tokuda, K., Kobayashi, T., Masuko, T., and Imai, S. (1994). Mel-generalized cepstral analysis - a unified approach to speech spectral estimation. In *Proc. of IEEE ICASSP 1994*.

Tyagi, V., McCowan, I., and Bourlard, H. (2003). Mel-cepstrum modulation spectrum (mcms) features for robust asr. In *Proc. of IEEE workshop on Automatic Speech Recognition and Understanding*, St. Thomas, US Virgin Islands.

Varga, A., Steeneken, H., Tomlinson, M., and Jones, D. (1992). The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, DRA Speech Research Unit, Malvern, England.

Varga, A. P. and Moore, R. K. (1990). Hidden markov model decomposition of speech and noise. In *Proc. of IEEE ICASSP*, pages 845–848.

Wesker, T., Meyer, B., Wagener, K., Anemuller, J., Mertins, A., and Kollmeier, B. (2005). Oldenburg logatome speech corpus (ollo) for speech recognition experiments with humans and machines. In *Proc. of ICSLP*, Lisbon, Portugal.

Widrow, B. (1975). Adaptive noise cancelling: Principles and applications. *Proceedings of IEEE*.

Young, S., Kershaw, D., J.Odell, Ollason, D., Valchev, V., and Woodland, P. (1995). *The HTK book*. Entropic.

Zeidler, J. R., Satorius, E. H., Chabries, D. M., and Wexler, H. T. (1978). Adaptive enhancement of multiple sinusoids in uncorrelated noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*.

Zhu, Q. and Alwan, A. (2000). Am-demodualtion of speech spectra and its application to noise robust speech recognition. In *Proceedings of ICSLP 2000*.

Zhu, Q., Chen, B., Morgan, N., and Stolcke, A. (2004). On using mlp features in lvcsr. In *In the Proc. of ICSLP, Jeju, Korea, 2004*.

Curriculum Vitae

Vivek Tyagi

Address:	Institute Eurecom
	2229, Route Des Cretes
	Sophia Antipolis, 06560, France
Home phone:	+33 49300 8170
email:	tyagi@eurecom.fr
Citizenship:	Indian

Education

August 2002 - till date	Docteur ès Sciences Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland Thesis: Novel Speech Processing techniques for robust speech recognition .
2001-2002	Pre-doctoral School in Computer Science Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland
1997-2001	Bachelor of Technology in Electrical Engineering Indian Institute of Technology, IIT Kanpur, India

Journal Papers

1. "On Variable-Scale Piecewise Stationary Spectral Analysis of Speech Signals for ASR," Vivek Tyagi, Herve Bourlard and Christian Wellekens, *Speech Communication, Volume 48, Issue 9, September 2006, pp 1182-1191*

2. "Least Squares Filtering of Speech Signals for Robust ASR," Vivek Tyagi and Christian Wellekens *Accepted for publication in Speech Communication. In Press*

Conference Papers

1. "Fepstrum and Carrier Signal decomposition of Speech Signals through Homomorphic Filtering," Vivek Tyagi and Christian Wellekens, *Invited paper in the special session, "Dealing with intrinsic speech variabilities in ASR", IEEE ICASSP 2006,* Toulouse, France.

2. "Fepstrum Representation of Speech," Vivek Tyagi and Christian Wellekens, *In the Proc. of IEEE ASRU, November 2005, Cancun Mexico.*

3. "A Variable-Scale Piecewise Stationary Spectral Analysis Technique Applied to the ASR," Vivek Tyagi, Christian Wellekens and Herve Bourlard, *In Springer Lecture Notes in Computer Science, 3869, Eds. Steve Renals and Samy Bengio,* January 2006.

4. "Least Squares Filtering of Speech Signals for Robust ASR," Vivek Tyagi and Christian Wellekens, *In Springer Lecture Notes in Computer Science, 3869, Eds. Steve Renals and Samy Bengio,* January 2006.

5. "On Variable-Scale Piecewise Stationary Spectral Analysis of Speech Signals for ASR," Vivek Tyagi, Herve Bourlard and Christian Wellekens, *In the Proc. of Eurospeech* September 2005, Lisbon, Portugal.

6. "On Desensitizing the Mel-Cepstrum to Spurious Spectral Components for Robust Speech Recognition," Vivek Tyagi and Christian Wellekens, *In the Proc. of IEEE ICASSP*, March 2005, Philadelphia, USA.

7. "Mel-Cepstrum Modulation Spectrum (MCMS) Features for Robust ASR," Vivek Tyagi, Iain McCowan, Herve Bourlard and Hemant Misra, *IN the Proc. of IEEE ASRU 2003*, December 2003, St. Thomas, US Virgin Islands.

8. "On Factorizing Spectral Dynamics for Robust Speech Recognition," Vivek Tyagi, Iain McCowan, Herve Bourlard and Hemant Misra, *In the Proc. of EUROSPEECH, September 2003* Geneva, Switzerland

9. "Entropy-Based Multi-Stream Combination," Hemant Misra, Herve Bourlard and Vivek Tyagi, *In the Proc. of IEEE ICASSP 2003,* Hong Kong.

Professional Service

1. Reviewer for IEEE Transactions on Speech and Audio Processing

2. Reviewer for IEEE Signal Processing Letters

3. Reviewer for Elsevier Signal Processing

Skills

1. C, C++, Matlab, Signal processing Toolbox of Matlab

2. Hidden Markov Model Toolkit (HTK)

3. Linux, Shell programming

4. Excellent proficiency in spoken/written English, fair proficiency in spoken French.