

NEW METHODS FOR ROBUST SPEECH RECOGNITION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF
ELECTRICAL AND ELECTRONICS ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Engin ERZİN
September 1995

Thesis
TK
7882
.S65
E79
1995

NEW METHODS FOR ROBUST SPEECH RECOGNITION

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND ELECTRONICS

ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCE

OF BİLKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

Engin Erzin

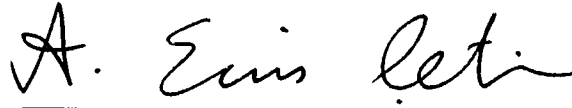
September 1995

Engin ERZİN
tarafından bağışlanmıştır

TK
7882
.S65
E79
1995

B032619

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.



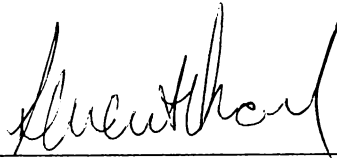
A. Enis Çetin, Ph. D. (Supervisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.



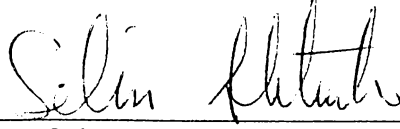
Orhan Arıkan, Ph. D.

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.



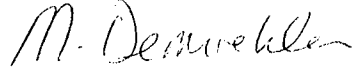
Levent Onural, Ph. D.

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.



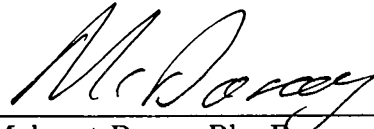
Selim Aktürk, Ph. D.

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.



Mübeccel Demirekler, Ph. D.

Approved for the Institute of Engineering and Science:



Mehmet Baray, Ph. D.

Director of Institute of Engineering and Science

Abstract

NEW METHODS FOR ROBUST SPEECH RECOGNITION

Engin Erzin

Ph. D. in Electrical and Electronics Engineering

Supervisor:

Assoc. Prof. Dr. A. Enis Çetin

September 1995

New methods of feature extraction, end-point detection and speech enhancement are developed for a robust speech recognition system.

The methods of feature extraction and end-point detection are based on wavelet analysis or subband analysis of the speech signal. Two new sets of speech feature parameters, SUBLSF's and SUBCEP's, are introduced. Both parameter sets are based on subband analysis. The SUBLSF feature parameters are obtained via linear predictive analysis on subbands. These speech feature parameters can produce better results than the full-band parameters when the noise is colored. The SUBCEP parameters are based on wavelet analysis or equivalently the multirate subband analysis of the speech signal. The SUBCEP parameters also provide robust recognition performance by appropriately deemphasizing the frequency bands corrupted by noise. It is experimentally observed that the subband analysis based feature parameters are more robust than the commonly used full-band analysis based parameters in the presence of car noise.

The α -stable random processes can be used to model the impulsive nature

of the public network telecommunication noise. Adaptive filtering are developed for α -stable random processes. Adaptive noise cancelation techniques are used to reduce the mismatch between training and testing conditions of the recognition system over telephone lines.

Another important problem in isolated speech recognition is to determine the boundaries of the speech utterances or words. Precise boundary detection of utterances improves the performance of speech recognition systems. A new distance measure based on the subband energy levels is introduced for endpoint detection.

Keywords : Speech recognition, linear prediction, endpoint detection, speech enhancement, subband decomposition, wavelet transform, line spectrum frequencies, α -stable distributions.

Özet

KONUŞMA TANIMA İÇİN GÜRÜLTÜYE DAYANIKLI YENİ YÖNTEMLER

Engin Erzin

Elektrik ve Elektronik Mühendisliği Bölümü Doktora

Tez Yöneticisi:

Doç. Dr. A. Enis Çetin

Eylül 1995

Dayanıklı konuşma tanıma sistemleri için öznitelik parametrelerinin elde edilmesi, sözcük sınırlarının belirlenmesi ve konuşma iyileştirilmesi alanlarında yeni yöntemler geliştirilmiştir.

Öznitelik parametrelerinin elde edilmesi ve sözcük sınırlarının belirlenmesi yöntemleri altbant analizine dayalı biçimde geliştirilmiştir. Altbant analizine dayalı iki yeni konuşma öznitelik parametre vektörü (SUBLSF ve SUBCEP) oluşturulmuştur. SUBLSF öznitelik parametrelerini elde etmek için konuşma işareti alt ve üst bantlara ayrılır. Her iki bant için doğrusal öngörü analizi yapılır ve bu analizlerden elde edilen Çizgisel Spektrum Frekansları birleştirilerek SUBLSF öznitelik vektörü oluşturulur. Diğer öznitelik parametre vektörü, SUBCEP, dalgacık analizi veya eşdeğer anlamda altbant analizi kullanılarak oluşturulur. SUBCEP parametreleri gürültülü bantları bastırarak dayanıklı bir başarımlı sağlamışlardır. Yapılan deneyler sonunda araç gürültüsü altında altbant analizi ile elde edilen öznitelik parametrelerinin yaygın olarak kullanılan tam-bant

parametrelerinden daha dayanıklı olduđu görülmüştür.

Telefon kanallarındaki gürültünün modellenmesinde α -kararlı rasgele süreçler kullanılabilir. α -kararlı rasgele süreçler için geliştirilen uyarlamalı süzgeçler gürültülü konuşmanın iyileştirilmesinde kullanılmış ve konuşma tanıma başarımı artırılmıştır.

Dayanıklı konuşma tanıma sistemlerinde bir diğeri önemli problem de sözcük sınırlarının belirlenmesidir. Sözcük sınırlarının hatasız belirlenmesi konuşma tanıma başarımını artırmaktadır. Sözcük sınırlarını belirlemeye yönelik, konuşma işaretinin altbant enerji değerlerine bağlı, yeni bir uzaklık ölçüsü sunulmuştur.

Anahtar sözcükler: Konuşma tanıma, doğrusal öngörü, sınır belirleme, konuşma iyileştirme, altbant ayrışımı, dalgacık dönüşümü, çizgisel spektrum frekansları, α -kararlı dağılımlar.

Acknowledgment

I would like to express my deepest gratitude to Dr. A. Enis Çetin for his supervision, guidance, suggestions and encouragement throughout the development of this thesis, and also for that he introduced me to the signal processing society in various international conferences from which I gained experience and motivation.

I would like to thank to Dr. Orhan Arıkan, Dr. Yasemin Yardımcı, and my project group in METU, especially to Prof. Dr. Mübeccel Demirekler, Burak Tüzün and Bora Nakipoğlu, for their valuable technical support and discussions.

I like to acknowledge to ASELSAN (Military Electronics Industry Inc.) for their financial supports to our research.

Thanks to all my friends who have been with me during my Ph.D. study, especially to Fatih, Cem, Aydın, Suat, and Ayhan for their morale support and friendship, and to Tuba, Zafer, and Esra for making me smile even if I feel so desperate and for their friendship.

Finally, it is a pleasure to express my sincere thanks to my family for their continuous morale support throughout my graduate study.

Contents

Abstract	iii
Özet	v
Acknowledgment	vii
List of Figures	xi
List of Tables	xiv
1 INTRODUCTION	1
1.1 Linear Predictive Modeling of Speech	4
1.1.1 Linear Prediction	7
1.1.2 Line Spectral Frequencies	11
1.2 Speech Recognition	12
1.2.1 Hidden Markov Model (HMM)	14
1.3 Speech Recognition Research in Turkey	24
2 LINE SPECTRAL FREQUENCIES IN SPEECH RECOGNITION	26

2.1	LSF Representation of Speech	27
2.2	LSF Representation in Subbands for Speech Recognition	29
2.2.1	Performance of LSF Representation in Subbands	31
2.2.2	Conclusion	32
3	WAVELET ANALYSIS FOR SPEECH RECOGNITION	34
3.1	Subband Analysis based Cepstral Coefficient (SUBCEP) Representation	35
3.1.1	Frequency Characteristics of the Iterated Filter Bank Structure	41
3.2	Simulation Studies	43
3.2.1	Experiments over the TI-20 Database	44
3.2.2	Performance over the Data Set Recorded in a Car	47
3.2.3	Conclusion	48
4	ADAPTIVE NOISE CANCELING FOR ROBUST SPEECH RECOGNITION	49
4.1	Adaptive Filtering for non-Gaussian Stable Processes	50
4.1.1	Adaptive Filtering for α -stable Processes	52
4.2	Adaptive Noise Canceling for Speech Recognition	55
4.3	Simulation Studies	58
5	WAVELET ANALYSIS FOR ENDPOINT DETECTION OF ISOLATED UTTERANCES	61
5.1	A New Distance Measure based on Wavelet Analysis	62
5.2	Simulation Examples	65

5.4	The energy measure of the noise free utterance “four” is plotted in (a). The new distance measure and the energy measure for different SNR levels are plotted in (b) and (c), respectively. The solid (dashed) [dash-dotted] {dotted}line corresponds to 30 dB (15 dB) [10 dB] {8 dB}noise level.	66
5.5	The energy measure of the noise free utterance “start” is plotted in (a). The new distance measure and the energy measure for different SNR levels are plotted in (b) and (c), respectively. The solid (dashed) [dash-dotted] {dotted}line corresponds to 30 dB (15 dB) [10 dB] {8 dB}noise level.	67
5.6	The recording of the Turkish word “hayır :/ h äyır/” (no) inside a car is plotted in (a), the new distance measure and the energy measure are plotted in (b) and (c), respectively.	68
5.7	(a) The new measure D_k for a utterance recorded inside a car, (b) the histogram of the new measure D_k .	69
5.8	Flowchart for the endpoint algorithm.	71
5.9	The distance measure D_k for the Turkish word “ <i>evet</i> ” and the end-point locations that are detected by the algorithm.	72
B.1	Transient behavior of tap weight adaptations in the NLMP, NLMAD, LMAD, LMP and LMS algorithms for AR(1) process. .	82
B.2	Transient behavior of tap weight adaptations in the NLMP, NLMAD, LMAD, LMP and LMS algorithms for AR(2) process. .	83
B.3	The adaptation performances of LMAD, NLMAD and NLMP algorithms with as a fourth order LPC synthesis filter.	84

List of Tables

2.1	Recognition rates of SUBLSF, MELCEP and LSF representations in percentage.	32
3.1	The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with Volvo noise recording.	45
3.2	The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with Skoda noise recording.	46
3.3	The performance evaluation of speaker independent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with Volvo noise recording.	46
3.4	The performance evaluation of speaker independent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with Skoda noise recording.	47
4.1	The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with α -stable noise ($\alpha = 1.95$).	59
4.2	The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with α -stable noise ($\alpha = 1.5$).	59

4.3	The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with α -stable noise ($\alpha = 2.0$).	60
A.1	<i>The angular offset factors which are used in simulations</i>	80
A.2	<i>Codebook sizes for each subvector at different rates</i>	80
A.3	<i>Spectral Distortion (SD) Performance of our method</i>	80
A.4	<i>Spectral Distortion (SD) Performance of the Vector Quantizers Atal et.al and Farvardin et.al.</i>	80

Chapter 1

INTRODUCTION

Speech is the central component of the field of telecommunications, as it is the most effective interface between human beings. How humans produce, process and recognize speech has been an active research area since the classical ages [1]. Advances in digital signal processing technology bring new challenges to the speech processing with wider application areas. Today the essential application areas of speech processing are compression, enhancement, synthesis and recognition [2].

The successful modeling of the vocal tract with the linear predictive (LP) analysis techniques brought simple and effective solutions to the speech coding and synthesis [3, 4]. Lower bit rates, acceptable quality and security are rather important objectives that are achieved in a satisfactory way with linear prediction of speech signals.

Although the speech production system is successfully modeled, the human perception and recognition system is not well understood [5]. The first experiments regarding human speech recognition was done by Harvey Fletcher in early 19th century. These recognition tests were in the form of an empirical probabilistic analysis of speech recognition scores obtained from a series of listening experiments. Fletcher used the word articulation in the perceptual context to mean the probability of correctly identifying nonsense speech sounds.

Human phone recognition rate for nonsense consonant-vowel-consonant syllables, under best conditions, was reported to be 98.5% [1, 6].

As the speech is becoming an important interface between human beings and machines, speech recognition stands as an important cornerstone of the speech processing. In the last 20 years, considerable progress has been achieved in this field. However, in terms of linguistic, natural language and understanding, speech recognition is a complicated application area, but when the dimensions are reduced, such as limited vocabulary and context, reliable recognition systems can be formed [7].

Automation of operator services, stock quotation services and voice control on mobile telephones can be considered as the current application areas of the limited vocabulary speech recognition systems.

In a recent review [8] on the challenges of spoken language systems by the leading authors in speech recognition field, the robustness in speech recognition is defined as *minimal graceful degradation in performance due to changes in input conditions caused by different microphones, room acoustics, background or channel noise, different speakers or other small systematic changes in the acoustic signal*. At present, one of the major drawback of speech recognition systems is the robustness. Their performance degrades suddenly and significantly with a minor mismatch between training and testing conditions [9]. Although signal processing strategies carry out some progress for more robust systems [10, 11, 12], the key point is to understand the many sources of variability in speech signal better for improving robustness. Variability is typically due to the speaker and the nature of the task, the physical environment, and the communication channel between user and the machine.

In this thesis, the sources of variability in car environment and telephone channels will be addressed and the methods on feature extraction, end-point detection and speech enhancement will be investigated for a robust speaker independent isolated word recognition system. In most cases, background noise is modeled as additive stationary perturbation which is uncorrelated to the speech signal. With this assumption, we tried to construct robust recognition systems in the presence of car noise and telephone channel noise. The current research

in robust speech recognition includes the car noise and the telephone channel environment due to the practical importance of these application areas [13, 14]. That is the main reason, these environments are considered in this thesis. In this chapter, vocal tract modeling and a speech recognition system based on the Hidden Markov Modeling (HMM) structure are reviewed.

In Chapter 2, robust feature extraction in subbands with LP analysis is investigated. Subband analysis is employed in order to separate the noisy and noise-free bands in the presence of car noise. The performances of the LSF representation and the cepstral coefficient representation of speech signals are comparable for a general speech recognition system [34, 35]. However, we obtain the SUBLSF feature parameters by first dividing the speech signal into low and high frequency bands, and then a linear predictive analysis will be performed on each subband. Consequently, the resulting two sets of LSF parameters are combined to form the SUBLSF feature parameters.

Multirate signal processing based feature extraction is presented in Chapter 3. The new feature set, SUBCEP, is obtained from the root-cepstral coefficients derived from the wavelet analysis of the speech signal [37]. The performance of the new feature representation is compared to the *mel* scale cepstral representation (MELCEP) in the presence of car noise. The SUBCEP parameters are realized in a computationally efficient manner by employing fast wavelet analysis techniques. Also, in the root cepstral analysis a fixed root value is used [13]. On the other hand the frequency bands are weighted by the proper selection of p_i values. This increases the robustness of the speech feature parameters against the colored environmental noise. Therefore, our main contributions to the derivation of SUBCEP parameters are the subband decomposition based time-frequency analysis, and the weighted root nonlinearity.

Chapter 4 will cover the adaptive filtering type enhancement strategies which are applied to speech recognition system. The α -stable random processes are used to model the impulsive nature of the public network telecommunication noise [45]. Recently, adaptive filtering techniques are developed for α -stable random processes [48, 49]. In Chapter 4, we mainly investigate some adaptive noise cancelation structures to eliminate the mismatch between training (noise-free training) and testing conditions of a speech recognition system for telephone

channel conditions.

In Chapter 5, the new distance measure for end-point detection based on the subband energy levels are presented. To determine the boundaries of the speech utterances or words is an important problem in isolated speech recognition. The energy measure and the zero-crossing rate are the widely used distances in endpoint detection algorithms [55]. They perform fairly well at high SNR levels, however their performance degrades drastically in many practical cases, such as car noise which is usually concentrated at low frequency bands. Instead of using the full-band energy a *mel*-frequency weighted energy measure is introduced in Chapter 5. In order to obtain the *mel*-scale frequency division in a computationally efficient manner, multirate signal processing based methods are employed.

Conclusions and discussions are given in Chapter 6.

In the following sections, the LP modeling of the speech signal is introduced and LP feature parameterization of the speech signal is given. Also, a general overview of a speech recognition system is given.

1.1 Linear Predictive Modeling of Speech

The speech production system has been successfully modeled by the linear predictive (LP) analysis which found large application areas that cover coding, synthesis and recognition. Computationally efficient and accurate estimation of the speech parameters, such as pitch, reflection coefficients, and formant frequencies, makes this method a widely used one [4].

The basic idea of linear predictive (LP) analysis is that the current speech sample, $s(n)$ at time instant n can be estimated as a linear combination of past p speech samples, such that

$$s(n) = \sum_{i=1}^p a_i s(n-i) + Gu(n), \quad (1.1)$$

where the predictor coefficients a_1, a_2, \dots, a_p are assumed constant over the

analysis frame and $u(n)$ is the excitation with gain G . By expressing Equation 1.1 in the z -domain, we write the relation

$$S(z) = \sum_{i=1}^p a_i z^{-i} S(z) + GU(z) \quad (1.2)$$

leading to the transfer function

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^p a_i z^{-i}} = \frac{1}{A(z)}. \quad (1.3)$$

Applying an error minimization criterion over a block of speech data one can come up with a set of predictor coefficients for an all-pole, linear, time-varying filter model.

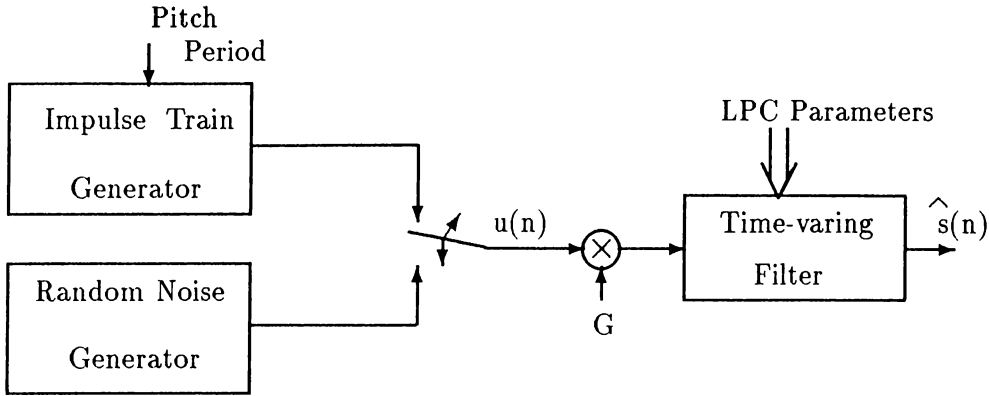


Figure 1.1: *LP Vocoder Synthesizer.*

A simple model of speech can be constructed using a linear predictor. For example, in LP vocoders (voice-coders) [4], the all pole linear predictor is excited by quasi-periodic pulses during voiced speech, and by random noise during unvoiced speech to generate synthetic speech.

The basic block diagram of an LP vocoder is shown in Figure 1.1. In LP vocoders the human vocal tract is modeled as an all-pole linear predictor as in Equation (1.3), where $A(z)$ is a minimum phase polynomial. In vocoders the coefficients of the filter are obtained by LP analysis which is carried out over speech frames of duration 20 to 30 msec. It is assumed that the statistical characteristics of the speech signal do not vary during a speech frame. Speech

frames are first characterized as voiced or unvoiced frames which are modeled differently. As it can be seen in Figure 1.2-(b) the speech signal is almost periodic in voiced regions. The “period” is called the pitch of the voiced frame. In unvoiced speech the time series exhibit such a pattern as shown in Figure 1.2-(c). It is assumed that the vocal tract or the linear filter is excited by an impulse train whose period is the same as the pitch value for voiced speech and by random noise for unvoiced speech. Voiced/unvoiced decision, pitch period calculation, can be carried out either in the time or the frequency domains.

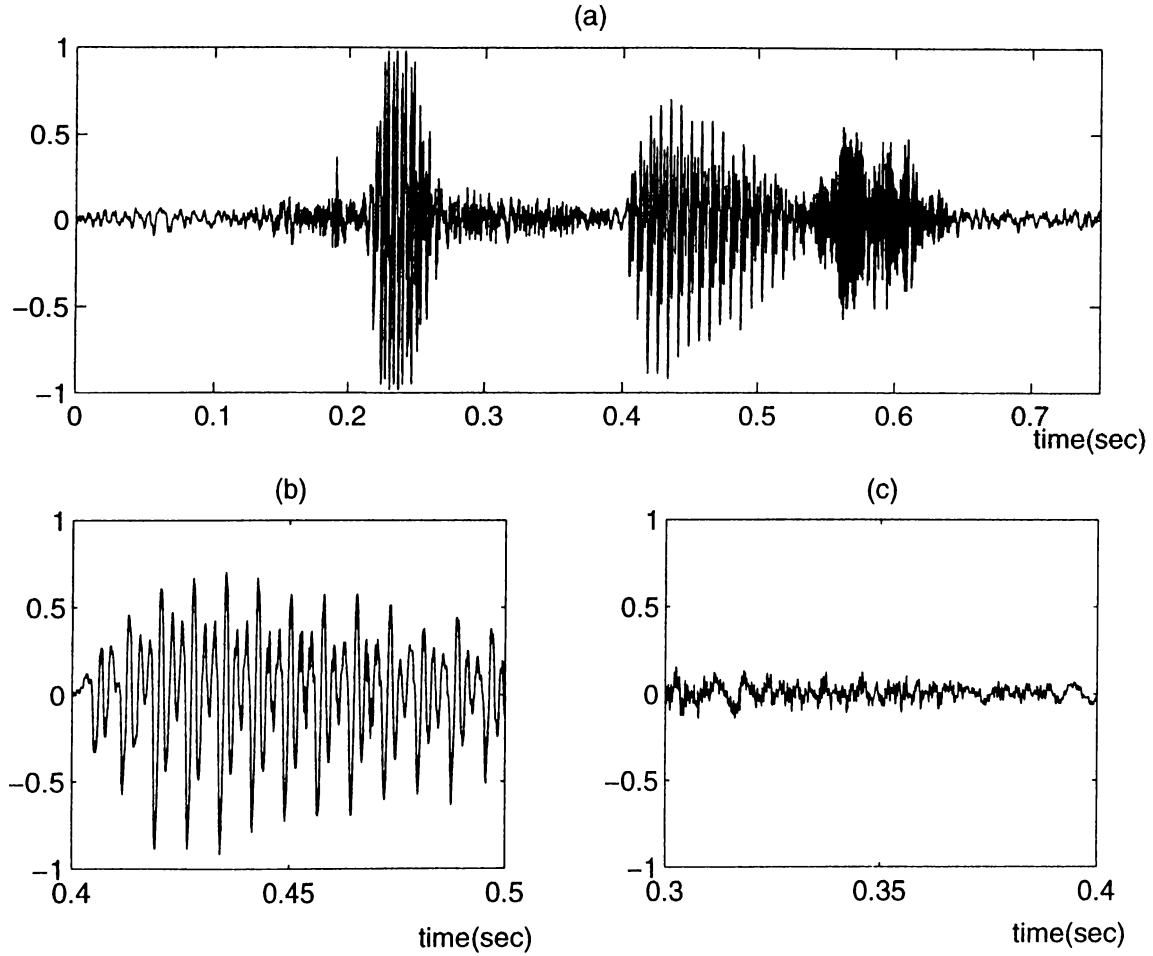


Figure 1.2: (a) An example of a speech signal corresponding to the Turkish utterance “sıfır”. (b) A voiced segment of the speech signal. (c) An unvoiced segment of the speech signal.

In the next section a commonly used LP analysis procedure is described for determining the filter coefficients, a_k 's.

1.1.1 Linear Prediction

The estimate $\hat{s}(n)$ of the speech signal $s(n)$ is defined as the linear combination of past speech samples

$$\hat{s}(n) = \sum_{k=1}^p a_k s(n-k), \quad (1.4)$$

by the speech model of Equation (1.3). Consequently the prediction error $e(n)$ is defined as,

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^p a_k s(n-k) \quad (1.5)$$

with the error transfer function

$$A(z) = \frac{E(z)}{S(z)} = 1 - \sum_{k=1}^p a_k z^{-k}. \quad (1.6)$$

If $s(n)$ were actually generated by a linear system as in Figure 1.1, then the prediction error, $e(n)$, would be equal to $Gu(n)$, the scaled excitation.

The basic problem of the LP analysis is to determine the predictor coefficients, $\{a_k\}$, directly from the speech signal. Since the spectral characteristics of the speech signal vary over time, the predictor coefficients at time instant, n , must be estimated from a short segment of the speech signal occurring around the same time instant. A set of predictor coefficients can be found by minimizing the mean-squared prediction error over a short speech frame.

Let us define short-time speech and error segments at time n as

$$s_n(m) = s(n+m) \quad (1.7)$$

$$e_n(m) = e(n+m), \quad m = 0, 1, 2, \dots, N-1, \quad (1.8)$$

and try to minimize the mean-squared error signal at time n

$$E_n = \sum_m e_n^2(m) \quad (1.9)$$

which can be written as

$$E_n = \sum_m \left[s_n(m) - \sum_{k=1}^p a_k s_n(m-k) \right]^2. \quad (1.10)$$

Equation (1.10) can be solved for the predictor coefficients by differentiating E_n with respect to each a_k and setting the results to zero,

$$\frac{\partial E_n}{\partial a_k} = 0, \quad k = 1, 2, \dots, p \quad (1.11)$$

giving

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^p \hat{a}_k \sum_m s_n(m-i)s_n(m-k). \quad (1.12)$$

By defining $\Phi_n(i, k) = \sum_m s_n(m-i)s_n(m-k)$ as the short-time covariance of $s_n(m)$, Equation (1.12) can be expressed in the compact form

$$\Phi_n(i, 0) = \sum_{k=1}^p \hat{a}_k \Phi_n(i, k), \quad i = 0, 1, 2, \dots, p-1 \quad (1.13)$$

which describes a set of p equations in p unknowns. A widely used method based on the autocorrelation analysis is described below in order to solve Equation (1.13) for optimum predictor coefficients.

By defining the limits on m from 0 to $N-1$ in the summation of Equation (1.10) it is inherently assumed that the speech segment, $s_n(m)$, is identically zero outside the interval $0 \leq m \leq N-1$. This is equivalent to assuming that the speech signal is multiplied by a finite length window $w(n)$

$$s_n(m) = \begin{cases} s(n+m)w(m), & 0 \leq m \leq N-1 \\ 0, & \text{otherwise} \end{cases} \quad (1.14)$$

Based on the Equation (1.14) the covariance, $\Phi_n(i, k)$, can be expressed as

$$\Phi_n(i, k) = \sum_{m=0}^{N-1+p} s_n(m-i)s_n(m-k), \quad (1.15)$$

$$= \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k), \quad (1.16)$$

$$= r_n(i-k), \quad 1 \leq i \leq p, \quad 0 \leq k \leq p \quad (1.17)$$

where the covariance function, $\Phi_n(i, k)$, reduces to the autocorrelation function $r_n(i - k)$. Since the autocorrelation function is symmetric, i.e., $r_n(k) = r_n(-k)$, the LP analysis equations can be represented as

$$\sum_{k=1}^p r_n(|i - k|) \hat{a}_k = r_n(i), \quad 1 \leq i \leq p, \quad (1.18)$$

and can be expressed in matrix form as

$$\begin{bmatrix} r_n(0) & r_n(1) & r_n(2) & \cdots & r_n(p-1) \\ r_n(1) & r_n(0) & r_n(1) & \cdots & r_n(p-2) \\ r_n(2) & r_n(1) & r_n(0) & \cdots & r_n(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_n(p-1) & r_n(p-2) & r_n(p-3) & \cdots & r_n(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(p) \end{bmatrix}. \quad (1.19)$$

The matrix of autocorrelation values in (1.18) form a $p \times p$ Toeplitz matrix (symmetric with all diagonal elements equal) and the predictor coefficients can be solved efficiently through one of the several well-known procedures. One such procedure that provides the predictor coefficients is the Levinson-Durbin algorithm which is described next:

Levinson-Durbin Algorithm:

In 1947, Levinson published an algorithm for solving the problem $Ax = b$ in which A is a Toeplitz positive definite matrix, and b is an arbitrary vector [15]. The Autocorrelation Equations (1.18) have this form, with b having a special relation to the elements of A . In 1960, Durbin extended Levinson's algorithm to this special case and developed an efficient algorithm [16]. Durbin's algorithm is widely known as the *Levinson-Durbin recursion*.

In Levinson-Durbin recursion the solution for the desired model order M is successively built up from lower order models. The basic steps of the recursion are as follows:

Initialization: $i = 0$

$$E^0 = r(0)$$

Recursion: For $i = 1, 2, \dots, M$

1. Compute the i th reflection coefficient (or the parcor coefficient),

$$k_i = \frac{1}{E^{i-1}} \left[r(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} r(|i-j|) \right], \quad (1.20)$$

2. Generate the order i set LP parameters,

$$\alpha_i^i = k_i \quad (1.21)$$

$$\alpha_j^i = \alpha_j^{i-1} - k_i \alpha_{i-j}^{i-1}, \quad j = 1, \dots, i-1 \quad (1.22)$$

3. Compute the error energy associated with the order i solution,

$$E^i = E^{i-1}(1 - k_i^2) \quad (1.23)$$

4. Return to step 1 with $i = i + 1$ if $i < M$.

With this recursion the final solution is given as

$$a_i = \text{LP coefficients} = \alpha_i^M \quad 1 \leq i \leq M \quad (1.24)$$

$$k_i = \text{Reflection coefficients.} \quad (1.25)$$

The reflection coefficients have some important properties:

- The coefficients $\{k_i\}$ uniquely determine the predictor coefficients $\{a_k\}$, or equivalently the linear predictor,
- the predictor filter can be implemented in lattice form [4] by the reflection coefficients without using the predictor coefficients $\{a_k\}$, and
- the reflection coefficients are bounded, $-1 \leq k_i \leq 1$ (whereas the filter coefficients can take any value).

Since the reflection coefficients form a model of the vocal tract, in 1960's they were used as a speech feature set in speech coding and recognition. In the LPC-10 speech vocoder standard, the predictor coefficients are represented through the

reflection coefficients [17]. However, speech coding results were not satisfactory [4].

In the next subsection another set of coefficients, which uniquely characterize the linear predictor, is described.

1.1.2 Line Spectral Frequencies

The LP filter coefficients can be represented by Line Spectral Frequencies (LSF's) which were first introduced by Itakura [18]. For a minimum phase m^{th} order LP polynomial, $A_m(z) = 1 + a_1 z^{-1} + \dots + a_m z^{-m}$ one can construct two $(m+1)^{st}$ order LSF polynomials, $P_{m+1}(z)$ and $Q_{m+1}(z)$, by setting the $(m+1)^{st}$ reflection coefficient to 1 and -1 in Levinson-Durbin algorithm as follows

$$P_{m+1}(z) = A_m(z) + z^{-(m+1)} A_m(z^{-1}), \quad (1.26)$$

and

$$Q_{m+1}(z) = A_m(z) - z^{-(m+1)} A_m(z^{-1}). \quad (1.27)$$

This is equivalent to setting the vocal tract acoustic tube model completely closed or completely open at the $(m+1)^{st}$ stage [4]. It is clear that $P_{m+1}(z)$ is a symmetric polynomial and $Q_{m+1}(z)$ is an anti-symmetric polynomial. There are three important properties of $P_{m+1}(z)$ and $Q_{m+1}(z)$:

- (i) all of the zeros of the LSF polynomials are on the unit circle,
- (ii) the zeros of the symmetric and anti-symmetric LSF polynomials are interlaced, and
- (iii) the reconstructed LP all-pole filter maintains its minimum phase property, if the properties (i) and (ii) are preserved during the quantization procedure.

As the roots of $P_{m+1}(z)$ and $Q_{m+1}(z)$ are on the unit circle, the zeros of $P_{m+1}(z)$ and $Q_{m+1}(z)$ can be represented by their angles which are called the line spectral frequencies. Therefore the LSF's are also bounded, $0 \leq f_i \leq 2\pi$ similar to the reflection coefficients.

It is also observed that the line spectral frequencies are closely related to the speech formants which are the resonant frequencies of the human vocal tract and hence the LSF's provide a spectrally meaningful representation of the LP filter.

Due to the above reasons, the LSF's are widely used in speech compression and recognition as speech feature parameters. For example it is possible to code coefficients of the LP filter by 24 bits for a speech frame of duration 20 msec without introducing any audible distortion. Details of the coding method [19, 20, 21] are described in Appendix A.

1.2 Speech Recognition

Speech recognition is mainly a pattern classification problem. The general recognition process can be divided into three sub-tasks, feature extraction, pattern classification and language processing, as shown in Figure 1.3.

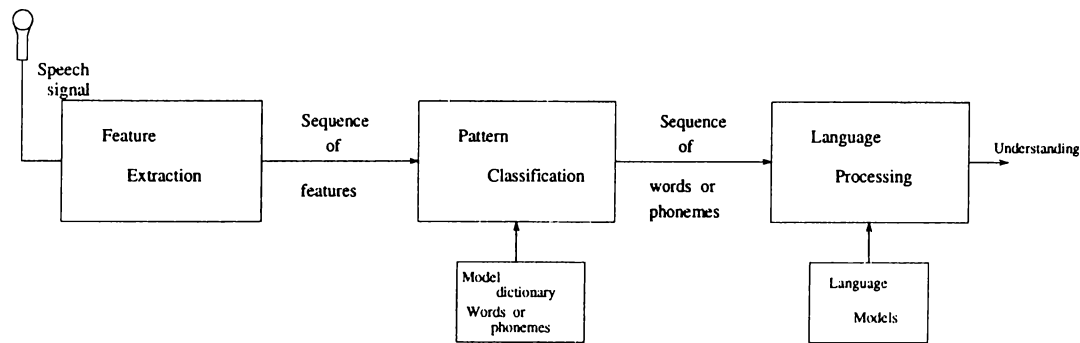


Figure 1.3: The block diagram of a speech recognition system.

The first step in speech recognition is to extract meaningful features from the speech signal. Feature extraction is usually based on the short-time spectral information of the speech signal. The spectral information is used to be modeled by either linear predictive analysis or Short-Time Fourier Transform (STFT) based features. LP filter derived or Short-Time Fourier Transform (STFT) based

feature sets are used to model the spectral information [7]. In this thesis subband analysis or equivalently wavelet based feature sets are constructed from the speech signal as an alternative to the commonly used Fourier domain feature sets. There are two important sub-tasks that should be done before and after the feature extraction process. Channel adaptation is a front-end process for feature extraction. The end-point detection should be performed by using the extracted features. In this thesis adaptive noise cancelation techniques are developed for channel adaptation, and a new distance measure based on the subband energy levels for the end-point detection is introduced.

The recognizer is first trained by a set of data containing the words in the vocabulary of the system. A dictionary of feature sequences is constructed from the training data corresponding to the words and phonemes in the vocabulary. Then task of pattern classifier is to match the observed sequence of feature sets with the sequences in the dictionary, and to chose the most likely dictionary entry. Hidden Markov Models (HMM), artificial neural-nets, template matching methods based on dynamic programming are the basic pattern classifier methods used in speech recognition. The HMM formalism is the most widely used one among these methods [5]. This statistical approach achieves quite satisfactory recognition rates [7, 22].

The last block of the speech recognition system is the language processing unit. The complexity of the language processing unit is application dependent. For simple applications, such as isolated word recognition, it is only a decision mechanism. For large vocabulary continuous speech recognition systems, it can be a set of rules verifying the grammar of the language [23].

Basic dimensions of the speech recognition problem can be considered as follows

- Degree of speaker independence : The performance of the speaker dependent recognition systems are superior.
- Size of vocabulary : Large vocabulary increase the error rates.
- Speaking rate and coarticulation : Recognition of isolated words are easier than that of continuous speech.

- Speaker variability and stress : Only for very simple cases, machines can handle these conditions as good as humans.
- Channel conditions : Noise strongly affects the performance of the recognition process.

Each of these dimensions should be considered before the planning of the applications, as the robustness of the speech recognizer will be determined according to the above parameters.

In this thesis, the feature extraction problem is considered. Speech feature parameters which are robust to variations in the original speech model assumptions are introduced. The performance of the new feature sets are evaluated using an isolated word recognition system based on an HMM structure. In the following subsection the HMM formalism is briefly reviewed.

1.2.1 Hidden Markov Model (HMM)

The pattern classification block is one of the key elements of a speech recognition system. In this block the speech feature vectors previously extracted from the speech signal vectors are processed. The recognition problem is to associate the sequence of feature vectors a word or a phoneme in the vocabulary.

Starting from the 1970s and mostly in 1980s, researchers working in the field of speech processing concentrated on the modeling of speech signal with stochastic approaches to address the problem of variability. The use of probabilistic modeling brought some advantages over the deterministic approaches. There are mainly two stochastic approaches in the speech recognition research. The first, Hidden Markov Model (HMM), is the stochastic finite state automaton, and the second is the Artificial Neural Networks (ANN). In this subsection the overview of the HMM formalism is given, since the HMM formalism is used as the pattern classifier in this thesis.

The HMM was first introduced to speech recognition field with the independent works of Baker at Carnegie-Melon University (1975) and Jelinek at IBM (1975). The HMM became popular in the last two decades and formed

the basis of many successful, large-scale speech recognition systems.

The HMM is a stochastic finite state automaton that is used to model the speech utterances such as word or subwords. In the HMM formalism the speech utterance is first reduced to a sequence of features. These speech features are also called observations,

$$\{y(1), y(2), \dots, y(T)\} \quad (1.28)$$

where $y(i)$ is the feature vector corresponding to the i -th speech frame and T is the total number of frames in the speech utterance. Entries of the vector, $y(i)$, may be the reflection coefficients, the LSF's or the new parameters described in Chapter 2-4 in this thesis. In this formalism, it is assumed that this sequence of observations is generated by a “hidden” finite state automaton. The classification problem is to determine the finite state automaton which produces the current observation sequence with the highest likelihood. It is important to note that every word or subword in the vocabulary is uniquely associated with an HMM, and in order to obtain a reliable speech recognition system the HMM's corresponding to the words in the vocabulary should be trained over a large population.

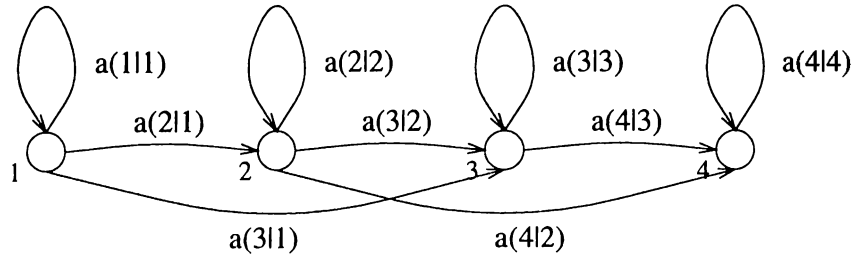


Figure 1.4: *Left to right HMM model.*

Although the finite state automaton can be in any configuration, in speech recognition usually the left-to-right model is used. The structure of a typical left-to-right HMM with four states is shown in Figure 1.4. The HMM can be characterized by a two stage probabilistic process. The first one is the allowable state transitions which are occurred in the model at each observation time. The

likelihoods of these transitions are stated as the state transition probabilities, where $a(i|j)$ is defined as the transition probability from state j to i as follows:

$$a(i|j) = P(\underline{x}(t) = i | \underline{x}(t-1) = j) \quad (1.29)$$

where $\underline{x}(t)$ is the state at time t . And the state transition matrix is given by

$$A = \begin{bmatrix} a(1|1) & a(1|2) & \cdots & a(1|S) \\ \vdots & \vdots & \ddots & \vdots \\ a(S|1) & a(S|2) & \cdots & a(S|S) \end{bmatrix} \quad (1.30)$$

where S is the total number of states in the model. An important property of this matrix is $\sum_j a(i|j) = 1$. That is, the entries of columns sum to 1 since, $a(i|j)$'s are probabilities. The state transition probabilities are assumed to be stationary in time, that is $a(i|j)$'s do not depend upon the time t . Due to this the random sequence is called a first order Markov process.

The second stage of the stochastic process is the generation of observations in each state. An observation probability density function $f_{\underline{y}|\underline{x}(t)}(\zeta|i)$ is assigned to each state i . Therefore the occurrence of the observations is governed by the pdf of each state. For the completeness of the HMM formalism, the state probability vector at time t is defined as follows,

$$\Pi(t) = \begin{bmatrix} P(\underline{x}(t) = 1) \\ P(\underline{x}(t) = 2) \\ \vdots \\ P(\underline{x}(t) = S) \end{bmatrix} \quad (1.31)$$

$$= A\Pi(t-1) \quad (1.32)$$

and

$$\Pi(t) = A^{t-1}\Pi(1). \quad (1.33)$$

In summary an HMM, Λ , is characterized by the following parameters

$$\Lambda = \{S, \Pi(1), A, \{f_{\underline{y}|\underline{x}}(\zeta|i), 1 \leq i \leq S\}\} \quad (1.34)$$

where S is the total number of states in the model, $\Pi(1)$ is the initial state probability matrix, A is the state transition matrix, and $f_{\underline{y}|\underline{x}}$'s are the observation probability density functions.

HMM Problems:

As indicated before, two key problems appear. These are the training and recognition problems.

The Training Problem:

Given a series of training observations for a given word, how do we construct an HMM for this word?

The Recognition Problem:

Given the trained HMM's for each word in the vocabulary, how do we find the HMM which produces the current observation sequence with the highest likelihood?

These questions will be examined in the following subsections.

The Discrete Observation HMM:

A multidimensional observation probability density function is associated with each state of the finite state automaton. When a particular state is reached an observation vector is generated according to this multidimensional pdf. The HMM is called the discrete (continuous) observation HMM when the observation pdf is a discrete (continuous) pdf. The discrete pdf's are usually constructed using vector quantization methods.

Let the observation pdf generate K distinct vectors for state i . In other words, the observation pdf for state i reduces to the form of K impulses on the real line for the discrete observation HMM. These observation probabilities are defined as,

$$b(k|i) = P(\underline{y}(t) = k | \underline{x}(t) = i), \quad k = 1, 2, \dots, K \quad (1.35)$$

and the observation probability matrix is formed as,

$$B = \begin{bmatrix} b(1|1) & b(1|2) & \dots & b(1|S) \\ \vdots & \vdots & \ddots & \vdots \\ b(K|1) & b(K|2) & \dots & b(K|S) \end{bmatrix}. \quad (1.36)$$

In the case of discrete observation HMM, the model, Λ , can be modified as,

$$\Lambda = \{S, \Pi(1), A, B, \{y_k, 1 \leq k \leq K\}\} \quad (1.37)$$

where $\{y_k, 1 \leq k \leq K\}$ represents the K possible observation vectors or equivalently the VQ codebook constructed in the training phase.

Recognition

Given the fully trained HMMs for each word in the vocabulary and the quantized observation sequence $\{y(1), y(2), \dots, y(T)\}$, it is tried to find the likelihood that each model could have produced the observation sequence in the recognition phase.

The partial observation sequences are defined for simplicity of the analysis as follows,

$$y_1^t = \{y(1), y(2), \dots, y(t)\} \quad (1.38)$$

$$y_{t+1}^T = \{y(t+1), y(t+2), \dots, y(T)\} \quad (1.39)$$

where y_1^t and y_{t+1}^T are called as forward and backward partial sequences of observations, respectively.

Forward-Backward (F-B) Approach:

An efficient way of computing the likelihood, $P(y|\Lambda)$, is the so-called forward-backward (F-B) algorithm of Baum et.al. [24]. It is necessary to define a forward-going and a backward-going probability sequence for this method. The joint probability of having generated the partial forward sequence y_1^t and having arrived at state i at time t , given HMM Λ , is defined as,

$$\alpha(y_1^t, i) = P(\underline{y}_1^t = y_1^t, \underline{x}(t) = i | \Lambda). \quad (1.40)$$

Similarly, the probability of generating the backward partial sequence y_{t+1}^T is defined as,

$$\beta(y_{t+1}^T, i) = P(\underline{y}_{t+1}^T = y_{t+1}^T | \underline{x}(t) = i, \Lambda). \quad (1.41)$$

Although, we only need the α sequence to compute the $P(y|\Lambda)$, we define β sequence for a future use in the training algorithm.

Suppose we are at time $t + 1$ and wish to compute $\alpha(y_1^{t+1}, j)$ for some state j . We should add up all the probabilities for all of the possible paths to state j at $t + 1$, that arising from all of the possible states at t . Then, clearly

$$\begin{aligned}\alpha(y_1^{t+1}, j) &= \sum_{i=1}^S \alpha(y_1^t, i) P(\underline{x}(t+1) = j | \underline{x}(t) = i) P(\underline{y}(t+1) = y(t+1) | x(t+1) = j) \\ &= \sum_{i=1}^S \alpha(y_1^t, i) a(j|i) b(y(t+1)|j).\end{aligned}\quad (1.42)$$

Hence, the α sequence can be computed with a lattice type structure and the recursion is initiated by setting,

$$\alpha(y_1^1, j) = P(\underline{x}(1) = j) b(y(1)|j) \quad 1 \leq j \leq S. \quad (1.43)$$

Similarly, the backward recursion can be derived for the β sequence:

$$\beta(y_{t+1}^T | i) = \sum_{j=1}^S \beta(y_{t+2}^T, j) a(j|i) b(y(t+1)|j) \quad (1.44)$$

where the recursion is initiated as,

$$\beta(y_{T+1}^T | i) = \begin{cases} 1 & \text{if } i \text{ is a legal final state} \\ 0 & \text{else} \end{cases} \quad (1.45)$$

Now, it is clear that,

$$P(y, \underline{x}(t) = i | \Lambda) = \alpha(y_1^t, i) \beta(y_{t+1}^T | i) \quad (1.46)$$

for any state i . Therefore,

$$P(y | \Lambda) = \sum_{i=1}^S \alpha(y_1^t, i) \beta(y_{t+1}^T | i). \quad (1.47)$$

Also, we can express the likelihood as,

$$P(y | \Lambda) = \sum_{\text{all legal final } i} \alpha(y_1^T, i). \quad (1.48)$$

The resulting algorithm for the recognition problem requires $O(S^2T)$ flops.

Training

Training a particular HMM to correctly represent its designated word is equivalent to finding the appropriate model matrices A and B , and the initial state probability vector $\Pi(1)$. The F-B algorithm provides an iterative estimation procedure for computing a model, Λ .

Let us define some new sequences to build up the estimation procedure,

$$\begin{aligned}\xi(i, j; t) &= P(\underline{u}(t) = u_{j|i} | y, \Lambda) \\ &= P(\underline{u}(t) = u_{j|i} | y, \Lambda) / P(y | \Lambda) \\ &= \frac{\alpha(y_1^t, i) a(j|i) b(y(t+1)|j) \beta(y_{t+2}^T | j)}{P(y | \Lambda)}, \quad 1 \leq t \leq T-1\end{aligned}\quad (1.49)$$

where $\underline{u}(t)$ represents the transition at time t , $u_{j|i}$ is the transition from state i to j and ξ is the probability of making a transition from state i to state j at time t ,

$$\begin{aligned}\gamma(i; t) &= P(\underline{u}(t) \in u_{\cdot|i} | y, \Lambda) \\ &= \sum_{j=1}^S \xi(i, j; t) \\ &= \frac{\alpha(y_1^t, i) \beta(y_{t+1}^T | j)}{P(y | \Lambda)}, \quad 1 \leq t \leq T-1\end{aligned}\quad (1.50)$$

where $u_{\cdot|i}$ is the set of transitions exiting state i and γ is the total probability of making transitions from state i to all other states at time t ,

$$\begin{aligned}\nu(j; t) &= P(\underline{x}(t) = j | y, \Lambda) \\ &= \begin{cases} \gamma(j; t) & 1 \leq t \leq T-1 \\ \alpha(y_1^t, j) & t = T \end{cases} \\ &= \frac{\alpha(y_1^t, j) \beta(y_{t+1}^T | j)}{P(y | \Lambda)}, \quad 1 \leq t \leq T\end{aligned}\quad (1.51)$$

where ν is the probability of being at state j at time t ,

$$\begin{aligned}\delta(j, k; t) &= P(\underline{y}_j(t) = k | y, \Lambda) \\ &= \begin{cases} \nu(j; t) & \text{if } y(t) = k \text{ and } 1 \leq t \leq T \\ 0 & \end{cases}\end{aligned}\quad (1.52)$$

where $\underline{y}_j(t)$ models the observation being emitted at state j and at time t , and δ is the probability of observing k -th observation vector at state j and at time t .

From these sequences four related key results can be computed:

$$\xi(i, j; \cdot) = P(\underline{u}(\cdot) = u_{j|\cdot} | y, \Lambda) = \sum_{t=1}^{T-1} \xi(i, j; t) \quad (1.53)$$

where $\xi(i, j; \cdot)$ is the total probability of making a transition from state i to state j ,

$$\gamma(i; \cdot) = P(\underline{u}(\cdot) \in u_{\cdot|i} | y, \Lambda) = \sum_{t=1}^{T-1} \gamma(i; t) \quad (1.54)$$

where $\gamma(i; \cdot)$ is the total probability of making a transition from state i to all possible states,

$$\nu(j; \cdot) = P(\underline{u}(\cdot) \in u_{j|\cdot} | y, \Lambda) = \sum_{t=1}^T \nu(j; t) \quad (1.55)$$

where $\nu(j; \cdot)$ is the total probability of making a transition from all possible states to state j , and

$$\delta(j, k; \cdot) = P((\underline{y})_j(\cdot) = k | y, \Lambda) = \sum_{t=1}^T \delta(j, k; t) \quad (1.56)$$

where $\delta(j, k; \cdot)$ is the total probability of observing the k -th observation while making a transition to state j .

With these interpretations, the following reestimation equations can be written,

$$\begin{aligned} \bar{a}(j|i) &= \frac{\xi(i, j; \cdot)}{\gamma(i; \cdot)} \\ &= \frac{\sum_{t=1}^{T-1} \alpha(y_1^t, i) a(j|i) b(y(t+1)|j) \beta(y_{t+2}^T|j)}{\sum_{t=1}^{T-1} \alpha(y_1^t, i) \beta(y_{t+1}^T|i)} \end{aligned} \quad (1.57)$$

$$\begin{aligned} \bar{b}(k|j) &= \frac{\delta(j, k; \cdot)}{\nu(j; \cdot)} \\ &= \frac{\sum_{t=1, y(t)=k}^T \alpha(y_1^t, j) \beta(y_{t+1}^T|j)}{\sum_{t=1}^T \alpha(y_1^t, j) \beta(y_{t+1}^T|j)} \end{aligned} \quad (1.58)$$

$$\begin{aligned} P(\underline{x}(1) = i) &= \gamma(i; 1) \\ &= \frac{\alpha(y_1^1, i) \beta(y_2^T|i)}{P(y|\Lambda)}. \end{aligned} \quad (1.59)$$

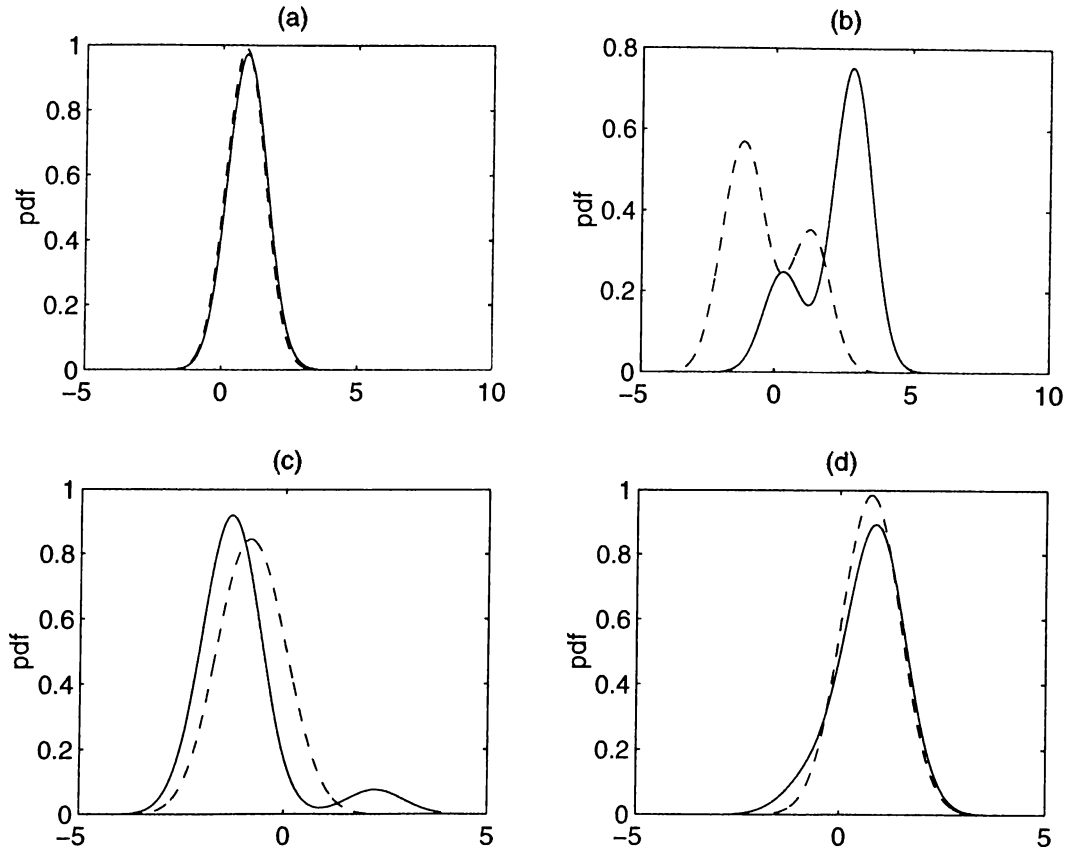


Figure 1.5: Observation probability densities of digit *zero* (line), and *four* (dotted) for states (a) one, (b) two, (c) three, and (d) four.

The Continuous Observation HMM:

In the more general case in which the observations are continuous, the distributions of observations within each state characterized by a multivariate pdf. In the recognition problem, the likelihood of generating observation $y(t)$ in state j is simply defined as

$$b(y(t)|j) = f_{\underline{y}|\underline{x}}(y(t)|j). \quad (1.60)$$

The most widely used form of observation pdf is the Gaussian mixture density, which is of the form

$$f_{\underline{y}|\underline{x}}(\zeta|i) = \sum_{m=1}^M c_{im} \mathcal{N}(\zeta; \mu_{im}, C_{im}) \quad (1.61)$$

where c_{im} is the mixture coefficient for the m -th component for state i , and $\mathcal{N}(\cdot)$ denotes a multivariate Gaussian pdf with mean μ_{im} and covariance matrix C_{im} . With this formulation the training problem can be solved with simple reestimation procedures [25]. A detailed description of the training problem can be found in [5, 22].

An illustrative example for the observation probability densities of a four-state HMM structure with 3 mixture densities are given in Figure 1.5. These observation densities correspond to the fourth element of the observation vector for two English digits *zero* and *four*. The state transition probabilities for these digits are given in Figure 1.6.

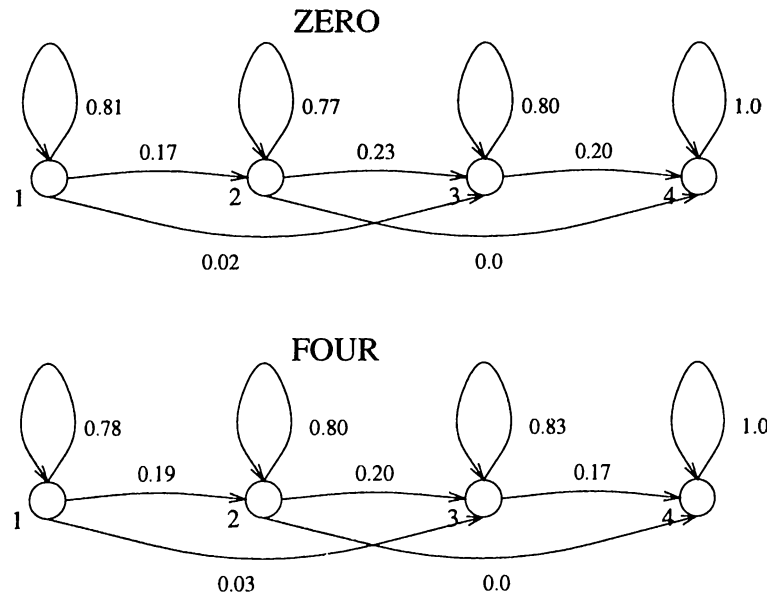


Figure 1.6: HMM structures for English digits *zero* and *four*.

In the isolated word recognition systems the elements of the vocabulary are discrete words. Consider a recognition system with a V -word vocabulary. First, for each word in the vocabulary an HMM model, Λ_v $1 \leq v \leq V$, is built using the training data as described in subsection 1.2.1. A graphical description of this procedure is depicted in Figure 1.7. During the recognition process the observation sequence $y = y(1), y(2), \dots, y(T)$ is extracted from the speech signal, and the probabilities $P(y|\Lambda_v)$ are evaluated for each word. The result

is determined according to the HMM word model which produces the highest likelihood [5].

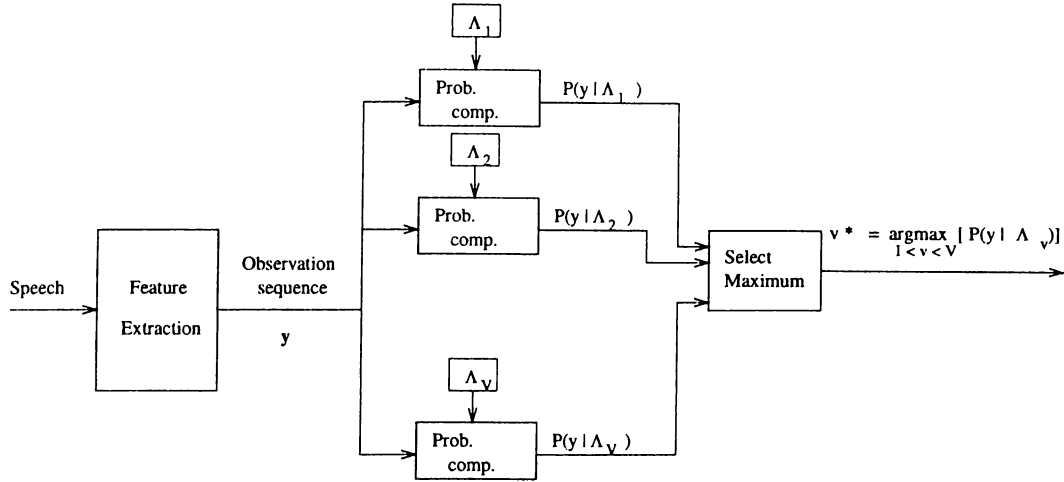


Figure 1.7: HMM recognizer.

The elements of the vocabulary may be phonemes as well. In many large vocabulary continuous speech recognition systems the continuous speech signal is first segmented into phonemes. Then the current phoneme is determined over the vocabulary containing all possible phonemes. Both in English and Turkish, the spoken language can be represented by about 40 phonemes. Given a good phoneme segmentation method, an HMM based recognition system can easily handle 40 phonemes when the speech is noise-free.

1.3 Speech Recognition Research in Turkey

During the last 10 years several researchers have developed speech recognition systems for Turkish. All of these systems consider isolated word recognition and vowel recognition [26, 27, 28, 29, 30]. These systems are described in a speech workshop that was held in Middle East Technical University (METU) on May 1995 [31]. The studies in Hacettepe University cover the isolated word recognition systems based on discrete density HMM structures with LP derived cepstral

features [31]. Isolated word and vowel recognition systems based on dynamic time warping and artificial neural networks are also investigated in Osmangazi University [27, 30, 31].

We recently finished voice dialing project for mobile telephones in cooperation with TÜBİTAK-BİLTEN [26, 31]. This project was also supported by ASELSAN, and some of the techniques developed in this thesis are successfully used in this project. The Turkish speech databases collected for this project are used in the simulation studies in Chapter 2 and Chapter 3.

One of the important result of the METU speech workshop was the lack of a speech corpus for Turkish speech recognition studies. In this workshop the need to construct a national speech corpus which is extremely important for speech recognition research was pointed out and TÜBİTAK-BİLTEN will prepare a speech corpus. Besides this, Hacettepe University, and Osmangazi University have their own recordings that include the isolated Turkish digits and some control words.

Chapter 2

LINE SPECTRAL FREQUENCIES IN SPEECH RECOGNITION

As indicated in Chapter 1, extraction of feature parameters from the speech signal is the first step in speech recognition. The feature parameters are desired to have perceptually meaningful parameterization and yet robust to variations in environmental noise. In this chapter, a new set of speech feature parameters based on the Line Spectral Frequency (LSF) representation in subbands, SUBLSFs, is introduced.

The LSF representation of speech is reviewed in Section 2.1. The new speech feature parameters, SUBLSFs are described in Section 2.2. The parameters are used in a speaker independent continuous density Hidden Markov Model (HMM) based isolated word recognition system operating in the presence of car noise. The simulation results are described in Section 2.2.1.

2.1 LSF Representation of Speech

Linear predictive modeling techniques are widely used in various speech coding, synthesis and recognition applications. The Line Spectral Frequency (LSF) representation of the linear prediction (LP) filter is introduced by Itakura [18]. LSFs are closely related to formant frequencies and they have some desirable properties which make them attractive to represent the Linear Predictive Coding (LPC) filter. The quantization properties of the LSF representation is recently investigated in [19, 20, 32].

Let the m -th order inverse filter $A_m(z)$,

$$A_m(z) = 1 + a_1 z^{-1} + \cdots + a_m z^{-m} \quad (2.1)$$

be obtained by the LP analysis of speech. As defined in Section 1.1.2, the LSF polynomials of order $(m+1)$, $P_{m+1}(z)$ and $Q_{m+1}(z)$, are constructed by setting the $(m+1)$ -st reflection coefficient to 1 or -1 . In other words, the polynomials, $P_{m+1}(z)$ and $Q_{m+1}(z)$, are defined as,

$$P_{m+1}(z) = A_m(z) + z^{-(m+1)} A_m(z^{-1}), \quad (2.2)$$

and

$$Q_{m+1}(z) = A_m(z) - z^{-(m+1)} A_m(z^{-1}). \quad (2.3)$$

The zeros of $P_{m+1}(z)$ and $Q_{m+1}(z)$ are called the Line Spectral Frequencies (LSFs), and they uniquely characterize the LPC inverse filter $A_m(z)$. $P_{m+1}(z)$ and $Q_{m+1}(z)$ are symmetric and anti-symmetric polynomials, respectively. Therefore, it is possible to decompose the power spectrum $|A_m(\omega)|^2$ as

$$|A_m(\omega)|^2 = \frac{1}{4} (|P_{m+1}(\omega)|^2 + |Q_{m+1}(\omega)|^2). \quad (2.4)$$

Roots of the LP polynomial corresponds to the peaks or the formants of the spectra of the speech signal. By examining Equation (2.4) it can be seen that the roots of $A_m(z)$ (or the formants) are related with the roots of the $P_{m+1}(z)$ and $Q_{m+1}(z)$. In order to illustrate this relationship between formants and the LSFs more clearly, the LP spectrum and the associated LSFs are plotted in Figure 2.1(a) for a vowel /a/ and Figure 2.1(b) for a fricative /sh/. It can be observed

that the cluster of (two to three) LSFs characterizes a formant frequency and the bandwidth of a given formant depends on the closeness of the corresponding LSFs. In addition, the spectral sensitivities of LSFs are localized. That is, a change in a given LSF produces a change in the LP power spectrum only in its neighborhood.

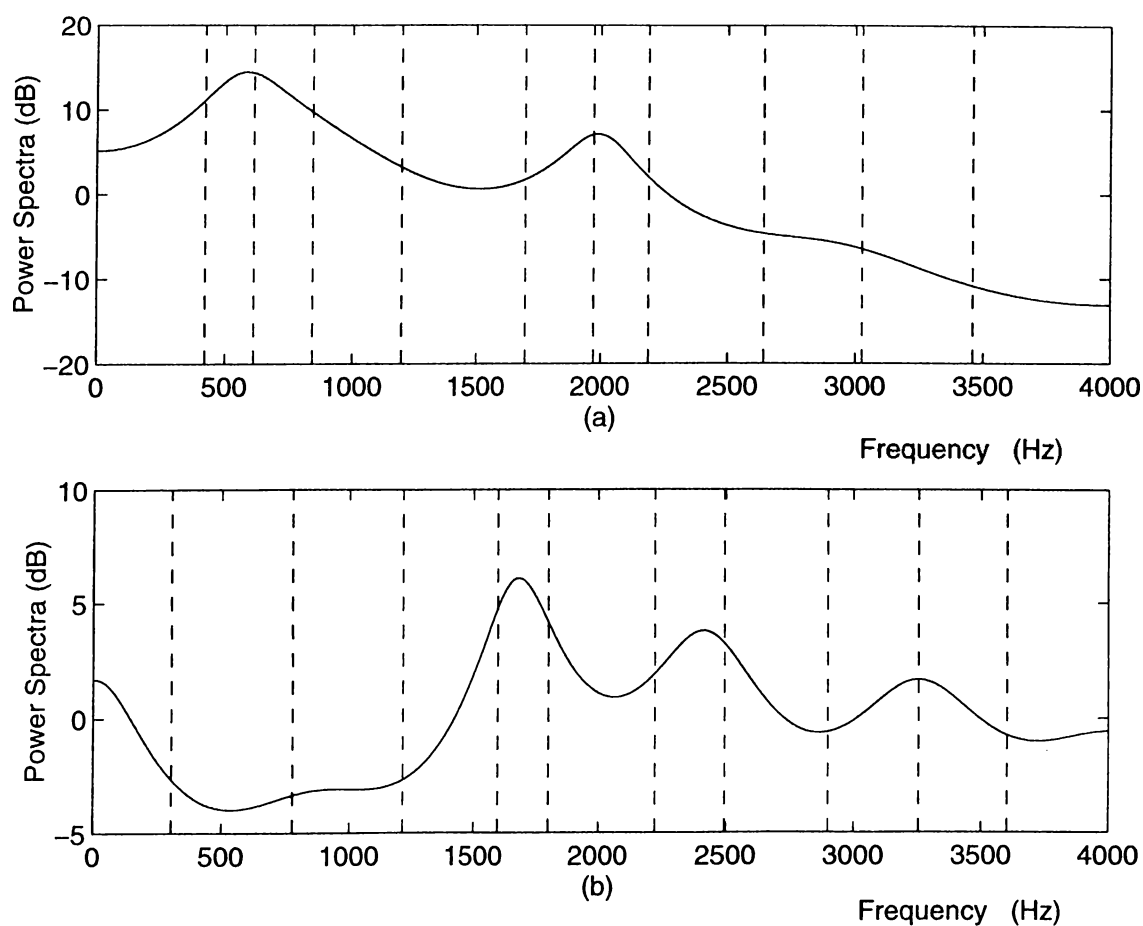


Figure 2.1: LP power spectrum and the associated LSFs for (a) vowel /a/ and (b) fricative /sh/.

2.2 LSF Representation in Subbands for Speech Recognition

In this section, a new set of speech feature parameters is constructed from subband analysis based Line Spectral Frequencies (LSFs) [33]. The speech signal is divided into several subbands and the resulting subsignals are represented by LSFs. The performance of the new speech feature parameters, SUBLSFs, is compared with the widely used *Mel* Scale Cepstral Coefficients (MELCEPs). SUBLSFs are observed to be more robust than the MELCEPs in the presence of car noise.

It is well known that LSF representation and cepstral coefficient representation of speech signals have comparable performances for a general speech recognition system [34, 35]. Car noise environments, however, have low-pass characteristics which may degrade the performance of a general full-band LSF or *mel* scaled cepstral coefficient (MELCEP) representations [5]. In this section, LSF based representation of speech signals in subbands is introduced.

The speech signal is filtered by a low-pass and a high-pass filter and the LP analysis is performed on the resulting two subsignals. Next, the LSFs of the subsignals are computed and the feature vector is constructed from these LSFs. More emphasis to the high-frequency band is given by extracting more LSFs from the high-band than the low-frequency band.

The *mel* scale is accepted as a transformation of the frequency scale to a perceptually meaningful scale, and it is widely used in feature extraction [36]. However the environmental noise may effect the performance of the *mel* scale derived features. It is experimentally observed that significant amount of the spectral power of car noise¹ is localized under 500 Hz. Figure 2.2 shows an utterance from the Volvo 340 recording, and the average spectra of this recording. Due to this reason the LP analysis of speech signal is performed in two bands, a low-band (0–700 Hz) and a high-band (700–4000 Hz). In this case the high-band can be assumed to be noise-free.

¹This noise is recorded inside a Volvo 340 on a rainy asphalt road by *Institute for Perception-TNO, The Netherlands*.

This kind of frequency domain decomposition can be generalized to cases in which the noise is frequency localized.

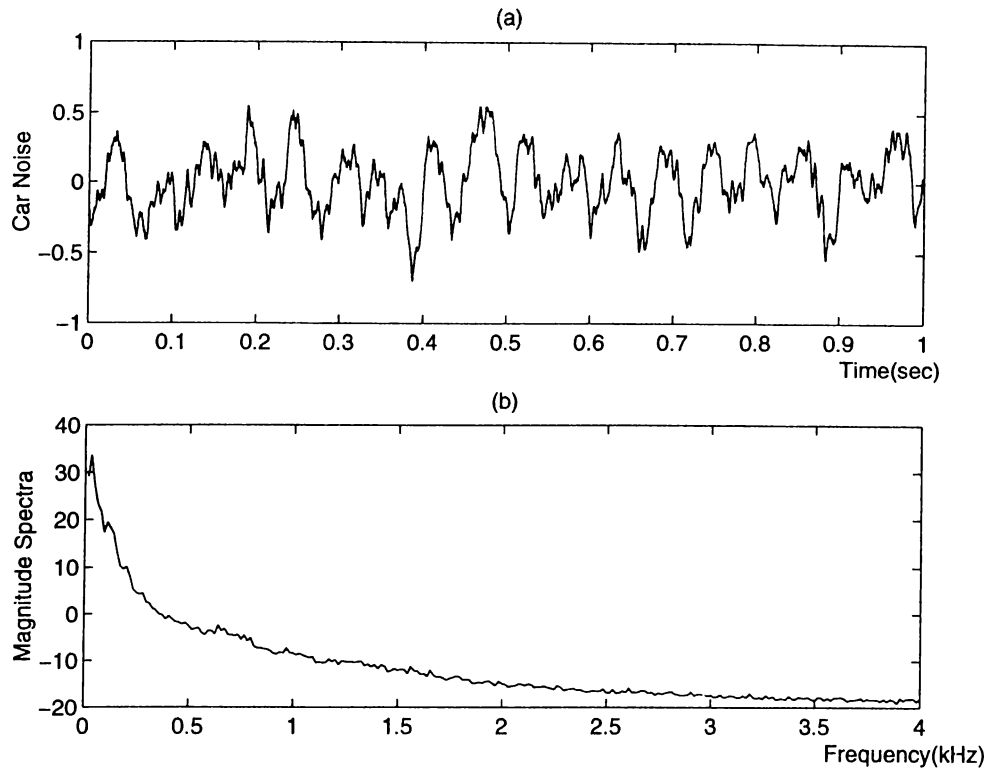


Figure 2.2: (a) Car noise time series, (b) Car noise spectra for Volvo 340 recording.

In simulation studies a continuous density Hidden Markov Model (HMM) based speech recognition system is used with 5 states and 3 mixture densities. Simulation studies are performed on the vocabulary of ten Turkish digits (0:sıfır, 1:bir, 2:iki, 3:üç, 4:dört, 5:beş, 6:altı, 7:yedi, 8:sekiz, 9:dokuz) from the utterances of 51 male and 51 female speakers. The isolated word recognition system is trained with 25 male and 25 female speakers, and the performance evaluation is done with the remaining 26 male and 26 female speakers. The speech signal is sampled at 8 kHz and the car noise is assumed to be additive. The low-pass and high-pass filters are chosen as 51-th order linear phase FIR filters. The magnitude responses of the low-pass and the high-pass filters are shown in Figure 2.3(a) and

Figure 2.3(b), respectively.

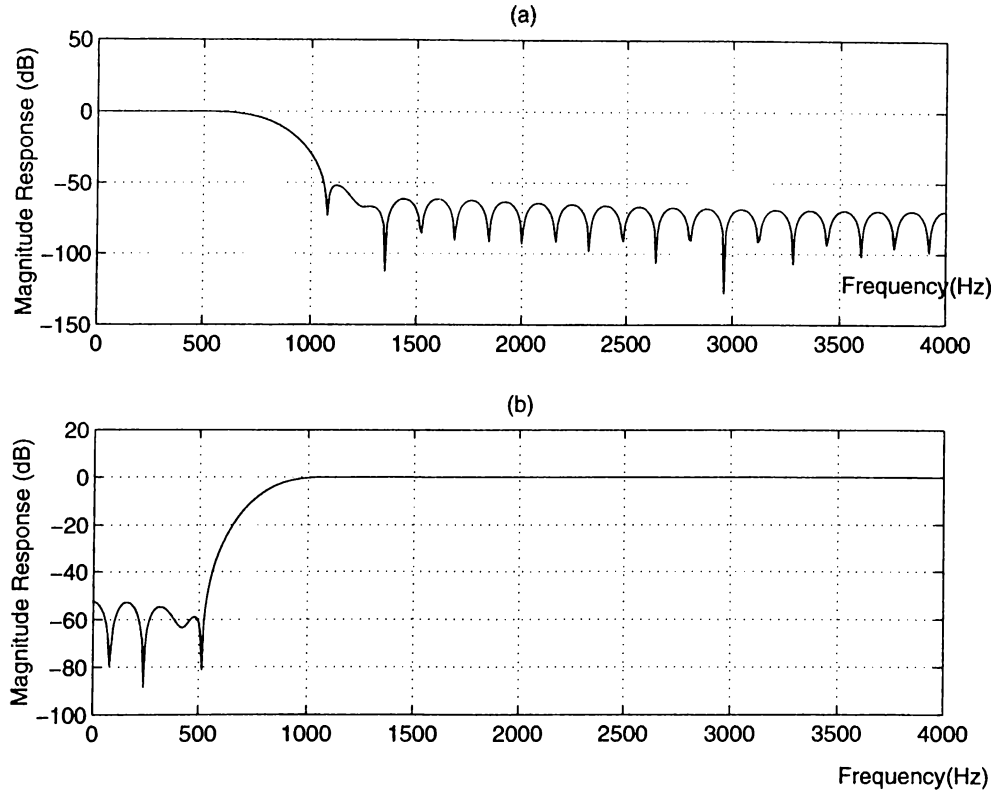


Figure 2.3: Magnitude responses of the (a) low-pass and the (b) high-pass filters.

2.2.1 Performance of LSF Representation in Subbands

Both 12-th and 20-th order LP analyses are performed on every 10 msec with a window size of 30 msec (using a Hamming window) for low-band (noisy band) and high-band (noise free band) of the speech signal, respectively. The first 5 LSFs of the low-band and the last 19 LSFs of the high-band are combined to form the Subband derived LSF feature vector (SUBLSF). The recognition rate of SUBLSFs are recorded in Table 1 under various SNRs.

The performance of SUBLSFs are compared with three other widely used

feature sets. The recognition rates of four feature sets, SUBLSF, LSF, LSF+DLSF, and MELCEP, for various SNR values are also given in Table 2.1.

Table 2.1: Recognition rates of SUBLSF, MELCEP and LSF representations in percentage.

SNR	SUBLSF	LSF	LSF+DLSF	MELCEP
16.0	86.54	85.00	84.81	85.00
11.0	86.73	84.04	85.00	84.40
7.0	85.00	80.96	80.96	83.70
5.0	84.04	80.19	79.23	82.90
3.0	83.46	78.84	76.73	82.10

In column 2 of Table 1 the full-band LSF representation is investigated. The size of the LSF vector is 24 which is obtained by a 24-th order LP analysis. The recognition rate of LSFs with their time derivatives, DLSFs, is also obtained. In this case 12-th order LP analysis is carried out to construct the 24-th order DLSF feature vector. The results are summarized in column 3 of Table 1.

In column 4 the results of MELCEP representation is given. Frequency domain cepstral analysis is performed to extract 12 *mel* scale cepstral coefficients and a 24-th order MELCEP feature vector is obtained from 12 *mel*-scale cepstral coefficients and their time derivatives.

In our simulation studies we observed that the SUBLSFs have the highest recognition rate.

2.2.2 Conclusion

In this chapter, a new set of speech feature parameters based on LSF representation in subbands, SUBLSFs, is introduced. It is experimentally observed that the SUBLSF representation provides higher recognition rate than the commonly used MELCEP, LSF, LSF+DLSF representations for speaker independent isolated word recognition in the presence of car noise. Since the car noise is concentrated in low frequencies, the high frequency bands can be

assumed to be noise free. This property is exploited in this chapter to achieve robustness against noise.

The computational complexity the SUBLSF scheme is clearly higher than the LSF and cepstral analysis methods, because two LP analyses are carried out in two subbands. In the next chapter a computationally efficient method for feature parameter extraction is introduced. The method is also based on subband analysis.

Chapter 3

WAVELET ANALYSIS FOR SPEECH RECOGNITION

In this chapter, a new set of feature parameters for speech recognition is presented. The new feature set is obtained from the root-cepstral coefficients derived from the wavelet analysis of the speech signal [37]. The performance of the new feature representation is compared to the *mel* scale cepstral representation (MELCEP) in the presence of car noise. Subband analysis based parameters are observed to be more robust than the commonly employed MELCEP representation.

The first step in automatic speech recognition is the extraction of acoustic features which characterize the speech signal in a perceptually meaningful manner. Many feature parameter sets including Linear Predictive Coding (LPC) parameters, Line Spectral Frequencies (LSF's), and cepstral parameters were used for speech recognition in the past [22, 34, 35, 38]. Among these methods the ones based on cepstral analysis are the most widely used in applications where noise is present [39]. This is due to the compatibility of some cepstral methods with the human auditory system. The new parameters are based on wavelet analysis or equivalently the multirate subband analysis which provides robust recognition performance by appropriately deemphasizing frequency bands that are corrupted by noise.

The new feature parameters are obtained via a two stage procedure. First, the frequency domain is partitioned into nonuniform regions using a basic half-band filter bank in a tree structured manner. As a result a family of subband signals are obtained. In order to incorporate the properties of the human auditory perceptual system the frequency range lower than 1kHz is uniformly divided whereas a logarithmic partition is applied above 1 kHz. The resulting frequency bands are very similar to the commonly used mel-scale division [36]. After this step, a set of cepstral-like parameters are obtained from the subband signals. This procedure is described in detail in Section 3.1.

The performance of the new Subband based Cepstral parameters, SUBCEPs, is tested in the presence of car noise. The SUBCEPs are compared to the short-time Fourier Analysis based methods. The SUBCEP parameters are observed to produce better recognition rates than other currently employed feature parameters. Simulation examples are presented in Section 3.2.

3.1 Subband Analysis based Cepstral Coefficient (SUBCEP) Representation

It is well known that the Fourier transform provides information about the frequency content of the signal, but it does not provide any temporal information due to the infinite extent of the Fourier basis functions. Therefore, in many applications including speech processing the short-time Fourier analysis is performed with overlapping time windows to obtain temporal information. On the other hand the wavelet transform is ideally suited for the analysis of signals with time varying characteristics. The new speech feature parameters described in this section are derived from the wavelet transform coefficients.

The wavelet analysis associated with a subband decomposition filter bank provides a fast and computationally efficient structure for decomposing the frequency domain along with the temporal information. The basic building block of a wavelet transform which is a multirate signal processing technique is realized by a subband decomposition filter bank as shown in Figure 3.1. The

filter bank structure consists of a lowpass and a highpass filter, and downsampling units. The passbands of the low and high pass filters are $[0, \pi/2]$ and $[\pi/2, \pi]$, respectively, so that the frequency domain is divided into two halfbands. In the subband decomposition filter bank structure the input signal is first filtered by the complementary low-pass and high-pass filters and then the outputs of the filters are passed to the downsampling units which drop every other sample of their inputs reducing the sampling rate by a factor of two. In this way two temporal subsignals containing the low and high frequency components of the original signal are obtained. The simple filter bank structure of Figure 3.1 provides a time-frequency decomposition of the original signal with the frequency regions, $[0, \pi/2]$ and $[\pi/2, \pi]$. Each of the subsignals can be further decomposed into two new subsignals using the same filter bank once again, and this analysis procedure can be repeated until the desired frequency domain decomposition is obtained. Resulting temporal subsignals constitute the so-called wavepacket representation uniquely characterizing the original signal [40]. The original signal in turn can be perfectly reconstructed (or synthesized) from the subsignals, if the lowpass and highpass filters are properly selected [40, 41].

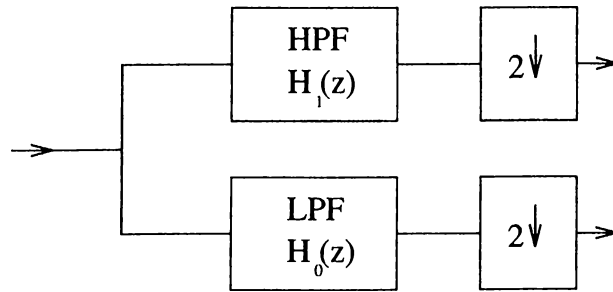


Figure 3.1: Basic block of subband decomposition.

In this work, a subband decomposition filter bank corresponding to a biorthogonal wavelet transform [41, 42] is used in simulation studies. The low-pass filter, $H_0(z)$, and the high-pass filter, $H_1(z)$, for this filter bank have the form

$$H_0(z) = \frac{1}{2}[1 + zA(z^2)], \quad (3.1)$$

and

$$H_1(z) = -z^{-1} + \frac{1}{2}B(z^2)(1 + zA(z^2)), \quad (3.2)$$

where $A(z^2)$ and $B(z^2)$ are polynomials of z^2 . The class of low pass filters satisfying (3.1) includes the maximally flat Lagrange filter family [41, 42]. For example, the 7-th order Lagrange filter have the transfer function

$$H_0(z) = \frac{1}{2} + \frac{9}{32}(z^1 + z^{-1}) - \frac{1}{32}(z^3 + z^{-3}) \quad (3.3)$$

which is a half-band linear phase FIR filter. Note that (3.3) can be easily put into the form of (3.1) with

$$A(z^2) = \frac{9}{16}(1 + z^{-2}) - \frac{1}{16}(z^2 + z^{-4}) . \quad (3.4)$$

The second polynomial, $B(z^2)$ in (3.2) is chosen as

$$B(z^2) = \frac{1}{2}(1 + z^{-2}), \quad (3.5)$$

This selection of $A(z^2)$ and $B(z^2)$ produces good low-pass and high-pass frequency responses for the filters $H_0(z)$ and $H_1(z)$ [41]. Consequently, the filter bank approximately divides the frequency domain into two half-bands, $[0, \pi/2]$ and $[\pi/2, \pi]$ as shown in Figure 3.2.

In our simulation studies, the speech signal, $s(n)$, is first decomposed into $L = 21$ subsignals, $\{s_l(n)\}_{l=1}^L$ using the basic half-band filter bank structure in a tree structured manner as shown in Figure 3.3. The corresponding frequency domain decomposition is similar to the *mel*-scale [5, 36] as shown in Figure 3.4. The frequency range lower than 1 kHz is uniformly divided whereas a logarithmic partition is applied above 1kHz.

After obtaining the subsignals cepstral analysis is performed to obtain the feature parameters. For each subsignal, a parameter e_l is defined over a time window of T_w seconds with an overlap of T_o seconds,

$$e_l = \frac{1}{N_l} \sum_{n=1}^{N_l} |s_l(n)|, \quad l = 1, 2, \dots, L \quad (3.6)$$

where N_l is the number of samples in the l -th band. For example, the window size, T_w , and the overlap interval T_o are selected as 48 msec and 32 msec, respectively,

at the sampling rate of 16 kHz. The Subband Cepstral (SUBCEP) parameters, $SC(k)$, are defined as root-cepstral [13, 43] coefficients as follows,

$$SC(k) = \sum_{l=1}^L (e_l)^{p_l} \cos\left(\frac{k(l-0.5)}{L}\pi\right), \quad k = 1, 2, \dots, 12 \quad (3.7)$$

where p_l is the root value for the l -th frequency band. In root cepstral analysis a fixed root value is used [13]. On the other hand the frequency bands can be weighted by the proper selection of p_l values. This increases the robustness of the speech feature parameters against the colored environmental noise. Our main contribution to the derivation of SUBCEP parameters are the subband decomposition based energy values, e_l , and the weighted root nonlinearity.

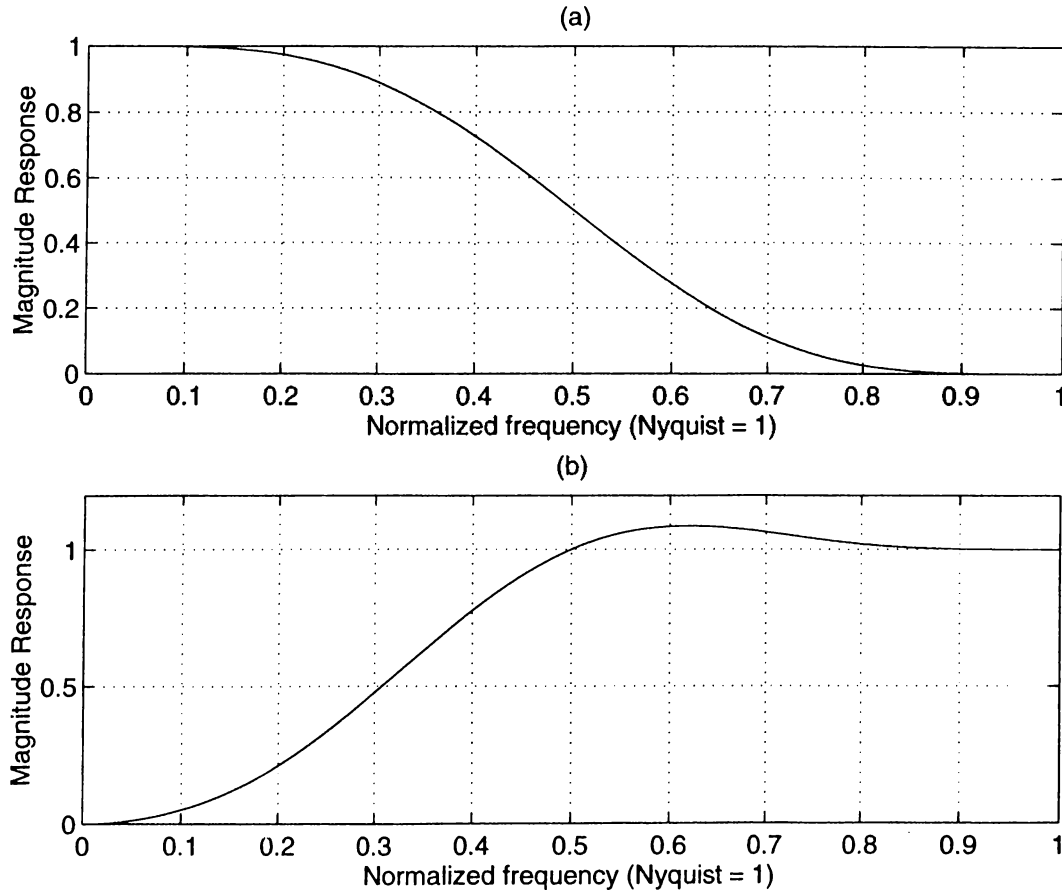


Figure 3.2: The magnitude response of the biorthogonal filter bank. (a) Low-pass filter, (b) High-pass filter.

The SUBCEP parameters can be obtained in a computationally efficient manner because the downsampling operation reduces the data size by a factor of two at every stage of the subband decomposition tree. Also, the biorthogonal filter bank structure of [41] can be implemented using integer arithmetic as all of the filters have rational coefficients. Furthermore, some wavelet transforms can be realized in $\text{Order}(N)$ multiplications where N is the data size [40].

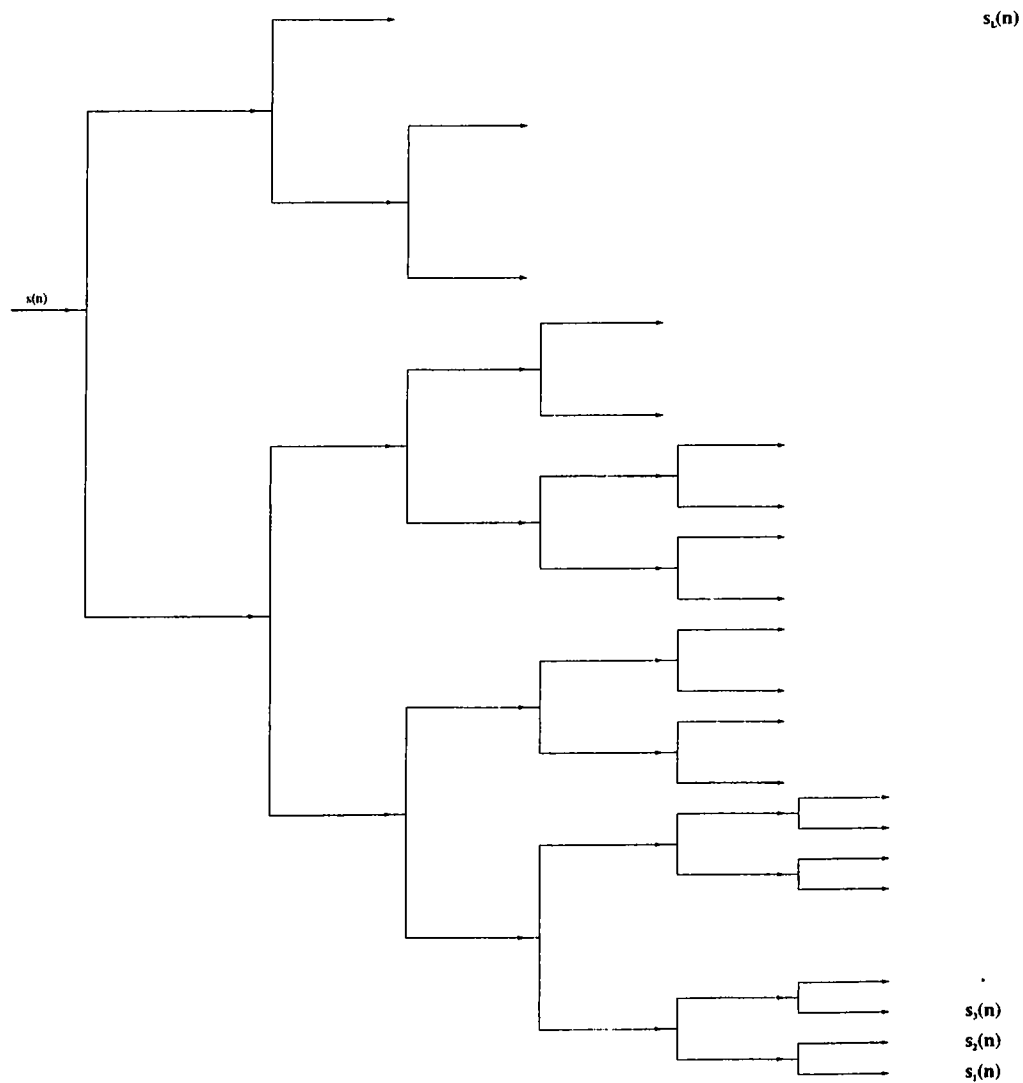


Figure 3.3: The tree-structure of the subband decomposition for TI20 database.

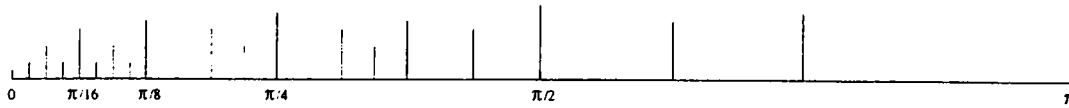


Figure 3.4: The subband decomposition of the speech signal.

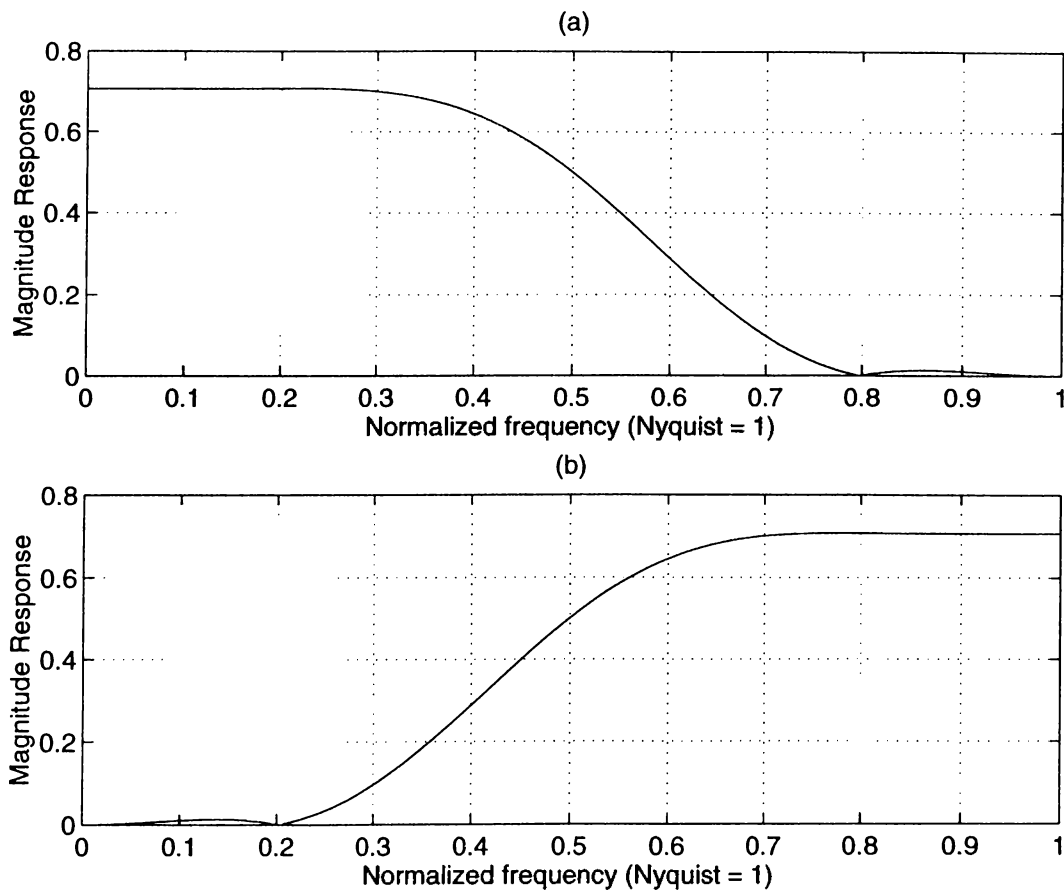


Figure 3.5: The magnitude responses of the orthogonal filterbank.

Commonly used MELCEP speech parameters are obtained either in time domain with critical band filter banks [36] or in frequency domain with critical band windowing of the speech spectrum [5]. Since multirate signal processing techniques are not employed in the design of the critical band filter bank, very large filter orders are necessary for narrow subbands. This results in a computationally expensive and memory intensive implementation. Critical band windowing, on the other hand, requires complex arithmetic.

Other filter-bank structures and wavelet transforms can also be used to achieve a similar frequency decomposition. In order to see the effects of the filter bank structure on speech recognition rate another filter bank by [44] which corresponds to an orthogonal wavelet transform is also used in simulation studies. This filter bank has the maximum vanishing moment property. Therefore it exhibits maximum decay among the filter banks with the same order. The magnitude responses of the orthogonal filterbank are shown in Figure 3.5. Simulation results are presented in the next section.

3.1.1 Frequency Characteristics of the Iterated Filter Bank Structure

Consider the multirate systems in Figure 3.6 and 3.7. In Figure 3.6, the input signal $x_1(n)$ is first filtered by $G_0(\omega)$, and the output is downsampled by a factor of two (every other sample is dropped). Let us call the output of the downsampler $x_2(n)$. The signal $y_1(n)$ is obtained as a result of filtering $x_2(n)$ by the filter $G_1(\omega)$.

In Figure 3.7, the signal $x_1(n)$ is first filtered by the filter $G_{eq}(\omega) = G_0(\omega)G_1(2\omega)$ and then the output is downsampled by a factor of two.

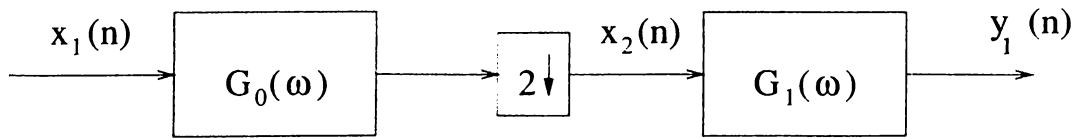


Figure 3.6: A multirate system with two similar cascaded blocks.

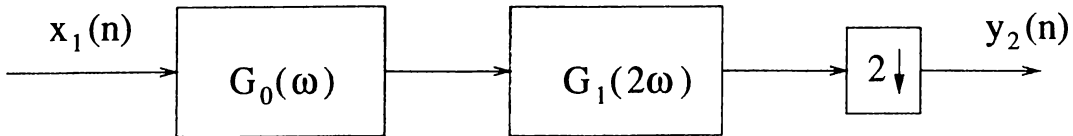


Figure 3.7: An equivalent multirate system.

Result 1 *The multirate systems in Figure 3.6 and Figure 3.7 are equivalent to each other.*

Proof:

Downsampling of a sequence $x(n)$ by an integer factor of N results in a sequence $y(n)$ given by

$$y(n) = x(nN). \quad (3.8)$$

In Fourier domain, the following relation exists between $X(\omega)$ and $Y(\omega)$,

$$Y(\omega) = \frac{1}{N} \sum_{k=0}^{N-1} X\left(\frac{\omega - 2\pi k}{N}\right). \quad (3.9)$$

Using Equation (3.9) it follows that downsampling the filtered signal by a factor of 2 results in

$$Y_2(\omega) = \frac{1}{2} \sum_{k=0}^1 X_1\left(\frac{\omega - 2\pi k}{2}\right) G_0\left(\frac{\omega - 2\pi k}{2}\right) G_1\left(2\left(\frac{\omega - 2\pi k}{2}\right)\right) \quad (3.10)$$

$$= \frac{1}{2} G_1(\omega) \sum_{k=0}^1 X_1\left(\frac{\omega - 2\pi k}{2}\right) G_0\left(\frac{\omega - 2\pi k}{2}\right) \quad (3.11)$$

$$= X_2(\omega) G_1(\omega) \quad (3.12)$$

$$= Y_1(\omega) \quad (3.13)$$

which is equivalent to filtering a downsampled version of $X_1(\omega)G_0(\omega)$ \square

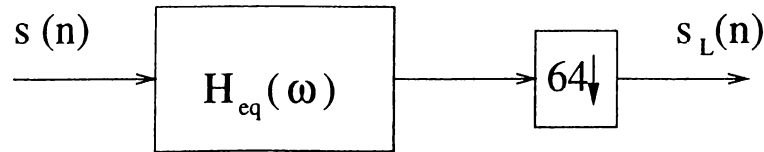


Figure 3.8: The equivalent block of the lowest, first, branch of the tree.

Using Result 1 the frequency characteristics of the iterated filter bank structure of Figure 3.3 can be determined. For example the lower-most branch of the tree is shown in Figure 3.8, and the equivalent filter is

$$H_{eq}(\omega) = H_0(\omega)H_0(2\omega)H_0(4\omega)H_0(8\omega)H_0(16\omega)H_0(32\omega). \quad (3.14)$$

The frequency response of the equivalent filter H_{eq} corresponding to the biorthogonal filterbank is plotted in Figure 3.9.

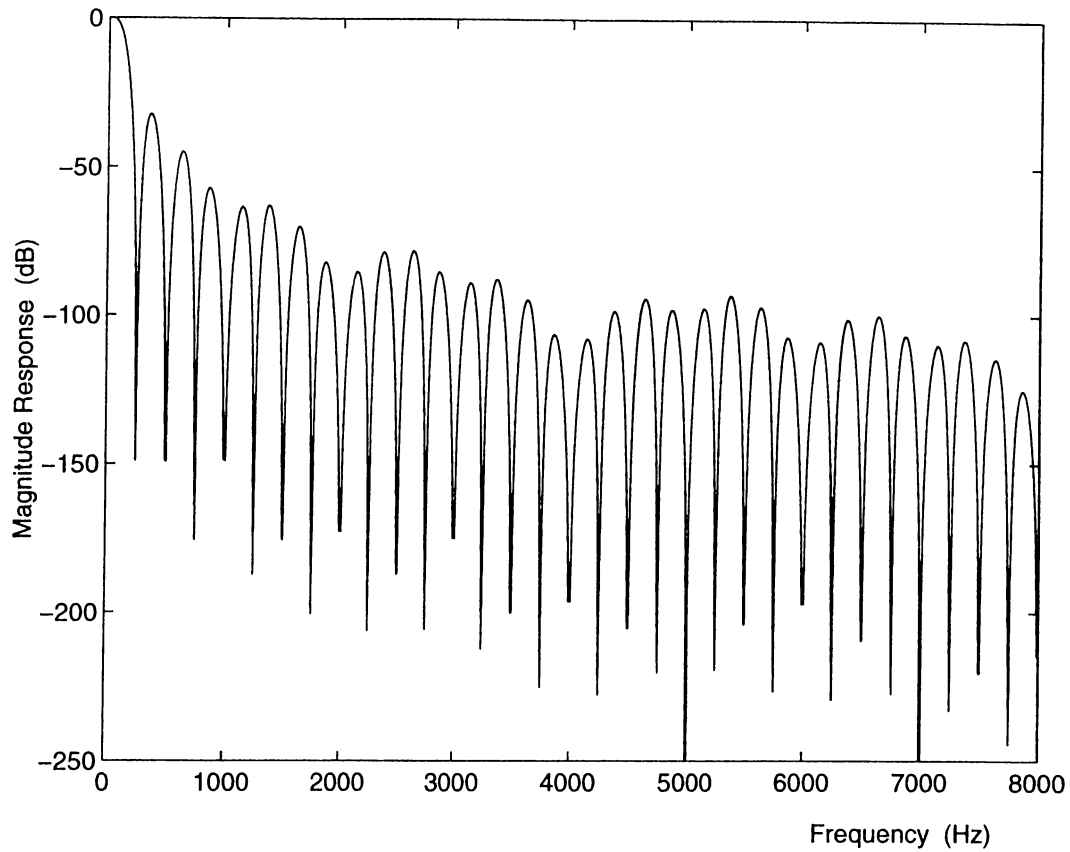


Figure 3.9: The frequency response of the equivalent filter H_{eq} .

3.2 Simulation Studies

A continuous density Hidden Markov Model (HMM) based speech recognition system with 5 states and 3 mixture densities is used in simulation studies. The recognition performances of the feature parameters are evaluated over two sets of data corresponding to artificial and real environments. The first set of experiments are carried out using the *TI-20* speech database of *TI-46 Speaker*

Dependent Isolated Word Corpus which is corrupted by various types of additive car noise. The second set of simulation experiments are carried out using a speaker dependent isolated word database which is recorded inside a car.

3.2.1 Experiments over the TI-20 Database

The *TI-20* vocabulary consists of ten English digits (0,1, 2, ..., 9), and ten control words (enter, erase, go, help, no, rubout, repeat, stop, start, yes) which are collected from 8 female and 8 male speakers. There are 26 utterances of each word from each speaker, where 10 designated as training tokens and 16 designated as testing tokens.

The noisy speech is obtained using two different car noise recordings, assuming that the noise is additive. The first noise recording is obtained inside a Volvo 340 on a rainy asphalt road by *the Institute for Perception-TNO, The Netherlands*. The second one is recorded inside a Skoda Favorit 135 L on an asphalt road at 90 km/hour speed. The filter bank structure of Figure 3.1 is applied to the speech signal in a tree structured manner as shown in Figure 3.3 (up to 6 levels) to achieve the subband frequency decomposition shown in Figure 3.4. This decomposition produces $L = 21$ subsignals. The window size is chosen as 48 msec (384 samples) with an overlap of 32 msec so that the subsignal with the smallest subband has 12 samples at 16 kHz sampling rate. The SUBCEP parameters are derived as in Equation (3.7) with the following root values,

$$\begin{aligned} \underline{P} &= [p_1 \ p_2 \ \cdots \ p_L] \\ &= [0.094 \ 0.281 \ 0.375 \ 0.375 \ \cdots \ 0.375]. \end{aligned} \quad (3.15)$$

We observed that the power spectral density of car noise is localized at low frequencies. Due to this reason low root-values are chosen for p_1 and p_2 which correspond to the subbands $[0, \pi/64]$ and $(\pi/64, \pi/32]$, respectively. This choice of p_l values deemphasized the effects of noisy bands. As it can be seen from the simulation studies this non-uniform weighting of the root values are very effective and provides a graceful degradation as the noise level increases.

The final feature vector is constructed from the SUBCEP parameters and

their time derivatives. The MELCEP parameters [13] are also extracted from the same signal by critical band windowing centered at the *mel*-scale frequencies. The time derivatives of the MELCEP parameters are added to the MELCEP feature vector as well.

The performance of the speech feature parameters are tested for both speaker dependent and speaker independent cases.

Table 3.1: The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with Volvo noise recording.

SNR (dB)	SUBCEP		MELCEP	
	orthogonal	biorthogonal	p_l (3.15)	uniform p_l
30	97.46	99.37	98.28	98.28
20	97.46	99.37	98.28	94.79
10	97.32	99.31	98.26	94.66
7	97.25	99.33	98.17	94.54
3	97.09	99.15	97.05	93.66
0	96.93	98.94	97.71	87.72
-3	95.97	98.11	96.58	70.11

In speaker dependent case, the models of the vocabulary is obtained from the training tokens of each speaker and evaluation is done with the testing tokens of the same speaker.

The average recognition rates of the 16 speakers for various SNR levels are given in Table 3.1 and Table 3.2 for Volvo and Skoda noise recordings, respectively. Each row represents the averaged recognition rate for the indicated SNR value, where the original (noise free) recording of the database has a 30 dB SNR and all of the recognition performances are evaluated subject to the training at 30 dB SNR level. The performances of SUBCEP representation with orthogonal and biorthogonal filter banks are given in the second and third columns, respectively. In the fourth column, the performance of the MELCEP representation is given, while in the Volvo noise recording the fifth column shows the performance of the uniform p_l values for MELCEP representation. The

performance of the uniform p_l values are not reported with the Skoda noise recording, as the performance is poorer with the uniform p_l values.

Table 3.2: The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with Skoda noise recording.

SNR (dB)	SUBCEP		MELCEP
	orthogonal	biorthogonal	
30	98.41	99.41	98.41
20	98.35	99.41	98.31
10	98.17	99.29	98.09
7	97.99	99.06	98.01
3	97.17	98.15	97.25
0	95.65	95.55	91.99
-3	84.28	71.00	60.88

Table 3.3: The performance evaluation of speaker independent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with Volvo noise recording.

SNR (dB)	SUBCEP		MELCEP	
	orthogonal	biorthogonal	p_l (3.15)	uniform p_l
30	93.03	91.56	91.25	91.25
20	92.93	91.51	91.25	90.41
10	92.72	91.41	90.99	89.73
7	92.46	91.41	90.46	89.42
3	90.78	90.73	89.52	85.07
0	89.37	89.26	88.47	74.43
-3	86.07	87.59	85.75	51.65

In the evaluation of speaker independent simulations, the models of the vocabulary is obtained from the training tokens of 5 male and 5 female speakers at noise free conditions as in the speaker dependent case, and the evaluation is done with the testing tokens of the remaining 3 male and 3 female speakers. The performances of the SUBCEP representation for orthogonal and biorthogonal filter banks and MELCEP representation are given in the second, third and

fourth columns of the Table 3.3 and Table 3.4. Also the performance of MELCEP representation for uniform p_l values are given in the last column of Table 3.3 with only Volvo noise recording, in order to present the effect of choosing uniform p_l values.

Table 3.4: The performance evaluation of speaker independent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with Skoda noise recording.

SNR (dB)	SUBCEP		MELCEP
	orthogonal	biorthogonal	
30	92.35	91.99	89.84
20	92.19	91.88	89.79
10	91.15	91.04	88.79
7	89.99	90.78	87.74
3	88.21	88.06	80.88
0	82.92	81.98	59.72

The SUBCEP representation exhibits more robust performance than MELCEP representation in both set of simulation studies for speaker dependent and speaker independent cases. Especially for speaker independent case, the performance gain with the SUBCEP representation is considerably high.

Also, the performance results for uniformly chosen p_l values showed the effectiveness of the choice of p_l values as in (3.15).

3.2.2 Performance over the Data Set Recorded in a Car

The second set of simulation studies are carried out over the data set recorded inside a car (Mazda 626) with a single electret microphone. The vocabulary of this database consist of ten Turkish digits { 0 (sıfır : /sıfır/), 1 (bir : /bēr/), 2 (iki : /ēkē/), 3 (üç : /üch/), 4 (dört : /dört/), 5 (beş : /besh/), 6 (altı : /ältı/), 7 (yedi : /yedē/), 8 (sekiz : /sekēz/), 9 (dokuz : /dokuz/)}, and yes (evet : /evet/), no (hayır : /hāyır/) which are collected from 4 female

and 5 male speakers. The average 9 utterances of each word from each speaker are recorded as training tokens while the the car was idle, and the average 10 utterances of each word from each speaker are recorded as testing tokens with a 90 km/hour speed.

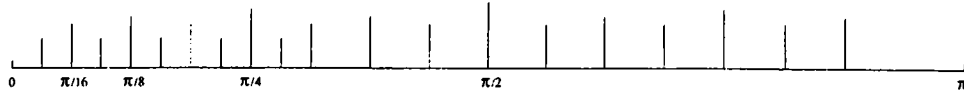


Figure 3.10: The subband decomposition of the speech signal for the car database.

The recording is digitized with a sampling rate of 8 kHz, and the window size is chosen as 48 msec with an overlap of 32 msec. In the subband decomposition a five level tree is used as shown in Figure 3.10, and the resulting frequency decomposition has $L = 20$ subbands. The recognition rates are obtained both with MELCEP and SUBCEP feature vectors by using the power values in 3.15. The SUBCEP representation achieved a recognition performance of 91.92% which is better than the recognition performance of MELCEP representation with 86.40%.

3.2.3 Conclusion

In this chapter, new speech feature parameters, SUBCEPs, are introduced. The SUBCEPs are based on wavelet analysis or equivalently the multirate subband analysis of the speech signal providing robust recognition performance by appropriately deemphasizing the frequency bands corrupted by noise. Furthermore SUBCEPs can be realized in a computationally efficient manner by employing fast wavelet analysis techniques.

It is experimentally observed that the SUBCEP representation produces better recognition rates for speaker dependent and independent isolated word recognition in the presence of car noise. SUBCEPs are promising candidates as feature parameters for large vocabulary recognition systems as well.

Chapter 4

ADAPTIVE NOISE CANCELING FOR ROBUST SPEECH RECOGNITION

As pointed out in Chapter 1, the goal of this thesis is to increase the robustness of a speech recognition system with respect to changes in the environment. Since the mismatches between training and testing environments lead to an important degradation in performance, it is necessary to improve the quality of the testing tokens or retraining of the system with the new environment. But, retraining of the system is not a reasonable solution if the environmental conditions are time varying.

It is well known that the speech signals in a public network telecommunication system are effected by impulsive noise which can be modeled as a symmetric α -stable random process [45]. Speech recognition over telephone channels has to be robust against the impulsive noise. In this chapter we mainly investigate some adaptive filtering techniques to eliminate the mismatch between training (noise-free training) and testing conditions of a speech recognition system for telephone channel conditions.

In Section 4.1, adaptive filtering techniques for non-Gaussian α -stable processes are introduced. In Section 4.2, the adaptive noise cancellation (ANC) system for speech enhancement is described. The enhanced speech signal is used in an isolated word recognition system, and simulation results are presented in Section 4.3.

4.1 Adaptive Filtering for non-Gaussian Stable Processes

In many signal processing applications the environmental noise is modeled as a Gaussian process. This assumption has been broadly accepted because of the Central Limit Theorem. However, a large class of physical observations exhibit non-Gaussian impulsive behavior, such as low frequency atmospheric noise, many types of man-made noise and telephone channel noise [45, 46]. There exists an important class of distributions known as α -stable distributions [47] which can be used to model this type of noise signals. These distributions have heavier tails than those of Gaussian distribution, and they exhibit sharp spikes or occasional bursts in their realizations. A random variable is called α -stable if its characteristic function has the following form:

$$\phi(t) = \exp\{iat - \gamma|t|^\alpha[1 + i\beta\text{sign}(t)\omega(t, \alpha)]\} \quad (4.1)$$

where

$$\omega(t, \alpha) = \begin{cases} \tan(\alpha\pi/2) & \text{for } \alpha \neq 1 \\ \frac{2}{\pi} \log |t| & \text{for } \alpha = 1. \end{cases} \quad (4.2)$$

and the parameters a , γ , α , and β satisfy the inequalities $-\infty < a < \infty$, $\gamma > 0$, $0 < \alpha \leq 2$, $-1 \leq \beta \leq 1$. Hence, a stable distribution is completely determined by the location parameter a , the scale parameter γ , the index of skewness β , and the characteristic exponent α .

There is no compact expression for the probability density function of these random variables except $\alpha = 1$ and 2 cases which correspond to the Cauchy and

Gaussian distributions, respectively. In Figure 4.1 the pdf's of α -stable random variables for $\alpha = 1$, $\alpha = 1.5$, $\alpha = 2$ are shown.

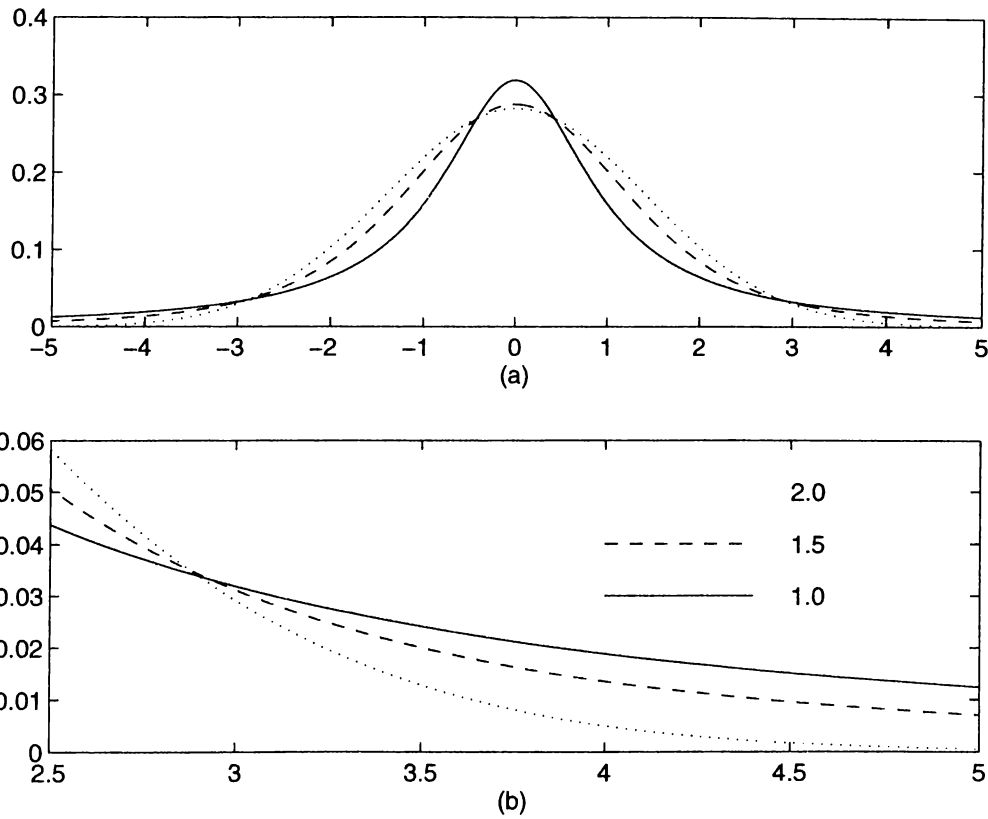


Figure 4.1: (a) Graphs of α -stable densities corresponding to the values $\alpha = 1.0$, $\alpha = 1.5$, and $\alpha = 2.0$, (b) detailed examination of the tails of the densities in (a).

Members of stable distributions also satisfy a generalized central limit theorem which states that if the sum of i.i.d. random variables converges then the limit distribution is a stable one. If individual distributions are of finite variance then the limit distribution is Gaussian. Tails of this type of distributions are characterized with the α parameter ($0 < \alpha \leq 2$) which is called as the characteristic exponent (α values close to 0 indicates impulsive nature and α values close to 2 indicates a more Gaussian type of behavior). With the Gaussian assumption, signals could be treated in a Hilbert space framework which would

allow the use of L_2 (or ℓ_2) norm in various optimization criteria. Whereas, the linear vector space generated by α -stable distributions is a Banach space when $(1 \leq \alpha < 2)$. In the linear space of stable processes only ℓ_p -norms exists for $p \leq \alpha$, hence, ℓ_2 norm cannot be used with an α -stable processes. Modeling α -stable processes under a Gaussian assumption leads to unacceptable results which are also experimentally verified in [47].

Let $\{u(n)\}$ be a family of i.i.d. α -stable random variables. Then

$$x(n) = \sum_{i=-\infty}^{\infty} a_i u(n-i) \quad (4.3)$$

defines a stationary α -stable random process if the coefficients, a_i 's are absolutely summable, $\sum_i |a_i| < \infty$ when $\alpha \geq 1$ [47]. The random process is also called a linear stable process with a moving average (MA) representation. Finite order autoregressive (AR), MA and autoregressive moving-average (ARMA) stable processes are other widely used examples of linear stable processes.

4.1.1 Adaptive Filtering for α -stable Processes

Recently, we introduced new algorithms for adaptive filtering under additive α -stable noise with finite mean corresponding to $1 \leq \alpha < 2$ [48, 49]. The objective for a general filtering application is to find an FIR filter of length N , \underline{w} , that relates the input, $x(n)$ to the desired signal $d(n)$:

$$\hat{d}(k) = \underline{x}(k)' \underline{w} \quad (4.4)$$

where $\hat{d}(k)$ is the estimate of the desired signal at time instant k , and

$$\underline{x}(k) = [x(k) \ x(k-1) \ \cdots \ x(k-N+1)]'. \quad (4.5)$$

Commonly used adaptive filtering algorithms utilize the Hilbert space framework. This allows the use of least squares cost function whose solution can be found either exactly as in Recursive Least Squares (RLS) algorithms or approximated by Least-Mean-Squares (LMS) type methods [50, 51]. However, in the existence of α -stable processes least squares cost function cannot be defined because the

variance of the error is not finite. Hence a new cost function other than least squares should be used.

In this work, we consider an adaptation algorithm for an FIR filter of length N . The problem is to adaptively update the tap weights of the FIR filter, \underline{w} , such that given an input sequence $x(n)$, the output of the filter is close to the desired response $d(n)$, both of which is assumed to be α -stable. In this case, it is appropriate to minimize the dispersion of the error function [47].

This adaptation problem can be solved asymptotically by using the stochastic gradient method with the motivation of the LMS algorithm [51]. Such an algorithm, least mean p -norm (LMP) algorithm, is proposed in [47]. This algorithm is a generalization of instantaneous gradient descent algorithm to α -stable processes, where the gradient of the p -norm of the error,

$$\begin{aligned} J &= E[|e(k)|^p] \\ &= E[|d(k) - \underline{w}(k)' \underline{x}(k)|^p], \quad 0 < p < \alpha \end{aligned} \quad (4.6)$$

is used instead of the commonly used ℓ_2 norm, and the tap weights, \underline{w} , are adapted at time step $k + 1$ as follows:

$$\underline{w}(k + 1) = \underline{w}(k) + \mu |e(k)|^{p-1} \text{sgn}(e(k)) \underline{x}(k) \quad (4.7)$$

where μ is the step size which should be appropriately determined. Note that, for $p = \alpha = 2$ the LMP algorithm reduces to the well-known LMS algorithm [51]. When p is chosen as 1, the LMP algorithm is called the Least Mean Absolute Deviation (LMAD) algorithm [47]:

$$\underline{w}(k + 1) = \underline{w}(k) + \mu \text{sgn}(e(k)) \underline{x}(k) \quad (4.8)$$

which is also known as the signed-LMS algorithm.

We introduced two normalized adaptation algorithms with the motivation of the Normalized-LMS algorithm in [48]. The first one, Normalized Least Mean p -Norm (NLMP) algorithm, uses the following update:

$$\underline{w}(k + 1) = \underline{w}(k) + \beta \frac{|e(k)|^{p-1} \text{sgn}(e(k))}{\|\underline{x}(k)\|_p^p + \lambda} \underline{x}(k) \quad (4.9)$$

where $\beta, \lambda > 0$ are appropriately chosen update parameters. In (4.9) normalization is obtained by dividing the update term by the p -norm of the input vector, $\underline{x}(k)$. The regularization parameter, λ , is used to avoid excessively large updates in case of an occasionally small inputs. For $p = 2$, NLMP (4.9) reduces to the Normalized-LMS algorithm [51].

The second algorithm, Normalized Least Mean Absolute Deviation (NL-MAD), corresponds to the case of $p = 1$ in (4.9) with the following time update:

$$\underline{w}(k+1) = \underline{w}(k) + \beta \frac{\text{sgn}(e(k))}{\|\underline{x}(k)\|_1 + \lambda} \underline{x}(k). \quad (4.10)$$

This adaptation scheme is especially useful when the characteristic exponent, α , of the α -stable random process either is unknown or varying in time. Among the stable distributions the heaviest tail occur for the Cauchy distribution, $\alpha = 1$. By selecting $p = 1$ the update term is guaranteed to have a finite magnitude for all $1 < \alpha \leq 2$. Due to the above reasons NLMAD is a safe choice for the adaptation.

The performances of these normalized algorithms are found to be superior than that of LMS and LMAD algorithms in simulation studies which are presented in Appendix B. Based on the experience gained in the simulation studies, it is observed that a safe choice of p value is 1 in the case of imprecise knowledge of α . This corresponds to the use of NLMAD algorithm in such cases.

Knowing that the telephone channel noise can be characterized by stable distributions with characteristic exponent close to 2 [45], ℓ_p -norm normalized algorithms can be used in adaptive noise canceling systems to reduce the mismatch between training and testing environments. In the next section, the new adaptive algorithms for α -stable processes will be combined with the adaptive noise canceler to construct a more robust speech recognition system.

4.2 Adaptive Noise Canceling for Speech Recognition

In this section, we consider speech enhancement techniques that improve the quality of speech by first estimating speech modeling parameters, and then resynthesizing the enhanced speech with the aid of adaptive filtering. The general technique of Adaptive Noise Canceling (ANC) has been applied successfully to a number of problems that include speech restoration, elimination of periodic interference, echo cancelation, and adaptive antenna theory.

Adaptive noise canceling refers to adaptive enhancement algorithms based on the availability of a primary input source and a secondary reference source. The primary source is assumed to be the speech plus additive noise

$$y(n) = s(n) + r_1(n) \quad (4.11)$$

where $s(n)$ is the speech signal and $r_1(n)$ is the noise, and they are assumed to be the realizations of stochastic processes, \underline{y} , \underline{s} and \underline{r}_1 . The secondary reference sequence is denoted by $r_2(n)$, which is a realization of stochastic process \underline{r}_2 that may be correlated with the additive noise \underline{r}_1 but not the speech signal \underline{s} .

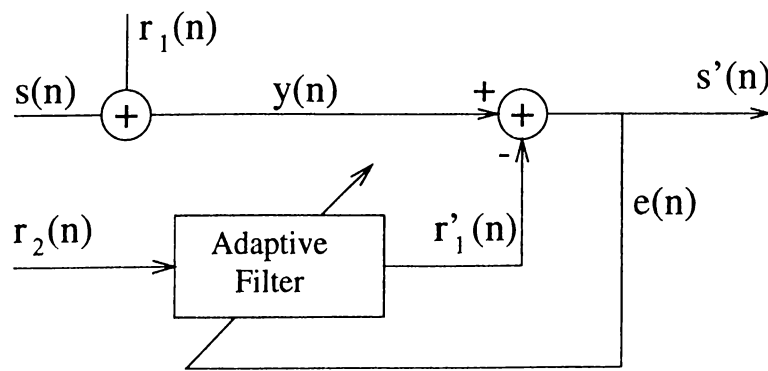


Figure 4.2: Adaptive noise canceler.

The adaptive noise canceler consists of an adaptive filter that acts on the reference signal $r_2(n)$ to produce an estimate of the noise $r_1(n)$, which is then

subtracted from the primary input $y(n)$. The output of the canceler is used in the adaptation of the coefficients of the adaptive filter, as shown in Figure 4.2.

In general, adaptive noise canceling can only be employed when a reference channel is available. But in most of the speech enhancement applications, a reference noise channel is not available. However, it is not so difficult to obtain a speech reference channel for some classes of speech. Due to the quasi-periodic nature of the voiced speech, a reference speech signal can be formed by delaying the primary sequence one or two pitch periods. This reference speech signal can then be used in the adaptive filtering in order to estimate the uncorrupted speech signal. This represents the basis for the ANC technique proposed by Sambur [52, 53]. Let us write the primary noisy speech signal $y_1(n)$,

$$y_1(n) = s(n) + r(n) \quad (4.12)$$

where $s(n)$ is the desired speech signal and $r(n)$ is the additive noise. The reference speech signal, $y_2(n)$, can be formed as the delayed version of the primary signal $y_1(n)$, with a delay of pitch period T_0 ,

$$y_2(n) = y_1(n - T_0) = s(n - T_0) + r(n - T_0). \quad (4.13)$$

Since the speech signal can be considered to be periodic, $s(n) \simeq s(n - T_0)$, and

$$y_2(n) \simeq s(n) + r(n - T_0). \quad (4.14)$$

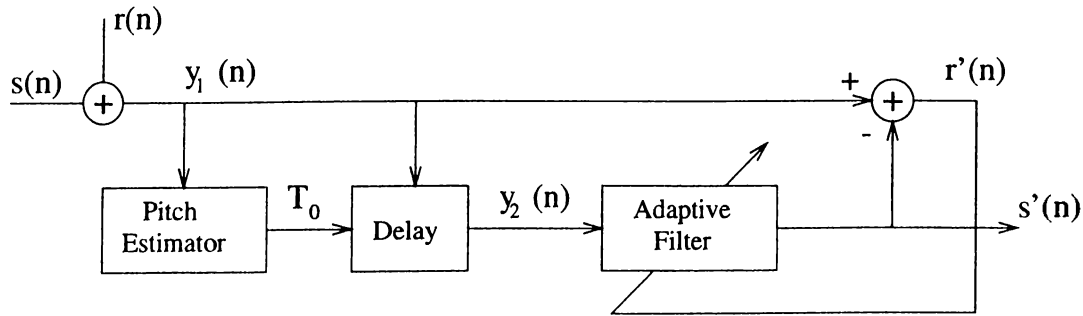


Figure 4.3: Single channel adaptive noise canceler.

The delayed speech signal $s(n - T_0)$ can be assumed to be highly correlated with the original speech signal $s(n)$ when T_0 is a multiple of the pitch period. Also,

the delayed noise $r(n-T_0)$ and the original noise $r(n)$ have a little correlation with the speech $s(n)$. The adaptive noise canceling structure for speech enhancement is shown in Figure 4.3. Since $y_1(n)$ and $y_2(n)$ are the main inputs to the ANC structure, the noise $r'(n)$ is primarily estimated. Consequently the restored speech signal is available at the output of the adaptive filter.

The pitch period was estimated using cepstral methods with frame processing [54]. Hence, the adaptation is performed only during the voiced segments of the speech signal and the adaptive filter is kept constant in the filtering process of unvoiced speech signal which does not have a periodic characteristics.

Telephone channel noise is considered to be additive, stationary over short periods (typically 1 second) and distinctly non-Gaussian [45]. The characteristic exponent, α , for telephone channel noise is reported as $\alpha = 1.95$ in [45]. Therefore, adaptive filtering techniques for non-Gaussian stable processes are appropriate to use in the adaptive noise canceling structure. In our simulation studies, the noisy speech signal is obtained by adding α -stable noise to the original speech signal at various SNR levels. The enhanced speech is obtained with the adaptive filtering structure using the LMAD algorithm. Considering the single channel adaptive filtering structure in Figure 4.3, adaptive filtering formulation can be written as follows,

$$s'(n) = \underline{\mathbf{w}}(n)' \underline{\mathbf{y}}_2(n) \quad (4.15)$$

and

$$\underline{\mathbf{w}}(n+1) = \begin{cases} \underline{\mathbf{w}}(n) + \mu \frac{\text{sgn}(r'(n))}{\|\underline{\mathbf{y}}_2(n)\|_1} \underline{\mathbf{y}}_2(n) & \text{for voiced speech} \\ \underline{\mathbf{w}}(n) & \text{for unvoiced speech} \end{cases} \quad (4.16)$$

where the $\underline{\mathbf{y}}_2(n) = [y_2(n)y_2(n-1) \cdots y_2(n-N+1)]'$ is the input speech vector, $\underline{\mathbf{w}}(n)$ is the N -th order tap weight vector, $s'(n)$ is the enhanced speech signal, $r'(n)$ is the enhanced noise, and μ is the adaptation parameter.

Next section covers the simulation studies of the speech recognition structure with the above adaptive filtering structure. This structure is used as a front-end block for a speech recognition system.

4.3 Simulation Studies

Simulation studies are carried out using the *TI-20* speech database of *TI-46 Speaker Dependent Isolated Word Corpus* which is corrupted by additive α -stable noise. As previously described in Section 3.2.1, *TI-20* vocabulary consists of twenty isolated words which are collected from 8 female and 8 male speakers. The noise-free training tokens are filtered with the ANC structure and the filtered utterances are used in the training phase of the recognition system. Testing tokens are obtained at different SNR levels by adding α -stable noise to the original testing tokens. The performance of the recognition system is obtained by the enhanced testing tokens of *TI-20* database.

A continuous density Hidden Markov Model (HMM) based speech recognition system with 5 states and 3 mixture densities is used in simulation studies. The adaptive filtering is carried out with a 20-th order NLMD structure, and the adaptation parameter is chosen as, $\mu = 0.1$. The recognition rates are evaluated using the feature parameters that are introduced in Chapter 3. Subband cepstral parameters, SUBCEPs, are both extracted with biorthogonal and orthogonal filter bank structures, and the commonly used *mel*-scale cepstral parameters, MELCEPs, are used in the performance evaluations.

The average recognition rates of the 16 speakers for various SNR levels are given in Table 4.1, Table 4.2 and Table 4.3 for characteristic exponent values $\alpha = 1.95$, $\alpha = 1.5$ and $\alpha = 2.0$, respectively. Each row represents the averaged recognition rate for the indicated SNR value, where the filtered original (noise free) recording of the database has a 30 dB SNR and all of the recognition rates are evaluated subject to the training at 30 dB SNR level. The performances of SUBCEP representation with orthogonal filter bank are given in the second and third columns for enhanced and noisy speech, respectively. Similarly, the performances of the biorthogonal filter bank are given in the forth and fifth columns. In the last two columns, the performance of the MELCEP representation is given in a similar way.

The front-end processing with the adaptive noise canceling structure exhibits better performance for all feature parameter sets. In particular, SUBCEP

parameters extracted with the biorthogonal filter bank outperforms the other parameter sets. When there is no speech enhancement, the performance of the recognition system degrades sharply below 14dB SNR.

Table 4.1: The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with α -stable noise ($\alpha = 1.95$).

SNR (dB)	SUBCEP				MELCEP	
	orthogonal		biorthogonal		ANC	noisy
	ANC	noisy	ANC	noisy		
30	96.75	98.41	97.68	99.41	97.28	98.31
20	96.46	98.13	97.34	99.09	96.83	98.03
17	96.08	97.82	96.97	98.70	96.46	97.40
14	93.80	89.57	95.89	95.38	94.75	91.50
12	92.15	77.43	94.82	89.91	92.54	84.69
11	89.43	55.55	93.53	77.78	89.98	72.61
10	84.79	36.00	91.81	61.11	90.75	55.25
7	55.41	-	76.60	-	76.00	-

Table 4.2: The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with α -stable noise ($\alpha = 1.5$).

SNR (dB)	SUBCEP				MELCEP	
	orthogonal		biorthogonal		ANC	noisy
	ANC	noisy	ANC	noisy		
30	96.79	98.41	97.70	99.41	96.42	98.31
20	96.42	98.09	97.26	99.00	95.76	97.52
17	95.75	97.48	96.87	98.33	95.54	94.81
14	88.86	77.74	91.89	89.67	86.11	59.84
12	84.55	57.10	88.69	78.37	81.38	-
11	77.16	-	83.35	60.58	74.59	-

Table 4.3: The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and MELCEP representations for various SNR levels with α -stable noise ($\alpha = 2.0$).

SNR (dB)	SUBCEP				MELCEP	
	orthogonal		biorthogonal		ANC	noisy
	ANC	noisy	ANC	noisy		
30	96.77	98.41	97.68	99.41	96.46	98.31
20	94.67	93.45	96.16	96.99	94.67	94.10
17	88.43	50.69	93.39	74.66	89.73	70.72
15	79.10	-	89.08	-	81.72	-
14	53.19	-	74.81	-	63.52	-

Chapter 5

WAVELET ANALYSIS FOR ENDPOINT DETECTION OF ISOLATED UTTERANCES

An important problem in speech recognition is to determine the boundaries of the speech utterances or words. The goal of the end-point detection is to separate the acoustic events of interest (speech to be processed) in a continuously recorded signal from other parts of the signal (background noise). Endpoint detection of an utterance is necessary and an important problem in speech recognition. Precise boundary detection of utterances improves the performance of speech recognition systems. In this chapter, a new distance measure for endpoint detection is introduced. The measure is based on the subband energy parameters obtained via the wavelet analysis of the speech signal. Its performance is tested in the presence of car noise.

The need for the end-point detection occurs in many applications in telecommunications. For example, a technique called time-assignment speech interpolation (TASI) is often used in analog multi channel transmission systems to take advantage of the channel idle time. An unused channel is only assigned

when speech is detected, in order to increase the channel capacity. The channel capacity gain with TASI is around a factor of 2.5.

A critical question in the end-point detection process is how can we achieve robustness against environmental noise? The robustness is directly related to the distance measure which is used in the separation of acoustic events. The energy measure and the zero-crossing rate are the widely used distances in endpoint detection algorithms [55]. They perform fairly well at high SNR levels, however their performance degrades drastically in many practical cases, such as car noise which is usually concentrated at low frequency bands.

In Section 5.1 the new distance measure based on the wavelet analysis is introduced, and in Section 5.2 the performance of the new distance measure is evaluated and compared with the widely used energy measure in the presence of car noise. Also an end-point detection algorithm is presented based on the new distance measure.

5.1 A New Distance Measure based on Wavelet Analysis

In this chapter, we use the wavelet analysis to tract the variations of subband energy levels. As described in Chapter 3, the wavelet analysis associated with a subband decomposition filter bank provides a fast and computationally efficient structure for decomposing the frequency domain along with the temporal information. The basic building block of a wavelet transform is realized by a subband decomposition filter bank shown in Figure 3.1. As described in Chapter 3, the filter bank structure consists of a lowpass and a highpass filter, and downsampling units. The passbands of the low and high pass filters are $[0, \pi/2]$ and $[\pi/2, \pi]$, respectively, so that the frequency domain is divided into two halfbands. In the subband decomposition filter bank structure the input signal is divided into two sub-signals containing the low and high frequency components of the original signal and then they are passed to the downsampling units which drop every other sample of their inputs reducing the sampling rate

by a factor of two. Each of the subsignals can be further decomposed into two new subsignals using the same filter bank once again, and this analysis procedure can be repeated until the desired frequency domain decomposition is obtained. Resulting temporal subsignals constitute the so-called wavepacket representation uniquely characterizing the original signal [40]. In this case the frequency domain is partitioned into nonuniform regions using a basic half-band filter bank in a tree structured manner as in Chapter 3. As a result, a family of subband signals are obtained. The resulting frequency bands are very similar to the commonly used *mel*-scale division [36].

A subband decomposition filter bank corresponding to a biorthogonal wavelet transform [41] is used in simulation studies. In our simulation studies, the speech signal, $s(n)$, is decomposed into L subsignals, $\{s_l(n)\}_{l=1}^L$ corresponding to the frequency domain decomposition shown in Figure 5.1. The same decomposition is used in the speech feature extraction in Chapter 3.

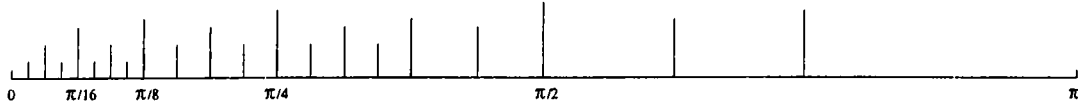


Figure 5.1: The subband decomposition of the speech signal.

For each subsignal, an energy parameter E_l^k for the k -th speech frame is defined over a time window of T_w seconds with an overlap of T_o seconds,

$$E_l^k = \frac{1}{N_l} \sum_{n=1}^{N_l} (s_l(n))^2, \quad l = 1, 2, \dots, L \quad (5.1)$$

where N_l is the number of samples in the l -th band. These energy parameters are used as the features for the new distance measure.

The new distance measure is defined as follows,

$$D_k = 10 \log \left[\frac{1}{L} \sum_{l=1}^L \frac{(E_l^k - \mu_l)^2}{\sigma_l^2} \right] \quad (5.2)$$

where μ_l and σ_l are the mean and variance of the background noise at the l -th band, respectively. The mean, μ_l and variance σ_l of the background noise are

estimated, a priori, from the utterance-free segments,

$$\mu_l = \frac{1}{N} \sum_{k=1}^N E_l^k, \quad (5.3)$$

$$\sigma_l^2 = \frac{1}{N} \sum_{k=1}^N (E_l^k - \mu_l)^2, \quad l = 1, 2, \dots, L \quad (5.4)$$

where the index k runs over background noise frames. The variability of the new measure D_k is decreased, as it is based on the ratio of the variances as suggested by Taguchi in the data analysis context [56]. Therefore, it is possible to set robust thresholds in the endpoint detection process.

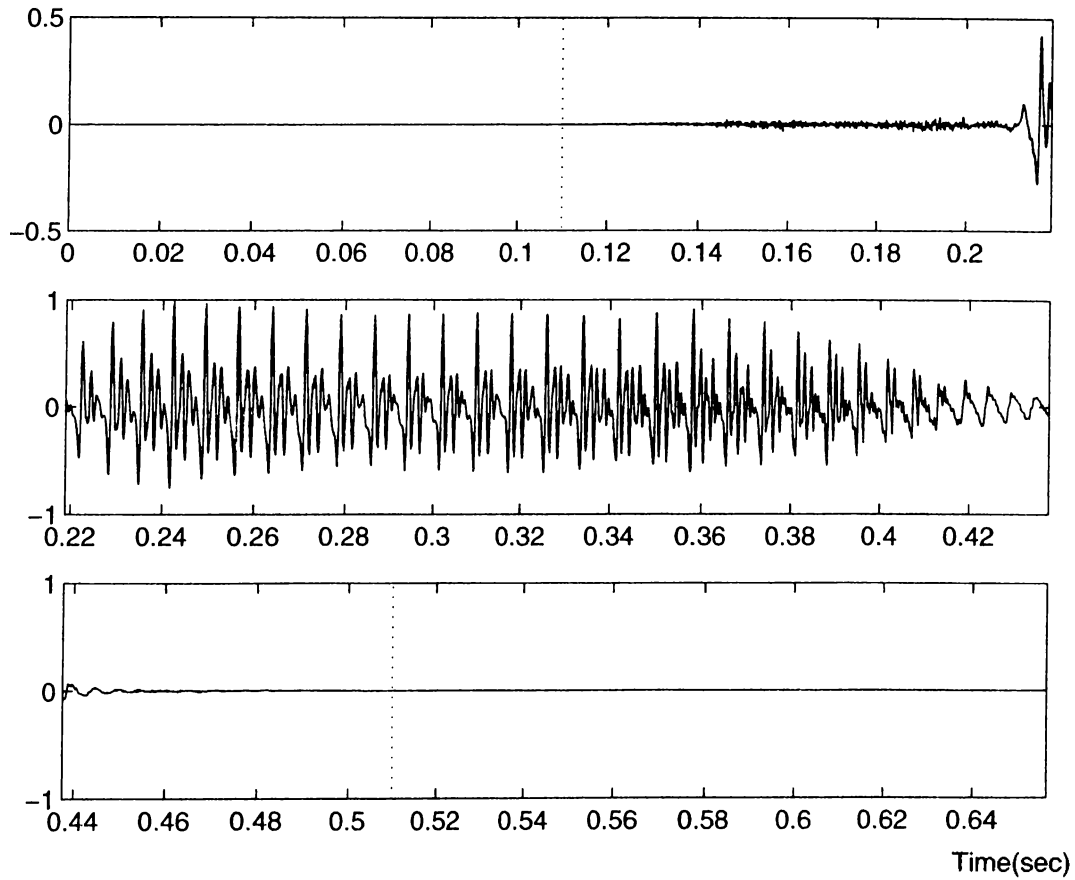


Figure 5.2: The noise free utterance “four” is plotted, where the waveform begins with the weak fricative /f/ on the top plot, and the utterance ends with the fricative /r/ on the bottom plot.

5.2 Simulation Examples

The performance of the new end-point detection measure is evaluated in situations where it is difficult to locate either the beginning or the end of an utterance. The broad categories of problems encountered can be classified as

- (i) weak fricatives (/f, th, h/) at the beginning or end of an utterance,
- (ii) weak plosive bursts (/p, t, k/), and
- (iii) final nasals.

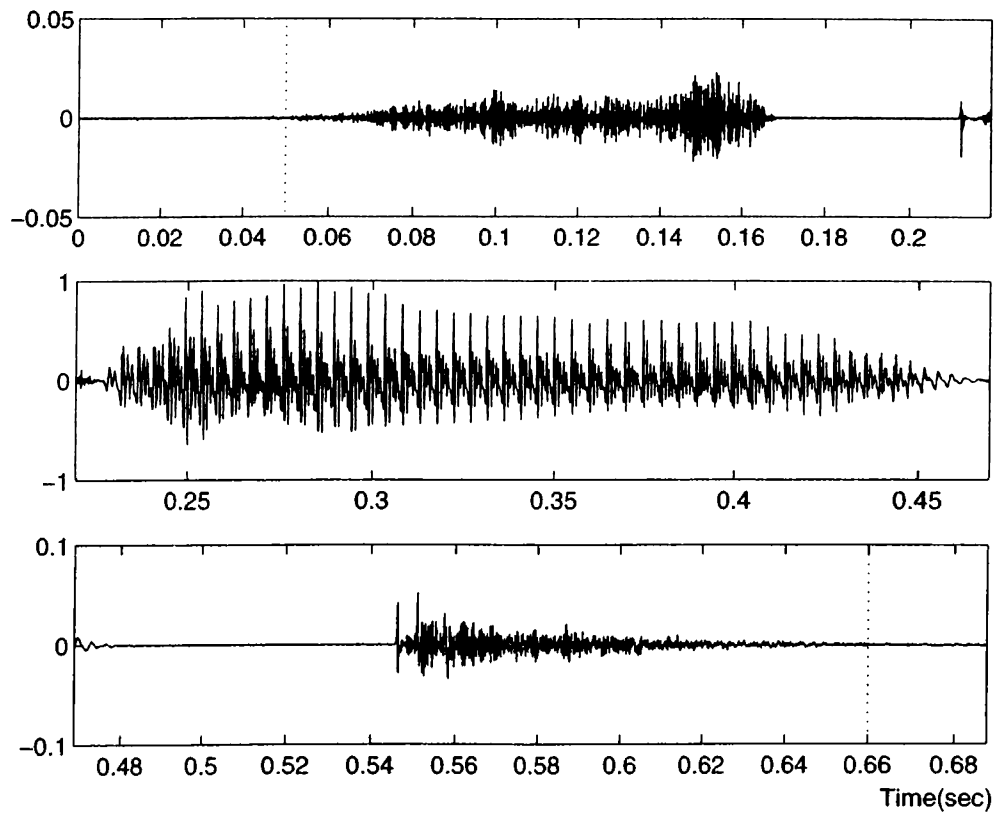


Figure 5.3: The noise free utterance "start" is plotted, where the fricative /s/ begins on the top plot, and the utterance ends with the plosive fricative /t/ on the bottom plot.

Figure 5.2 shows the waveform of the utterance "four". This utterance begins with the weak fricative /f/. The speech energy at the beginning of the utterance

is not radically higher than the background energy, however weak fricative /f/ begins at time instant $t = 0.085$ sec, approximately.

Figure 5.3 shows another example in which the waveform of the utterance “start” is plotted. This utterance begins with the fricative /s/, and ends with the plosive fricative /t/. These fricatives are slightly apart at the beginning and end of the utterance. Hence in the presence of environmental noise, it is hard to locate the end-points.

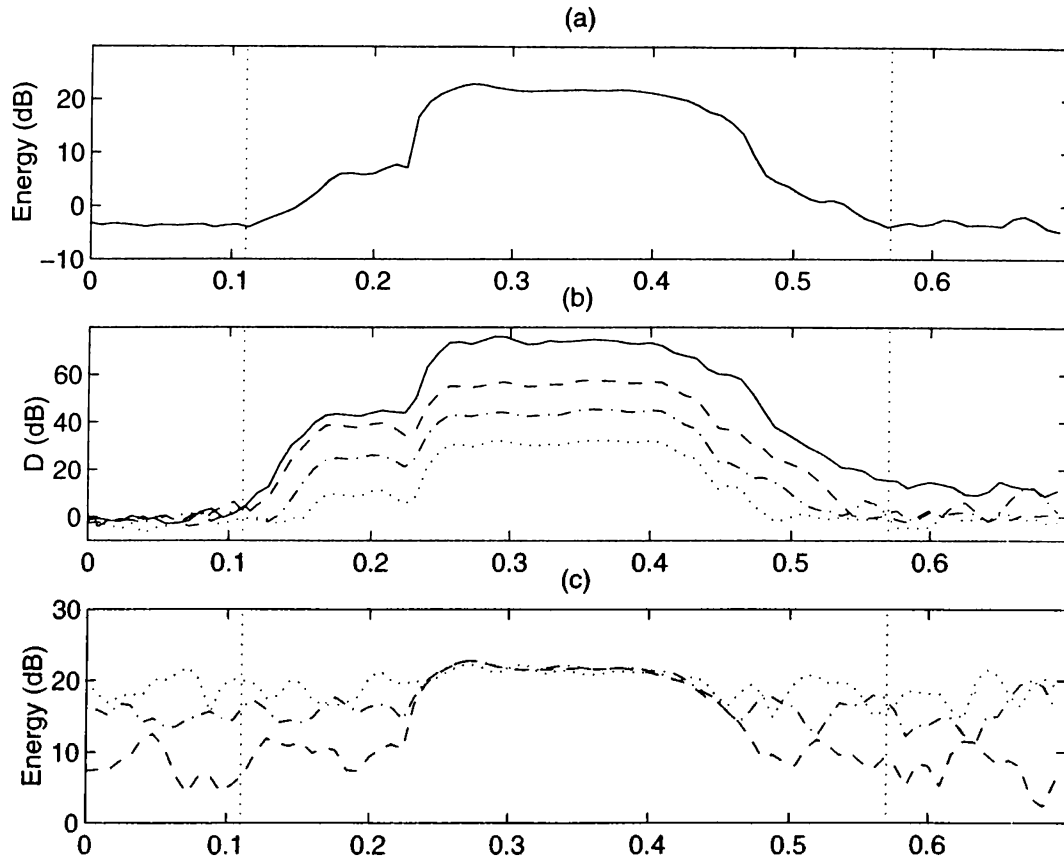


Figure 5.4: The energy measure of the noise free utterance “four” is plotted in (a). The new distance measure and the energy measure for different SNR levels are plotted in (b) and (c), respectively. The solid (dashed) [dash-dotted] {dotted} line corresponds to 30 dB (15 dB) [10 dB] {8 dB} noise level.

The performance of the new distance measure D_k is compared to the energy measure in the presence of car noise. The speech signal is sampled at 16 kHz

and a 16 msec window with an overlap of 8 msec is used. In Figure 5.4(a) the widely used energy measure of the noise-free utterance “four” is plotted. Figure 5.4(b) and 5.4(c) depict the new distance measure D_k and the widely used energy measure for various SNR levels in the presence of additive car noise, respectively. When the SNR gets lower, it is almost impossible to locate the weak fricative /f/ at the beginning.

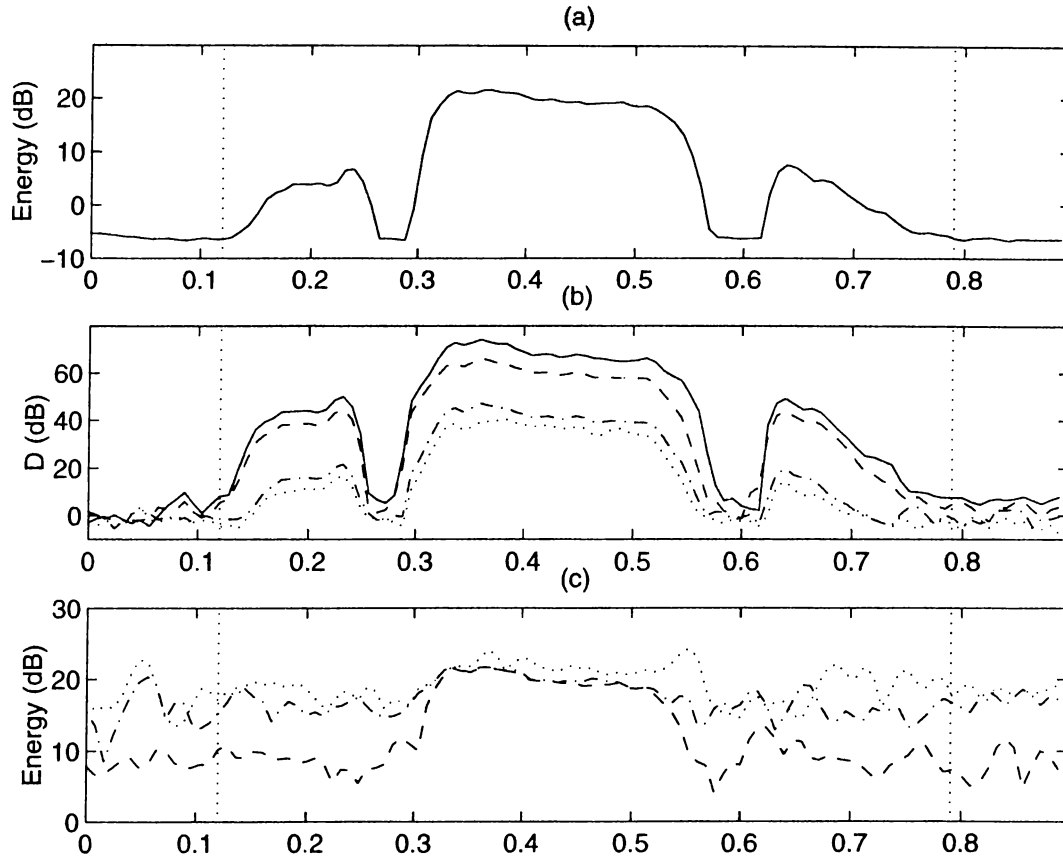


Figure 5.5: The energy measure of the noise free utterance “start” is plotted in (a). The new distance measure and the energy measure for different SNR levels are plotted in (b) and (c), respectively. The solid (dashed) [dash-dotted] {dotted} line corresponds to 30 dB (15 dB) [10 dB] {8 dB} noise level.

In Figure 5.5(a) the energy measure of the noise-free utterance “start” is plotted. Similarly, Figure 5.5(b) and Figure 5.5(c) depict the new distance

measure D_k and the widely used energy measure for various SNR levels.

The original utterances that are recorded inside a car (Mazda 626) traveling 90 km/hour are used in the performance evaluations of the new distance measure. As described in Chapter 3 the recording is done with an electret microphone and digitized with 8 kHz sampling rate. In the subband decomposition a five level tree is used as shown in Figure 3.10.

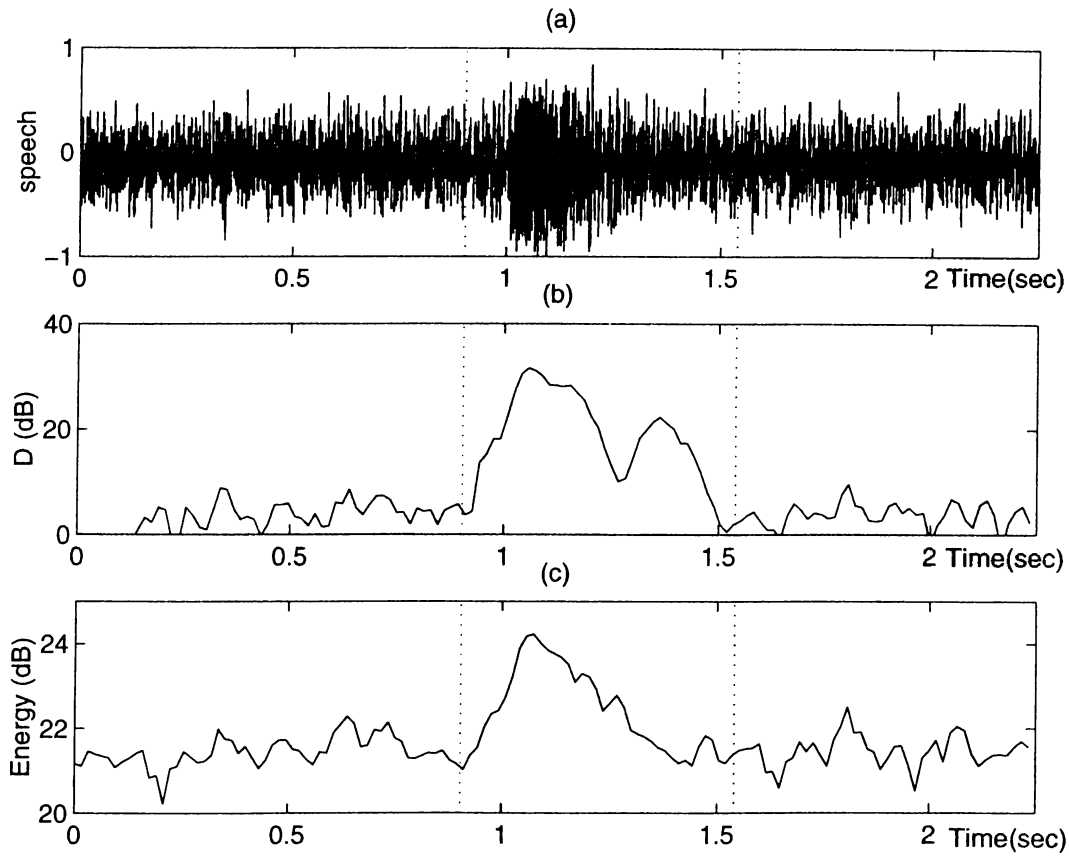


Figure 5.6: The recording of the Turkish word “hayır :/häyır/” (no) inside a car is plotted in (a), the new distance measure and the energy measure are plotted in (b) and (c), respectively.

In Figure 5.6(a) the word “hayır” as recorded inside the car is plotted. Figure 5.6(b) and 5.6(c) depict the new distance measure D_k and the widely used energy measure, respectively. The new distance measure D_k clearly determines the weak

fricative /h/ at the beginning of the utterance. Hence, it is experimentally observed that the new distance measure is more robust than the widely used energy measure.

In Figure 5.7(a) the new distance measure D_k is plotted for a continuous recording inside a car. The end-points of the isolated words can be clearly detected. In Figure 5.7(b) the histogram of the new distance measure is depicted for the set of data shown in Figure 5.7(a). By examining the histogram, a good threshold that separates the background noise and speech is found as $D = 12$ dB. In the next subsection, a new automatic end-point detection algorithm is presented based on the new distance measure D_k .

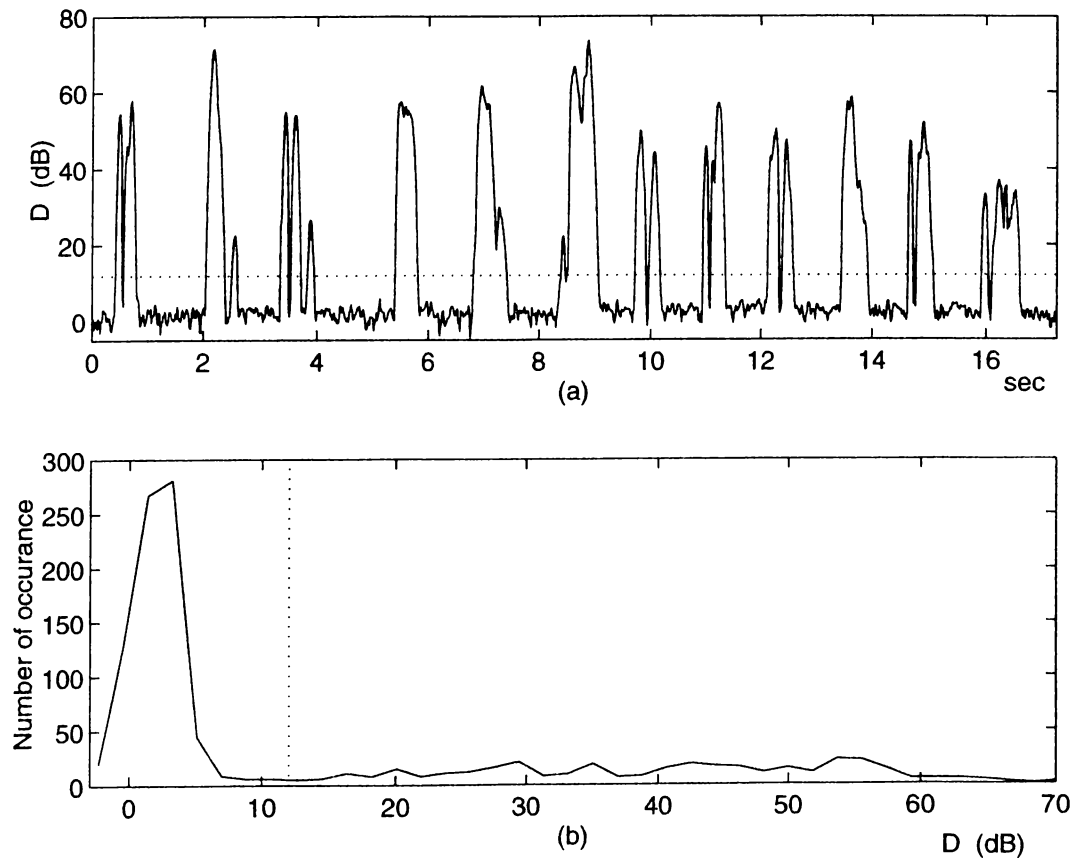


Figure 5.7: (a) The new measure D_k for a utterance recorded inside a car, (b) the histogram of the new measure D_k .

5.2.1 End-point Detection Algorithm

In this section a new end-point detection algorithm using the new distance measure D_k is presented. Figure 5.8 shows the flowchart of the end-point detection algorithm. It is assumed that during the first 200 msec of the recording interval the speech is not present, so that the mean, m_l and the variance, σ_l values of the environmental noise can be estimated during this interval. In the flowchart, T_b represents the beginning threshold value, and T_l and T_h represent a lower-threshold value and a higher-threshold value, respectively. The lower and higher threshold values are determined in terms of the beginning threshold T_b as follows,

$$T_l = \frac{3}{4}T_b, \quad (5.5)$$

and

$$T_h = \frac{3}{2}T_b. \quad (5.6)$$

Other parameters in the algorithm are N_b , N_e , and N_f . The parameters N_b and N_e correspond to the index of the beginning and end frames of the word in consideration. For a given word the distance measure D_k fall below the threshold value T_b between the phonemes. The maximum number of frames that D_k stays below T_b is denoted by N_f .

The algorithm scans the utterance and it labels the frame in which the threshold T_b is first exceeded as N_b , unless the distance measure D_k falls below the threshold value T_l before it rises above the threshold value T_h . The ending frame N_e is determined when the distance measure D_k falls below the threshold value T_b for longer than N_f frames.

These threshold values are predetermined by examining the histogram of the distance measure D_k for the current application environment. For example under car noise environments, the beginning threshold is selected as $T_b = 12$ dB, and consequently the lower and the higher threshold values are chosen as $T_l = \frac{3}{4}T_b = 9$ dB and $T_h = \frac{3}{2}T_b = 18$ dB, respectively. The index N_f is chosen as 10.

Figure 5.9 shows an example of how the algorithm worked on a typical isolated word. The distance measure D_k is plotted for the Turkish word “*evet*” recorded inside the car. The end-points were correctly labeled and the final plosive /t/

was not missed by the algorithm.

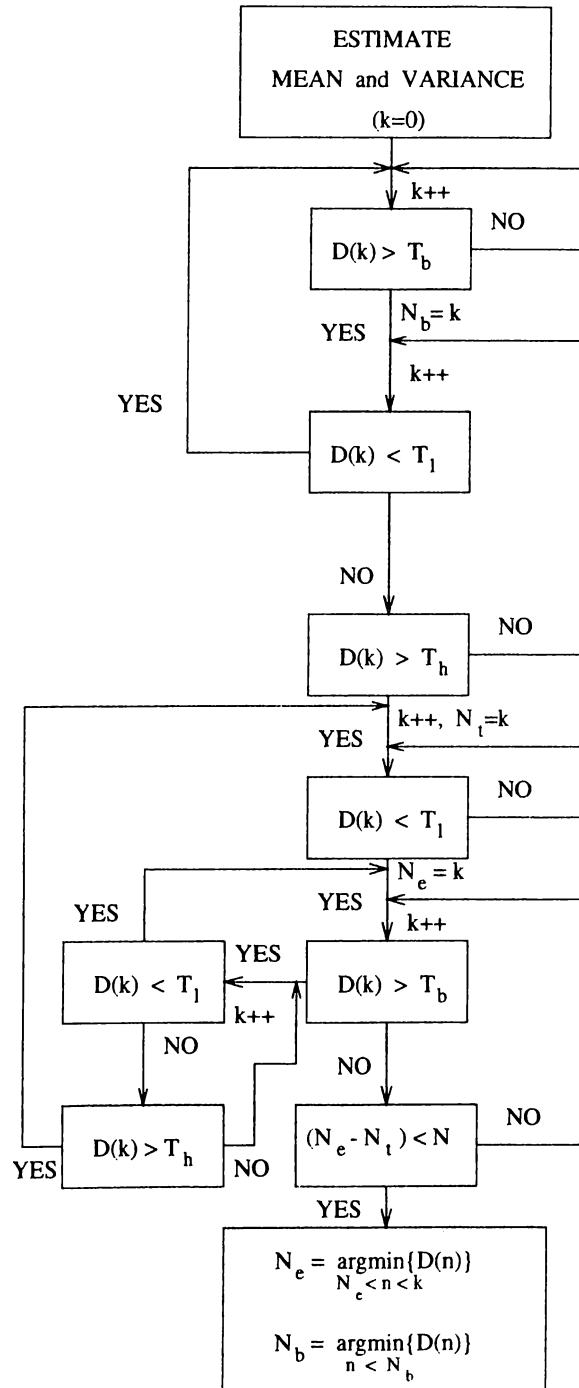


Figure 5.8: Flowchart for the endpoint algorithm.

5.2.2 Conclusion

The isolation of speech from noisy background is a challenging problem. In this chapter, a subband analysis based distance measure is introduced, and it produces better results than the widely used energy measure. Also, an end-point detection algorithm based on the new distance measure is presented. A similar end-point detection algorithm is developed for the industry¹, and it is implemented successfully in a prototype.

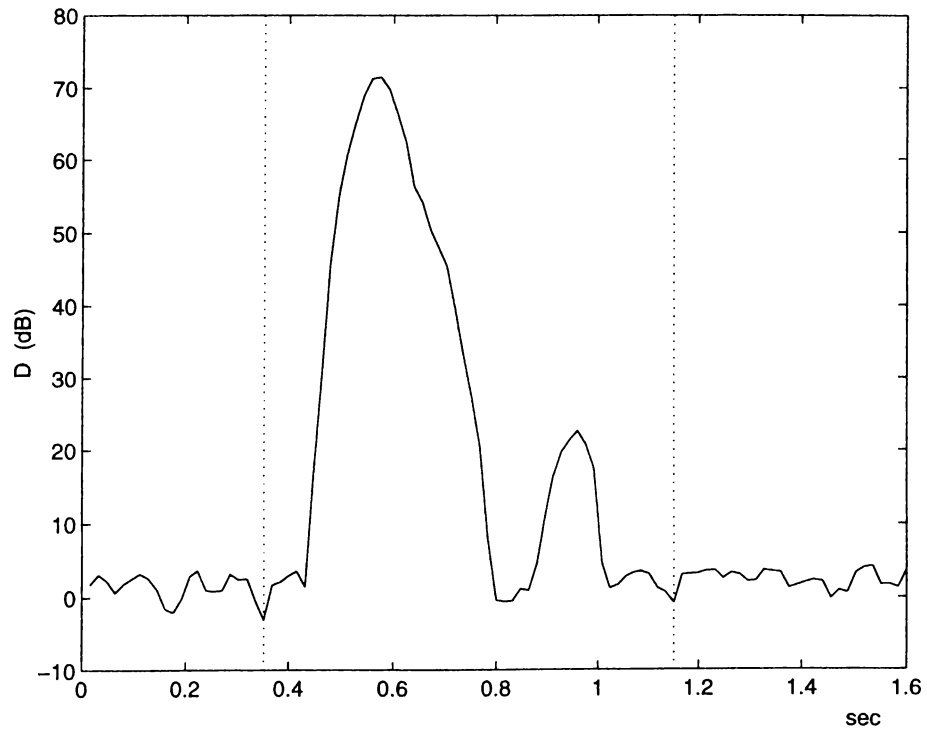


Figure 5.9: The distance measure D_k for the Turkish word “*evet*” and the end-point locations that are detected by the algorithm.

¹*Mobil Telefonlarda Sözcük Tanıma (Voice dialing for mobile telephones)* by BİLTEN-TÜBİTAK. This project was supported by ASELSAN (Military Electronics Industry Inc.), Turkey.

Chapter 6

CONCLUSION

In this thesis, new methods of feature extraction, end-point detection, and speech enhancement are developed for a speech recognition system working in a noisy environment.

Two new sets of speech feature parameters, SUBLSF's and SUBCEP's, are introduced. Both parameter sets are based on subband analysis. If the noise is colored then these speech feature parameters can produce better results than the full-band parameters by appropriately emphasizing the noise-free bands.

To obtain the SUBLSF feature parameters the speech signal is first divided into low and high frequency bands and linear predictive analysis is performed on each subband. Consequently two sets of LSF parameters are obtained, and they both form the SUBLSF vector. By extracting fewer number of LSF parameters from the more noisy band robustness against noise is achieved. The SUBLSF parameters are tested in car noise which is colored and they yield superior results in an Hidden Markov Model (HMM) based isolated speech recognition system compared to the widely used full-band LSF parameters. The computational cost of obtaining SUBLSF parameters are higher than the full-band LSF parameters because linear predictive analysis has to be performed twice.

The other set of feature parameters developed in this thesis are the SUBCEP

parameters and they are based on wavelet analysis or equivalently the multirate subband analysis of the speech signal. The SUBCEP parameters also provide robust recognition performance by appropriately deemphasizing the frequency bands corrupted by noise. In this case the frequency domain is first divided into many nonuniform subbands determined according to the *mel*-scale division which is compatible with the human auditory perception system. Then, a set of cepstrum coefficients are obtained from the subsignals corresponding to the subbands. These cepstrum coefficients form the SUBCEP parameter set. The SUBCEP parameters can be realized in a computationally efficient manner by employing fast wavelet analysis techniques. It is experimentally observed that the SUBCEP representation produces better recognition rates than both the commonly used feature parameters and the SUBLSF parameters for speaker dependent and independent isolated word recognition in the presence of car noise. Therefore SUBCEP parameters are promising candidates as feature parameters for large vocabulary recognition systems as well.

Another important problem considered in this thesis is the enhancement of speech signals for a speech recognition system. It is well known that the speech signals in a public telecommunication network are effected by impulsive noise which can be modeled as a symmetric α -stable random process. Adaptive noise cancelation techniques are developed for α -stable random processes. The noisy speech signal is first enhanced by the new adaptive filtering techniques and then fed to the speech recognition system. In this way, higher recognition rates are obtained for speech signals embedded in impulsive α -stable noise.

The last problem considered in this thesis is the detection of the boundaries of the speech utterances or words in isolated speech recognition. If the boundaries are not properly determined then the recognizer may incorrectly interpret the word or utterance. A commonly used measure for the endpoint detection is the energy of the signal. Instead of using the full-band energy a *mel*-frequency weighted energy measure is introduced in this thesis. In order to obtain the *mel*-scale frequency division in a computationally efficient manner multirate signal processing based methods are employed. It is experimentally observed that the new energy measure determines the end-points more accurately compared to the widely used full-band energy measure in the presence of car noise.

The techniques developed in this thesis are promising candidates for continuous and large vocabulary speech processing systems. For example, SUBCEP and SUBLSF parameters can be used as feature parameters in a large vocabulary system. they can also be used in speaker identification and verification, and *mel*-frequency weighted energy measure can be employed to determine the phoneme boundaries in continuous speech recognition system. Further simulation studies will be performed for these application areas in the future.

Appendix A

INTERFRAME DIFFERENTIAL VECTOR CODING OF LSFs

Vocoders achieve data compression by synthesizing speech based on the current frame linear predictive model of the speech signal. Therefore the LP filter should be coded with as few bits as possible.

In this section, we present the new LSF coding method for vocoders. The key idea of our scheme is to estimate the LSF's of the current speech frame by using *both the LSF's of the previous frame and some of the LSF's of the current frame*. The prediction error vector between the true LSF's and the predicted LSF's is vector-quantized.

In both the LPC-10 Standard [17] and the Code Excited Linear Prediction (CELP) vocoders the predictor order is chosen around 10. Let $A_{10}^k(z)$ be the 10-th order LPC filter of the k^{th} speech frame. Corresponding to $A_{10}^k(z)$, 10 LSF's are defined by Equations (1.26) and (1.27). Let us denote the i^{th} LSF of the k^{th} frame by f_i^k , $i = 1, 2, \dots, 10$. Our differential vector coding scheme estimates the current LSF, f_i^k , from i^{th} LSF of the $(k - 1)^{th}$ frame, f_i^{k-1} , and $(i - 1)^{th}$ LSF

of the k^{th} frame, f_{i-1}^k . In this way, we not only take advantage of the relation between neighboring LSF's but the relation between the LSF's of the consecutive frames as well. The estimate, \hat{f}_i^k , of the LSF, f_i^k , is predicted as follows,

$$\hat{f}_i^k = \begin{cases} c_i^k \Delta_i + b_i^k f_i^{k-1} & i = 1 \\ c_i^k (f_{i-1}^k + \Delta_i) + b_i^k f_i^{k-1} & i = 4, 8 \\ b_i^k f_i^{k-1} & \text{otherwise} \end{cases} \quad (\text{A.1})$$

where c_i^k 's and b_i^k 's are the predictor coefficients and Δ_i is an offset factor which is the average angular difference between the i^{th} and $(i-1)^{th}$ LSF's. The parameter, Δ_i , is experimentally determined. The set of offset factors that are used in our simulation examples are listed in Table A.1. Predictor coefficients c_i^k 's and b_i^k 's are adapted by the Least Mean Square (LMS) algorithm as follows,

$$\begin{bmatrix} c_i^k \\ b_i^k \end{bmatrix} = \begin{bmatrix} c_i^{k-1} \\ b_i^{k-1} \end{bmatrix} + \alpha_i^{k-1} \begin{bmatrix} f_{i-1}^{k-1} + \Delta_i \\ f_i^{k-2} \end{bmatrix} d_i^{k-1} \quad (\text{A.2})$$

where d_i^{k-1} is the quantized error value between the true LSF, f_i^{k-1} , and the predicted LSF, \hat{f}_i^{k-1} , and the adaptation parameter, α_i^{k-1} is given as:

$$\alpha_i^{k-1} = \frac{\lambda_i}{(f_{i-1}^{k-1} + \Delta_i)^2 + (f_i^{k-2})^2}, \quad 0 < \lambda_i < 2. \quad (\text{A.3})$$

The parameters, λ_i 's, are also experimentally determined.

The error vector whose entries are, $f_i^k - \hat{f}_i^k$, $i = 1, 2, \dots, 10$, is divided into three subvectors containing the first three LSF's, the middle four LSF's and the last three LSF's, respectively. We experimentally observed that choosing the LSF subvectors with the above partition produces better results than any other grouping. Due to the fact that there are three subvectors, only quantized f_{i-1}^k , $i = 4, 8$, are available in the predictor. Therefore, the predictor described in Equation (A.1) uses only f_{i-1}^k , $i = 4, 8$. This intraframe information improves the performance of the predictor.

Each subvector is quantized using different vector quantizers which are designed using simulated annealing based methods [57, 58] and it is possible to code the LP filter by 24 bits for each speech frame of duration 20 msec

without introducing any audible distortion. The codebook sizes that are used in simulation examples are shown in Table A.2. For example, in the coding of the first (second) [third] subvector a codebook of size 128 (1024) [128] is used for 24 bits/frame case.

A weighted Euclidean distance measure [32] is used in quantizer design. The weights (ω_i) are proportional to the value of LPC power spectra at a given LSF, f_i^n :

$$\omega_i = [P(f_i^n)]^r \quad (\text{A.4})$$

where $P(f)$ is the LPC power spectra of the n -th frame and r is an empirical constant which is chosen to be equal to 0.15 in our simulation examples.

In simulation studies, we compare our results to other LSF coding schemes, including the vector quantizer based methods of Atal [32] and Farvardin [59].

The weighted M.M.S.E quantizers are trained in a set of 15000 speech frames containing six male and six female persons. The performance of the interframe LSF coding scheme is measured in a set of 11000 speech frames obtained from utterances of three male and three female persons. Lowpass filtered speech is digitized at a sampling rate of 8 kHz. A 10-th order LPC analysis is performed by using stabilized covariance method with high frequency compensation [60]. During the analysis a 30-msec Hamming window is used with a frame update period 20 msec. In order to avoid sharp spectral peaks in the LPC spectrum, a fixed bandwidth of 10 Hz is added uniformly to each LPC filter by using a fixed bandwidth-broadening factor, 0.996.

In our simulation examples we use the following spectral distortion measure

$$d'(A(\omega), A'(\omega)) = \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \left[10 \log \frac{|A'(\omega)|^2}{|A(\omega)|^2} \right]^2 d\omega \right]^{1/2} \quad (\text{A.5})$$

which is also used in [32] and [59]. The methods described in [32] and [59] reach 1.0 dB spectral distortion and an acceptable percent of outliers (less than 2% outliers with spectral distortion greater than 2 dB, [32]) at 24 and 25 bits/frame, respectively. Our method also reaches this LPC quantization level at 24 bits/frame. Our simulation results and the results of [32] and [59] are summarized in Table A.3 and Table A.4, respectively. Although we use different

evaluation data sets than [32] and [59] (the sets used in [32] and [59] are also different from each other), we observe that our method produces comparable results to [32].

Interframe differential vector coding of LSF's is more advantageous than direct vector quantization of LSF's. Since the overall codebook size of our coder is much smaller than the ones used in [61] and [32] (e.g., 6.4 times lower than [32]), our method is computationally more efficient than [61] and [32], and it requires smaller storage space.

Previously, other interframe differential coding schemes are also described in [19] and [62]. In [19] the scalar quantization is used and the prediction coefficients are fixed. In [62] the predictor does not utilize the angular offset factor, Δ_i , and a 1900 bits/sec (with a comparable distortion level) transmission rate is reported. In this work better results than [19] and [62] are obtained. With our coding scheme a transmission rate of 1200 bits/sec with 1 dB average spectral distortion can be achieved. This is because of the fact that in this work an adaptive predictor is used and the difference vector resulting from the prediction is vector quantized.

Table A.1: The angular offset factors which are used in simulations

Δ_1	Δ_2	Δ_3	Δ_4	Δ_5
0.22	0.12	0.24	0.37	0.32
Δ_6	Δ_7	Δ_8	Δ_9	Δ_{10}
0.26	0.37	0.23	0.29	0.28

Table A.2: Codebook sizes for each subvector at different rates

Rate	Codebook Sizes		
bits/frame	first	middle	last
	3	4	3
22	128	256	128
23	128	512	128
24	128	1024	128

Table A.3: Spectral Distortion (SD) Performance of our method

Rate bits/frame	Av. SD (dB)	outliers > 2dB
22	1.02	3.66%
23	0.92	2.26%
24	0.86	1.55%

Table A.4: Spectral Distortion (SD) Performance of the Vector Quantizers Atal et.al and Farvardin et.al.

Rate	[32]		[59]	
bits per frame	Av. SD (dB)	outliers > 2dB	Av. SD (dB)	outliers > 2dB
22	1.17	2.73%	-	-
23	1.10	1.60%	-	-
24	1.03	1.03%	1.11	1.50%
25	0.96	0.61%	1.02	0.20%

APPENDIX B

PERFORMANCE

EVALUATION OF ADAPTIVE

FILTERING ALGORITHMS

In [47] the performance of the LMAD algorithm is compared with that of the LMS algorithm for a first order α -stable AR process and it was observed that LMAD outperforms LMS especially for low α values. In this appendix, we compare the performances of the new algorithms with the LMAD, LMP and LMS algorithms.

In simulation studies we consider $AR(N)$ α -stable processes, which are defined as follows,

$$x(n) = \sum_{i=1}^N a_i x(n-i) + u(n) \quad (\text{B.1})$$

where $u(n)$ is a α -stable sequence of i.i.d random variables. The common distribution of $u(n)$ is chosen to be an even function ($\beta = 0$), and the gain factors are all set to one ($\gamma = 1$) without loss of generality. It can be shown that $x(n)$ will also be a α -stable random variable with the same characteristic exponent when $\{a_i\}$ is an absolutely summable sequence [47, 63].

Two sets of simulation studies are performed. In the first set, the adaptation

algorithms are compared for the cases of first and second order α -stable AR processes with a fixed characteristic exponent, $\alpha = 1.2$. In the second set the performances of LMAD, NLMAD and NLMP algorithms are compared when a fourth order α -stable AR process with different values of the characteristic exponent is used. For both sets, the tap weights are obtained by averaging 100 independent trials of the experiment and for each trial, a different computer realization of the process $\{u(n)\}$ is used. To get a fair comparison between algorithms the step size of the LMS is adjusted as large as possible while ensuring the convergence. Then the step sizes of other algorithms are chosen in such a way that they all had a comparable steady-state error. In the first simulation

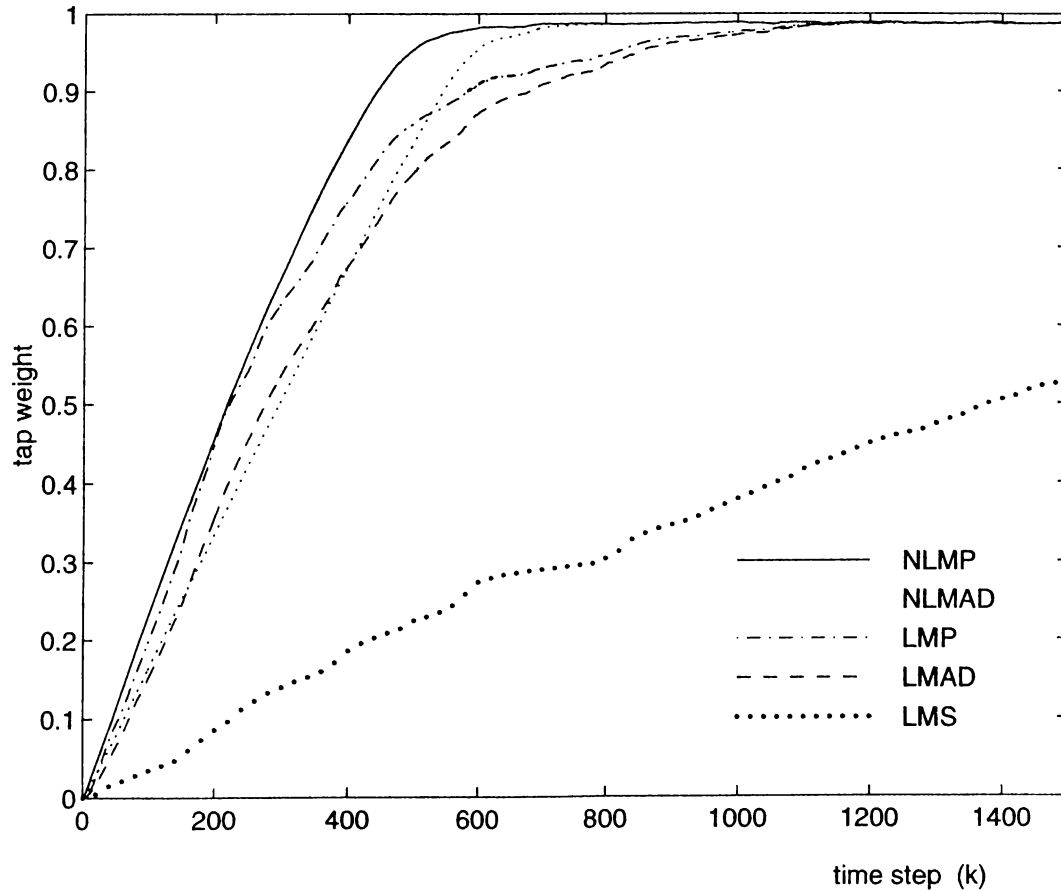


Figure B.1: Transient behavior of tap weight adaptations in the NLMP, NLMAD, LMAD, LMP and LMS algorithms with $\alpha = 1.2$ for $AR(1)$ process. AR parameter is chosen as, $a_1 = 0.99$.

set, tap weight adaptation is performed for $AR(1)$ and $AR(2)$ processes with first and second order LMP, LMAD, NLMP, NLMAD and LMS algorithms, respectively. The coefficient of $AR(1)$ process is chosen as, $a_1 = 0.99$. In Figure B.1, the transient behaviors of the tap weight adaptations for $AR(1)$ process are plotted. The NLMP and NLMAD algorithms introduced in this study has a better convergence behavior than the LMAD [47], LMP [47] and the LMS algorithms.

For $AR(2)$ process, the coefficients are chosen as, $a_1 = 0.99$ and $a_2 = -0.1$. In Figure B.2, the transient behaviors of the tap weight adaptations for $AR(2)$ process are plotted. In this case, the NLMP and NLMAD algorithms again show faster convergence.

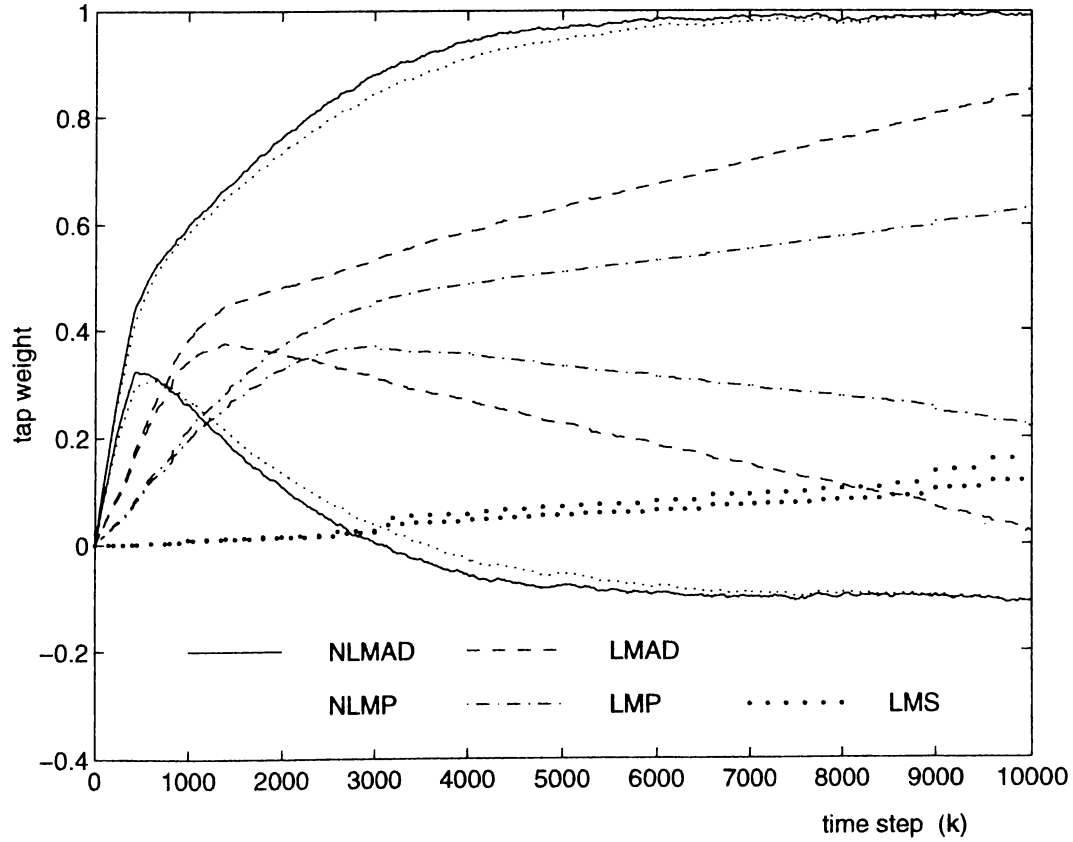


Figure B.2: Transient behavior of tap weight adaptations in the NLMP, NLMAD, LMAD, LMP and LMS algorithms with $\alpha = 1.2$ for $AR(2)$ process. AR parameters are chosen as, $a_1 = 0.99$ and $a_2 = -0.1$.

Second simulation set tests the performances of the LMAD, NLMD and NLMP algorithms for α values 1.2, 1.5, and 1.9 with an $AR(4)$ process. The $AR(4)$ process is chosen as a fourth order LPC synthesis filter of a voiced speech frame [5], $A(z) = 1.323z^{-1} - 0.152z^{-2} - 0.097z^{-3} - 0.115z^{-4}$. The adaptation performances are plotted in Figure B.3. The NLMD and NLMP algorithms

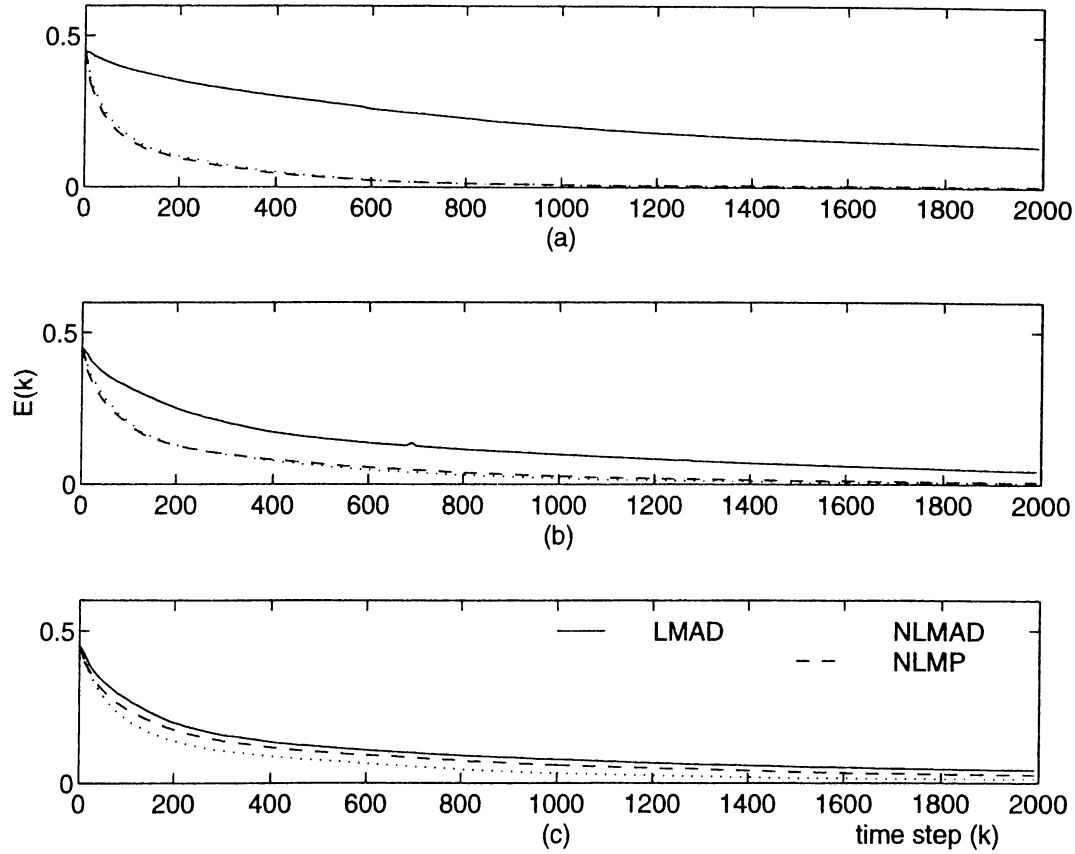


Figure B.3: Tap weight error powers for (a) $\alpha = 1.2$, (b) $\alpha = 1.5$, and (c) $\alpha = 1.9$ in LMAD, NLMD and NLMP algorithms. $E(k) = \|\underline{w}(k) - \underline{w}^*\|^2$ where $\underline{w}(k)$ and \underline{w}^* are the current tap weight and optimal solution vectors, respectively.

have comparable performances for small α values both of which converge faster than the LMAD algorithm.

The performance of these normalized algorithms are found to be superior than that of LMS and LMAD algorithms in simulation studies. Based on the

experience gained in the simulation studies, it is observed that a safe choice of p value is 1 in the case of imprecise knowledge of α . This corresponds to the use of NLMAD algorithm in such cases.

Bibliography

- [1] J. B. Allen “How do humans process and recognize speech,” *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 567–577, October 1994.
- [2] L.R. Rabiner “Applications of voice processing to telecommunications,” *Proc. of the IEEE*, vol. 82, February 1994.
- [3] B.S. Atal, V.Cuperman, and A. Gersho. *Advances in Speech Coding*. Kluwer Academic Publishers, 1991.
- [4] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.
- [5] J.R. Deller, J.G. Proakis, and J.H.L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan, 1993.
- [6] H. Fletcher and R. H. Galt “Perception of speech and its relation to telephony,” *J. Acoustic. Soc. Amer.*, vol. 22, pp. 89–151, March 1950.
- [7] J.W. Picone “Signal modeling techniques in speech recognition,” *Proc. of the IEEE*, vol. 81, pp. 1215–1247, Sept. 1993.
- [8] R. Cole, L. Hirschman, L. Atlas, M. Beckman, A. Biermann, M. Bush, M. Clements, J. Cohen, O. Garcia, B. Hanson, H. Hermansky, S. Levinson, K. McKeown, N. Morgan, D. G. Novick, M Ostendorf, S. Oviatt, P. Price, H. Silverman, J. Spitz, A. Waibel, C. Weinstein, S. Zahorian, and V. Zue “The challenge of spoken language systems: Research directions for the nineties,” *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 1–21, January 1995.

- [9] V. W. Zue "The use of speech knowledge in automatic speech recognition," *Proc. of IEEE*, pp. 1602–1615, 1985.
- [10] B. A. Hanson and H. Wakita "Spectral slope distance measures with linear prediction analysis for word recognition in noise," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 35, pp. 968–973, 1987.
- [11] Y. Ephraim "Gain-adaptive hidden markov models for recognition of clean and noisy speech," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 40, pp. 1303–1316, June 1992.
- [12] D. Mansour and B. H. Juang "A family of distortion measures based upon projection for robust speech recognition," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1988 (ICASSP '88)*, pp. 36–39, 1988.
- [13] P. Alexandre and P. Lockwood "Root cepstral analysis: A unified view. application to speech processing in car noise environments," *Speech Communication*, vol. 12, pp. 277–288, 1993.
- [14] L. G. Neumeyer, V. V. Digalakis, and M. Weintraub "Training issues and channel equalization techniques for the construction of telephone acoustic models using a high quality speech corpus," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 590–597, Oct. 1994.
- [15] N. Levinson "The Weiner RMS (root mean square) error criterion in filter design and prediction," *Journal of Matematical Physics*, vol. 25, pp. 261–278, 1947.
- [16] J. Durbin "Efficient estimation of parameters in moving-average models," *Biometrika*, vol. 46, pp. 306–316, 1959.
- [17] Military Agency for Standardization. "NATO stardardization agreement, STANAG 4196, parameters and coding characteristics that must be common to assure interoperability of 2400 bps linear predictive encoded digital speech,".
- [18] F. Itakura "Line spectrum representation of linear predictive coefficients of speech signals," *Journal of Acoust. Soc. Am.*, p. 535a, 1975.

- [19] E. Erzin and A. E. Çetin "Interframe differential coding of Line Spectrum Pairs," *IEEE Trans. on Speech and Audio Processing*, April 1994. Also presented in part at *Twenty-sixth Annual Conference on Information Sciences and Systems*, Princeton, NJ. March 1992.
- [20] E. Erzin and A.E. Çetin "Interframe differential vector coding of Line Spectrum Frequencies," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1993 (ICASSP '93)*, vol. II, pp. 25-28, April 1993.
- [21] E. Erzin and A.E. Çetin "On the use of interframe information of Line Spectral Frequencies in speech coding," in *NATO-ASI, New Advances and Trends in Speech Recognition and Coding*, Bubion (Granada), June-July 1993.
- [22] L.R. Rabiner "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, pp. 257-285, Feb. 1989.
- [23] D.B. Roe and J.G. Wilpon "Whither speech recognition: The next 25 years," *IEEE Communications Magazine*, vol. 31, pp. 54-62, November 1993.
- [24] L.E. Baum "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1-8, 1972.
- [25] B.H. Juang, S.E. Levinson, and M.M. Sondhi "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. on Information Theory*, vol. 32, pp. 307-309, March 1986.
- [26] B. Tüzün, E. Erzin, M. Demirekler, T. Memişoğlu, S. Uğur, and A.E. Çetin "A speaker independent isolated word recognition system for Turkish," in *NATO-ASI, New Advances and Trends in Speech Recognition and Coding*, Bubion (Granada), June-July 1993.
- [27] S. Alkan, R. Edizkan, O. Parlaktuna, and A. Barkana "Kişiden bağımsız ses tanımda temel özelliklerin kullanılması," *Sinyal İşleme ve Uygulamaları Kurultayı*, vol. B, pp. 1-6, 1995.
- [28] M. Altekin and A. Daloğlu "Konuşmacı eğitilmiş ayırık ses tanıma," *Sinyal İşleme ve Uygulamaları Kurultayı*, vol. B, pp. 7-12, 1995.

- [29] A. Kaderli and A. S. Kayhan "Zaman-sıklık yöntemlerinin söz tanımayaya uygulanması," *Sinyal İşleme ve Uygulamaları Kurultayı*, vol. B, pp. 13-17, 1995.
- [30] C. Çetinkaya, Ü. Künkçü, and A. Barkana "Çoklu ayırma analizi yardımı ile oluşturulan entropi ağlarla Türkçe grupların birbirinden ayrılması," *Sinyal İşleme ve Uygulamaları Kurultayı*, vol. B, pp. 31-36, 1995.
- [31] M. Demirekler, ed. *Konuşma işleme çalıştayı*. METU, May. 1995.
- [32] K.K. Paliwal and B.S. Atal "Efficient vector quantization of LPC parameters at 24 bits/frame," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1991 (ICASSP '91)*, pp. 661-664, May 1991.
- [33] E. Erzin and A.E. Çetin "Line spectral frequency representation of subbands for speech recognition," *Signal Processing*, vol. 44, June 1995.
- [34] F. S. Gürgen, S. Sagayama, and S. Furui "Line spectrum frequency based distance measures for speech recognition," *Proc. Int. Conf. Spoken Language Processing, Kobe, Japan*, pp. 521-524, 1990.
- [35] K.K. Paliwal "On the use of Line Spectral Frequency parameters for speech recognition," *Digital Signal Proc. A Review Jour.*, vol. 2, pp. 80-87, April 1992.
- [36] Zwicker and E. Terhardt "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.*, pp. 1523-1525, December 1980.
- [37] E. Erzin, A.E. Çetin, and Y. Yardımcı "Subband analysis for robust speech recognition in the presence of car noise," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1995 (ICASSP '95)*, May 1995.
- [38] B. S. Atal "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, June 1974.
- [39] S. B. Davis and P. Mermelstein "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE*

- Trans. on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357–366, August 1980.
- [40] I. Daubechies. *Ten Lectures on Wavelets*. SIAM Press, Philadelphia, 1992.
- [41] C. W. Kim, R. Ansari, and A. E. Çetin “A class of linear-phase regular biorthogonal wavelets,” *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1992 (ICASSP '92)*, vol. IV, pp. 673–677, 1992.
- [42] S. M. Phoong, C. W. Kim, P.P. Vaidyanathan, and R. Ansari “A new class of two-channel biorthogonal filter banks and wavelet bases,” *IEEE Trans. on Signal Processing*, pp. 649–665, 1995.
- [43] J. Lim “Spectral root homomorphic deconvolution system,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, June 1989.
- [44] G. Beylkin, R. Coifman, and V. Rokhlin. “Fast wavelet transforms and numerical algorithms 1,”. Technical report. YALBU/DCS/RR-696, 1989.
- [45] B.W. Stuck and B. Kleiner “A statistical analysis of telephone noise,” *Bell System Tech.*, vol. 53, pp. 1263–1320, 1974.
- [46] S.S. Pillai and M. Harisankar “Simulated performance of a ds spread spectrum system in impulsive atmospheric noise,” *IEEE Trans. Electromagnetic Compat.*, vol. 29, pp. 80–82, 1987.
- [47] M. Shao and C.L. Nikias “Signal processing with fractional lower order moments: Stable processes and their applications,” *Proc. of IEEE*, vol. 81, pp. 986–1010, July 1993.
- [48] O. Arikan, A.E. Çetin, and E. Erzin “Adaptive filtering for non-Gaussian processes,” *IEEE Signal Processing Letters*, vol. 1, pp. 163–165, Nov. 1994. Also presented in part at “Twenty-eighth Annual Conference on Information Sciences and Systems”, Princeton, N.J., March 1994.
- [49] O. Arikan, M. Belge, A.E. Çetin, and E. Erzin “Adaptive filtering approaches for non-gaussian processes,” *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1995 (ICASSP '95)*, May 1995.

- [50] J. M. Cioffi "An unwindowed RLS adaptive lattice algorithm," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 365–371, March 1988.
- [51] B. Widrow and S.D. Stearns. *Adaptive Signal Processing*. Prentice Hall, NJ, 1985.
- [52] M. R. Sambur "LMS adaptive filtering for enhancing the quality of noisy speech," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1978 (ICASSP '78)*, pp. 610–613, Apr. 1978.
- [53] M. R. Sambur "Adaaptive noise canceling for speech signals," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 26, pp. 419–423, Oct. 1978.
- [54] E. Erzin. "Low bit rate speech coding and a new interframe differential coding of Line Spectrum Pairs,". Master's thesis, Bilkent, 1992.
- [55] L.R. Rabiner "An algorithm for determining the endpoints of isolated utterances," *Bell System Tech.*, pp. 297–315, Feb. 1975.
- [56] D. C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, 1991.
- [57] A. E. Çetin and V. Weerackody "Design of vector quantizers using simulated annealing," *IEEE Trans. on Circuits and Systems*, vol. 35, p. 1550, 1988.
- [58] K. Zeger and A. Gersho "Stochastic relaxation algorithm for improved vector quantiser design," *Electronics Letters*, vol. 25, pp. 896–898, July 1989.
- [59] N. Phamdo, R. Laroia, and N. Farvardin "Robust and efficient quantization of LSP parameters using structured vector quantizers," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1991 (ICASSP '91)*, pp. 641–645, May 1991.
- [60] B. S. Atal "Predictive coding of speech at low bit rates," *IEEE Trans. on Communications*, vol. COM-30, pp. 600–614, April 1982.
- [61] M. Yong, G. Davidson, and A. Gersho "Encoding of LPC spectral parameters using switched adaptive interframe vector prediction," *Proc.*

- of the Int. Conf. on Acoustics, Speech and Signal Processing 1988 (ICASSP '88)*, pp. 402–405, 1988.
- [62] C.C. Kuo, F.R. Jean, and H.C. Wang “Low bit rate quantization of LSP parameters using two-dimensional differential coding,” *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1992 (ICASSP '92)*, pp. 97–100, 1992.
- [63] Y. Hosoya “Discrete-time stable processes and their certain properties,” *Ann. Prob.*, vol. 6, pp. 94–105, 1978.

Vita

Engin Erzin was born in Kemalpaşa, İzmir, Turkey, in 1967. He received his B.Sc. degree and M.Sc. degree from the Bilkent University, Ankara, Turkey, in 1990 and 1992, respectively, both in electrical and electronics engineering. His current research interests include speech coding in low bit rates, speech recognition, text-to-speech synthesis and adaptive signal processing.