LARGE VOCABULARY SPEECH RECOGNITION IN NOISY ENVIRONMENTS

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCES

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Firas Jabloun

July 1998

LARGE VOCABULARY SPEECH RECOGNITION IN NOISY ENVIRONMENTS

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCES

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Firas Jabloun

July 1998

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

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 thesis for the degree of Master of Science.

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 thesis for the degree of Master of Science.

Mübeccel Demirekler, Ph. D.

Approved for the Institute of Engineering and Sciences:

Prof. Dr. Mehmet Baray
Director of Institute of Engineering and Sciences

ABSTRACT

LARGE VOCABULARY SPEECH RECOGNITION IN NOISY ENVIRONMENTS

Firas Jabloun
M.S. in Electrical and Electronics Engineering
Supervisor: A. Enis Çetin, Ph. D.
July 1998

A new set of speech feature parameters based on multirate subband analysis and the Teager Energy Operator (TEO) is developed. The speech signal is first divided into nonuniform subbands in mel-scale using a multirate filter-bank, then the Teager energies of the subsignals are estimated. Finally, the feature vector is constructed by log-compression and inverse DCT computation. The new feature parameters (TEOCEP) have a robust speech recognition performance in car engine noise which has a low pass nature.

In this thesis, we also present some solutions to the problem of large vocabulary speech recognition. Triphone-based Hidden Markov Models (HMM) are used to model the vocabulary words. Although the straight forward parallel search strategy gives good recognition performance, the processing time required is found to be long and impractical. Therefore another search strategy with similar performance is described. Subvocabularies are developed during the training session to reduce the total number of words considered in the search process. The search is then performed in a tree structure by investigating one subvocabulary instead of all the words.

Keywords : Speech recognition, Multirate subband analysis, Teager Energy Operator, Nonlinear speech modeling, Triphones, Tree structure search strategy.

ÖZET

KONUŞMA TANIMA

Firas Jabloun
Elektrik ve Elektronik Mühendisliği Bölümü Yüksek Lisans
Tez Yöneticisi: Doç. Dr. A. Enis Çetin
Temmuz 1998

Altbant analizi ve Teager Enerji Operatörüne (TEO) dayal yeni bir konuşma öznitelik parametresi seti geliştirildi. Konuşma işareti önce *mel*skala içinde düzgün olmayan altbantlara bölündü. Sonra, alt-işaretlerin Teager Enerji kestirimleri yapıldı. Son olarak *log* sıkıştırma ve ters DCT hesaplamasiyla öznitelik vektörleri oluşturuldu. Yeni öznitelik parametreleri (TEOCEP), döşük geçiren bir yapısı olan araba motor sesine karşı görbüz tanıma performansına sahiptir.

Bu tezde ayrica geniş kelime hazneli konuşma problemine ilişkin çözümler sunuldu. Kelimeler üçlü-fon temelli HMM ile modellendi. İşlem zamanını azaltmak için, öğrenme süresinde alt kelime hazneleri geliştirilip, ağaç yapılı arama işlemi gerçekleştirildi.

Keywords : Konuşma tanıma, altbant analizi, Teager Enerji Operatörü, doğrusal olmayan konuşma modeli, üçlü-fonler, Ağaç yapılı arama stratejisi.

ACKNOWLEDGMENTS

I would like to express my deep gratitude to my supervisor Dr A. Enis Çetin for his guidance, suggestions and valuable encouragement throughout the development of this thesis.

I would like to thank Dr Orhan Arıkan and Dr Mübeccel Demirekler for reading and commenting on the thesis and for the honor they gave me by presiding the jury.

I am also indebted to my family for their continuous morale support throughout my graduate study.

Sincere thanks are also extended to all my friends and to Sevinç who encouraged me during the development of this thesis.

To my family

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech is the most efficient and natural means of communication among human beings. While this has been true since the dawn of civilization, the invention and widespread use of the telephone, audio-phonic storage media, radio and TV has given even further importance to speech communication and hence to speech processing [1, 2]. Further, the way humans produce and perceive speech sounds has also become a stimulating area of research, aiming to create machines that can receive spoken information and act appropriately upon that information or even answer and be able to discuss [3, 4]. In the world of science fiction, computers have always listened and spoken to us exactly as humans do. However, in reality, the speech technology is still not as sophisticated as the dream itself. Fortunately, the advances in digital signal processing technology brought new robust methods for both speech recognition and synthesis [5, 6]. Nonetheless, the objective of a robust, intelligent, fluent conversant machines remains a distant goal.

Three main areas can be distinguished in the field of speech processing, although they overlap considerably; coding, recognition and synthesis. Speech enhancement and compression are also useful for both recognition and coding. In this thesis the problem of speech recognition is considered and new methods which enhance the recognition ability are introduced.

Speech recognition is a rather challenging problem whose dimensions of difficulty are usually viewed in three different ways. The size of the vocabulary can be considered as the first dimension. The performance of the recognizer usually degrades as the vocabulary

size of the system increases. Speech recognition systems are generally classified with small (2-99 words), medium (100-999 words) or large (more than 1000 words) vocabularies [5].

The second dimension of difficulty is the problem of speaker dependency. A speaker dependent recognizer uses the utterances of a single speaker to train the models which characterize the system. This system, then, should be used specifically for recognizing the trainer's speech. Accordingly, the recognizer will yield relatively high recognition rates compared with a speaker independent recognition system. The latter is trained by multiple speakers and used to recognize many speakers including those who may be outside the training population.

The most complex systems perform continuous speech recognition, i.e., the user utters the message in the most natural manner used in real life [7,8]. The difficulty here is to be able to detect the boundaries in the acoustic signal, and to perform well in spite of all the co-articulatory effects and sloppy articulation (including insertions and deletions) that accompany flowing speech. To get rid of all these difficulties, simpler systems use isolated-word recognition which is a relatively much easier problem. Pauses between words simplify the recognition process because endpoints become easier to identify in addition to minimizing the co-articulation between words. Sometime the application makes it unnecessary to use continuous speech though achieving a high recognition rate with continuous speech stays as the ultimate goal of the speech recognition research.

In this thesis, the problem of large vocabulary speech recognition in noisy environments is investigated. In most cases background noise is modeled as additive stationary perturbation which is uncorrelated with the speech signal. With this assumption we construct a robust speech recognition system in the presence of car noise. Applications inside automobiles have a very practical importance and that is why car noise environment is examined in this thesis.

In Chapter 2, general concepts of speech recognition are introduced. Namely, the different modules required by the speech recognition system based on a Hidden Markov Model (HMM) are reviewed.

In Chapter 3, the wavelet analysis of speech signals or equivalently the multirate subband analysis is introduced and inspected. Subband cepstral coefficients show a more robust performance for speech recognition than the commonly used mel-scale cepstral representation. Chapter 3 also introduces the nonlinear modeling of speech signals. Experimental results show the existence of important nonlinear phenomena during the

speech production that cannot be accounted for by the linear model [9,10]. Each speech resonance is modeled with an AM-FM modulation signal and the total speech signal as a superposition of such AM-FM signals. In [11–15] a new energy operator (Ψ) was successfully used to separate the modulation energies.

In Chapter 4, the Ψ energy operator is used with a multirate subband decomposition approach. The new energy is used as a substitute for the traditional energies to benefit from its robustness against colored car noise. Experimental results show that the car noise Ψ-energy is negligible compared to that of the speech signal, if the resonance frequency of the latter falls within the current analysis frequency band.

A large vocabulary speech recognition system is designed in Chapter 5. A triphone based Hidden Markov Model (HMM) is used to model the phonetic content of one phoneme together with its left and right neighbors. The use of triphone subwords compensate for the lack of a large training database. Furthermore, as the vocabulary size increases, the recognition time becomes important since in speech recognition a fast response is necessary in order to have a practical system. A guided strategic search approach is designed to reduce the effective vocabulary code book.

Conclusions and discussions are given in Chapter 6.

Chapter 2

The speech recognition system

The speech recognition system consists of several modules each of which constitutes a fundamental stage for the recognition process. In the case of continuous speech, however, the endpoint detection module becomes meaningless while other more complex manipulations are performed usually during the recognition phase using the word models themselves. Each of the modules shown in Figure (2) is described below in detail.

2.1 Endpoint Detection

The endpoint detection module is an important step in the isolated word speech recognition systems. Obviously, a robust endpoint detection algorithm contributes significantly to improve the overall performance of the recognition system . In the presence of noise (such as car noise) the traditional methods for endpoint detection which use the energy and the zero-crossing rate [16] , degrade drastically and usually fail to detect some phonemes such as the weak fricatives. Instead, an alternative measure which takes into account the statistics of the environmental noise, is used.

In this thesis, sub-band decomposition [17–19] plays a fundamental role in the acoustic modeling of speech, and it will be discussed in detail in Chapter 3. Nonetheless, it is introduced here since it is needed to describe the endpoint detection algorithm.

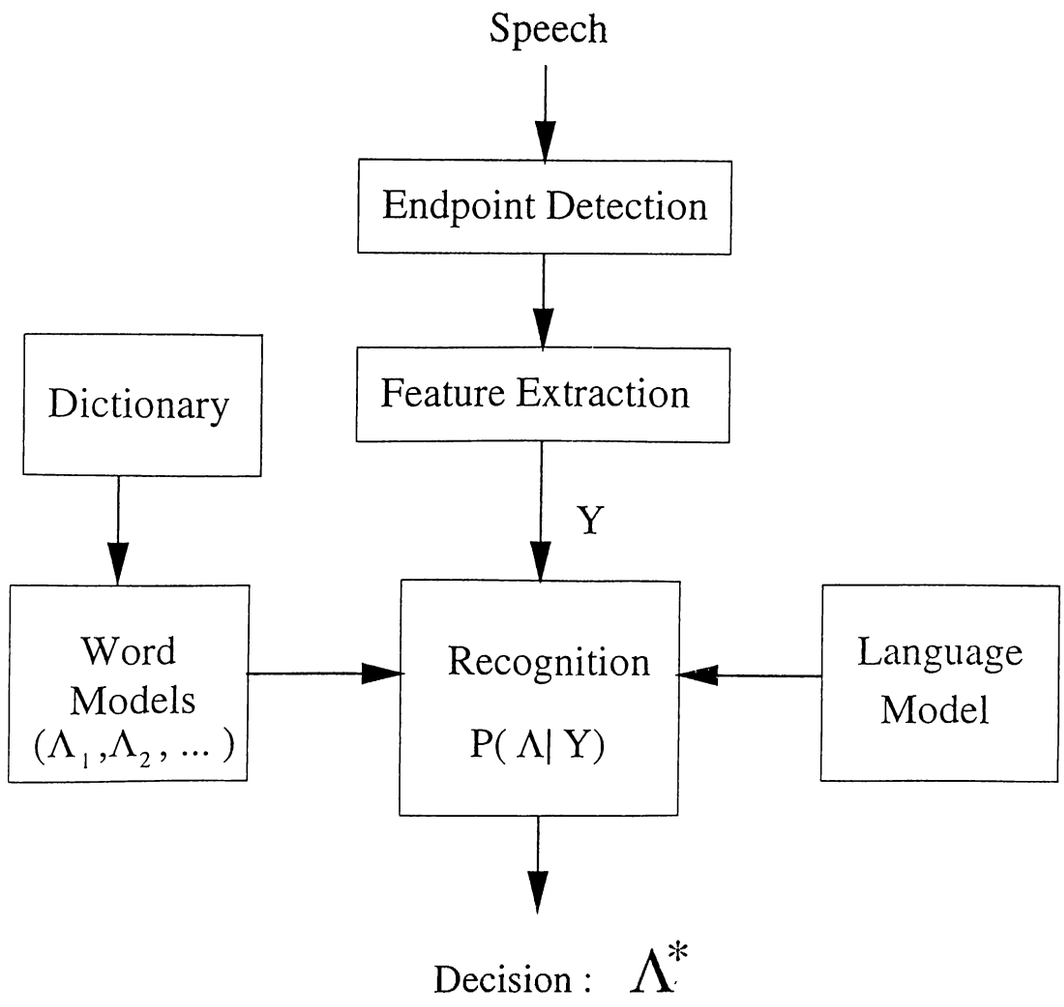

Figure 2.1: The speech recognition system

Suppose that a filter bank is used to decompose the speech signal into several sub-signals each of which is associated with one of the bands in the frequency domain, the following energy parameter E_l^k is defined for the k^{th} speech frame and the l^{th} sub-band.

$$E_l^k = \frac{1}{N_l} \sum_{n=1}^{N_l} s_l^2[n] \qquad l = 1, \ldots, L \tag{2.1}$$

Consequently the distance measure used is defined as follows :

$$D_k = 10log\left[\frac{1}{L} \sum_{l=1}^{L} \frac{(E_l^k - \mu_l)^2}{\sigma_l^2}\right] \tag{2.2}$$

where μ_l and σ_l are the mean and variance of the background noise at the l^{th} band,

respectively. They are estimated a-priory from the utterance free segments as :

$$\mu_l = \frac{1}{N}\sum_{k=1}^{N} E_l^k \tag{2.3}$$

$$\sigma_l^2 = \frac{1}{N}\sum_{k=1}^{N}(E_l^k - \mu_l)^2, \quad l = 1, 2, \ldots, L \tag{2.4}$$

Using this distance measure an efficient algorithm for endpoint detection can be designed and is described in detail in [20].

2.1.1 Feature Extraction

A fundamental assumption in most speech recognition systems is that the speech signal can be considered as stationary over an interval of a few milliseconds. This assumption is a legitimate one because the human speech production system is a physical system which can not change very quickly with time. Thus speech can be divided into frames each of which is considered as a stationary signal. The spacing between the frames is typically in the order of 10 msec, and the blocks are usually overlapping providing a longer analysis window, typically around 25 msec long. Within each of these windows, some features to characterize the given frame are computed. In literature, several parameters were used. The Linear Prediction (LP) coefficients are the earliest ones used for both speech recognition and coding [21–24]. LP coefficients do not show good performance so other features like the Line Spectral Frequencies (LSF's) are proposed in [25–28]. Recently, Melcep coefficients [7, 29, 30] and sub-band cepstrum coefficients [17–20], become the most widely used features for speech recognition. In this thesis, new features which offer robustness against noise with high recognition rates are developed and are discussed in detail in Chapter 4.

2.1.2 Recognition

After detecting the endpoints of the utterance and extracting the features from the speech signal, the recognition phase starts. Assuming that we have a model for each word in the vocabulary, we want to find the model which has the highest probability of producing the given sequence of feature vectors. In other words, suppose we have a

sequence of vectors $Y = \{y_1, y_2, \ldots, y_T\}$, where T is the total number of frames, and y_i is the i^{th} feature vector. A word W^* is claimed to be the uttered word if

$$W^* = \underset{W}{arg\,max} \quad P\{\Lambda_W|Y\} \tag{2.5}$$

where Λ_W is the model characterizing the word W. This problem, however, is rather a difficult one. The solution is to solve the dual problem of finding the probability of having the feature vector Y given the model Λ_W or $P\{Y|\Lambda_W\}$. This can be achieved using the Bayesian rule [31]:

$$W^* = \underset{W}{arg\,max} \quad P\{\Lambda_W|Y\} = \underset{W}{arg\,max} \quad \frac{P\{Y|\Lambda_W\}P\{\Lambda_W\}}{P\{Y\}} \tag{2.6}$$

The probability $P\{Y\}$ is of no importance since Y is given and its probability can be considered to be one. $P\{\Lambda_W\}$ can be computed through a language model. The simplest case is to use a uniform distribution where the verification process follows a parallel strategy which includes all the words of the vocabulary as shown in Figure (2.1.2). This technique, while satisfactory for small vocabularies, is not practical for medium and large vocabulary applications in which more complex search strategies must be used.

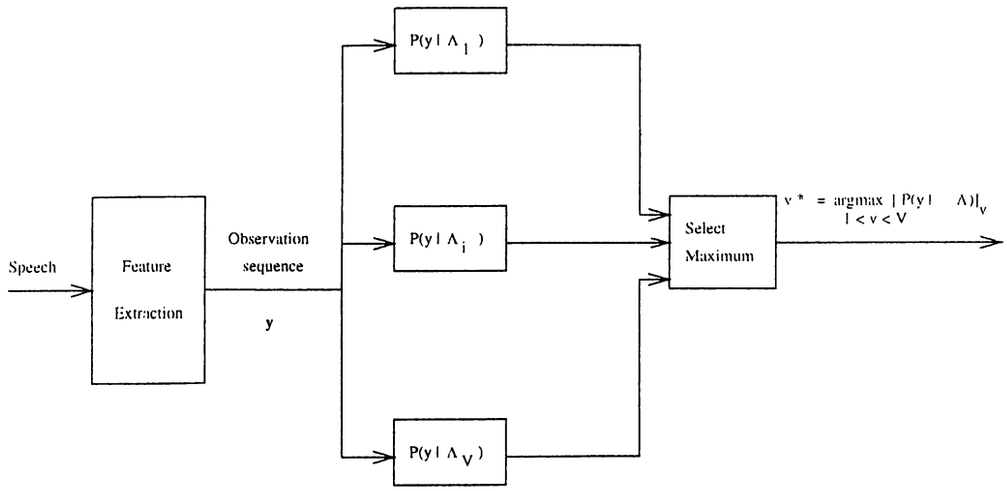

Figure 2.2: The parallel strategy recognition procedure

2.1.3 Word Models

There are several methods to model the words of the dictionary. These include the neural network approach [5], a stochastic model using **H**idden **M**arkov **M**odels (HMM's) [5, 6, 32]. In this thesis, the HMM stochastic model is used.

The idea is to represent a word by a finite state machine with a left to right structure as shown in Figure (2.1.3).

The HMM is assumed to generate observation sequences by jumping from state to state and emitting a feature vector upon arrival at each successive state. Each model is characterized by a set of parameters. These are the state transition probabilities $a(i|j) = P\{x_t = i|x_{t-1} = j\}$, where x_t is the state at time t (or t^{th} frame).

The matrix of state transition probabilities is given by

$$\mathbf{A} = \begin{bmatrix} a(1|1) & & a(1|S) \\ & a(i|j) & \\ a(S|1) & & a(S|S) \end{bmatrix} \tag{2.7}$$

Where S is the total number of states in the model. The state transition probabilities are assumed to be stationary in time, so that $a(i|j)$ does not depend on the time t at which the transitions occurs. Note that any column of $\mathbf{A}$ must sum to unity, since it is assumed that a transition takes place with certainty at every time instant.

The state probability vector at time t is defined as

$$\pi(t) = \begin{bmatrix} P(x(t) = 1) \\ \\ P(x(t) = S) \end{bmatrix} \tag{2.8}$$

So that $\pi(t) = \mathbf{A}\pi(t-1)$ and by recursion $\pi(t) = \mathbf{A}^{(t-1)}\pi(1)$

Therefore the state transition matrix and the initial state probability vector, fully identify the probability of residing in any state at any time.

Upon arrival at one of the states, the model emits an observation vector $y(t)$ with some probability

$$b_j(y(t)) = P\{y(t)|x(t) = j\} \tag{2.9}$$

This probability can be discrete and computed using vector quantization from a priorly defined fixed code book. However the precision of this approach (though computationally very efficient) is very limited with the noise introduced by quantization. Therefore it is not used anymore since its continuous counterpart has shown a much better performance.

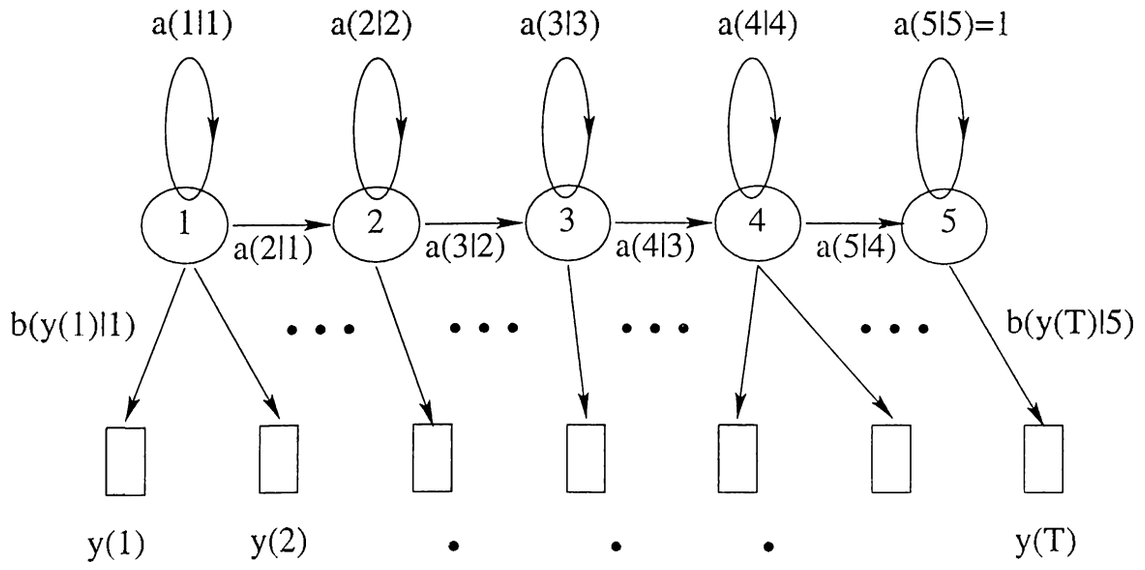

Figure 2.3: A five-state left to right HMM model with the feature vectors each being generated by one state

Modern systems use parametric continuous-density output distributions that model the acoustic vector directly. This distribution together with the initial state probability and the state transition matrix are necessary and sufficient to completely model the words of the vocabulary. Formally this is written as :

$$\mathcal{M} = \{\pi(1), A, f_{y|x}(\zeta|x) \; ; \; 1 \leq i \leq S\} \tag{2.10}$$

Usually $f_{y|x}(\zeta|x)$ is chosen to be a multivariate mixture normal distribution [33], that is a linear combination of Gaussian pdf's :

$$f_{y|x}(\zeta|x = i) = \sum_{m=1}^{M} c_{mi} \, \mathcal{N}(\underline{\mu}_{mi}, \mathbf{C}_{mi}) \tag{2.11}$$

Where c_{mi}, $\underline{\mu}_{mi}$ and $\mathbf{C}_{mi}$ are the weight, mean vector and covariance matrix, respectively, of the m^{th} distribution of the i^{th} state. Note that $\sum_{m=1}^{M} c_{mi} = 1$ for all $i = 1, \ldots S$.

This stochastic model presents three major problems which shall be solved. These three problems are the recognition, training and the state sequence followed. Fortunately in literature several algorithms to solve these problems were used, the most widely used are the Forward-Backward [5,6,34,35] algorithm and the Viterbi algorithm [5,6,36–38]. While the first method can only calculate the likelihood $P(y|\Lambda)$ regardless of the path followed, the second one tries to calculate the same likelihood through the best path and thus provides information about the best state sequence. In this thesis, the Viterbi algorithm was used in for recognition while a mixture of both approaches was used for training.

Chapter 3

Speech Processing Techniques

In this chapter, the wavelet analysis associated with a sub-band decomposition technique is introduced. The subband analysis decomposes the frequency domain of the speech signal into several subsignals. The average energy of each subsignal is computed. The final feature vector is obtained by applying the log compression and the cosine transform to these energies. In Section (3.2) a new energy measure which accounts for the nonlinear characteristics of the speech signal, is presented.

3.1 Sub-band Analysis

Many speech recognition systems use the mel-frequency cepstral coefficients (**MFCC**'s) as features to characterize the speech signal [7,29,30]. Briefly, the MFCC's are computed by smoothing the Fourier transform spectrum by integrating the spectral coefficients within triangular bins arranged on a non-linear scale called the mel-scale. This scale tries to imitate the frequency resolution of the human auditory system which is linear up to 1kHz and logarithmic thereafter. In order to make the statistics of the estimated speech power spectrum approximately Gaussian, log compression is applied to the filter bank output. Finally, the Discrete Cosine Transform (DCT) is applied in order to compress the spectral information into the lower-order coefficients. Moreover the DCT de-correlates these coefficients allowing the subsequent statistical modeling to use diagonal covariance

matrices.

Obviously, this approach operates on the frequency domain which can be a computationally costly task. Therefore the wavelet analysis associated with a sub-band decomposition technique was proposed and was widely discussed in literature [17, 20, 39–42]. It provides a fast structure for decomposing the frequency domain along with the temporal information. The implementation of a wavelet transform can differ according to the application, but the easiest seems to be the tree structure which uses a single basic building block repeatedly until the desired decomposition is accomplished.

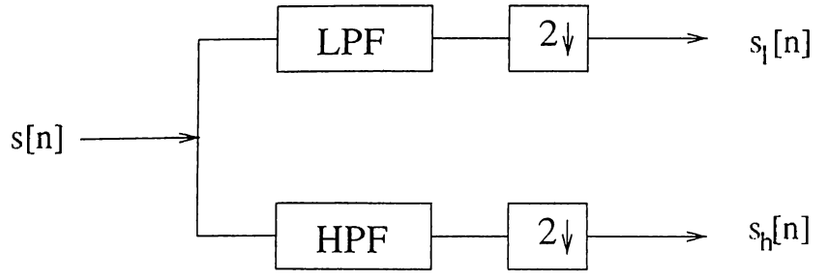

Figure 3.1: Basic block of sub-band decomposition

This basic unit uses techniques of multi-rate signal processing and consists of a low and a high pass filter followed by a down-sampling unit Figure 3.1. The pass-bands of the low and high pass filters are $[0, \pi/2]$ and $[\pi/2, \pi]$, respectively. These two filters divide the frequency range into two half-bands and the down-sampling units reduces the lengths of the obtained signals by two which consequently reduces the computational cost. Each of the sub-signals $s_l[n]$ and $s_h[n]$ can be further decomposed into two new sub-signals using the same filter bank once more, and this procedure can be repeated until the desired frequency decomposition is achieved. Here it is worth mentioning that the process of high-pass filtering and down-sampling inverts the frequency spectrum [40]. Thus while further decomposing the high-pass sub-signals, the low-pass and high-pass filters must be switched in order to get the designed frequency decomposition as shown in Figure (3.2).

These sub-signals make up the so-called wave-packet representation uniquely characterizing the original signal [39]. Perfect reconstruction can be achieved with a proper choice of the low-pass and high pass filters [43]. This sub-band analysis was recently proposed for speech coding purposes where it produces satisfactory results [44].

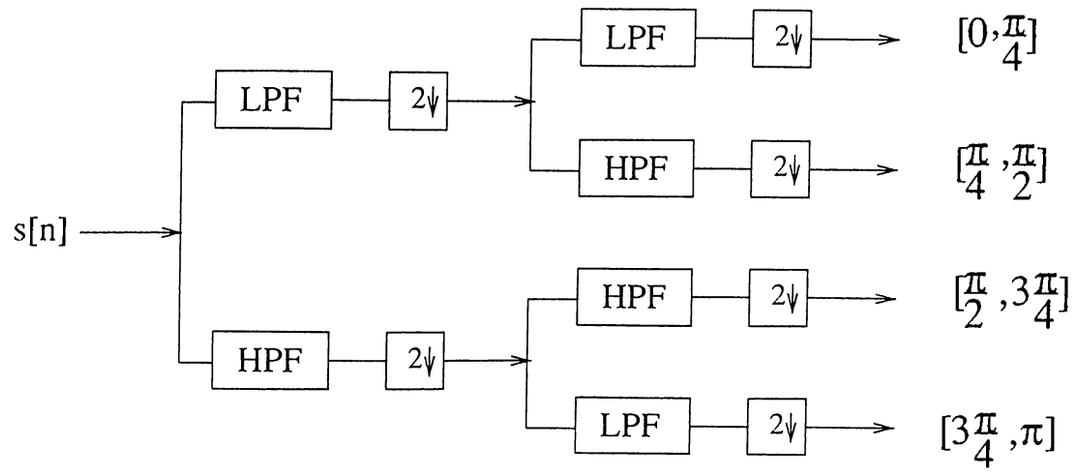

Figure 3.2: A two stage tree structure subband decomposition design

It is shown in [17, 20] that the use of a sub-band decomposition filter bank corresponding to a biorthogonal wavelet transform [43, 45] (rather than an orthogonal one) offers better results for speech recognition purposes especially for speech contaminated with noise. One of the possible choices for the low-pass filter is the 7^{th} order Lagrange filter having the transfer function

$$H_l(z) = \frac{1}{2} + \frac{9}{32}(z^{-1} + z^{1}) - \frac{1}{32}(z^{-3} + z^{3}). \tag{3.1}$$

The corresponding high-pass filter has the transfer function

$$H_h(z) = \frac{1}{2} - \frac{9}{32}(z^{-1} + z^{1}) + \frac{1}{32}(z^{-3} + z^{3}). \tag{3.2}$$

For simulation purposes, the speech signal, $s(n)$, is decomposed into $L = 21$ sub-signals, $\{s_l(n)\}_{l=1}^{L}$ using the tree structured filter bank, Figure 3.4. The corresponding frequency domain decomposition is similar to the *mel-scale* [30] described above and shown in Figure 3.1.

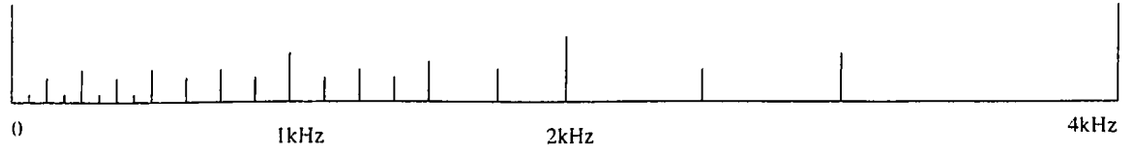

Figure 3.3: The sub-band frequency decomposition of the speech signal

After obtaining the sub-signals cepstral analysis is performed to obtain the feature

parameters. For each sub-signal, an energy parameter e_l is defined

$$e_l = \frac{1}{N_l} \sum_{n=1}^{N_l} |s_l[n]| \qquad ; \qquad l = 1, \ldots, L \tag{3.3}$$

where N_l is the number of samples in the l^{th} band. Note that each of the parameters is defined over a finite size window of length W and overlap T. Log compression and DCT transformation is then applied to obtain the sub-band cepstrum coefficients or SUBCEP's :

$$SC(k) = \sum_{l=1}^{L} \log(e_l) \cos[\frac{k(l-0.5)\pi}{L}] \qquad ; \qquad k = 1, \ldots, N. \tag{3.4}$$

where N is the number of coefficients considered ($N = 12$).

In [17,20] instead of the Log compression the sub-band cepstral parameters are defined as root-cepstral coefficients [46] as follows,

$$SC(k) = \sum_{l=1}^{L} (e_l)^{p_l} \cos[\frac{k(l-0.5)\pi}{L}] \qquad ; \qquad k = 1, \ldots, N. \tag{3.5}$$

p_l is the root value for the p^{th} band, with each band being properly weighted according to the choice of p_l values. This increases the robustness of the speech feature parameters in case of speech contaminated with colored environmental noise [20]. For instance, for speech recognition under automobile noise, it is observed that the car noise has very little high frequency content. Therefore lower root values are used for the first two bands. In [20] the following root values were used

$$\begin{aligned} \underline{P} &= [p_1 \; p_2 \; \ldots \; p_L] \\ &= [0.094 \; 0.281 \; 0.375 \; 0.375 \; \ldots \; 0.375] \end{aligned} \tag{3.6}$$

However, the spectrum of the speech signal itself may have important information in these same lower bands, so the recognition rate may degrade with bad choices of the root values.

Usually, each acoustic vector is supposed to be uncorrelated with its neighbors. However this is a rather poor assumption since the physical constraint of the human vocal system ensures that there is continuity between successive spectral estimates [7]. Thus appending the first-order differentials to the initial feature vector will greatly reduce the problem. So if the first 12 subcepstrum coefficients are taken with additional 12 coefficients from the derivative, a final feature vector with dimension 24 is obtained and is used for training and recognition.

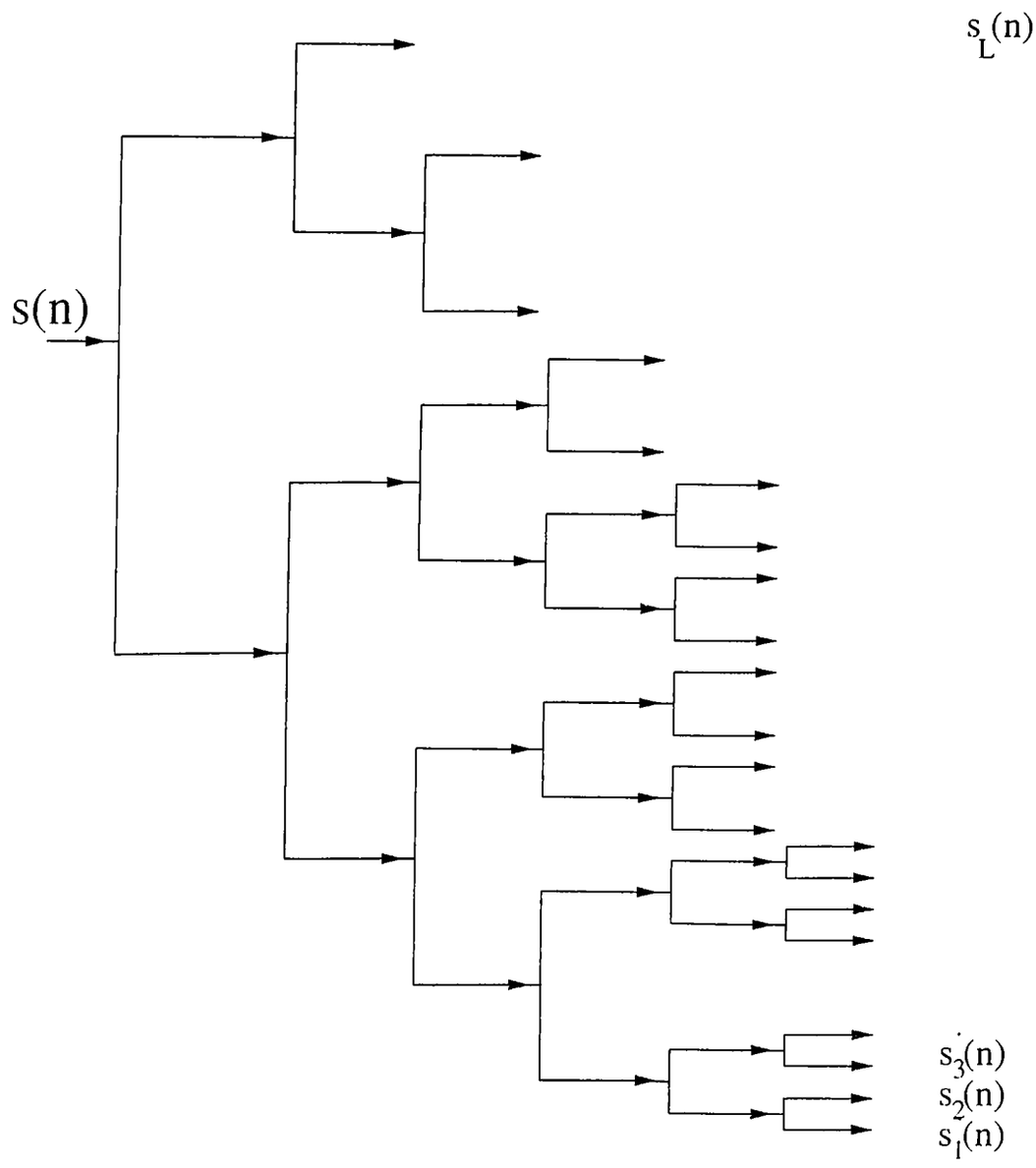

Figure 3.4: The tree structure sub-band decomposition

3.2 Non-linear Properties of Speech Signals

In speech processing, the vocal tract is traditionally modeled by a linear filter. The actual non-linear characteristics of speech production are approximated by the standard assumptions of linear acoustics and the 1-D plane wave propagation of the sound in the vocal tract. The well-known linear prediction model had some success in several

applications such as speech coding, recognition and synthesis. However, it was proven both theoretically and experimentally that there exist important non-linear phenomena during the speech production process which cannot be accounted for by the linear model [9, 10].

One of the non-linear properties that has been proven to characterize the human speech production is the existence of some kind of amplitude modulation (**AM**) and frequency modulation (**FM**) in speech resonance signals. This fact makes the amplitude and resonance frequency vary instantaneously within one pitch period [12, 14]. Here speech resonances loosely refer to the oscillator systems formed by local cavities of the vocal tract emphasizing certain frequencies and de-emphasizing others during speech production. In linear speech modeling, speech resonances, also called formants, are characterized by the poles of the transfer function of the linear filter modeling the vocal tract.

Motivated by these and other evidences, Maragos, Quatieri and Kaiser [11, 12, 14] proposed to model each speech resonance with an AM-FM signal

$$x(t) = a(t)\ \cos[\phi(t)] = a(t)\ \cos[\int_0^t \omega(\tau)d\tau + \phi(0)] \tag{3.7}$$

where $a(t)$ and $\phi(t)$ are the instantaneous amplitude and instantaneous frequency respectively and $w(t) = d\phi(t)/dt$. The total speech signal is assumed to be a superposition of such AM-FM signals, one for each formant. This approach starts by isolating individual resonances by bandpass filtering the speech signal around its formants. Next, the amplitude and frequency signals are to be estimated based on an "energy tracking" operator. Teager has developed several tools for non-linear speech processing [10] such as the continuous time energy operator

$$\Psi_c[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t) \tag{3.8}$$

where $\dot{x} = \frac{dx}{dt}$.

If $x(n)$ is a sampled version of a continuous-time signal then the discrete version of the Teager Energy Operator (TEO) is obtained by approximating derivatives $\dot{x}$ with the 2-sample backward (or forward) difference $[x(n) - x(n-1)]/T$ where T is the sampling period. Without any loss of generality, T can be set to one which is the case for signals initially defined in a discrete context. Then the continuous-time energy operator reduces (up to one sample shift) to the following discrete version

$$\Psi_d[x(n)] = x^2(n) - x(n+1)x(n-1) \tag{3.9}$$

Both Ψ_c and Ψ_d are nonlinear and translation invariant and were shown to track the energy of simple harmonic oscillators. Namely, if $x(t) = A\cos(w_c t + \theta)$ then

$$\Psi_c[x(t)] = (A\omega_c)^2. \tag{3.10}$$

The idea that Ψ is an energy measure was motivated by the fact that an undamped oscillator consisting of a mass m and a spring of constant k has a displacement $x(t) = A\cos(\omega_0 t + \theta)$, with $\omega_0 = \sqrt{k/m}$. The instantaneous energy E_0 of this undamped oscillator is the sum of its kinetic and potential energies and equals the constant

$$E_0 = \frac{m}{2}(A\omega_0)^2 \tag{3.11}$$

Thus the energy of the linear oscillator is proportional to $\Psi_c[x(t)]$ in Equation (3.10).

This result can also be observed using the discrete Ψ operator. Let $x_n = A\cos(\Omega n + \phi)$, where Ω is the digital frequency in radians/sample and is given by $\Omega = 2\pi f/f_s$ where f is the analog frequency and f_s is the sampling frequency. ϕ is an arbitrary initial phase in radians. To calculate the discrete Teager Energy of this sequence, three adjacent equally spaced samples of x_n are given

$$\begin{aligned} x(n) &= A\cos(\Omega n + \phi) \\ x(n+1) &= A\cos[\Omega(n+1) + \phi] \\ x(n-1) &= A\cos[\Omega(n-1) + \phi] \end{aligned} \tag{3.12}$$

using the last two samples with the appropriate trigonometric identities

$$\begin{aligned} x(n+1)x(n-1) &= \tfrac{A^2}{2}[\cos(2\Omega n + 2\phi) + \cos(2\Omega)] \\ &= A^2\cos^2(\Omega n + \phi) - A^2\sin^2(\Omega) \end{aligned} \tag{3.13}$$

Notice that the first term of the right-hand side of the previous equation is simply x_n squared so

$$\Psi_d[x(n)] = x(n+1)x(n-1) - x^2(n) = A^2\sin^2(\Omega) \tag{3.14}$$

Though the result is not proportional to $(A\Omega)^2$, this result is still an interesting one and can easily lead to solve for A and Ω. Moreover, for small values of Ω, $\sin(\Omega) \approx \Omega$ and then

$$\Psi_d[x(n)] \approx A^2\Omega^2 \tag{3.15}$$

The TEO has some useful properties which are worth mentioning at this stage

$$\Psi_c[x(t)y(t)] = x^2(t)\Psi_c[y(t)] + y^2(t)\Psi_c[x(t)] \tag{3.16}$$

The discrete version also has a similar property

$$\Psi_d[x(n)y(n)] = x^2(n)\Psi_d[y(n)] + y^2(n)\Psi_d[x(n)] - \Psi_d[x(n)]\Psi_d[y(n)] \tag{3.17}$$

Now using this property and applying the TEO to a general real-valued AM signal

$$x(t) = a(t)cos(\omega_c t + \theta) \tag{3.18}$$

we have

$$\Psi[x(t)] = \underbrace{a^2(t)\omega_c^2}_{D(t)} + \underbrace{\cos^2(\omega_c t + \theta)\Psi[a(t)]}_{E(t)} \tag{3.19}$$

Note that no subscript was used with Ψ since from now on both subscripts c and d will be dropped as it will be clear from the context whether the continuous or discrete-time Ψ is used.

Here we are interested in the envelope contained in $D(t)$ so $E(t)$ is viewed as an error term. If $a(t)$ is band limited with highest frequency w_a such that $w_a << w_c$ then the error term $E(t)$ will be negligible [13]. So Equation (3.19) becomes $\Psi[x(t)] \approx a^2(t)\omega_c^2$, which has the same form as when $a(t)$ is constant.

All the above examples provide some motivation to the use of the Teager Energy Operator in nonlinear speech processing. An Energy Separation Algorithm (ESA) was developed to demodulate the signal by tracking the physical energy implicit in the source producing the observed acoustic resonance signal and separating it into amplitude and frequency components [11, 13, 14]. In this case $w(t)$ and $a(t)$ can be estimated as follows

$$w(t) \approx \sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \quad ; \quad |a(t)| \approx \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \tag{3.20}$$

Simulation results show that the ESA performs pretty well and thus suggests that the Teager Energy Operator Ψ "contains" useful information about the speech signal. This is subject to modeling speech resonances as a superposition of AM-FM signals instead of the usual approach were the speech resonances are viewed as the poles of the transfer function of the linear model of the vocal tract. The information embedded in the Teager Energy Operator (TEO) can potentially outperform the traditional energy measures used in some speech recognition applications. In Chapter 4, we will use the TEO in sub-bands and construct feature vectors based on the Ψ-energy measure.

Chapter 4

The Teager Energy Operator

In this chapter, a new set of feature parameters for speech recognition is presented. The new feature set is obtained from the cepstral coefficients derived from the wavelet analysis of the speech signal or equivalently the multirate sub-band analysis. While in [17, 20] an ordinary energy measure is used to compute the cepstral coefficients, in this chapter a new energy measure based on the Teager Energy Operator (TEO) is described.

The performance of the new feature representation is compared to the Sub-band Cepstral coefficients (SUBCEP) in the presence of car noise. The new features are observed to be more robust than the SUBCEP parameters which were shown to outperform the commonly employed MELCEP representation [17, 20].

4.1 Properties of the Teager Energy Operator

As was mentioned in Section (3.2) the TEO is an efficient tool for nonlinear speech processing as long as the speech resonances are modeled as a superposition of AM-FM signals. In this chapter more attention will be paid to the discrete version of the TEO

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1). \tag{4.1}$$

Clearly, this operator makes successive samples exchange information between each other instead of being treated independently as in the commonly used instantaneous energy

$\xi[x(n)] = x^2(n)$[1]. Moreover, as will be discussed later, the Ψ-energy of the colored noise is negligible compared to that of the speech signal (especially in voiced frames) which makes it very efficient for speech recognition in noisy environments.

To be able to examine the properties of the TEO we calculate its mean, variance and the autocorrelation function in terms of the statistics of the original signal. Taking the expectation of $\Psi[x(n)]$ or simply $\Psi_x(n)$

$$\begin{aligned} E\{\Psi_x(n)\} &= E\{x^2(n)\} - E\{x(n+1)x(n-1)\} \\ &= R_x(0) - R_x(2) \end{aligned} \tag{4.2}$$

where $R_x(k)$ is the autocorrelation function of $x(n)$ defined by

$$R_x(k) = E\{x(n+k)x(n)\} \tag{4.3}$$

Similarly the autocorrelation function $R_\Psi(k)$ can be found

$$\begin{aligned} R_\Psi(k) &= E\{\Psi(n+k)\Psi(n)\} \\ &= E\{x^2(n+k)x^2(n)\} \\ &\quad - E\{x^2(n+k)x(n+1)x(n-1)\} \\ &\quad - E\{x^2(n)x(n+k+1)x(n+k-1)\} \\ &\quad + E\{x(n+1)x(n-1)x(n+k+1)x(n+k-1)\} \end{aligned} \tag{4.4}$$

To simplify the above equation we assume that $x(n)$ is a zero mean WSS jointly Gaussian random process. Hence using the Isserlis formula [31, 47] and after some computation $R_{\Psi_x}(k)$ is found to be

$$\begin{aligned} R_{\Psi_x}(k) &= [R_x^2(0) + R_x^2(2) - 2R_x(0)R_x(2)] \\ &\quad - 4R_x(k-1)R_x(k+1)] + [3R_x^2(k) + R_x(k+2)R_x(k-2) \end{aligned} \tag{4.5}$$

Consequently, the variance of $\Psi_x(k)$ is

$$Var\{\Psi_x(n)\} = 3R_x^2(0) - 4R_x^2(1) + R_x^2(2) \tag{4.6}$$

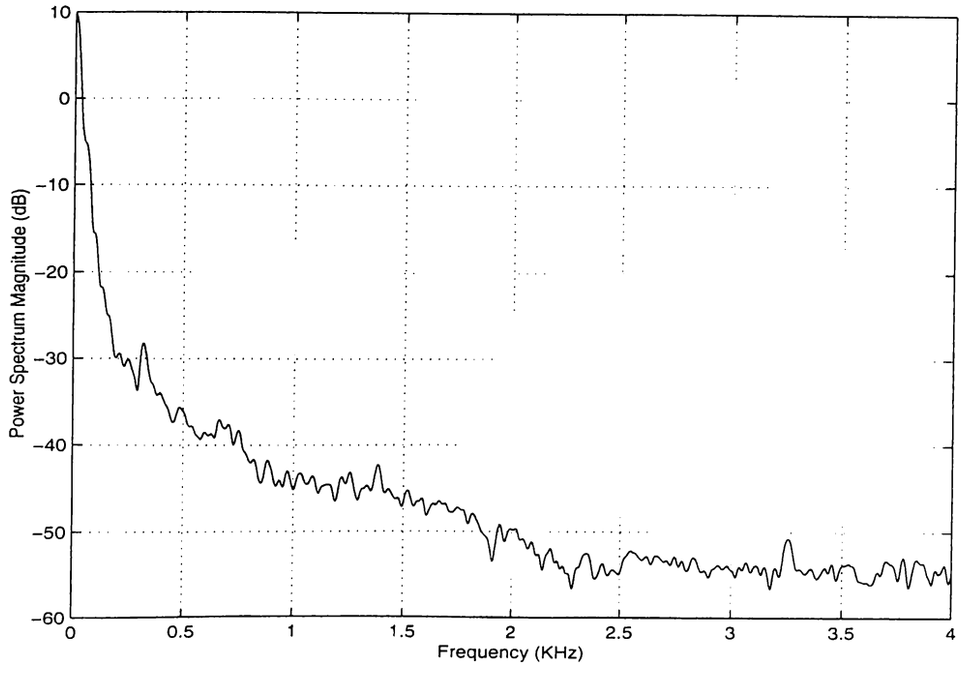

Figure 4.1: Power Spectrum Density of the car noise signal

4.2 Car Noise

Speech recognition applications inside a car can be important especially after the huge developments of the cellular mobile telephones. For instance, due to security reasons, voice dialing applications can be of great interest. Nonetheless, the noise coming from the car engine will greatly degrade the recognition performance unless perfectly handled in the initial system design.

The car noise is a colored noise were its spectrum is mostly concentrated in low frequencies as shown in Figure (4.1). Thus, its correlation function varies very smoothly and it is almost flat near the origin for several lags. Consequently, with little error, we can assume that $R_v(0) \approx R_v(1) \approx R_v(2)$. Experimental results show that

$$\begin{aligned} R_v(1) &= 0.9997\, R_v(0) \\ R_v(2) &= 0.9991\, R_v(0) \end{aligned} \tag{4.7}$$

[1] In this chapter, $\xi[.]$ is used to denote the instantaneous energy of a signal $x(n)$, $\xi[x(n)] = x^2(n)$

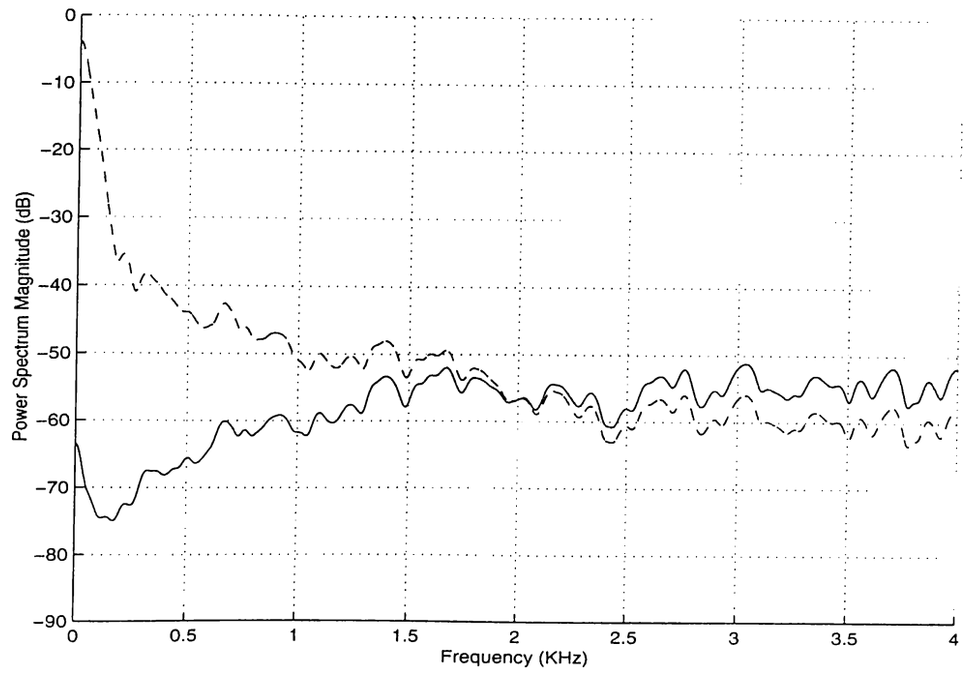

Figure 4.2: Spectrum of car noise energy $\xi[v(n)]$ (dashed line) and the spectrum of $\Psi[v(n)]$ (continuous line)

On the other hand for a speech signal $s(n)$ of the vowel /a/, we have

$$\begin{aligned} R_s(1) &= 0.7415\ R_s(0) \\ R_s(2) &= 0.4584\ R_s(0) \end{aligned} \tag{4.8}$$

For this property, the mean and variance of the Ψ-energy of the noise signal

$$\begin{aligned} E\{\Psi_v(n)\} &= R_v(0) - R_v(2) \\ Var\{\Psi_v(n)\} &= 3R_v^2(0) - 4R_v^2(1) + R_v^2(2) \end{aligned} \tag{4.9}$$

are negligible compared to the speech signal, if the resonance frequency of the latter falls within the frequency range of the current analysis band [12]. This property is experimentally verified using several noise recordings. In Figure (4.2), it can be seen that the high level of the Power Spectrum Density (PSD) in the low frequency ranges of the car noise is sustained by the PSD of $\xi[v(n)]$ whereas the TEO canceled it out and thus the PSD of $\Psi[v(n)]$ has an almost constant low level for the entire frequency range.

4.3 The Cross Ψ-Energy

Let $x(n)$ and $y(n)$ be two discrete signals, their cross Ψ-energy is defined as

$$\tilde{\Psi}[x(n), y(n)] = x(n)y(n) - \frac{1}{2}[x(n+1)y(n-1) + x(n-1)y(n+1)] \tag{4.10}$$

Obviously we have $\tilde{\Psi}[x(n), x(n)] = \Psi[x(n)]\tilde{\Psi}[x(n), x(n)]$.

Noting that the cross-correlation function of $x(n)$ and $y(n)$, which are supposed to be jointly WSS, is $R_{xy}(k) = E\{x(n+k)y(n)\}$, we can study the statistics of the cross Ψ-energy.

The expected value of $\tilde{\Psi}[x(n), y(n)]$ is

$$E\{\tilde{\Psi}[x(n), y(n)]\} = R_{xy}(0) - R_{xy}(2) \tag{4.11}$$

and the variance is

$$\begin{aligned} Var\{\tilde{\Psi}[x(n), y(n)]\} &= \tfrac{1}{2}[3R_x(0)R_y(0) - 4R_x(1)R_y(1) + R_x(2)R_y(2)] \\ &\quad + \tfrac{1}{2}[3R_{xy}^2(0) - 4R_{xy}^2(1) + R_{xy}^2(2)] \end{aligned} \tag{4.12}$$

Note that if $x(n)$ and $y(n)$ were independent and zero mean then we would have

$$E\{\tilde{\Psi}[x(n), y(n)]\} = 0 \tag{4.13}$$

$$\mathrm{Cov}\{\Psi[x(n)], \tilde{\Psi}[x(n), y(n)]\} = \mathrm{Cov}\{\Psi[y(n)], \tilde{\Psi}[x(n), y(n)]\} = 0 \tag{4.14}$$

The reason for defining this operator is just to facilitate the analysis as it does not have an important physical significance. Namely, it helps handling the cross terms in the computation of the Ψ energy of the sum of two signals because of the nonlinearity of the TEO. This energy is found to be

$$\Psi[x(n) + y(n)] = \Psi[x(n)] + \Psi[y(n)] + 2\tilde{\Psi}[x(n), y(n)] \tag{4.15}$$

More properties of the cross Ψ-energy are discussed in Appendix A.

4.4 Speech in car noise

Now suppose that $x(n) = s(n) + v(n)$ where $s(n)$ is the noise free speech signal and $v(n)$ is a colored zero mean additive noise (car noise for instance). Since $v(n)$ and $s(n)$ are

supposed to be independent, the autocorrelation function of $x(n)$ is

$$R_x(k) = R_s(k) + R_v(k) \tag{4.16}$$

Now since

$$\Psi[x(n)] = \Psi[s(n)] + \Psi[v(n)] + 2\tilde{\Psi}[s(n), v(n)] \tag{4.17}$$

then based on the assumptions discussed in Section (4.2) about the car noise, and using the properties of the Ψ and cross-Ψ energies, we have

$$\begin{aligned} E\{\Psi_x(n)\} &= E\{\Psi_s(n)\} + E\{\Psi_v(n)\} \\ &\approx E\{\Psi_s(n)\} \end{aligned} \tag{4.18}$$

Equation (4.18) shows that the noise bias is negligible. Similarly, the variance of $\Psi[x(n)]$ is

$$\begin{aligned} Var\{\Psi_x\} &= Var\{\Psi_s\} + Var\{\Psi_v\} + 4Var\{\tilde{\Psi}[s, v]\} \\ &= [3R_s^2(0) + R_s^2(2) - 4R_s^2(1)] \\ &+ [3R_v^2(0) + R_v^2(2) - 4R_v^2(1)] \\ &+ 2[3R_s(0)R_v(0) + R_s(2)R_v(2) - 4R_s(1)R_v(1)] \end{aligned} \tag{4.19}$$

The first term of Equation (4.19) is related to the speech signal. The second term, $Var\{\Psi_v\}$ would be negligible as discussed in the previous section and can be considered as a small error term. However the cross terms at the end which are due to the nonlinear nature of the TEO, do not cancel out. Their effect, however, is reduced compared to the case when $\xi[x(n)]$ is used.

The expected value and variance of $\xi_x(n) = x^2(n)$ are as follows

$$E\{\xi_x(n)\} = R_x(0) = R_s(0) + R_v(0) \tag{4.20}$$

$$Var\{\xi_x(n)\} = 3R_x^2(0) = 3[R_s^2(0) + R_v^2(0) + 2R_s(0)R_v(0)] \tag{4.21}$$

As the noise energy increases so does the total energy of the feature parameter which is not a desirable phenomenon because it reduces the recognition quality.

This result can be checked by considering a simple example. Since we use a filter bank which divides the speech signal into narrowbands we can assume that there is only one formant within one band and consequently the noise free speech signal $s(n)$ can be assumed to be of the following simple form:

$$s(n) = a\cos(\Omega n + \theta) \tag{4.22}$$

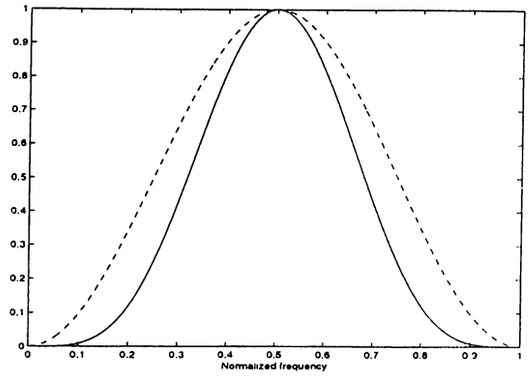

Figure 4.3: Plot of the function $f(\Omega) = \frac{a^4}{4}[3 + \cos^2(2\Omega) - 4\cos^2\Omega]$ (continuous line), and $g(\Omega) = a^2 \sin^2\Omega$ (dashed line) for $\Omega \in [0, \pi]$ with $a = 1$.

where $\Omega \in [0, \pi]$ and a are a constant amplitude and frequency, respectively. The parameter $\boldsymbol{\theta}$ is a random phase with a uniform distribution, $\boldsymbol{\theta} \sim \mathcal{U}[-\pi, \pi]$. Thus we have $R_s(k) = \frac{a^2}{2}\cos(\Omega k)$. Consequently,

$$\begin{aligned} E\{\Psi_s\} &= a^2 \sin^2\Omega \\ Var\{\Psi_s\} &= \tfrac{a^4}{4}[3 + \cos^2(2\Omega) - 4\cos^2\Omega] \end{aligned} \tag{4.23}$$

The curves of $E\{\Psi_s\}$ and $Var\{\Psi_s\}$ are shown in Figure (4.3) as a function of Ω for $a = 1$. Both functions have a bell-shaped curve with a wide band-width.

Now consider a noisy signal $x(n) = s(n) + v(n)$ so we have

$$\Psi[s(n) + v(n)] = \Psi[s(n)] + \Psi[v(n)] + 2\tilde{\Psi}[s(n), v(n)] \tag{4.24}$$

where $v(n)$ is a zero mean car noise having the properties discussed in Section (4.2). Let us select $E\{v^2(n)\} = \sigma^2$ and $a = \sqrt{2}\sigma$ in order to make the signal to noise ratio be SNR = 0 dB. Also, assume that $R_v(2) \approx R_v(1) \approx R_v(0) = \sigma^2$. As discussed in previous sections of this chapter,

$$E\{\tilde{\Psi}[s(n), v(n)]\} = 0 \tag{4.25}$$

and

$$E\{\Psi[v(n)]\} \approx 0 \tag{4.26}$$

so that

$$E\{\Psi[s(n) + v(n)]\} \approx E\{\Psi[s(n)]\} \tag{4.27}$$

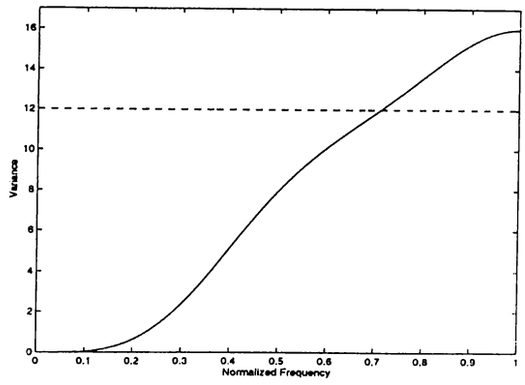

Figure 4.4: $Var\{\Psi_x\}$ (continuous line) and $Var\{\xi_x\}$ (dashed line) in function of Ω for $\Omega \in [0, \pi]$. Here $\sigma^2 = 1$

For the variance we have

$$\begin{aligned} Var\{\Psi_x\} &\approx Var\{\Psi_s\} + 4Var\{\Psi[s(n), v(n)\} \\ &= \sigma^4[6 + (\cos^2(2\Omega) + \cos(2\Omega)) - 4(\cos^2\Omega + \cos\Omega)] \end{aligned} \tag{4.28}$$

Similarly for $\xi\{s(n) + v(n)\}$

$$\begin{aligned} E\{\xi_x\} &= E\{\xi_s\} + \sigma^2 \\ Var\{\xi_x\} &= 3(\tfrac{a^4}{4} + \sigma^4 + a^2\sigma^2) \\ &= 12\sigma^4 \end{aligned} \tag{4.29}$$

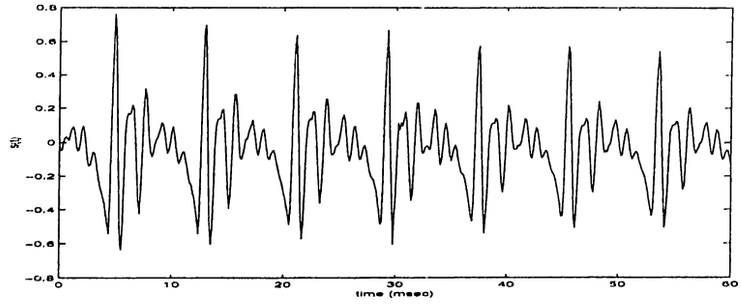

Figure 4.5: Plot of 60 msec of the vowel /a/

From Figure (4.4) it can be seen that the variance of Ψ_x is smaller than that of ξ_x for most of the frequency spectrum. The difference is clearly seen especially at low

frequency bands were the noise energy is mostly concentrated as shown in Figure (4.2). At 0.7π angular frequency the variance of Ψ_x starts to exceed that of ξ_x. However at this frequency the speech formants have a relatively low amplitude for most of the phonemes and they usually do not have a significant influence on the discrimination between the phonemes [5,6]. In the case of the colored car noise, the noise power is negligible at such high frequencies. It can also be noticed that $E\{\xi_x\}$ has a bias term proportional to the noise energy whereas this is not the case for $\Psi[x(n)]$.

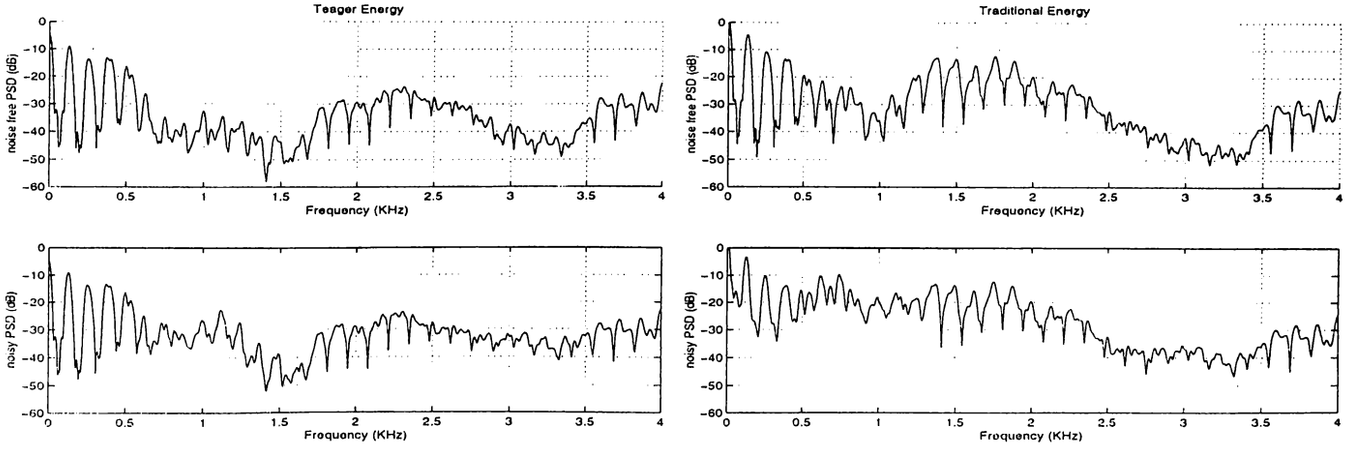

Figure 4.6: Power spectrum of the Ψ-energy (left) and ξ-energy (right) of the vowel /a/ in noise free (upper plot) and noisy (bottom plot) conditions with SNR= 0dB.

To verify this result for a real speech signal, the spectrum of $\Psi[s(n)]$, where $s(n)$ is a noise free recording of the vowel /a/, is shown in the upper plot of Figure (4.6). The signal $s(n)$ is corrupted by car noise at 0 dB SNR level. The spectrum of the resulting Ψ-energy is shown in the bottom plot of the same figure. The difference between both spectra is negligible compared to the relatively big difference shown in the plots on the right hand side of Figure (4.6). These plots are obtained by performing another experiment, under the same conditions, where the common ξ-energy was used. It can be seen that, in this case, the spectrum was largely affected especially at frequencies around 1kHz. In Figure (4.7) the same power spectrum densities are shown for just one tenth of the frequency range for zooming purposes. The noisy spectrum, shown in dashed line, almost overlaps the noise free spectrum of the Ψ-energy. However, this is not the case with the spectra of the ξ-energy.

The same experiment is performed on the unvoiced fricative (/s/) at SNR=-5 dB. Since the unvoiced speech power is low, the gain achieved by using the Ψ-energy is also

low compared to the gain achieved with voiced speech. However, while the ξ-energy fails to cancel the effect of noise at low frequencies, the Ψ-energy reduces this effect as seen in Figure (4.8). Note that both energy measures fail to cancel the effect of noise at high frequency bands because of the lack of high frequency harmonics in the speech signal.

Finally, we mention that in case of white noise, both energy measures have the same performance because $R_s(k) = 0$ for $k \neq 0$. Actually, in both cases the recognition degrades quickly as the SNR gets smaller. In Section (4.6) simulation studies in both car noise and white noise environments will be presented. The performance of the TEOCEP features, which are defined in Section (4.5), is compared to the performance of SUBCEP's, introduced in Section (3.1). In the SUBCEP's the traditional energy measure is used.

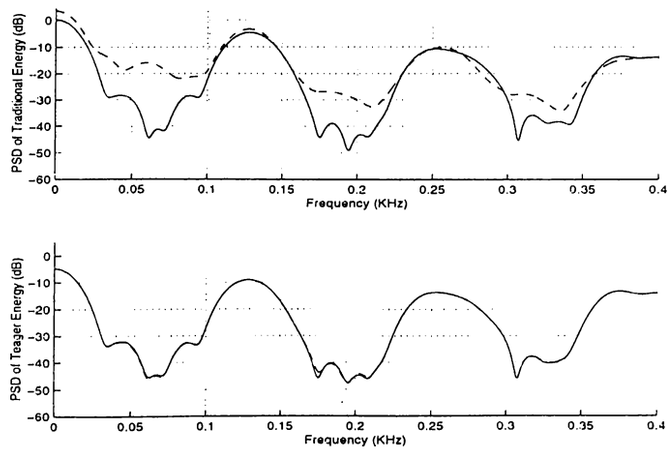

Figure 4.7: Power spectrum of the ξ-energy (up) and Ψ-energy (down) of the vowel /a/ in noise free (continuous line) and noisy (dashed line) conditions with SNR=0 dB. Just the first 1/10 of the spectrum is shown here.

4.5 The TEOCEP Feature Vector

The TEOCEP features just differ from the SUBCEP's [17, 20] in the energy measure used. The feature extraction procedure is actually introduced in Section (3.1) but it is repeated here for convenience.

Subband analysis is used to decompose the speech signal of the current frame into $L = 21$ sub-signals, $s_l(n)$ for $l = 1, \ldots, L$. For every sub-signal, the average Ψ-energy e_l is found

$$e_l = \frac{1}{N_l} | \sum_{n=1}^{N_l} s_l(n)^2 - s_l(n-1)s_l(n+1) | \quad ; \quad l = 1, \ldots, L \tag{4.30}$$

where N_l is the number of samples in the l^{th} band.

Although it is possible that the instantaneous Ψ-energy have negative values in very rare circumstances, the average value e_l is a positive quantity [13]. Nonetheless, the magnitude of the Ψ energy is used to be sure that e_l is positive.

Log compression and DCT transformation is then applied to obtain the TEO based cepstrum coefficients or TEOCEP's.

$$TC(k) = \sum_{l=1}^{L} \log(e_l) \cos[\frac{k(l-0.5)\pi}{L}] \quad ; \quad k = 1, \ldots, N. \tag{4.31}$$

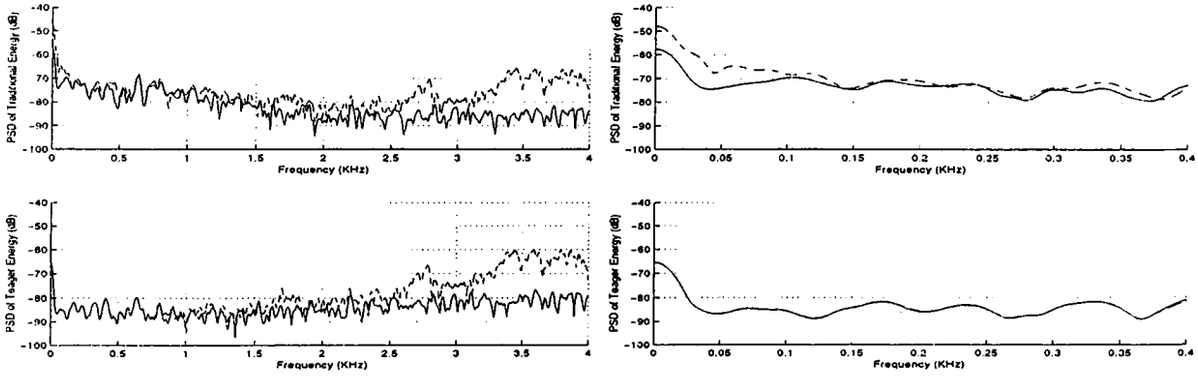

Figure 4.8: Power spectrum of the ξ-energy (up) and Ψ-energy (down) of unvoiced phoneme /s/ in noise free (continuous line) and noisy (dotted line) conditions with SNR=-5 dB. The plots on the right hand side show the same spectra zoomed to the frequency range 0 Hz to 500 Hz

the first 12 TEO cepstrum coefficients are used to form the feature vector. Twelve more coefficients obtained from the first-order differentials are also appended. A final feature vector with dimension 24 is obtained and is used for training and recognition.

4.6 Simulation Results

A continuous density Hidden Markov Model based speech recognition system with 5 states and 3 mixture densities is used in simulation studies. The recognition performances of the TEOCEP feature parameters are evaluated 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 *TI-20* vocabulary consists of ten English digits $(0, 1, \ldots, 9)$ and ten control words ("enter", "erase", "go", "help", "no", "rubout", "repeat", "stop", start", "yes"). The data is collected from 8 male and 8 female speakers. There are 26 utterances of each word from each speaker, where 10 designated as training tokens and 16 designated as testing tokens. The data was recorded in a low noise sound isolation booth, using an Electro-Voice RE-16 cardoid dynamic microphone, positioned two inches from the speaker's mouth and out of the breath stream.

The speech signal is corrupted by additive car noise at various SNR levels, with car noise. The noise recoding was obtained inside a Volvo 340 on a rainy asphalt road by the *Institute for Perception-TNO, The Netherlands.* Simulation results in white noise environments are also presented.

The filter bank of Figure (3.1) is applied to the speech signal in a tree structured manner as shown in Figure (3.4) to achieve the sub-band decomposition shown in Figure 3.3. However to get the correct frequency resolution switching of the basic unit is done at the output of every high-pass filter as shown in Figure 3.2. A decomposition of $L = 21$ sub-signals is achieved eventually. The window size is chosen as 48 msec with an overlap of 32 msec so that the sub-signal with the smallest sub-band has 12 samples at 16 kHz sampling rate. The final feature vector is constructed from the **TEOCEP** parameters and their time derivatives as explained in Section (4.5).

The SUBCEP parameters are also extracted using the log version as shown in Equation (3.4) rather than the root cepstral version of Equation (3.5) for comparison purposes

because the Log compression is used in the new TEO cepstrum coefficients. The time derivatives of the SUBCEPS are added also to the feature vector, Section (3.1).

Table 4.1: The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and TEOCEP representations for various SNR levels with Volvo noise recording.

SNR (dB)	TEOCEP	SUBCEP
30	99.66	99.15
10	99.26	99.05
7	99.37	97.98
5	99.05	97.02
3	98.84	96.41
0	98.17	95.14
-3	97.83	93.12
-5	96.86	90.62

Table 4.2: The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and TEOCEP representations for various SNR levels with white noise.

SNR (dB)	TEOCEP	SUBCEP
20	97.79	98.37
10	87.07	87.7
7	86.12	85.17
5	82.97	81.70
3	79.83	79.50

Speaker dependent recognition performance is presented in Table (4.1). The models of the vocabulary are obtained from the training tokens of each speaker and evaluation is done with the testing tokens of the same speaker. The average recognition rates are shown in Table (4.1). Each row of Table (4.1) represents the averaged recognition rate for the indicated SNR value, where the original (noise free) recording of the database has a 30 dB SNR. All of the recognition rates are obtained according to the training at 30 dB SNR level. The new feature parameters TEOCEP showed a more robust performance with car noise than the SUBCEP features presented in [20]and [25]. The gain in the recognition rate becomes more important at low SNR values.

In the case of white noise the TEOCEP's slightly outperform the SUBCEP's only

Table 4.3: The average recognition rates of speaker independent isolated word recognition system with SUBCEP and TEOCEP representations for various SNR levels with Volvo noise recording.

SNR (dB)	TEOCEP	SUBCEP
30	91.22	91.25
10	91.13	90.96
7	90.74	89.94
3	89.10	88.40
0	87.13	86.63
-3	85.26	80.17

at SNR$<$ 7. Both feature sets fail to resist to high white noise corruption and the recognition rate gets very low as the SNR decreases as shown in Table (4.2).

The TEOCEP's are also checked in a speaker independent speech recognition case. The utterances of five men and five women were used for training. The utterances of the rest speakers are used to test the performance of the system. The simulation results are shown in Table (4.3). Here also, the TEOCEP's show better performance than the SUBCEP's in the presence of car noise.

4.7 Other TEO-based feature parameters

Other simulation studies are also carried out using different feature parameters based on TEO. These parameters are obtained by mixing the sub-cepstrum coefficients and the teo-cepstrum coefficients in various manners in order to improve the recognition performance. Two sets of features are investigated.

The first one, which we refer to by TEOSUB1, are obtained using the first twelve TEOCEP coefficients and the first order differential of the first twelve SUBCEP coefficients. The second set, called TEOSUB2, is formed by log compressing the 21 average Teager energies of the subband signals. and appending three subband energies (after log compression) corresponding to the third, fourth and fifth subbands. A feature vector is directly formed by using these 24 coefficients without inverse DCT computation.

Speaker dependent simulations in Volvo car noise are performed. Table (4.4) show

Table 4.4: The average recognition rates of speaker dependent isolated word recognition system with TEOSUB1 and TEOSUB2 features for various SNR levels with Volvo noise recording.

SNR (dB)	TEOSUB1	TEOSUB2
30	95.33	93.12
10	93.72	91.44
5	91.85	90.3
3	90.04	88.21
0	88.31	86.62
-3	86.14	83.10

the recognition rates for these features with various SNR values. Their recognition performance is unfortunately not as good as the TEOCEP's or the SUBCEP's. They can perhaps provide better results, if they are further studied and improved.

4.8 Conclusion

In this chapter, new feature parameters for speech recognition are introduced. The new features are based on the Teager Energy Operator and the multirate sub-band analysis providing a robust recognition performance under car noise. The performance of the new features are compared to the performance of the SUBCEP's introduced in [20] and are shown to give better recognition results. The achieved improvment is mainly significant at low SNR levels which is the case of real life applications. A typical SNR inside a running car is about 0 dB, so the new feature set is expected to provide an important improvment in real applications. The Teager Energy is potentially able to provide robust distance measure for endpoint detection because the teager energy of the noise is very low compared to that of the speech region.

Chapter 5

Large Vocabulary Speech Recognition

The problem of large vocabulary speech recognition does not normally differ from the small vocabulary recognition problem. It is still possible to model each word with a different finite state HMM model. However, as the vocabulary size increases the memory requirements and the processing cost increase. Thus, other modeling techniques are proposed in literature and most of them try to use sub-word modeling such as syllables, phonemes or triphones [5–7]. A triphone is an efficient sub-word for speech recognition because it models the phoneme together with the phonetic context it appears in.

5.1 Triphone-Based Markov Models

An utterance is theoretically formed by a collection of finite mutually exclusive sounds. Once a speaker has formed a thought to be communicated to a listener, he makes up a word accordingly from this collection of sounds already existing in his memory. The basic theoretical unit for describing how speech conveys linguistic meaning is called a phoneme. In most languages there are about forty phonemes on the average. Each phoneme can be considered as a code that consists of a unique set of articulatory gestures. These

articulatory gestures include the type and location of sound excitation as well as the position or movement of the vocal tract articulators.

In speech processing, many recognition systems choose phonemes as the smallest unit for recognition. Theoretically, a single state in the HMM model is assigned to each phoneme. In reality, however, we shall clearly distinguish between the theoretical phoneme and the actual sound produced, the phone. In addition to the coarticulatory effects, the phoneme will sound differently according to the context in which it appears. In other words, the previous and next phonemes affect the current phoneme significantly. So modeling the word phones should give better results than trying to model the phoneme independently of its phonetic context. A good phonetic discrimination can be achieved by modeling each phoneme together with its left and right neighbors which makes what is usually referred to as a triphone. For instance suppose a phoneme /y/ appears with the phonemes /x/ on the left and /z/ on the right. The corresponding triphone for /y/ is /x-y+z/ [7]. The word "*speak*", for instance, is represented with the following triphones

/sil-s+p/ /s-p+iy/ /p-iy+k/ /iy-k+sil/

where */sil/* stands for silence.

The triphone model is successfully used in many speech recognition systems including the Hidden Markov Model Tool Kit (HTK). The HTK is a triphone-based speech recognition system designed by Young for the *Entropic Research Company* [7].

5.2 Recognition

In triphone based speech recognition systems, a word lexicon is required to provide a transcription of every word in the vocabulary in terms of its triphone sub-words. That is, this lexicon provides the pronunciation of each word.

Suppose we have a vocabulary of N words W_i for $i = 1, \ldots, N$, where each word consists of p_i triphones

$$W_i = U_{i1} U_{i2} \ldots U_{ip_i} \tag{5.1}$$

At this point, we assume that for each triphone a three-state HMM model has been computed in the training phase. The probability of staying in the last state is not one as

this will be the case just for triphones occuring at the end of a word. The model of each word is formed by cascading the models of its triphones together to obtain a composite model or a composite finite-state network as shown in Figure (5.2) for the Turkish word "*bir*". Hence, for every word W_i a $3p_i$-state model is obtained. The correct word is eventually selected using Equation (2.5) with the Viterbi algorithm for example.

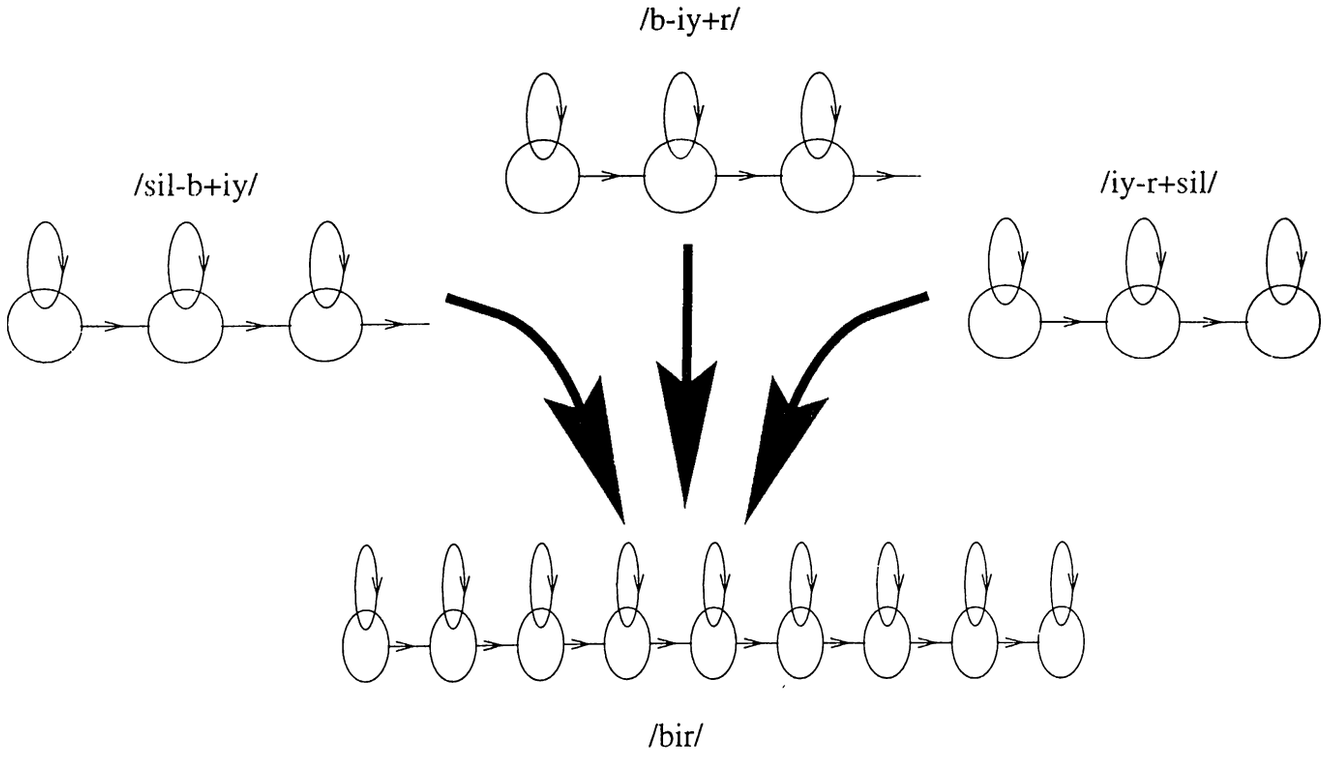

Figure 5.1: Cascading the HMM models of the three triphones forming the Turkish word "*bir*" to form the final word model.

5.3 The Best State Sequence

The Viterbi algorithm provides, in addition to computing the likelihood, the best state sequence followed by the observation vectors. So it can provide information about the best distribution of the feature vectors around the different states of the model. Since every three states of the model correspond to a unique triphone, the best state sequence can offer information about the phonetic segmentation of the available utterance.

The Viterbi algorithm is a kind of Dynamic Programming solution to the optimization problem. It solves the problem of speech recognition by finding the best path (state sequence) with the *highest* probability. This property will be very useful in the training process which will be discussed in the next section.

5.4 The Training Problem

Usually, in order to train the different words in the vocabulary, a fixed model topology is assigned to all of them. Surely, the model parameters differ from one word to the other. These parameters are approximated using one of the available training solutions. However, in the approach adopted in this thesis, the size of the word model is fixed according to a lexical information fed to the training system. Therefore, every word is assigned its own HMM model with a different number of states. As discussed previously, every three successive states correspond to one triphone so the number of states is decided accordingly. On the other hand, the number of mixtures per state is kept fixed along all the states for all the words.

Once the model topology is defined, the training process is performed to approximate the parameters of Equations (2.10) and (2.11) which completely define the word model. Eventually, all the parameters of every three successive states are stored independently in memory. As explained in Section (5.3), each of these three states correspond to a specific triphone. Therefore in the recognition process words sharing the same triphone also share the same model corresponding to the common triphone.

All training algorithms actually need to start with an initial guess. A random guess is a possible choice but it may lead to a local maximum solution. A better guess is to distribute the feature vectors uniformly on the states and use the K-means clustering algorithm to classify the vectors according to the number of mixtures used. Some improvement can be achieved if instead of the uniform distribution, more vectors are assigned to the states corresponding to vowels and semi-vowels than those corresponding to the plosive consonants or the weak fricatives for instance. There is no additional cost in doing so as the lexical information is already available.

If the word being trained has some triphones which are already trained by other

words, then their models will exist in memory. These models are used by the corresponding three states as an initial point, and their parameters are further adjusted by the new word. With this approach some words will have a bigger "influence" on the shared triphones because of the initialization strategy. Simulation results show that some of the words are completely dominated by other words and during the recognition process these words will almost always be wrongly declared to be their dominant words. A straight forward solution to this problem is to perform several training sessions instead of just one where each session deals with all the words from the beginning. Thus, starting from the second training session, all the triphones will have an initial starting point already stored in memory. Moreover, changing the training order of words from one session to the other can help improving the recognition rate. The simulation results show that the training process converges just after a few sessions.

5.5 The Subvocabulary Based Search Strategy

As the vocabulary size increases, the processing time to recognize a word also increases. It can reach a level which makes the recognition system very slow and hence impractical. Usually language models are used to reduce the search field [7]. The purpose of the language model is to provide a mechanism for estimating the probability of some word, w_k, in an utterance given the preceding words $W_1^{k-1} = w_1 \dots w_{k-1}$. To do this, an efficient way is to use N-grams, in which it is assumed that w_k depends only on the preceding $N-1$ words.

Nonetheless, it is desirable that the search field is reduced regardless of the previous words in the utterance. This can be done by observing that any large vocabulary contains many words sharing similar phonetic information and some words differ just with one or two phonemes. Also, these words cause frequent confusion for the recognition system and are usually the reason of recognition errors, as expected. This suggests that these phonetically similar words be grouped in some manner and a pre-classification process is performed before searching into the determined group. For every group, a shared HMM model, trained by words of the same group, can be found. However, inaccurate models are achieved if the classification procedure is not restricted. Restricted classification, on the other hand, leads to a large number of groups which makes the computation time gain insignificant. In addition, since the vocabulary size is large, an automatic

classification of the words is needed.

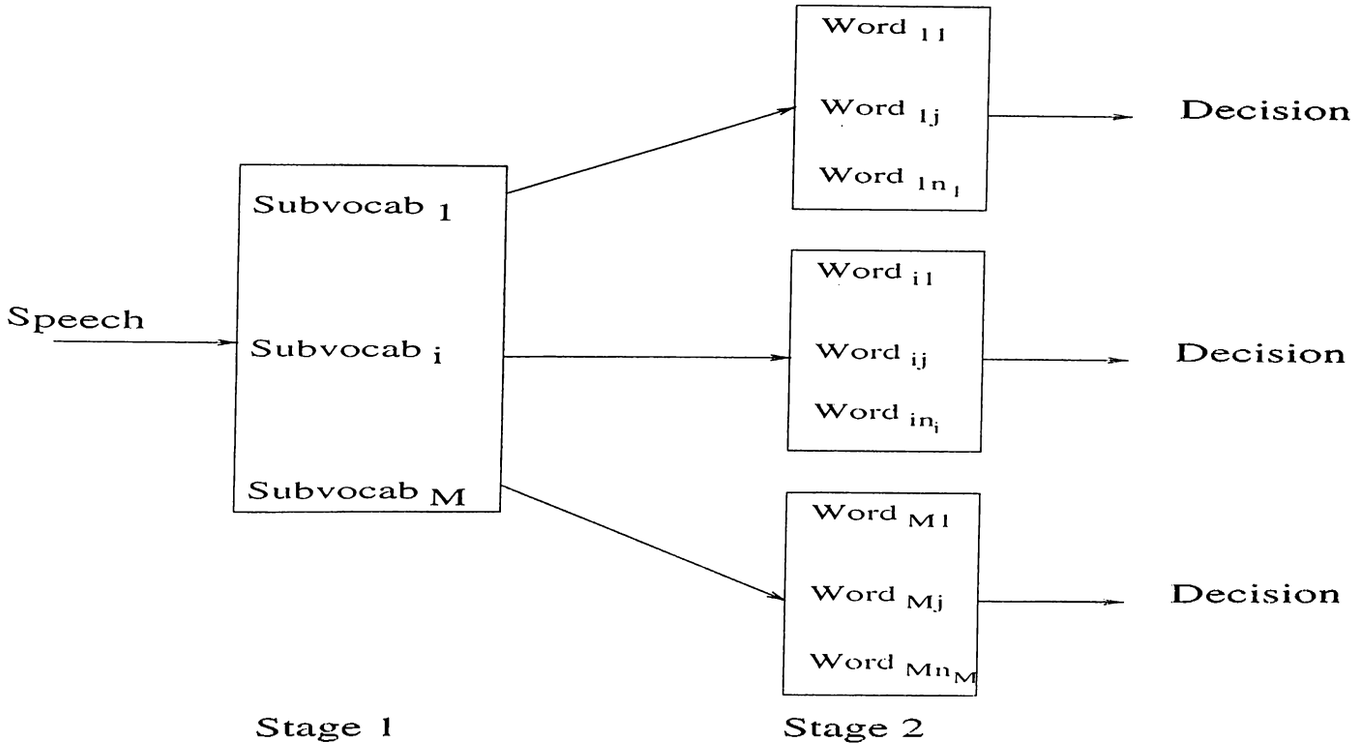

Figure 5.2: A two stage subvocabulary search strategy.

Assume that there is no "junk" word in the vocabulary. For example, let the word "mad", which is not in the vocabulary, be uttered. The system, in this case, selects a word which is the phonetically most similar to "mad". For instance, it can select the word "bad". The key idea here is to form subvocabularies in order to reduce the computational cost of the recognition problem. In our classification approach, each subvocabulary is presented by a group head which is a member of the subvocabulary itself.

The recognition process is carried out in two stages as shown in Figure (5.5). In the first stage, a reduced vocabulary containing the group heads, is searched. Since the reduced vocabulary just contains the group heads, all the other words should be recognized to be the group head of the vocabulary they belong to. According to the result of the first stage, the search subvocabulary of the second stage is determined. The final word is found in the subvocabulary used in the second stage. The number of recognition stages can be more than two according to the size of the vocabulary and the phonetic properties of the words. Note that within every stage, the parallel search strategy shown in Figure (2.1.2) is carried out using the available subvocabulary.

1) Initialization
    1.1 Choose the group heads
    1.2 Choose an initial classification
2) Use the rank ordered probabilities to compute the new recognition rate
3) While the recognition rate is not sufficiently close to the maximum possible
    ⋆ For all words in database
        3.1 Remove it from its current group(s)
        3.2 Assign it to another group if a recognition improvement occurs
        3.3 If no improvement, don't make any changes
    ⋆ For all groups
        3.4 Change the group head if it leads to an improvement
    ⋆ Compute the new recognition rate

Table 5.1: The classification Algorithm

The classification of the words into subvocabularies is made automatically using the algorithm described in Table (5.1). After computing the triphone model parameters using the training database, a recognition verification is performed. However, the same utterances used for training are used for recognition. The recognition rate is not important here since it would be as high as 100% . What is important is to keep track of the descending order of the vocabulary words according to their probabilities. The rank ordered words for every utterance are then stored in memory. Correspondingly, the recognition decision will be the first word in that rank ordered word sequence. If that word is omitted from the vocabulary book then the following word (which has the next highest probability) is chosen. Again, if this second word is also omitted, then the system chooses the third one etc.

The algorithm in Table (5.1) benefits from this decision logic and uses the rank ordered words to classify the vocabulary words in the most optimum way subject to the total recognition error. Step (1.1) can be implemented by choosing the group heads arbitrarily. Since this can lead to a local optimum, choosing the group heads manually according to some phonetic logic can give better results. This is not a difficult task because the number groups is usually small relative to the vocabulary size. Just the group heads are sufficient to initialize the algorithm, so Step (1.2) is not necessary though it can improve the solution. This algorithm is an iterative method which can converge to a local minimum. Therefore, when the algorithm stops before some previously defined threshold value, the group heads can randomly be substituted by a word from the same

subvocabulary and the iterations restart. This is done even if the change may result in an increase in the number of errors. As the number of utterances per word increase, the classification becomes more robust and handles more possible scenarios that can happen in application involving recordings outside the classification data.

5.6 Simulation Results

Two databases are used to train and to evaluate the performance of the system. The first is used for training where three utterances of every word were recorded. The second database is used to test the system and has ten utterances per word. The vocabulary contains 100 Turkish words containing the ten digits $(0,\ldots,9)$, the words *"evet"*(yes) and *"hayır"*(no) and 88 words chosen from the Turkish stock market list.

The two databases were recorded with background noise such that the SNR$\approx$ 25 dB. The recordings were sampled at 8 kHz and the window size is chosen as 32 msec with an overlap of 8 msec. Both the SUBCEP features in Equation (3.4) and the TEOCEP features were used to model the speech signals. A three-state triphone based HMM model with 3 mixtures per state was used to model the vocabulary words. The experiments are performed for a speaker dependent case.

Training Session	SUBCEP	TEOCEP
1st	94.6	.95
2nd	96.4	96.6
3rd	96.6	96.8

Table 5.2: Recognition rates for Stock Market Database after several training sessions

As can be seen in Table (5.2) the recognition rate increases as the number of the training sessions increase. However starting from the third session, the improvement in the recognition performance is not significant. Also note that there is not much difference between the use of TEOCEP's or SUBCEP's since the SNR is high in this case.

Table (5.3) shows the recognition rates achieved when the algorithm in Table (5.1) is used to classify the words into subvocabularies. The triphone models used here are obtained after three training sessions. The row "Parallel Search" shows the recognition rates obtained via a parallel search strategy through the whole vocabulary. The second

row shows the recognition rates of the subvocabulary based search strategy. The third row shows the percentage of the computation gain achieved by the subvocabulary based search versus the parallel search.

To define the computation gain, we assume that there are M subvocabularies obtained from a total number of N words. The i^{th} subvocabulary contains n_i words, for $i = 1, \ldots, M$. To recognize a word, the M grouped heads will be searched plus the number of words in the determined subvocabulary. So the number of words to be searched is on the average $\overline{n} = M + \frac{1}{M} \sum_{i=1}^{M} n_i$.

The computation Gain is then defined to be

$$Gain = \frac{N - \overline{n}}{N} \times 100 \tag{5.2}$$

In Table (5.3), it can be seen that a gain of around 70% is achieved by both feature sets with almost similar recognition rates. The TEOCEP's are slightly better than the SUBCEP's in terms of recognition rate and average processing time.

	SUBCEP	TEOCEP
Parallel Search	96.6	96.8
Subvocab. Search	96.4	96.5
Computation Gain	69.61	71.1

Table 5.3: The recognition performance : the parallel search versus the subvocabulary based search

5.7 Conclusion

In this chapter, a large vocabulary isolated word speech recognition system is designed. Triphone based HMM's are used to model the vocabulary words. With these models, a 96.6% recognition rate is achieved using the TEOCEP feature parameters. A novel tree structure search strategy is used to reduce the processing time. The recognition performance degrades by just 0.2 % , but the the gain in processing time is approximated by about 71.1% . The techniques described here are also potentially good candidates to present solutions to the continuous speech recognition problem.

Chapter 6

Conclusion

In this thesis a large vocabulary speech recognition system in car noise environments proposed.

A new set of speech feature parameters, TEOCEP's, are introduced. These parameters are based on wavelet analysis or equivalently the multirate subband analysis of the speech signal. The frequency domain is first divided into many nonuniform subbands determined according to the *mel*-scale division which imitates the human auditory perception system. For each of the subsignals obtained, the average Ψ-energy based on the Teager Energy Operator (TEO) is computed. Logarithmic compression and cosine transformation is applied to these average energy values to form what we call the TEOCEP's. Simulation results show that the new feature set outperforms the SUBCEP's which are derived from the traditional energies in colored car noise environments. Experiments performed in white noise conditions show no advantage of the TEOCEP's over the SUBCEP's. It is also theoretically shown that TEOCEP's do not provide any improvement compared to SUBCEP's in white noise.

These new features are used to build a large vocabulary isolated word speech recognition system. Triphone Hidden Markov Models (HMM) are used to model the vocabulary words. In the thesis, the proposed algorithm was described in detail, and simulations show satisfactory recognition performance. In order to reduce the processing time, the search field is reduced by classifying the vocabulary words into subvocabularies according to their phonetic content. The recognition process is carried out in two stages. The first

determines the subvocabulary to be used in the second stage using the group heads. The second stage finds the solution of the recognition problem from the reduced subvocabulary search group. An algorithm to provide a good classification solution is described. The gain in the recognition time is more than 70% on the average with just a slight degradation in the recognition performance. This approach provides a good solution to the problem and can handle vocabularies of any size.

The features and algorithms presented in this thesis can be extended to cover large vocabulary continuous speech recognition systems which are the ultimate target of the speech recognition technology.

Appendix A

Properties of the cross-Ψ energy

In this appendix some of the properties of the cross Ψ-energy are presented. For convenience it is worth mentioning the definition of the cross-Ψ energy here. Given two signals $x(n)$ and $y(n)$, their cross-Ψ energy is defined as

$$\tilde{\Psi}[x(n), y(n)] = x(n)y(n) - \frac{1}{2}[x(n+1)y(n-1) + x(n-1)y(n+1)] \tag{A.1}$$

We have then the following properties

Property 1 :
Given three signals $x_1(n)$, $x_2(n)$ and y(n)

$$\tilde{\Psi}[x_1(n) + x_2(n), y(n)] = \tilde{\Psi}[x_1(n), y(n)] + \tilde{\Psi}[x_2(n), y(n)] \tag{A.2}$$

Proof:

$$\begin{aligned} \tilde{\Psi}[x_1(n) + x_2(n), y(n)] &= [x_1(n) + x_2(n)]y(n) \\ &\quad -\tfrac{1}{2}[x_1(n-1) + x_2(n-1)]y(n+1) \\ &\quad -\tfrac{1}{2}[x_1(n+1) + x_2(n+1)]y(n-1) \end{aligned}$$

Therefore

$$\begin{aligned}\tilde{\Psi}[x_1(n)+x_2(n),y(n)] &= x_1(n)y(n)-\tfrac{1}{2}[x_1(n-1)y(n+1)+x_1(n+1)y(n-1)] \\ &\quad + x_2(n)y(n)-\tfrac{1}{2}[x_2(n-1)y(n+1)+x_2(n+1)y(n-1)] \\ &= \tilde{\Psi}[x_1(n),y(n)]+\tilde{\Psi}[x_2(n),y(n)]\end{aligned}$$

■

Property 2 :

The cross Ψ energy is defined to handle the Ψ energy of the sum of two signals

$$\Psi[x_1(n)+x_2(n)] = \Psi[x_1(n)]+\Psi[x_2(n)]+2\tilde{\Psi}[x_1(n),x_2(n)] \tag{A.3}$$

The proof of *Property 2* is straight forward.

Property 3 :

In general we have

$$\Psi[\sum_{i=1}^{N} x_i(n)] = \sum_{i=1}^{N}\Psi[x_i(n)]+2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\tilde{\Psi}[x_i(n),x_j(n)] \tag{A.4}$$

This property can be proven easily by induction.

Proof: Using **A.3**

$$\Psi[\sum_{i=1}^{N+1} x_i(n)] = \Psi[\sum_{i=1}^{N} x_i(n)]+\Psi[x_{N+1}]+2\tilde{\Psi}[[\sum_{i=1}^{N} x_i(n),x_{N+1}]$$

Assuming **A.4** is true and using **A.2** we have

$$\Psi[\sum_{i=1}^{N+1} x_i(n)] = \sum_{i=1}^{N+1}\Psi[x_i(n)]+2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\tilde{\Psi}[x_i(n),x_j(n)]+2\sum_{i=1}^{N}\tilde{\Psi}[x_i(n),x_{N+1}(n)]$$

Merging the last two terms we have

$$\Psi[\sum_{i=1}^{N+1} x_i(n)] = \sum_{i=1}^{N+1}\Psi[x_i(n)]+2\sum_{i=1}^{N}\sum_{j=i+1}^{N+1}\tilde{\Psi}[x_i(n),x_j(n)]$$

This completes the proof. ■

Property 4 :

For any two real numbers a and b,

$$\widetilde{\Psi}[ax(n), by(n)] = ab\widetilde{\Psi}[x(n), y(n)] \tag{A.5}$$

Property 5 :

If $\Psi[y(n)] = 0$, where $y(n)$ is an arbitrary signal, and a is a constant real number then

$$\widetilde{\Psi}[x(n) + ay(n), y(n)] = \widetilde{\Psi}[x(n), y(n)] \tag{A.6}$$

Bibliography

[1] L. Rabiner, "Applications of voice processing to telecommunications," *Proc. of the IEEE*, vol. 82, February 1994.

[2] 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.

[3] J. B. Allen, "How do humans process and recognize speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 567–577, October 1994.

[4] 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.

[5] J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals.* Macmillan, 1993.

[6] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition.* Prentice-Hall, Englewood Cliffs, NJ, 1993.

[7] S. Young, "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Processing Magazine*, Sept. 1996.

[8] 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.

[9] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. on Speech and Audio Processing*, October. 1980.

[10] H. M. Teager and S. M. Teager, "Evidence for nonlinear speech production mechanisms in the vocal tract," *NATO Advanced Study Institute on Speech Production and Speech Modelling, Bonas, France*, July 1989.

[11] P. Maragos, "Modulation and Fractal Models for Speech Analysis and Recognition," *Proceedings of COST-249 Meeting*, Feb. 1998.

[12] A. C. Bovik, P. Maragos, and T. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *IEEE Trans. on Signal Processing*, vol. 41, pp. 3245–3265, December 1993.

[13] P. Maragos, T. Quatieri, and J. F. Kaiser, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. on Signal Processing*, vol. 41, pp. 1532–1550, April 1993.

[14] P. Maragos, J. F. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. on Signal Processing*, vol. 41, pp. 3025–3051, October 1993.

[15] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1990 (ICASSP '90)*, pp. 381–384, April 1990.

[16] L. Rabiner, "An algorithm for determining the endpoints of isolated utterances," *Bell System Tech.*, pp. 297–315, Feb. 1975.

[17] E. Erzin, A. 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.

[18] R. Sarikaya and J. N. Gowdy, "Subband Based Classification of Speech Under Stress," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1998 (ICASSP '98)*, vol. 1, pp. 596–572, 1998.

[19] R. Sarikaya, B. L. Pellom, and J. H. Hansen, "Wavelet Packet Transform Features with Application to Speaker Identification," *NORSIG'98*, pp. 81–84, 1998.

[20] E. Erzin, *New methods for robust speech recognition.* PhD thesis, Bilkent University, 1995.

[21] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, 1975.

[22] B. Atal and S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, pp. 637–655, 1971.

[23] 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.

[24] 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, no. 7, pp. 968–973, 1987.

[25] E. Erzin and A. Çetin, "Line spectral frequency representation of subbands for speech recognition," *Signal Processing*, vol. 44, June 1995.

[26] P. Kabal and R. Ramachandran, "The computation of Line Spectral Frequencies using chebyshev polynomials," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 1419–1426, Dec 1986.

[27] 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.

[28] 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.

[29] 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.

[30] Zwicker and E. Terhardt, "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.*, vol. 68, pp. 1523–1525, December 1980.

[31] A. Papoulis, *Probability, Random Variables and Stochastic Processes.* McGraw-Hill International Editors, 3 ed., 1991.

[32] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, pp. 257–285, Feb. 1989.

[33] B. Juang, S. Levinson, and M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. on Information Theory*, vol. 32, pp. 307–309, March 1986.

[34] L. E. Baum and J. A. Eagon, "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology," *Bulletin of the American Mathematical Society*, vol. 73, pp. 360–363, 1967.

[35] L. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.

[36] A. J. Viterbi, "Error bounds for Convolutional Codes and an asymptotically optimal decoding algorithm," *IEEE Trans. on Information Theory*, vol. 13, pp. 260–269, April 1967.

[37] A. J. Viterbi and J. Omura, *Principles of Digital Communication.* New York : McGraw-Hill, 1979.

[38] G. D. Forney, "The Viterbi Algorithm," *Proceedings of IEEE*, vol. 16, pp. 268–278, Mar 1973.

[39] I. Daubechies, *Ten Lectures on Wavelets.* SIAM Press, Philadelphia, 1992.

[40] L. Rabiner and R. E. Crochiere, *Mutirate Digital Signal Processing.* Prentice-Hall, Englewood Cliffs, NJ, 1973.

[41] A. N. Akansu and M. J. T. Smith, *Subband and Wavelet Transforms, Design and Applications.* Kluwer Academic Publishers, 1996.

[42] G. Beylkin, R. Coifman, and V. Rokhlin, "Fast wavelet transforms and numerical algorithms 1," tech. rep., YALBU/DCS/RR-696, 1989.

[43] 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.

[44] B. Atal, V.Cuperman, and A. Gersho, *Advances in Speech Coding.* Kluwer Academic Publishers, 1991.

[45] S. M. Phoong, C. W. Kim, 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.

[46] 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.

[47] L. Isserlis, "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables," *Biometrica*, vol. 12, pp. 134–139, 1918.