ISTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY

SPEECH RECOGNITION
BASED ON
PATTERN COMPARISON TECHNIQUES

Msc THESIS

Computer Eng. Osman MERAL

Date of Submission: May 29, 96

Date of Approval : June 14, 96

Supervisor : Asc.Prof. Bülent ÖRENCİK

Other Members : Prof.Dr. Erdal PANAYIRCI

Asc.Prof. Muhittin GÖKMEN

JUNE 96

FOREWORD

This thesis has been prepared to get degree of master of the science in Computer Engineering at Istanbul Technical University.

I wish to express my gratitude and sincere thanks to Doç.Dr.Bülent Örencik, my thesis supervisor, for his very kind interest and guidance in the accomplishment of this work.

I would also like to thank Fahrettin Başaran for his tolerance, Hakan Göcümoğlu for his help to edit text and very dear Fehmi Abime.

This thesis is dedicated to in memory of dear E.H.M.

I will be very pleased when the thesis is useful for speech recognition applications in the industry.

Osman MERAL

June, 1996

CONTENTS

SUMMARY

This study describes pattern comparison method that recognizes speaker dependent isolated spoken words. Word training samples are obtained from samples male-speech. LPC coefficients are used as a features that describes speech. A low pass filter is used for antialiasing and to limit frequency of signal in voice band. Autocorrelation analysis is made to calculate LPC filter coefficients. Durbin's recursive algorithm is used to obtain coefficients by using output of autocorrelation analysis. Endpoint detection algorithm suppresses background noise and uses energy of the signal to find accurate beginning and end of the signal. In the training phase extracted parameters (LPC coefficients,cepstral coefficients) are strored and called reference templates of speech. In the recognition phase extracted parameters of the spoken word(test pattern) are compared with the pre-stored reference patterns. Time registration of the test and reference may not be perfectly aligned, so warping process is used to find how much test and reference patterns are similar to each other. Dynamic time warping algorithm and log likelihood as a distance measure and ceptral distance measures are used together to find distances between test and reference pattern. By using a decision rule, spoken word is found by taking the reference pattern that has minimum error (error value is below than reject error threshold) is selected as a recognized word or rejected if there is no minimum error. True recogniziation ratio of the words is nearly 95 % for the available library reference patterns.

ÖZET

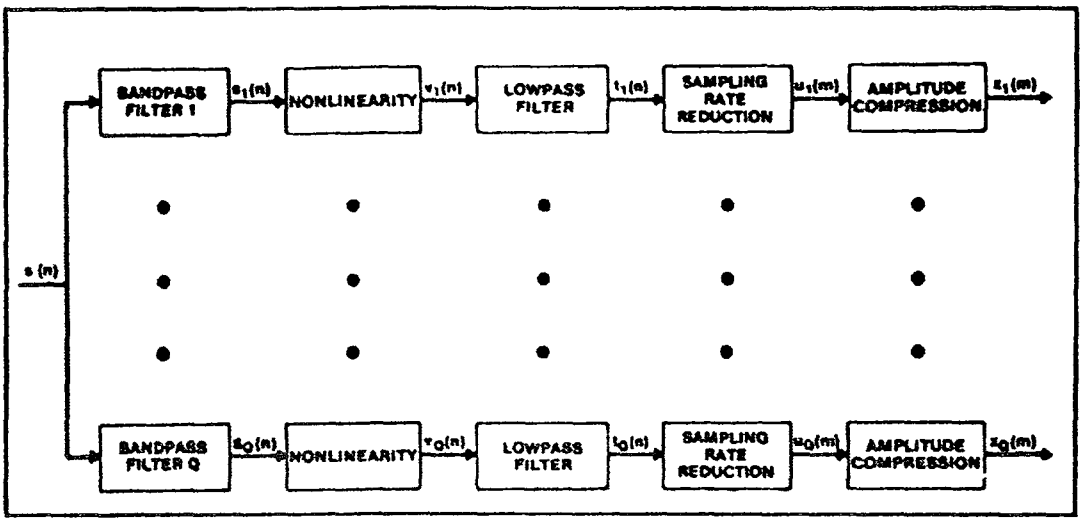
Bu tez çalışmasında şablon karşılaştırma (pattern comparison) metodu ile konuşmacıya bağımlı sözcük tanıma işlemi yapılmıştır.

Tez 6 bölümden oluşmaktadır. Her bölüm sözcük tanıma işinin ayri fazlarına karşılık düşer. Bölüm 1'de konuya giriş yapılmıştır.

Bölüm 2'de konuşma tanıma teknolojisindeki kavramlar nelerdir sorusuna yanıt aranmış, günümüze kadar mevcut konuşma tanıma yöntemleri özetlenmiştir. Konuşma tanımada gerçeklenmesi düşünülen sistem için birçok parametre vardır. Sözcüklerin birbirinden ayrık şekilde söylendiğinde mi tanınacağı yoksa sürekli konuşma içinden mi tanınacaği, tek bir kullanıcı için düzenlenmiş yani kullanıcı bağımlı(speaker dependent) veya bütün kullanıcılara açık kullanıcıdan bağımsız mı (speaker independent) tanınacagı en önemli parametrelerdir. Bunun yanında sözcük sayısının az sayıda (1-20 sözcük), orta sayıda (20-100) sözcük ve büyük sayıda (100' büyuk) olmasına bağlı olarak yöntem değiştirmek gerekebilir. Konuşma ortamının gürültüden izole veya gürültülü olması, transmisyon ortamının santralden geçmesi, yakın konuşma gibi parametreler de konuşma tanıma başarısını etkileyen parametrelerdir. Bu tezde gerçeklenen yöntemi yukarıda bahsedilen parametrelere göre sınıflandırırsak şunları söylemek mümkün. Konuşma tanıma yöntemi kullanıcı bağımlıdır, ayrık sözcükleri tanımaktadır, orta sayıda(20-100) sözcük grubu için geçerlidir ve gürültüsüz ortamda (SNR > 20 dB) çalışmaktadır. Yine bölüm 2'de konuşma tanıma yöntemlerinden olan okustik-fonetik yaklaşım, şablon karşılaştırma yaklasımı, ve yapay zeka yöntemi kısaca anlatılmıştır. Konuşma tanıma ile ilintisi sebebiyle konuşmacının tanınması(speaker recognition), sürekli konuşma içinden herhangi bir sözcüğü tanıma (word-spotting) kavramları ile ilgili kısa bilgiler de bu bölümde verilmiştir.

Bölüm 3'de konuşma tanıma için analiz yöntemleri anlatılmıştır. Konuşma tanımada, konuşma işaretinin karakteristiğini ifade eden parametrelerin sayısal işaret işleme yöntemleri ile elde edilmesi en önemli aşamalardan biridir. Konuşma işaretinin bir çok parametrik gösterimi mevcuttur. İşaretin kısa zamanlı enerjisi, sifir geçiş oranı sesli-sessiz olması vb. bilgiler ses işareti için parametrik gösterimlerdir. Bunların yanında en önemli parametrik gösterim, ses işaretinin kısa zamanlı spektral zarfidır. Bu gösterim işaret ile ilgili en önemli bilgiyi vermektedir. Bu nedenle spekral analiz, konuşma tanıma sisteminde sayısal işaret işleme ön bloğunun çekirdeği sayılır. Süzgeç bankası spektrum analizi ve doğrusal öngörü (linear prediction) analizi, spectral analiz için en çok kullanılan iki yöntemdir.

Süzgeç bankası spektral analiz yönteminde telefon şebekesi kalitesindeki işaretler 100-4000 hz arası bandla sınırlandıktan sonra Q adet band geçiren süzgeçten geçirilir. Band geçiren süzgeçlerin çıkışı ses işaretinin kısa zamanlı spectral bilgisini verir. Böyle bir analiz için blok şema aşağıda verilmiştir.



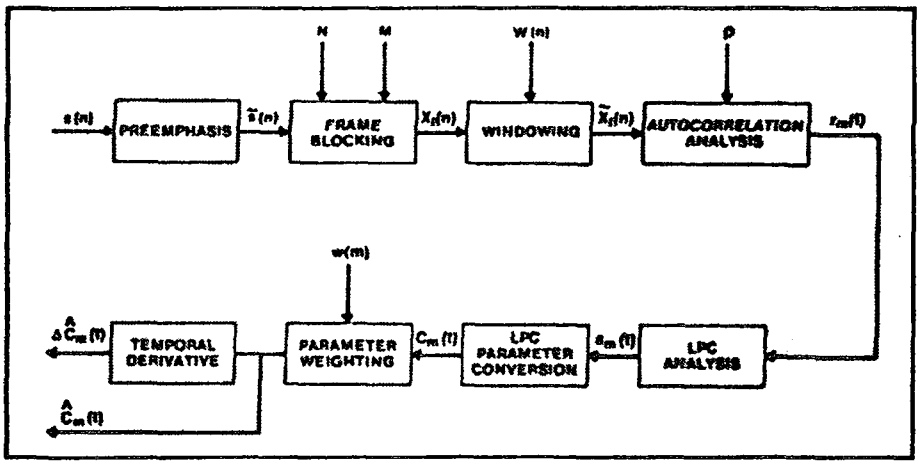
Süzgeç bankası analizinde en çok kullanılan süzgeç tipi eşit bandlı(uniform) süzgeç bankası yöntemidir. Bütün spektrum Q eşit banda bölünür. Genelde süzgeçlerin band genişlikleri birbirinin biraz içine girer(overlapping). Bu durum hem fiziksel gerçeklenmenin getirdiği bir sınırlamadır hem de çoğu zaman tercih edilen bir yoldur. Alternatif yöntem ise eşit bandlı olmayan (non-uniform) süzgeç bankası yöntemidir. Bu yöntemde her bir süzgecin band genişliği ve merkez frekansı bir kritere göre belirlenir. Bu çoğu zaman logaritmik frekans sıkalasına karşılık gelir. Bunun yanında insanın sesi algıladığı frekanslarla ilgili kritik band sikalası da (mel scale) popüler spectral analiz yöntemidir.

Süzgeç bankası spektral analizinde bazı işlemlerin süzgeçleme isleminden önce ve sonra yapılması gerekebilir. Süzgeçleme öncesi işlemler , işaretin vurgulanması, gürülteden arındırma, işaretin iyileştirilmesidir. Süzgeçleme işleminden sonra yapılacak işlemler ise süzgeç çıkışındaki işaretin zamanda yumuşatılması, süzgeç çıkışının normalizasyonudur.

Ses işaretinin analizinde kullanılan diğer yöntem ise doğrusal öngörüdür. Bu analiz sıkça kullanılır. Bunun sebeplerine gelince:Doğrusal öngörü ses işaretinin iyi modellenmesine olanak verir. Özellikle sözcüklerin sesli bölgelerini iyi modeller, sessiz bölgelerde ise kabul edilebilir sınırlar içerisindedir. Doğrusal öngörü matematiksel olarak basit ve yazılım ve donanım olarak gerçeklenmesi kolay bir yöntemdir. Süzgeç analizinden daha az işlem gerektirir. Doğrusal öngörü, süzgeç bankası ile karşılaştırıldığında, konuşma tanıma işlemi için gayet iyi performans göstermiştir.

Aşağıda konuşma tanımada kullanılan doğrusal öngörü ön işlemcisi için blok şema verilmiştir. İşaret önce konuşma bandında sınırlandırılmıştır. Konuşma bandı 100-3200 Hz olarak kabul edilmiştir. Sınırlandırılan bu işaret sabit frekansla örneklenmiştir. Bu çalışmada örnekleme frekansı 6.67 kHz seçilmiştir. Örnekleme işleminden sonra işaretin gerekirse daha anlamlı hale gelmesi için işaret eğer gürültülü ise bu gürültünün yok edilmesi, işaretin iyileştirilmesi işlemleri gerekebilir. Bunlar için literatürde birçok yöntem mevcuttur. Spektral çıkartma (spectral subtraction)

yöntemi gürültünün olduğu frekansların tespit edilmesi ve gürültünün spektrumunun işaretin spektrumundan çıkartılması ile gürültüden arınmış işaretin elde edilmesini sağlar. Bu tezde bu işlemler gerekmemiştir,cünkü tezin amacı gürültülü ortamlarda konuşma tanımak değildir.



Doğrusal öngörü yöntemi yüksek frekanslı işaretler için çok iyi modelleme yapamaz ve yüksek frekansları zayıflatır.Bundan dolayi ön vurgulama(preemphasis) bloğunda işaretin yüksek frekanslı kısımları daha belirginleştirilir ki bu bilgiler analiz sırasında kaybolmasın. Bunun için birinci dereceden süzgeç kullanılır. Vurgulanan işaret daha sonra çerçevelere ayrılır(frame blocking). Analiz çerçevesi genelde 10 ile 50 ms arası zamana karşılık gelir. Tezde analiz çerçevesi 45 ms olarak alınmıştır bu 300 örneğe karsılık düşmektedir. Her analiz çerçevesi 30 ms diğerinin içine geçmiştir (ovelapping). Bu durumda her 15 ms de bir analiz edilen bilgiler elde edilmektedir. Bir sonraki adım ise işaretin bir çerçeveden sonraki çerçeveye geçerken meydana gelebilecek kesintileri önlemek için pencereleme(windowing) işlemidir. Pencere fonksiyonu olarak Hamming penceresi seçilmiştir. Doğrusal öngörü katsayılarının elde edilmesi için otokorelosyon yöntemi kullanılmıştır. Bunun için her çerçevenin otokorelasyon değerleri hesaplanmıştır. Bu hesaplanan değerler daha sonrasında doğrusal öngörü (LPC analysis) bloğunda kullanılarak doğrusal öngörü katsayıları Levinson- Durbin rekursif metodu yardımı ile elde edilir. Elde edilen bu doğrusal öngörü katsayılarından kepstral(cepstral) katsayılar bulunur. Kepstral katsayılar konuşma tanıma amacıyla kullanıldığında doğrusal öngörü katsayılarından daha iyi sonuç verdiği görülmüştür. Bütün bu analiz sonucunda ses işaretin karakteristiğini veren özellikler(feautures) elde edilmiştir. Artık bundan sonra bu özellliklerin karşılaştırılması problemiyle uğraşılacaktır.

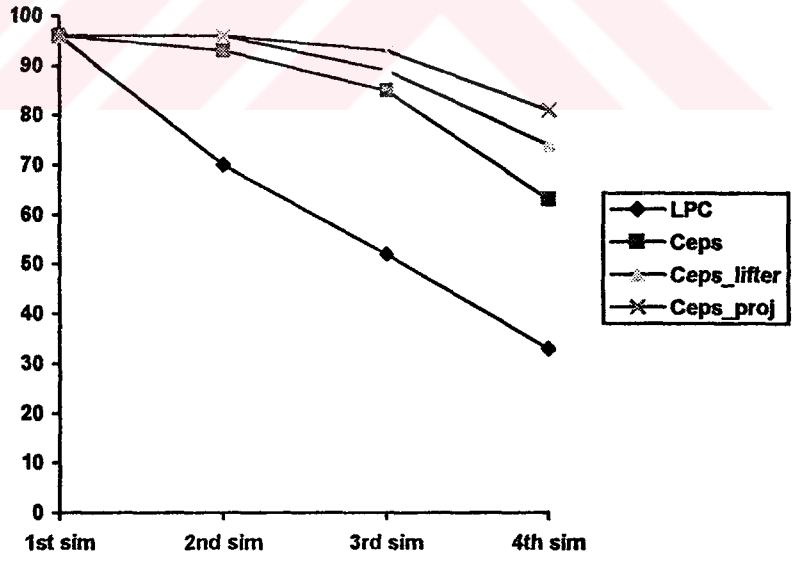
Bölüm 4'de şablonların karşılaştırılması teknikleri anlatılmıştır. En çok kullanılan şablon karşılaştırma tekniği, analiz sırasında elde edilen özellik vektörlerinin zaman sekanslarını kullanmaktır. Test şablonu, T, konuşma süresince elde edilen spektral sekans aşağıdaki gibi ifade edilir.

$$T = \{t_1, t_2, \cdots, t_I\}$$

Her t_i işaretin i zamanındaki spektral vektöre karşılık düşer. Başvuru şablonları(reference patterns) kümesi ise şöyle tanımlanır.

yakınındaki diğer vektörlerle de karşılaştırmayı kapsar. Böylece zamanda kayma kompanze edilmiş olur. Zamanda düzeltmeyi yapmak için zamanda dinamik eşleme yöntemi kullanılır(dynamic time warping). Aslında problem test vektörleri ile başvuru vektörlerinin karşılaştırılması sırasında en az hatalı yolu bulma (yol karşılaştırılan vektörlerin indisinin kordinatlari ile belirlenir) problemidir. Zamanda dinamik eşleme yöntemi bu problemin basit ve etkin çözümüne olanak sağlar.

Bölüm 5'de gerçekleştirilen sistem anlatılmıştır. Öğrenme(training) aşamasında doğrusal öngörü katsayıları ve kepstral katsayılar bulunmuştur. Tanıma fazında bulunan bu katsayılar spektral domende karşılaştırılarak ne kadar birbirine benzedikleri belirlenmiştir. Bu karşılaştırma sonucunda elde edilen hata(yada benzerlik ölçüsü) belli bir eşik değerin altında ise karşılaştırılan sözcük tanınmış olur. İsimler komutlar ve rakamları içine alan 27 sözcük 5 kez tekrarlanmak suretiyle başvuru kümesi oluşturulmuştur. Itakura'nın logaritmik benzerlik ölçüsü, kepstral uzaklık ölçüsü, lifterli kepstral uzaklık ölçüsü, lifterli kepstral projeksiyon uzaklık ölçüsü spektral vektörlerin karşılaştırılması için kullanılmıştır. Karşılaştırma sonucunda en yakın 10 sözcük benzerliklerine göre sıralanmıştır. Beklenen; karşılaştırılan sözcükle sıralamada en çok benzeyen 5 sözcüğün aynı kümeden olmasıdır; çünkü aynı sözcükten 5 kez tekararlandı, en cok benzeyenler de o kümeden olması gerekir. Bu durum kullanılan uzaklık ölçüsünün dayanıklılığının bir göstergesi olarak kabul edilir. Bu anlamda en dayanıklı (robust) uzaklık ölçüsü lifterli kepstral projeksiyondur. Eğer karşılaştırma sonucunda elde edilen sıralamada, en benzer sözcük tanınan sözcük olarak kabul edildiğinde, tanıma algoritmasının başarısı %96 dir. Bahsedilen karşılaştırma şekli çok guvenilir değildir. Karşılaştırma başarımında ençok benzer 5 sözcük eğer o sözcüğün öğrenme kümesindeki sözcükler ise en başarili tanıma yapılmış olur. Bu kritere göre başarım yüzdesi aşağıda verilmiştir.



Şekilden görüleceği gibi en çok benzeyen 4 sözcüğün(kendi dışında) aynı kümeden olma yüzdesi lifterli kepstral projeksiyon ölçüsü için %85'ler civarındadır. Itakura'nın logaritmik benzerlik ölçüsünun en kötü spektral karşılaştırma ölçüsü olduğu görülmüştür. Bu calışmalar sonucunda görülmüştür ki kepstral uzaklık ölçülerinin başarımı daha yüksektir ve kepstral uzaklık ölçüleri daha dayanıklı(robust) uzaklık ölçülerdir.

$$\{R^1, R^2, ..., R^V\}.$$

Her başvuru şablonu , R^j, spektral vektörlerdir . $R^j = \{r_1^j, r_2^j, ..., r_J^j\}$

Şablon karşılaştırmanın amacı test şablonu ile başvuru şablonları arasındaki benzerliği bulmaktır. Bunun için aşağıdaki problemlerin dikkate alınıp çözülmesi gerekmektedir.

- T and R^j genellikle eşit olmayan uzunluklardadır. Bu durum özellikler için konuşma hızının farklı olmasından kaynaklanır.

- Spektral vektörleri karşılaştırmak için yerel benzerlik ölçüsüne gerek vardır.

Bütün bunlar yerel(local) benzerlik ölçüsü ve tüm (global) benzerlik ölçüsünün tanımlanmasını gerektirmektedir.

Yerel benzerliklerin bulunması için spektral karşılaştırma yöntemleri kullanılır. Logaritmik spektral uzaklık ölçüsü spektral vektörlerin karşılaştırılmasında en çok kullanılan ölçüdür. Karşılaştırılacak iki spektrum $S(\omega)$ ve $S'(\omega)$ verilmişse , bu iki spektrumun farkı asağıdaki gibi tanımlanır

$$V(\omega) = \log S(\omega) - \log S'(\omega)$$

$S(\omega)$ ve $S'(\omega)$ arasındakı uzaklık ölçüsü ise aşağida tanımlanmıştır.

$$d(S,S')^p = (d_p)^p = \int_{-\pi}^{\pi} |V(\omega)|^p \frac{d\omega}{2\pi}$$

Bu formüller yardımıyla kepstral uzaklık ölçüleri için formüller çıkarılmıştır. Kepstral uzaklık ölçülerinde kepstral katsayıların bazılarının ağırlığının değiştirilmesi konuşma tanıma yüzdesini artırmaktadır. Büyük indeksli kepstral katsayılar doğrusal öngörü analizinin sınırlamalarından kaynaklanan sebeplerden daha fazla etkilenirler. Düşük indeksli kepstral katsayılar ise transmisyon ortamından, sesin periyodundan, kullanıcının karakteristiklerinden daha fazla etkilenirler. Eğer karşılaştırma sırasında düşük ve yüksek indeksli kepstral katsayıların ağırlığı toplam hata üzerinde azaltılır diğer katsayıların ağırlığı artırılırsa daha sağlıklı bir karşılaştırma yapmak mümkün olacaktır. Kepstral katsayılar üzerindeki bu işleme liftering(liftering) denir. Kepstral katsayılarda ilk 20 katsayıyı almak yeterlidir.

Yukarıda anlatılan kepstral uzaklık ölçüleri yanında gürültülü ortamda konuşma tanıma performansını artırmak için daha dayanıklı(robust) uzaklık ölçüleri kullanmak mümkündür.

Aynı sözcüğün farklı hızlarda söylenmesi spektral vektörlerin zamanda kaymasına sebep olur. Test şablonu ile başvuru şablonunu karşılaştırırken zamandaki kaymanın kompanze edilerek düzeltilmesi işlemi zamanda sıraya dizme(time alignment) işlemi olarak adlandırılır. Aslında bu işlem test ve başvuru vektörlerini karşılaştırıp bir uzaklık ölçüsü bulurken zamanda kayma olacağını da varsayarak

CHAPTER 1. INTRODUCTION

Speech recognition , perception, and understanding have been active research fields since 1950's. However, in the mid 1960's the introduction of the general purpose led to the discipline we now call automatic speech recognition (ASR). Initially, speech recognition focused on recognizing a few words from a single user (speaker dependent recognition) ,however, now ASR systems handles recognition of connected large-vocabulary and speaker independent words.

Speech signals convey information about who spoke what message in what manner and in what environment The task of the speech recognizer is to determine automatically the message (i.e. ,what was said) regardless of the variability's introduced by speaker identity , manner of speaking, and environmental conditions.

There are different methods of recognizing words regarding to restrictions. If one method is good for one restriction, may be bad for other restriction. Methods are chosen regarding restrictions of the recognition.

Dynamic Time Warping (DTW) is one of the oldest methods for speech recognition. The signal model in DTW is deterministic speech signal model. Deterministic models generally exploit some known specific properties of the signal. The methods uses pattern matching algorithm by warping reference and test patterns. Test and reference patterns are speech parameters like LPC, zerocrossing, power, spectra of the speech signal. This method is practical for especially speaker dependent and independent recognition of isolated words.

Hidden Markov Model is used for speaker dependent and independent recognition of isolated and continues words. The signal model in HMM is the set of statistical properties of the signal. The underlying assumption of the statistical model is that the signal can be well characterized as a parametric random process, and the parameters of the stochastic can be determined (estimated) in a precise, well-defined manner. For the application of the interest, namely speech processing, both deterministic and stochastic signal model have had a good success. Today's many speech recognition system uses HMM[2].

Neural network model is a new method and has great potential in areas such as speech and image recognitions where many hypothesis are pursued in parallel, high computation rates required. Nowadays researches in artificial neural network science is growing up.

CHAPTER 2. SPEECH TECHNOLOGY

2.1 AUTOMATIC SPEECH RECOGNITION SYSTEMS AND CLASSIFICATION OF SPEECH RECOGNITION SYSTEMS

The first ASR systems which appeared in the market performed very constrained speech recognition tasks. These systems now, developed so that large-vocabulary multi-speaker discrete word recognition, speaker-independent continuous speech recognition become available. Examples of these options are below[1]:

-Type of speech:

Isolated word recognition ,involves recognition of single words. The speaker pronounces these words in isolation, namely there is enough pauses between spoken words. Connected speech recognition, in which each word is clearly articulated and there is no pauses between utterated words.

-Number of speakers:

Speaker dependent systems are trained with one talker and templates of speech parameters is optimized for that talker and perform satisfactory results for that talker. Speaker independent recognizers are capable of identifying words regardless of the talkers.

-Type and amount of training and speaking environment:

No training ,fixed training, or continuous training may be necessary. Soundproof booth, computer room, ,public places, noise isolated rooms.

-Vocabulary size:

Small vocabulary (1-20 words), medium vocabulary (20-100), large vocabulary (more than 100 words).

-Transmission systems:

High quality microphone, close talking, using standard telephone handset, public or private exchanges.

2.2 APPROACHES TO AUTOMATIC SPEECH RECOGNITION BY MACHINE.

There are three approaches to speech recognition, namely[15]:

1. the acoustic-phonetic approach
2. the pattern recognition approach
3. the artificial intelligence approach

The acoustic-phonetic approach is based on the theory of phonetics that postulates that there exist finite, distinctive phonetic units in spoken language and these units are broadly characterized by a set of properties that are manifest in the speech signal, or its spectrum, over time. The first step in the acoustic-phonetic approach to speech recognition is called a segmentation and labeling phase because it involves segmenting the speech signal into discrete (in time) regions where the acoustic properties of the signal are representative of one phonetic units, and then attaching one or more phonetic labels to each segmented region according the to the acoustic properties.

The pattern-recognition approach to speech recognition is basically one in which the speech patterns are used directly without explicit feature determination and segmentation. As in most pattern-recognition approaches, the method has two steps, training of speech patterns, and recognition of patterns via pattern comparison. Speech knowledge is brought into the system via training procedure.

The pattern-recognition approach is the method of choice for speech recognition for three reasons:

1. Simplicity of the use. The method is easy to understand, it is rich in mathematical and communication theory justification for individual procedures used in training and decoding, and is widely used and understood.

2. Robustness and invariance to different speech vocabularies, users, feature sets, pattern comparison algorithm and decision rule.

3. Proven high performance.

The so-called artificial intelligence approach to speech recognition is a hybrid of the acoustic-phonetic approach and pattern-recognition approach in that it exploits ideas and concepts of both methods. The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person applies its intelligence in visualizing, analyzing, and finally making a decision on the measured acoustic features.

The use of neural networks could represent a separate structural approach to speech recognition or be regarded as an implementational architecture that may be incorporated in any of the above three classical approach. The concepts and ideas of applying neural networks to speech-recognition problems are relatively new[15].

2.2.1 Acoustic-Phonetic Approach to Speech Recognition.

Figure 2.2.1.1 shows a block diagram of the acoustic-phonetic approach to speech recognition. The first step in the processing is the speech analysis system(feature measurement method) , which provides an appropriate(spectral) representation of the characteristics of the time-varying speech signal. The most common techniques of spectral analysis are the class of filter bank methods and the class of linear predictive coding methods.

The next step in the processing is the feature detection stage. The idea here is to convert the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. Among the features proposed for

recognition nasality(presence or absence of nasal resonance), frication(presence or absence of random excitation in the speech), formant locations(frequencies of the first three resonances), voiced-unvoiced classification, and ratios of high and low-frequency energy.

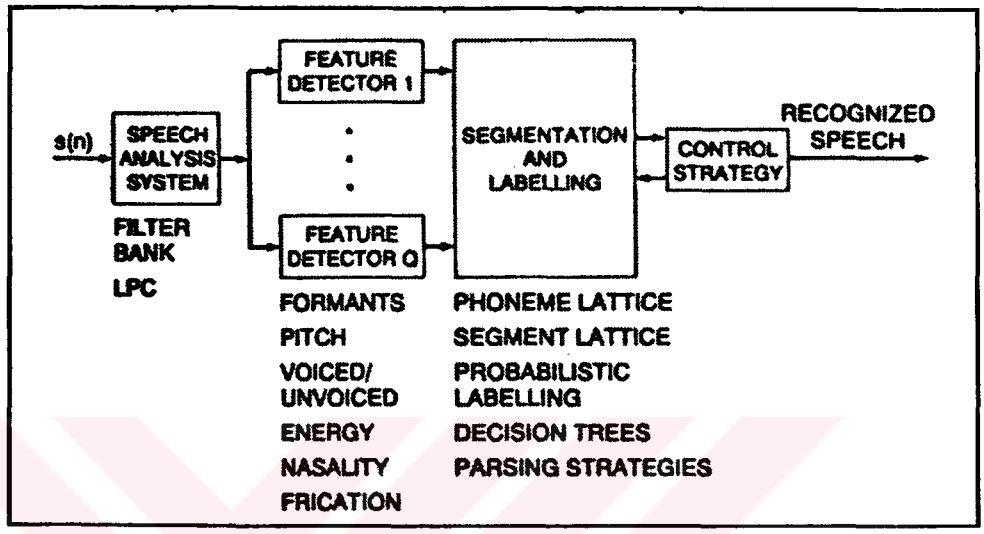


Figure 2.2.1.1 Block diagram of acoustic-phonetic speech recognition system

The third step in the procedure is the segmentation and labeling phase whereby the system tries to find stable regions(where the features change very little over the region) and then labeling segmented region according to how well the features within that region match those of individual phonetic units. This stage is the heart of the acoustic-phonetic recognizer.

The result of the segmentation and labeling step is usually a phoneme lattice from which a lexical access procedure determines the best matching word or sequence of words. The final output of the recognizer is the word or word sequence that best matches[15].

2.2.2 Statistical Pattern-Recognition Approach to Speech Recognition.

A block diagram of a pattern recognition approach to speech recognition has been shown in Figure 2.2.2.1. The pattern recognition method has four steps:

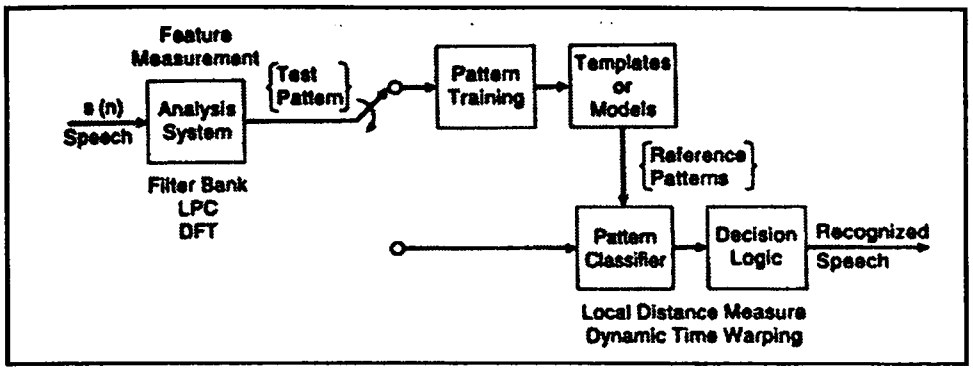


Figure 2.2.2.1. Block diagram of pattern-recognition speech recognizer.

1. Feature measurement. For speech signals the feature measurements are usually the output of some type of spectral analysis technique, such as filter bank analyzer, a linear predictive coding analysis, or a discrete Fourier transform analysis.

2. Pattern training, in which one or more test patterns corresponding to speech sounds of the same class are used to create representation of the features. The resulting pattern generally called a reference pattern or called template.

3. Pattern classification, in which the unknown test pattern is compared with each class reference pattern and a measure of similarity(distance) between the test pattern and reference pattern is computed. To compare speech patterns, it is required both local distance measure(that is spectral distance between two spectral vectors) and global time alignment procedure[15][16].

4. Decision logic is used to decide whether test pattern is valid in reference patterns or not by using similarity scores and thresholds.

The general strength and weaknesses of the pattern recognition model include the following:

1. The performance of the system is sensitive to the amount of training data available. Generally the more training, the higher the performance of the system.

2. The reference patterns are sensitive to the speaking environment and transmission characteristics of the medium. The reason for this is that because the speech spectral characteristics are effected by transmission and background noise.

3. The computational load for both pattern training and pattern classification is generally linearly proportional to the number of reference patterns.

4. Because the system is insensitive to sound class, the basic technique are applicable to a wide range of speech sounds, including phrases, whole words, and subword units.

5. It is relatively straightforward to incorporate syntactic constraints directly into the pattern-recognition structures, thereby improving recognition accuracy[15][16].

2.2.3 Artificial Intelligence Approach to Speech Recognition.

The basic idea of the artificial intelligence approach to speech recognition is to compile and incorporate knowledge from a variety of knowledge sources and to bring it to bear to the problem at hand. Thus, for example, the AI approach to segmentation and labeling would be augment the generally used acoustic knowledge with phonetic knowledge, lexical knowledge, syntactic knowledge, and even pragmatic knowledge. Definitions of the knowledge sources are below.

- acoustic knowledge: evidence of the which sounds are spoken on the basis of the spectral measurements.
- lexical knowledge: the combination of acoustic evidence so as to postulate words as specified by a lexicon that maps sound into words
- syntactic knowledge: the combination of the words to form grammatically correct strings (according to a language model) such as sentences or phrases.
- semantic knowledge: understanding of the task domain so as to be able to validate sentences that are consistent with the task being performed.
- pragmatic knowledge: inference ability necessary in resolving ambiguity of meaning based on the ways in which words are generally used[15][16].

2.2.4 Neural Networks and Their Application to Speech Recognition.

The system of Figure 2.2.4.1 shows the model of the human speech understanding system. The auditory analysis is based loosely on our understanding of the acoustic processing in the ear. The various feature analysis represent processing at various levels of the neural pathways to the brain. The short- and long-term memory provide external control of the neural processes in ways that are well understood. The overall form of the model that of a feed forward connectionist network- that is , a neural net.

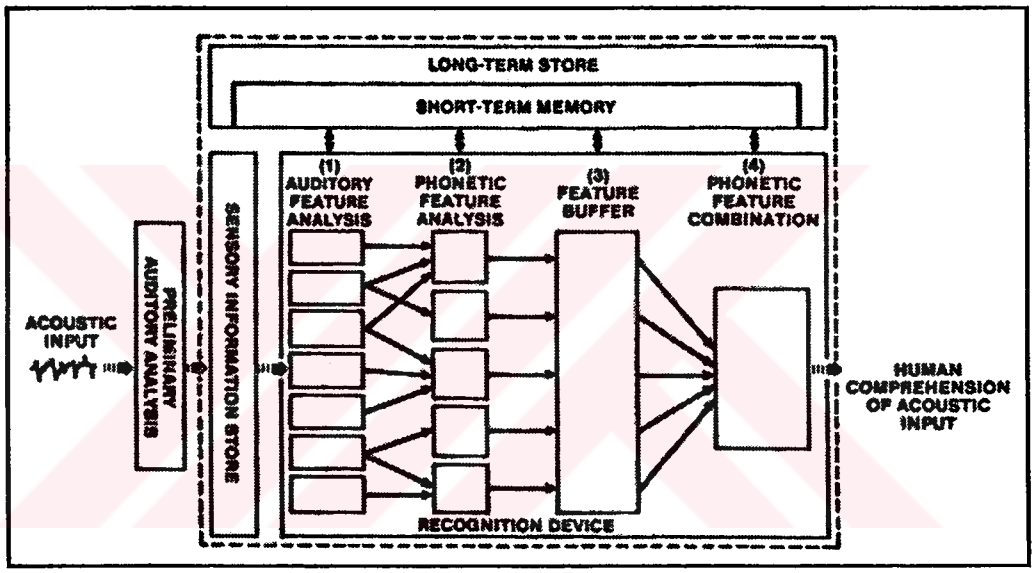


Figure 2.2.4.1. Block diagram of a human speech understanding system.

The details of the neural net is not described in here but the advantages are explained following:

1. They can readily implement a massive degree of parallel computation. Because a neural net is highly parallel structure of simple, identical, computational elements. It should be clear that they are prime candidates for massively parallel(analog or digital) computation.

2. They intrinsically possess a great deal of robustness or fault tolerance. Since the "information" embedded in a neural network is a "spread" to every computational

elements. This structure is inherently among the least sensitive of networks to noise or defects within the structure[15].

3. The connection weights of the network need not to be constrained being fixed; they can be adapted in real time to improve performance[15].

2.3 DIFFERENCES OF SPEAKER VERIFICATION, SPEAKER IDENTIFICATION, SPEECH RECOGNITION, AND WORD-SPOTTING.

There are two distinct subareas of speaker recognition- speaker verification and speaker identification. For speaker verification, an identity is claimed by the user, and the decision required of the verification system is strictly binary; i.e. to accept or reject the claimed identity. In order to make this decision , a set of features designed to retain essential information about speaker identity is measured from one or more utterances of the speaker, and resulting measurements are compared to a set of reference patterns for the claimed speaker. Thus for speaker verification only a single comparison between the set of measurement, and the reference pattern is required to make the final decision to accept or reject the claimed identity[1].

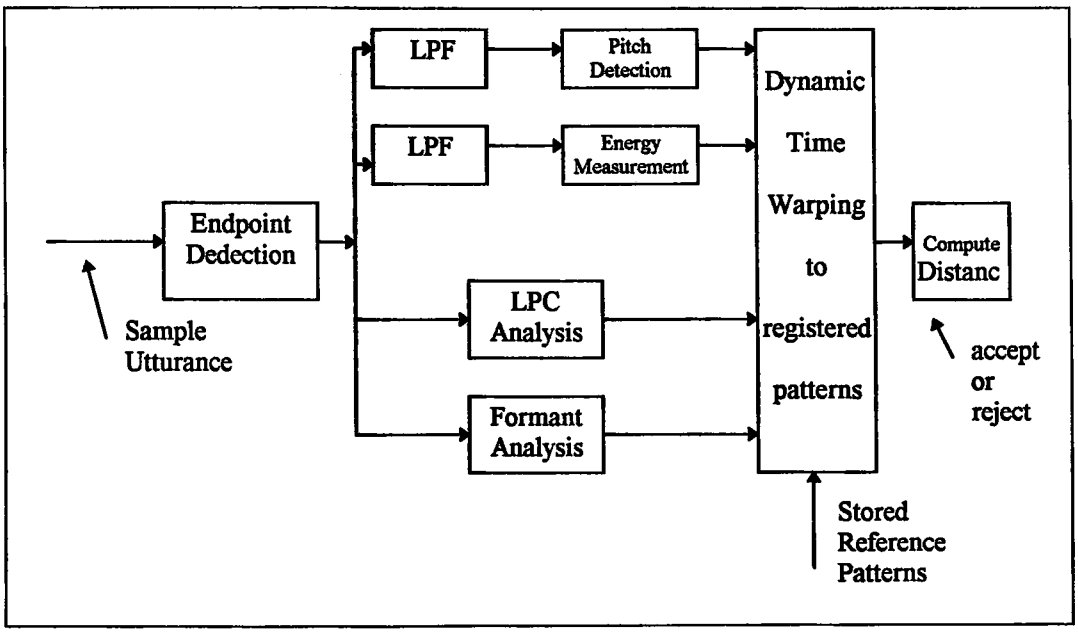


Figure 2.3.1. Signal processing aspects of the speaker verification systems.

Figure 2.3.1 shows a block diagram of signal processing aspects of the on line speaker verification system. The person wishing to be verified first enters his claimed identity, and then request from the verification system utters is verification phrase, and request some transaction to be made in the event he is verified.

Endpoint detector is used to find beginning and end of the utterance. In particular, a pitch detector is used to measure the pitch contour of the utterance; a short-time energy measurement is made to give energy contours, and LPC analysis is used to give predictor parameter contours, and finally estimate of the formant locations is made. The DTW function is to warp the time scale t of a reference utterance and test patterns for pattern matching. Decision algorithm will find the best candidate by using computed distances. Details of the blocks will be explained in the following chapters[1].

The problem of speaker identification differs significantly from the speaker verification problem. In this case the system is required to make an absolute identification among the N speaker in the user population. Thus instead of a single comparison between a set of measurements and a stored reference pattern, N complete comparisons are required.

There are similarities between the problems of speaker verification and identification. In terms of signal processing aspects , the two problems are treated almost identically in the most of the processing shown in Figure 2.3.1. The main differences in processing involve the choice of parameters used to make distance measurements, and the necessity of making N distance measurement than one.

For speech recognition , as in speaker recognition, digital speech processing techniques are applied to obtain a pattern which is then compared to stored reference patterns. In speech recognition the goal is to determine what word, phrase , or sentences was spoken.

For word spotting, the goal is to recognize a single word in a continuous speech. For example talking occurs like this ,"xxxxx NIXON xxxx WATERGATE xxxx". Spotted word may be Nixon and Watergate. Another examples, assume that you want a machine response you, when you called its name while you are talking. The machine spots its name from the other spoken words and response you[3].

2.4 SIGNAL PROCESSING NEED FOR SPEECH RECOGNITION

Speech recognition uses digital signal processing theory. Most of the digital signal processing methods are applicable for speech processing. Digital filters are used for preprocessing, decimation and interpolation are used for resampling.

2.5 GENERALIZED SPEECH RECOGNITION MODEL

General speech recognition model has been shown in Figure 2.4.1. This is the primary model by which most researchers in speech recognition have typically structured their solutions to ASR problem. The major components of this model is explained below.

-Signal processing module is used to obtain necessary signal information of the speech signal by preprocessing.

-Feature extraction module is used to obtain the key components of this information by eliminating redundant information.

-Pattern classification module is used to classify the signal via pattern matching(e.g., template matching with dynamic programming) or , via statistical modeling (such as Markovian model), neural networks.

-Language modeling for selecting a final word string.

The modular partition has permitted researchers to focus on specific components of the model. components of the model explained below.

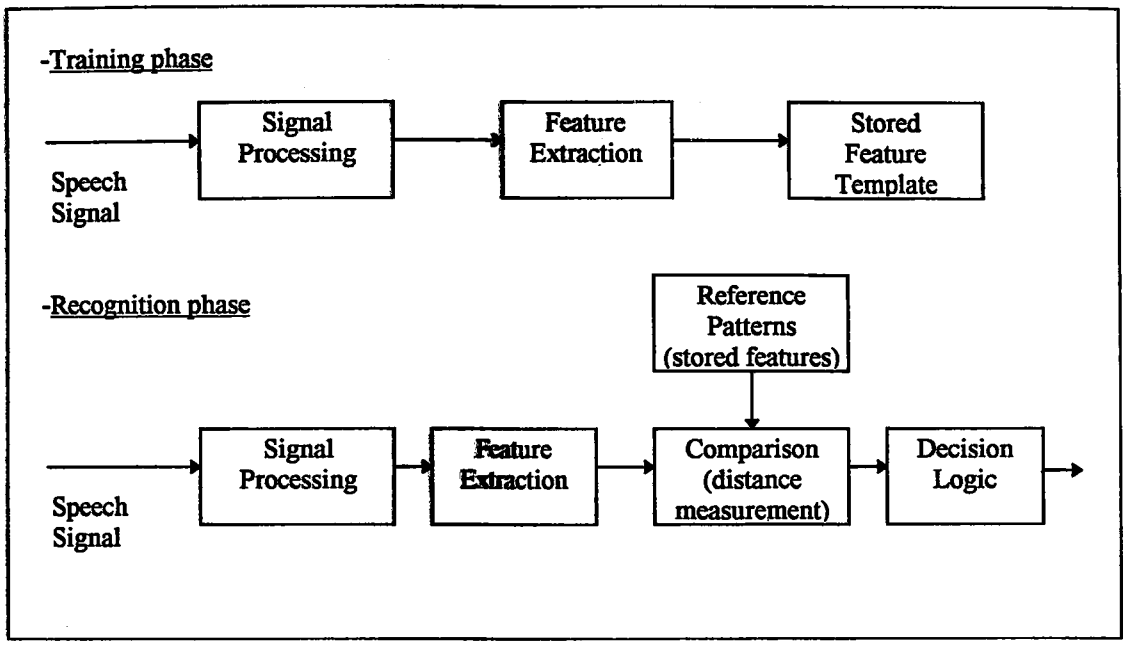


Figure 2.4-1 Generalized speech recognition model .shows training and recognition phase of utterated words.

The purpose of the signal processing module is to process the sampled speech signal and produce a representation which is independent of talker stress, or noise introduced by channel or environment.

Feature extraction module computes a set of parameters that characterize the spoken word or phoneme. These parameters are called features of speech and generally computed at a fixed interval.

In pattern classification the sequence of the features (called test pattern) is compared previously stored features (called reference patterns). A decision rule is applied to recognize spoken word. As mentioned before , there are three basic classes of the solutions to the pattern classification problem: pattern matching (template matching with dynamic programming), statistical modeling (such as HMM), and neural networks.

The final stage of the recognition process consist of a language processing module to resolve the possible words by using language specific constraints. For small vocabulary recognizer, a set of phonetically dissimilar words can be selected without

language processing module. However, for large vocabulary speech recognition systems, language processing module is necessary. This module uses semantic and syntactic knowledge of the language to decide correct word which are phonetically close.

CHAPTER 3. ANALYSIS METHODS FOR SPEECH RECOGNITION

3.1 INTRODUCTION

Speech recognition system comprises a collection of algorithms drawn from a wide variety of disciplines such as, statistical pattern recognition, communication theory, signal processing, mathematics, linguistics and others. Perhaps the greatest common denominator of all recognition systems is the signal processing front end, which converts the speech waveform to some type of parametric representation for analysis.

A lot of possibilities exists for parametically representing the speech signal. These include the short time energy, zero crossing rates and others. Probably the most important parametric representation is the short time spectral envelope. Spectral analysis methods therefore generally considered as the core of the signal processing front end in ASR systems. Mainly , two dominant methods is used for spectral analysis: The filter-bank spectrum analysis method, and linear predictive coding (LPC) spectral analysis model. Vector quantization is another way of encoding speech spectral information via using a finite codebook of spectral shapes.

3.2 FILTER-BANK FRONT-END PROCESSOR FOR RECOGNITION.

The most common choices of signal processing front-end for speech recognition are filter bank model and LPC model. The structure of filter banks has been given in Figure 3.2.1 The speech signal (digitized signal) is passed through a bank of Q bandpass filters whose frequency ranges are 100-3000 Hz for telephony quality signal and 100-8000 Hz for broadband signal. The bandpass filters generally do overlap in frequency as shown Figure 3.2.1. The output of the bandpass filters is the short time spectral representation of the speech signal s(n).

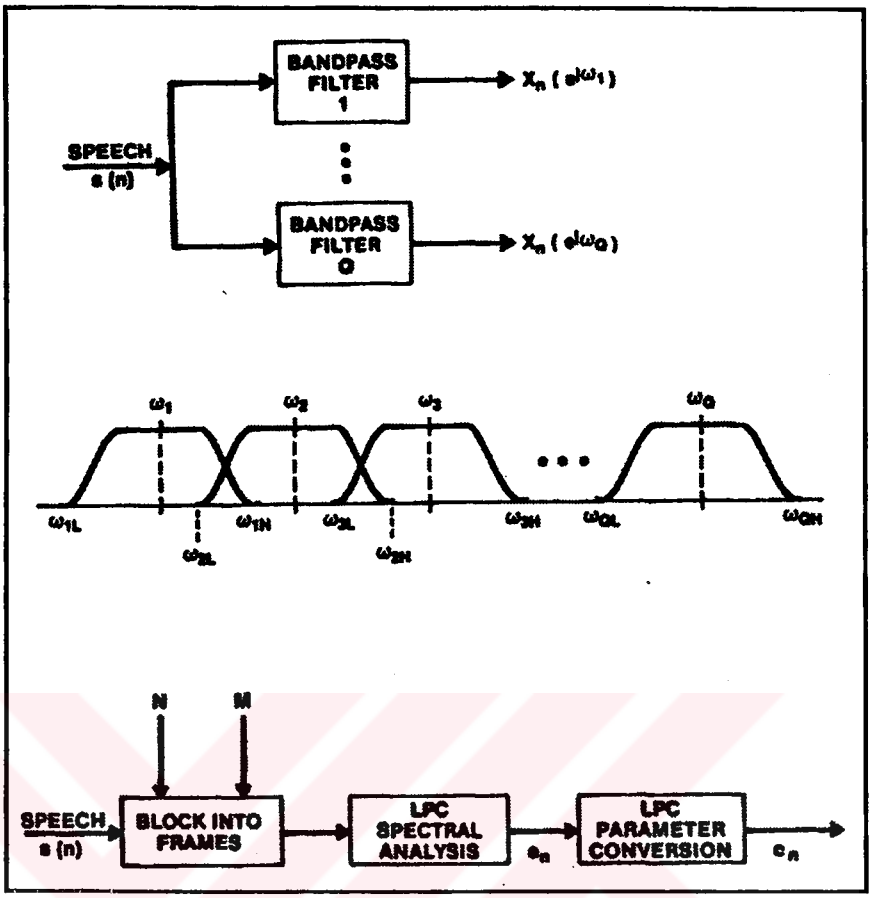


Figure 3.2.1 Bank of filters analysis method and LPC analysis method.

A block diagram of the canonic structure of a complete front end filter analyzer has been given in Figure 3.2.2The sampled signal, *s(n)*, is passed through a bank of Q bandpass filters giving the signals.

$$s_i(n) = s(n) * h_i(n), \qquad 1 \le i \le Q \qquad (3.2.1.a)$$

$$= \sum_{m=0}^{M_i-1} h_i(m) * s(n-m), \qquad (3.2.1.b)$$

where we assumed that the impulse response of the i^{th} bandpass filter is $h_i(m)$ with a duration of M samples. Since the purpose of the filter bank analysis is to obtain the energy of the speech signal in a given frequency band, each of the bandpass signals, $s_i(n)$, is pass through a nonlinearity, such as a full wave rectifier. The nonlinearity shifts the bandpass signal spectrum to low frequency band as well as creates high frequency images. A lowpass filter is used to eliminate the high frequency images.

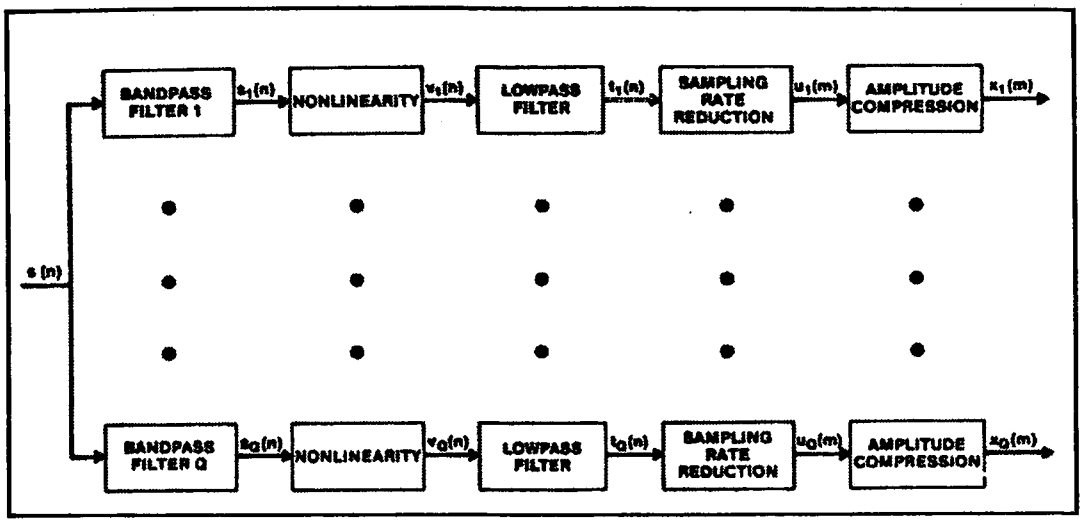


Figure 3.2.2 Complete bank of filters analysis method.

If a full-wave rectifier is used as the nonlinearity the signals will shown as following.

$$v_i(n)=s_i(n).w(n) \qquad (3.2.2)$$

$$\begin{aligned} w(n)&=+1 \quad \text{if} \quad s_i(n)\geq 0 \\ w(n)&=-1 \quad \text{if} \quad s_i(n)<0 \end{aligned} \qquad (3.2.3)$$

The effect of the lowpass filter is to retain the DC component and to filter out the higher frequencies due to nonlinearity. Because of the time-varying nature of the speech signal, the spectrum of the lowpass signal is not a pure DC impulse, but instead the information in the signal is contained in a low-frequency band around DC. Figure 3.2.3 illustrates typical waveform of *s(n)*, $s_i(n)$, *w(n)* and $v_i(n)$ for a 20 ms section of the voice speech by a narrow bandwidth channel with center frequency of 500 Hz. It can be seen that $|S_i(e^{jw})|$ has the most energy around 500 Hz, whereas $|W_i(e^{jw})|$ approximates and odd harmonic signal with peaks at 500,1500,2500 Hz. The resulting spectrum $|V_i(e^{jw})|$ shows desired low-frequency concentration of energy. The role of the final lowpass filter is to eliminate the undesired spectral peaks. The sampling rate reduction box after lowpass filter box is used to resample the signal at the rate on the order of 40-60 Hz for economic representation and the signal dynamic range is compressed using a amplitude compression like logarithmic encoding.

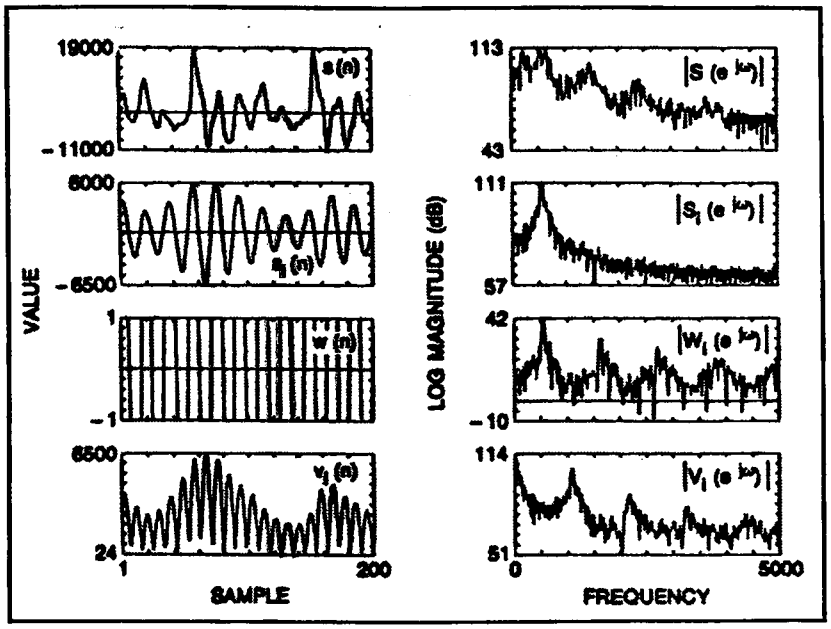


Figure 3.2.3 Typical waveforms and spectra of a voice speech signal.

3.2.1 Types of Filter Bank Used for Speech Recognition.

The most common type of filter used for speech recognition is the uniform filter bank for which the center frequency , f_i, of the i^{th} bandpass filter is defined as

$$fi = \frac{F_s}{N} i, \qquad 1 \le i \le Q, \qquad (3.2.1.1)$$

where Fs is the sampling rate of the speech signal, and N is the number of uniformly spaced filters required to span the frequency range of the signal. The bandwidth, b_i, of the i^{th} filter, generally satisfies the relation below.

$$b_i \ge \frac{F_s}{N}$$

Equality means that there is no frequency overlapping between adjacent filter channels, inequality means that there is overlap. Figure 3.2.1.1 shows a set of Q ideal ,non-overlapping bandpass filters whose bandwidths are equal. Below figure a more realistic set of Q overlapping filters covering approximately the same range.

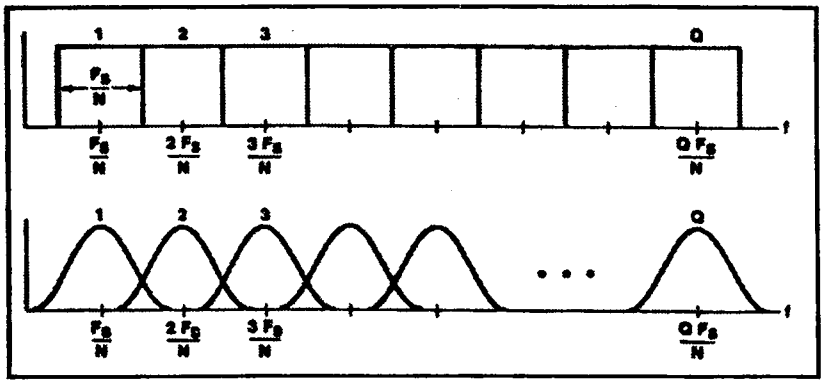


Figure 3.2.1.2. Ideal (a) and realistic(b) set of filter responses of a Q channel filter bank.

The alternative to uniform filter banks is nonuniform filter banks designed according to some criterion. One commonly used criterion is to space filters uniformly along a logarithmic frequency scale (A logarithmic frequency scale is often justified from a human auditory perception point of view). To design filter banks following formulas can be given,

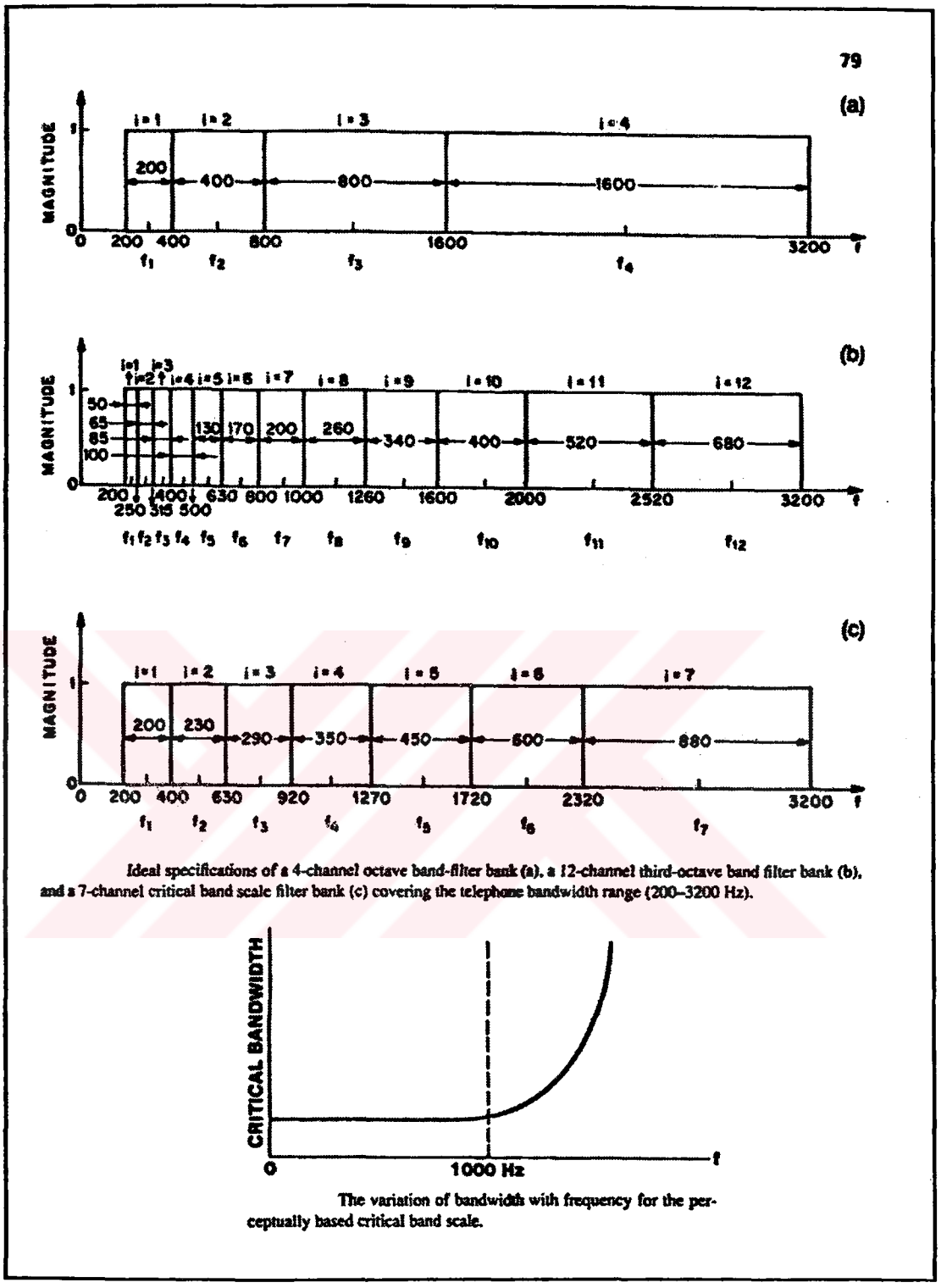
$$b_i = C \qquad (3.2.1.2)$$

$$b_i = \alpha b_{i-1}, \qquad 2 \le i \le Q, \qquad (3.2.1.3)$$

$$f_i = f_1 + \sum_{j=1}^{i-1} b_j + \frac{(b_i - b_1)}{2} \qquad (3.2.1.4)$$

The most common used values of α=2 , which gives an octave band spacing of adjacent filters, $\alpha = 4/3$ which gives a 1/3 octave filter spacing. An example of 4 band ,octave, non-overlapping filter bank design, and 12 band 1/3 octave ideal filter bank design specification covering 200-3200 Hz has been given in Figure 3.2.1.2.

An alternative criterion for designing a non uniform filter bank is to use the critical band scale directly. The spacing of filters along the critical band is based on perceptual studies and is intended to choose the bands that give equal articulation. The general shape of the critical band scale has been given in Figure 3.2.1.3. Mel scale critical band filter bank composition is used widely for extraction of mel cepstrum coefficients of the speech signal[15][16].



Ideal specifications of a 4-channel octave band-filter bank (a), a 12-channel third-octave band filter bank (b), and a 7-channel critical band scale filter bank (c) covering the telephone bandwidth range (200–3200 Hz).

The variation of bandwidth with frequency for the perceptually based critical band scale.

Figure 3.2.1.3. Band pass filter bandwidth selection scheme.

3.2.2 Implementation of Filter Banks.

A filter bank can be implemented in several ways, depending on the method used individual filters. Design methods for digital filters fall into two broad classes:

Finite impulse response (FIR) and infinite impulse response (IIR). Generally the most efficient is to realize each individual bandpass filter as a cascade or a parallel structure.

In filter bank design some operations should be do done before filtering and after filtering. This operations are:

-preprocessing

- signal preemphasis
- noise elimination,
- signal enhancement,

-postprocessing

- temporal smoothing of filter bank output vectors
- frequency smoothing of filter bank output vectors
- normalization of filter output
- thresholding and quantization

Name of the some methods in the literature to implement filter banks has been given below. Frequency domain interpretation of the short time Fourier transform, FFT implementation of short time Fourier transform, Nonuniform FIR filter bank implementation, FFT based nonuniform filter banks, tree structure realization of nonuniform filter banks using perfect reconstruction criterion and wavelet decomposition[22][29].

3.3 LINEAR PREDICTION ANALYSIS

The theory of linear predictive coding (LPC) or linear prediction (LP), has been well understood for many years. In here, how LPC analysis can be used in speech recognition will be explained. For this purpose mathematical formulas will be given, details of formulas can be found in reference papers and literature[17][16][15].

Before describing a general LPC front-end processor for speech recognition, it will be better to review the reasons why LPC has been so widely used. Here is the reasons:

- LPC provides a good model of the speech signal. These is especially true for the voiced regions of speech signal. All-pole model of LPC provides a good approximation to the vocal tract spectral envelope. During unvoiced and transient regions, LPC model is less effective but it still provides an acceptably useful model for speech recognition.

- LPC is an analytically tractable model. The method of LPC is mathematically precise and is simple and straightforward to implement in either software or hardware. The required computation for LPC analysis is considerably less than required computation for filter-bank analysis.

- The LPC model works well in speech recognition applications. Its performance is higher than recognizers based on filter-bank front ends[15][16].

3.3.1 The LPC Model.

The basic idea behind the LPC model is that given speech sample at time n, $s(n)$, can be approximated as a linear combination of the past p speech samples, such that

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p), \qquad (3.3.1.1)$$

where the coefficients $a_1, a_2, \ldots, a_p$ are assumed constant over the speech analysis frame. Eq. (3.3.1.1) can be written as follows when an excitation term , $Gu(n)$ added

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n), \qquad (3.3.1.2)$$

where $u(n)$ is a normalized excitation and G is the gain of the excitation. In z-domain the relation

$$S(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + GU(z), \qquad (3.3.1.3)$$

leading to transfer function

$$H(z)=\frac{S(z)}{GU(z)}=\frac{1}{1-\sum_{i=1}^{p}a_i z^{-i}}=\frac{1}{A(z)} \qquad (3.3.1.4)$$

Figure 3.3.3.1 shows a normalized excitation source, *u(n)*, being scaled by the gain, *G*, and acting as input to the all-pole system , *H(z)=1/A(z)*, to produce the speech signal *s(n)*. It is known that actual excitation is a pulse train for the voiced speech sounds or a random noise source for the unvoiced sounds which is shown in Figure 3.3.1.2 The excitation source is selected by the switch according to voice characteristic, and the gain of the source is estimated from the speech signal and scaled source is used for the filter *H(z)*. Thus the parameters of this model are unvoiced/voiced classification, pitch period for voiced sounds, the gain of excitation source, and the coefficients of the filter *H(z)*. These parameters all vary slowly with time and it is accepted that these parameters are constant in frame duration.

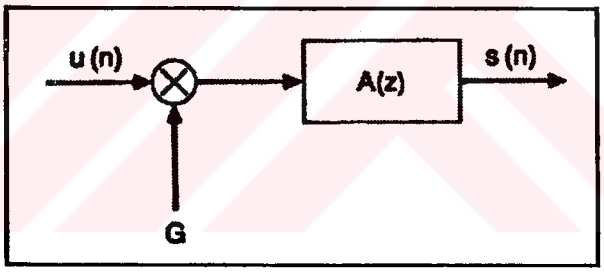


Figure 3.3.3.1. Linear prediction model of speech

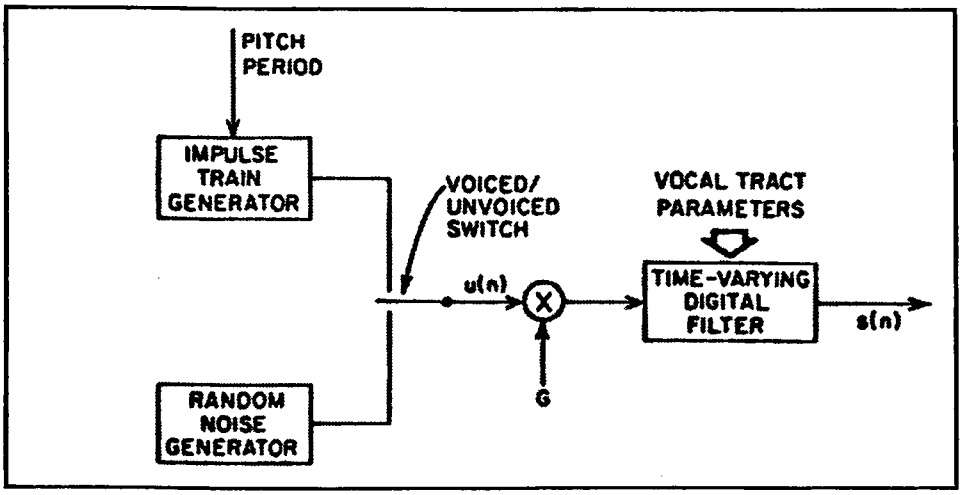


Figure 3.3.3.2. Speech synthesis model based on LPC model.

3.3.2 LPC analysis Equations.

The speech signal is formulated in Eq. (3.3.1.2) and the linear combination of past speech signal is used for estimated speech signal $\tilde{s}(n)$ defined

$$\tilde{s}(n)=\sum_{k=1}^{p} a_k s(n-k) \qquad (3.3.1.5)$$

The task of minimizing the prediction error, E_n, is to choose $a(k)$ such that

$$\begin{aligned} E_n &= \sum_{m} e_n^2(m) \\ E_n &= \sum [s_n(m)-\tilde{s}_n(m)]^2 \\ E_n &= \sum_{m} [s_n(m)-\sum_{k=1}^{p} a_k s_n(m-k)]^2 \end{aligned} \qquad (3.3.1.6)$$

is minimized. There are a lot of methods to solve this equation such as autocorrelation method, covariance method, classical least squares problem solution, lattice structure based on the L-D recursion, Itakura-Saito (parcor) lattice, Burg lattice and etc. According to [16][15], to solve this equations, the autocorrelation method with Levinson-Durbin recursion is the most efficient method when compared with required MIPS and memory. Levinson-Durbin recursion method will be explained later.

To illustrate some of the properties of the signal gathered with LPC analysis, Figure 3.3.2.1 shows a waveform for a wovel (a), prediction error signal(b), the signal spectrum(FF.-based) (c) fitted by an LPC log spectrum and the log spectrum of the error signal(d). The error signal is a almost 4 times smaller in magnitude than the speech signal and has a much flatter spectral trend than the speech signal. This the important characteristic of the LPC analysis whereby the error signal spectrum is approximately a flat spectrum signal representing the source characteristics rather than

those of the vocal tract. It can be seen that fairly close matches exist between the peaks of the FFT-based signal spectrum and the LPC spectrum.

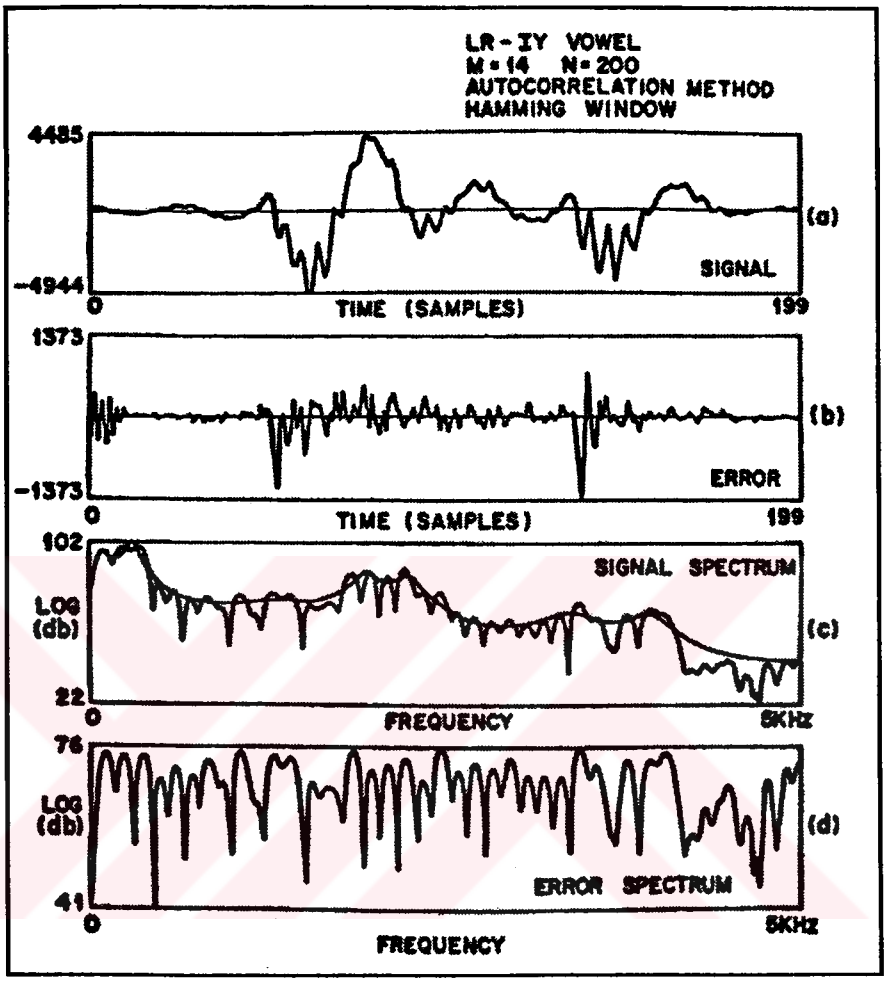


Figure 3.3.2.1. Typical signals and spectra for LPC autocorrelation method.

Figure 3.3.2.2 shows the effect of the prediction order p , on the RMS prediction error normalized by the signal energy. A sharp decrease in normalized prediction error occurs for small values of p such as p=1-4, beyond this value prediction error decreases more slowly. Normalized prediction error for the unvoiced signal significantly higher than for voiced signal because of the nonperiodic nature of the unvoiced signal.

Figure 3.3.2.3 shows the effect of prediction order, p, on the all-pole spectrum. It is clear that as p increases , more of the detailed properties of the signal spectrum are preserved in the LPC spectrum. It is seen in the figure that, LPC spectrum is the

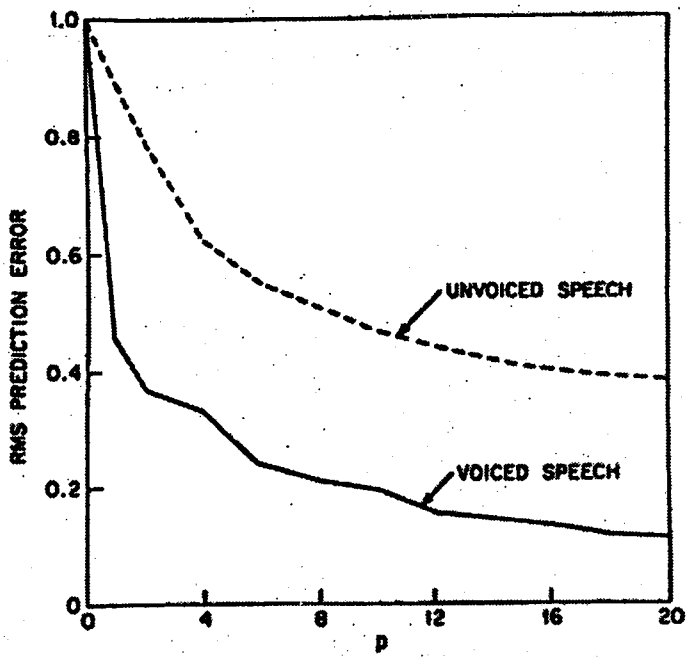


Figure 3.3.2.2. Variation of the RMS prediction error with the number of prediction error.

smoothed representation of the short time spectrum and gives more robust information about the spectrum of the signal for speech recognition applications due to good representation of the spectrum. On the extensive experimental evaluations, the order of the predictor, p, can be choosen 8-10 and it is reasonable for most speech recognition applications.

3.3.3 An Example of LPC Front-End Processor for Speech Recognition.

The details of example LPC front-end processor is explained in following texts. Figure 3.3.3.1 shows a block diagram of a such a processor. The basic steps and details of the block is described below.

3.3.3.1 Band pass filter and A/D.

The analog input signal must be band pass filtered between voice band. The begin and end frequency of the band pass filter is 100 and 3200 Hz for most of speech recognition system. The signal which band pass filtered is digitized by sampling at a fix frequency. The sampling frequency may be changed by decimation or interpolation methods, if necessary. Sampling frequency is selected at 6.67 Hz in this study[6].

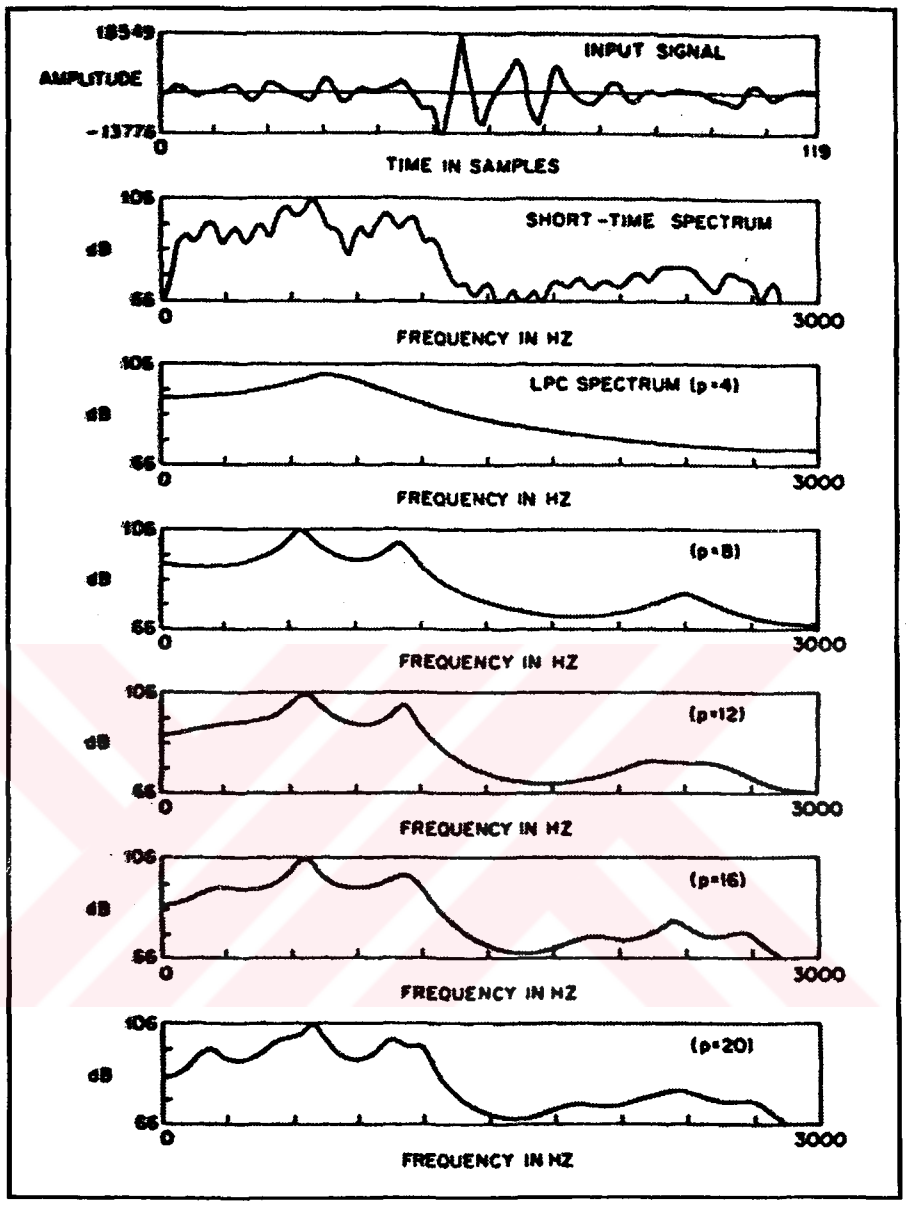


Figure 3.3.2.3. Spectra for a vowel sound for several values of predictor or p.

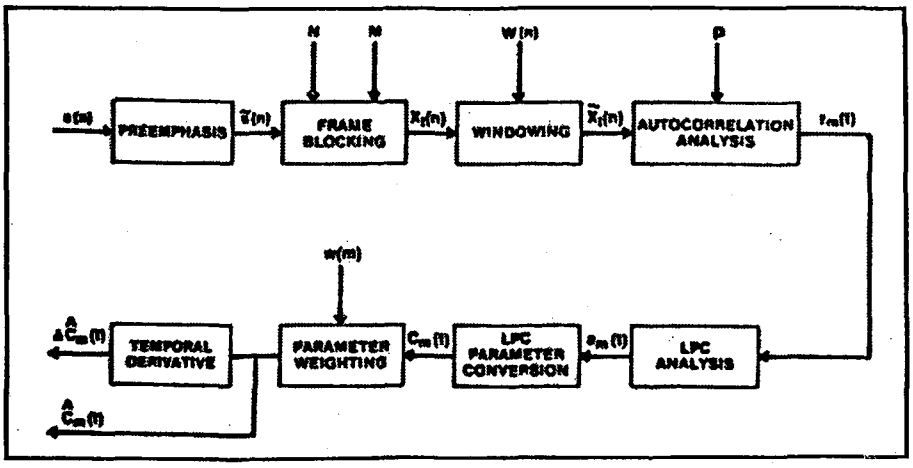


Figure 3.3.3.1. Block diagram of LPC front-end processor for speech recognition.

3.3.3.2 Pre-process of the sampled signal

The sampled signal is raw, unrefined signal. For robust speech recognition the sampled signal must be refined because the speech recognition rate is highly dependent to the noise in the speech. The kind of the noise may be a additive noise ,a car noise , musical noise or impulsive noise. Recognition rate may decrease dramatically when signal to noise ratio is low or in case of a presence of noise in the speech. A lot of classical and new techniques for noise suppression and speech enhancement available in the literature and new papers are briefly described below[7].

Spectral subtraction is well-known method to enhance speech corrupted by additive wideband noise. In this method, the "clean" signal is approximated by subtracting a noise estimate from the spectrum of the corrupted signal. Methods based on SS to suppress "musical" noise without affecting speech are available on the literature.[8].

Smoothed spectral subtraction (SSS) technique is an improved version of the well-known spectral subtraction algorithm. On the other side generalized Wiener filtering may be used for noise suppression. A new method vector quantization of line spectral frequencies (VQ-LSF) have a good success for the simulated noise and actual car noise[9].

Noise estimation techniques is used to estimate the noise spectra or the noise characteristics for noisy speech. Past noisy segment of the speech is needed for the estimation. The algorithm is able to quickly adapt to slowly varying noise levels[10].

Speech enhancement based on temporal processing like FIR-Wiener filters are applied to time trajectory of the cubic-root compressed short-term power spectrum of noisy speech. Informal listening indicate that the technique brings noticeable improvement to the quality of processed noisy speech while not causing any significant degradation to clean speech[11].

3.3.3.3 Preemphasis.

The digitized speech signal *s(n)* is put through a low order digital system (typically a first order FIR filter) , to flatten the signal spectrally.

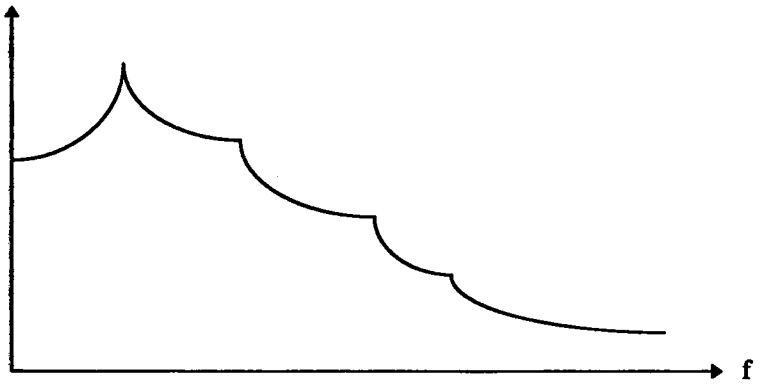


Figure 3.3.3.3.1 Typical spectral envelope of a voiced sound

Figure 3.3.3.3.1 shows a typical spectral envelope of the speech signal for a voiced sound. It seems that the spectrum rolls of in high frequencies. If this is permitted to happen , the LPC model will satisfactorily approximate the low frequencies, but it will do a poor job in high frequencies. To prevent this , the speech signal is passed through a filter which transfer function $1- az^{-1}$ called a preemphasis filter, which emphasizes the high frequencies before processing. Viewing it another way, the spectral rolloff is caused by the radiation effects of the sound from the mouth, and preemphasis filter is used to remove these effects. Typical values of coefficient *a* are around 0.9, with a usual choice a=15/16=0.9373 or a=0.95. Then if s(n) the input signal ,the preempahisized signal is below[12].

$$\tilde{s}(n) = s(n) - 0.95\, s(n-1) \qquad (3.3.3.3.1)$$

If preemphasis is used in speech synthesis the processed signal must be de-emphasized by using below formulation.

$$\tilde{s}(n) = s'(n) + 0.95\, s(n-1) \qquad (3.3.3.3.2)$$

In the speech recognition de-emphasis is not used because there is no synthesis. The pre-emphasis is necessary in speech recognition not to lose high frequency speech information during LPC feature extraction.

3.3.3.4 Frame Blocking.

In this step the preemphasized speech signal , $\tilde{s}(n)$, is blocked into frames of N samples, with adjacent frames being separated by M samples. The first frame consists of first N speech samples. The second frame begins M samples after the first frame, and overlaps it by N-M samples. Similarly, the third frame begins $2M$ samples after the first frame. If adjacent frames overlap, the resulting LPC spectral estimates will be correlated from frame to frame if $M \le N$ and if $M \ll N$ than spectral estimates from frame to frame will be quite smooth. Thus lth. frame of the speech , $\tilde{x}$, is given by

$$x_l(n) = \tilde{s}(Ml + n), \quad n = 0,1,....,N-1 \qquad l = 0,1,....,L-1 \qquad (3.3.3.4.1)$$

Typical values for N and M are 300 and 100 when the sampling rate is 6.67 kHz. These corresponds to 45 ms frames separated by 15 ms frames.

3.3.3.5 Windowing.

The next step in the processing is to window each individual frame so as to minimize signal discontinuities at the beginning and end of the each frame. The window function is defined as $w(n)$, the resulting windowed signal for each frame is

$$\tilde{x}_l(n) = w(n).x_l(n) \qquad 0 \le n \le N-1 \qquad (3.3.3.5.1)$$

A typical window used for the autocorrelation method of LPC is the Hamming window, which has the form

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}), \qquad N = 300 \qquad (3.3.3.5.2)$$

3.3.3.6 Autocorrelation Analysis.

Each frame of windowed signal is next autocorrelated to give

$$R_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m); \qquad m = 0,1,.....,8 \qquad (3.3.3.5.2)$$

where the highest autocorrelation value, p, is the order of the LPC analysis. Typically, values of p from 8 to 16 can be used, in this thesis p is selected 8. A side benefit of the autocorrelation analysis is that the zeroth autocorrelation is the energy of the current frame(analyzed frame).

3.3.3.7 LPC analysis.

The next processing step is the LPC analysis, which converts each frame of p+1 autocorrelations into an LPC parameter set in which set might be LPC coefficients, the reflection coefficients, the log area ratio coefficients, or cepstral coefficients. The formal and mostly used method of converting autocorrelation coefficients to an LPC parameter set is known Levinson-Durbin's method or Durbin's recursion. Durbin's recursion is then applied to calculate a set of PARCOR coefficient, k_i, $i = 1,2,.....,8$ and a prediction residual from the $R_l(m)$ for each frame as follows (the frame index l is suppressed)[6][15].

$$E^{(0)} = R'(0)$$

For i=1,2,...,8 do Eqs. (3.3.3.7.1) through (3.3.3.7.3):

$$k_i = \frac{[R'(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R'(i-j)]}{E^{(i-1)}} \qquad (3.3.3.7.1)$$

$$\alpha_j^{(i)} = \alpha_j^{i-1} - k_i \alpha_{(i-j)}^{(i-1)}; \qquad (j = 1,2,.....,i-1; i \neq 1) \qquad (3.3.3.7.2)$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \qquad (3.3.3.7.3)$$

Extract final residual, E, and LPC coefficients a_j :

$$E = E^{(8)} \qquad (3.3.3.7.4)$$

$$a_j = \alpha_j^{(8)} \qquad (3.3.3.7.5)$$

3.3.3.8 LPC Parameter Conversion to Cepstral Coefficients.

A very important LPC parameter set , which can be derived from the LPC coefficient set, is the LPC cepstral coefficients, *c(m)*, The recursion used is

$$c_o = ln\sigma^2 \qquad (3.3.3.8.1)$$

$$c_m = a_m + \sum_{k=1}^{m-1} (k/m) c_k a_{m-k}, \qquad 1 \leq m \leq p \qquad (3.3.3.8.2)$$

$$c_m = \sum_{k=1}^{m-1} (k/m) c_k a_{m-k}, \qquad m > p \qquad (3.3.3.8.3)$$

where σ^2 is the gain term in the LPC model. The cepstral coefficients, which are the Fourier transform representation of the log magnitude spectrum, have been shown to be a more robust, reliable feature set for speech recognition than the LPC coefficients, A more detailed description will be given next in the chapter that contains robust recognition and distance measures subjects.

CHAPTER 4. PATTERN COMPARISON TECHNIQUES

4.1 INTRODUCTION

Mostly used techniques for pattern comparison for speech recognition is to use time sequence of vectors obtained front-end spectral analysis. If test pattern, T, as the sequence of spectral frames over the duration of the speech, such that

$$T = \{t_1, t_2, \ldots, t_I\}$$

where each t_i is the spectral vector of the input speech at time i, and I is the total number of frames of speech. In a similar manner, a set of reference patterns can be defined as $\{R^1, R^2, \ldots, R^V\}$ where each reference pattern, R^j, is also a sequence of spectral frames, such that

$$R^j = \{r_1^j, r_2^j, \ldots, r_J^j\}$$

The goal of the pattern comparison stage is to determine dissimilarity or distance T to each of the R^j, $1 \le j \le V$, in order to identify the reference pattern that has the minimum dissimilarity, and to associate the spoken input with this pattern.To find the global similarity of T and R^j, the following problems should be considered.

- T and R^j generally are unequal lengths in time duration. This is because of different speaking rates.
- T and R^j need not line up in time. This is because different sounds can not be varied in duration to the same degree. Thus the vowels are easily lengthened ; however, most consonants can not changed dramatically in duration.
- A way to compare pairs of spectral vectors is needed. A local dissimilarity or distance measure and a global similarity measures is needed.

The concept of a spectral dissimilarity measure is equally applicable to both the feature based approach and template-based approach. For template based systems, the dissimilarity measure that are used have strong physical interpretations in terms of spectral difference or spectral distortion between the pair of spectra.

The other problem should be solved by defining a global dissimilarity between the pairs of patterns. A side problem is to find begin and end of utterated speech signal. This problem is called endpoint-detection problem.

4.2 SPEECH ENDPOINT DETECTION.

The goal of speech endpoint detection is to separate events of interest in a continuously recorded signal from other parts of the signal such as background noise. The need for speech detection occurs in many application in telecommunications. In speech recognition applications, endpoint detection should be accurate for the test speech signal and reference speech signals. By changing the begin and end of utterances by manually, the effect of endpoint detection can be seen. An experiment was done in AT&T [18] on a large database, it has been seen that if error is ±60 ms (4 frames) than %3 reduction occurs in digit accuracy. As the endpoints are moved farther away from the reference endpoints , recognition accuracy decreases uniformly. While the results are undoubtedly dependent on the particular recognizer implementation and used endpoint detection algorithm.

Carefully articulated and spoken speech signal endpoint detection problem is a simple problem in relatively noise-free environments. But this is not usually the case. In practice one or more problems usually make accurate endpoint detection difficult. For example, during articulation the talker often produces sound artifacts, including lip smacks, heavy breathing, mouth clicks and pops. High performance endpoint detector must handle this kind of artifacts.

Another factor in the robust endpoint detection is the environmental conditions in which speech is produced. The ideal environment is quite room , but this

is not practical, so, endpoint detector must consider background noise due to running machine, fans, irregulated road noise etc.

Many methods have been proposed for speech detection to use in speech recognition. These methods can be explicit approach, the implicit approach, and hybrid approach.

The explicit approach is based on the premise that speech detection can be accomplished independent of other pattern matching operations in the next stage. The detected speech then send to the pattern-comparison algorithm. For signals with stationary and low level noise background, the approach has reasonable good detection accuracy. The approach fails when the environments is noisy.

The implicit approach considers the speech detection problem simultaneously with the pattern matching and recognition decision process. Endpoint candidates are tested with pattern comparison and the best match endpoint candidate is accepted as endpoints of the speech. The computational load of the implicit approach is heavily high but offers a potentially higher detection accuracy in noisy conditions.

In this thesis, explicit method has been used because of its simplicity. In the following text a simple version of Lamel [19] endpoint detector is described.

In Figure 4.2.1 ,

$$E(l)=10\log_{10} R_l(0), \qquad l=1,2,...,\mathrm{NF}$$

where NF is the total number of frames in the recording interval, $R_l(0)$ is the zeroth order correlation coefficients, that is, the energy of the signal. The next step in the processing (called adaptive-level equalization) is a normalization of the energy contour to compensate for the mean background noise level. First E_{min} is computed as

$$E_{min} = \min_{1\le l\le NF} (E(l))$$

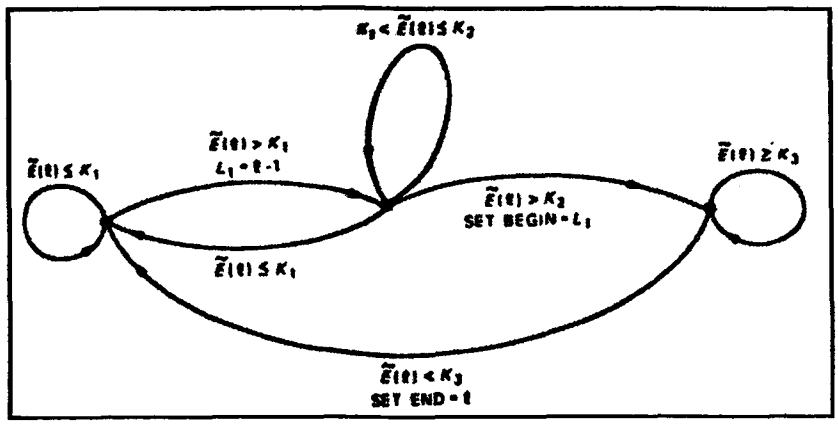


Figure 4.2.1. State diagram of endpoint detector.

$\tilde{E}(l)$ is then formed as the difference between $E(l)$ and E_{min}

$$\tilde{E}(l) = \tilde{E}(l) - E_{min} \qquad l = 1,2,...,NF$$

The remaining blocks of the detector are energy pulse detector and endpoint ordering procedure. Pulse combining rules are used to eliminate short pulses and combine close pulses. Several parameters are listed below with their definitions.

1. *K1, K2 and K3* are energy thresholds used in determining the word boundaries (3,10,5 dB)
2. IT1 and IT2 are frame counter thresholds for determining the presence or absence of any breath noises at the boundary points of a detected utterances (5,5 frames).
3. IT3 is the minimum length for a detected pulse (5 frames).
4. NFMIN is the minimum length in frames for an utterance (10 frames).

Figure 4.2.1 shows a state representation of the operation of the energy pulse detector. The normalized energy of the signal is scanned from left to right. If $\tilde{E}(l)$ rises first above *K1*, then above *K2* (without falling below *K1*) a beginning pulse marker is assigned to frame *l*. Similarly , when the energy dips below *K3*, ending marker is assigned. The beginning IT1 and ending IT2 frames are checked for breath-type noises.

This speech endpoint detector is reasonable for sample utterated words in this thesis and good results obtained by using simple version of Lamel's detector.

4.3 DISTORTION MEASURES

Pattern recognition algorithms uses a measurement of dissimilarity between two feature to find how much they are similar to each other. This measurement of dissimilarity can be handled with mathematical methods. A dissimilarity measurement must satisfy following equations.(x and y defined vector)

(a) $0 \le d(x,y) \le \infty$ for x, y and $d(x,y) = 0$ if and only if $x = y$

(b) $d(x,y) = d(y,x)$ for x, y

(c) $d(x,y) \le d(x,z) + d(y,z)$ for x, y, x

A mathematical measure of distance, to be useful is speech processing, has to have a high correlation between its numerical value and the subjective distance judgment, as evaluated on real speech signals. For example, a distortion $d(x,y) = (x-y)^2$ measure may be a bad distortion measure for most kind of the speech features. Most of distance measures are related with feature type. The same mathematical formulas mostly can not be used for different type of features.

4.4 SPECTRAL DISTORTION MEASURES.

Measuring the difference between two speech patterns in terms of average or accumulated spectral distortion is one of the reasonable way of comparing patterns. Psychoacoustic characteristic of the sound varies with the spectral characteristic, so spectral distortion measures can be used for dissimilarity of the comparing patterns.

4.4.1 Log Spectral Distance.

Given two spectra $S(\omega)$ and $S'(\omega)$, the difference between the two spectra on a log magnitude versus frequency scaled is defined by

$$V(\omega) = \log S(\omega) - \log S'(\omega) \qquad (4.4.1.1)$$

Distortion measure between $S(\omega)$ and $S'(\omega)$ is the set of L^p norms define by

$$d(S,S')^p = (d_p)^p = \int_{-\pi}^{\pi} |V(\omega)|^p \frac{d\omega}{2\pi} \qquad (4.4.1.2)$$

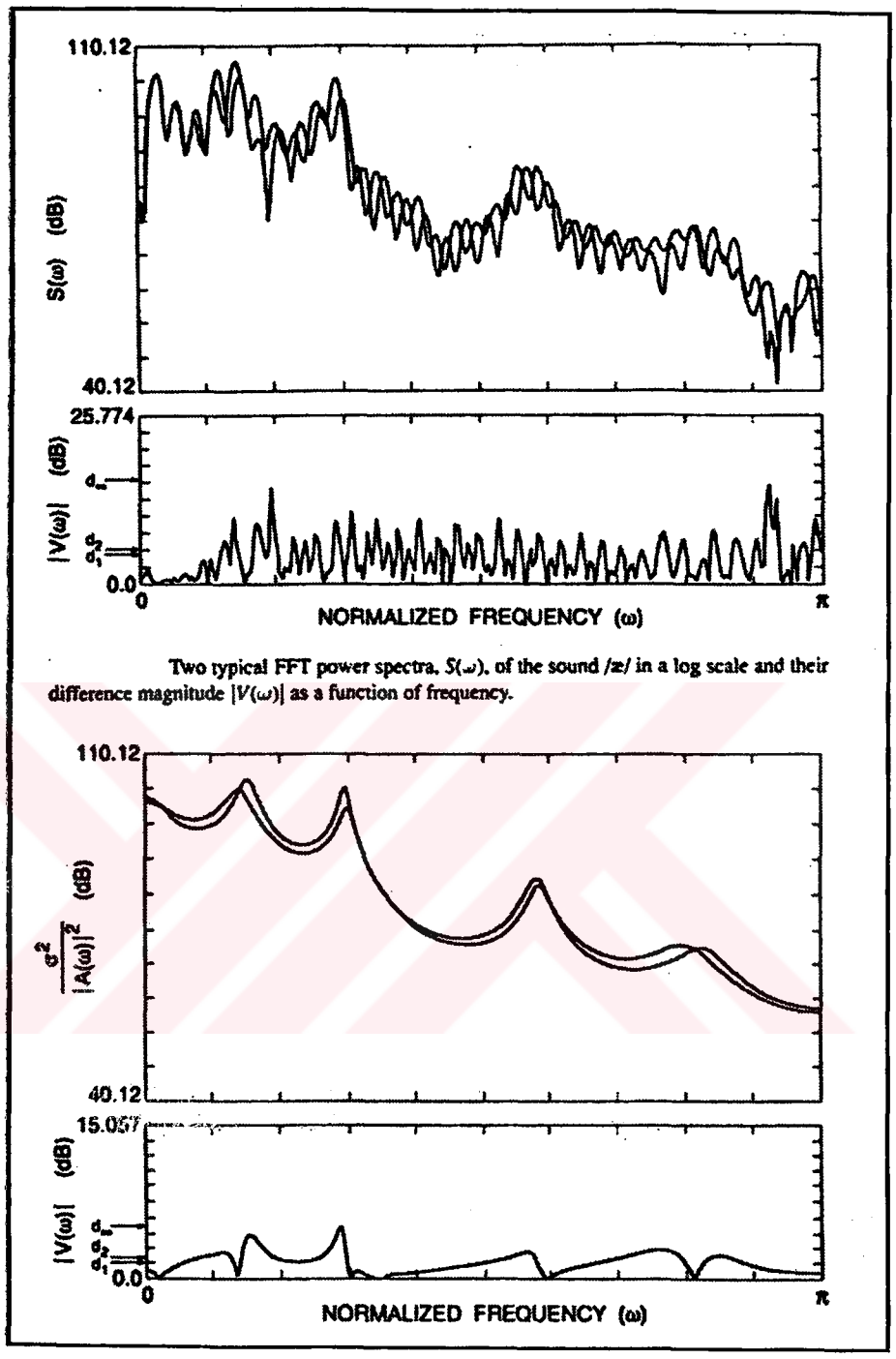
For p=1 ,Eq. (4.4.1.2) defines the mean absolute log spectral distortion. For p=2, Eq. (4.4.1.2) defines the rms log spectral distortion that has found applications in many speech processing system[15].

Representative example of log spectral distortions of the typical speech spectra are shown in Figure 4.4.1.1. The log spectral differences $|V(\omega)|$, as computed from the FFT of the data sequence, is shown in the figure to be a quite irregular. A major part of the variations comes from the difference in the fundamental frequency, which have a little effect on the perceived phonetic content of the steady-state vowel. The log spectral difference between the two all-pole models of the spectra $\left| \log \frac{\sigma^2}{|A(e^{j\omega})|^2} - \log \frac{\sigma'^2}{|A'(e^{j\omega})|^2} \right|$ is much smoother than the FFT counterpart. The smooth spectral difference function allows a closer examination of the properties of the distortion measure. L^p distortion measure do not take the known and well-understood perceptual masking effect, so is not very efficient for speech recognition.

4.4.2 Cepstral Distances.

The complex cepstrum of a signal is defined as the Fourier transform of the log of the signal spectrum. The Fourier series representation of $\log S(\omega)$ can be expressed as

$$\log S(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{-jn\omega}, \qquad (4.4.2.1)$$



Two typical FFT power spectra, $S(\omega)$, of the sound /æ/ in a log scale and their difference magnitude $|V(\omega)|$ as a function of frequency.

Figure 4.4.1.1 Two typical FFT spectra and two typical LPC model spectra and their difference magnitude.

where $c_n = c_{-n}$ are real and often referred to as the cepstral coefficients. For a pair of the spectra , $S(\omega)$ and $S'(\omega)$, it can be related cepstral distance of the spectra to the rms. log spectral distance :

$$d_2^2 = \int_{-\pi}^{\pi} \left| \log S(\omega) - \log S'(\omega) \right|^2 \frac{d\omega}{2\pi} = \sum_{n=-\infty}^{\infty} (c_n - c_n')^2 \qquad (4.4.2.2)$$

where c_n and c_n' are the cepstral coefficients of the $S(\omega)$ and $S'(\omega)$ respectively[15].

4.4.3 Weighted Cepstral Coefficient and Liftering.

The usefulness of the cepstral distances is simple and efficient method for estimating the rms. log spectral distance. Several other properties of the cepstrum, when properly applied , are beneficial for speech recognition applications.

It can be shown that [20] cepstral coefficients have zero means and variances essentially inversely proportional to the square of the coefficient index. If the n^2factor (n is the index of the coeff.) is incorporated into the cepstral distance so as to normalize(by the variance inverse) the contribution from each cepstral term, the distance of Eq. (4.4.2.2) becomes

$$d^2{}_{2W} = \sum_{n=-\infty}^{\infty} n^2 (c_n - c_n')^2 \qquad (4.4.3.1)$$

This distance measure is more robust than Eq. (4.4.2.2) [15][20].

Variability of higher cepstral coefficients are more influenced by inherent artifacts of the LPC analysis (due to all-pole constraint, analysis window position and etc.). The variability of low cepstral coefficients, while less effected by the analysis mechanism, is primarily due to variations in transmission, speaker characteristics, vocal efforts, and other factors[20]. For example, the effect of differences in frequency response roll off in various telephone channels is usually most prominent in the first few cepstral coefficients. The speaker's glottal shape and vocal cord duty

cycles effects mainly the first few cepstral coefficients. For recognition of speech, independent of talker's identity, these sources of the variability, that are not essential in representing the phonetic content of the sound, need to be de-emphasized.

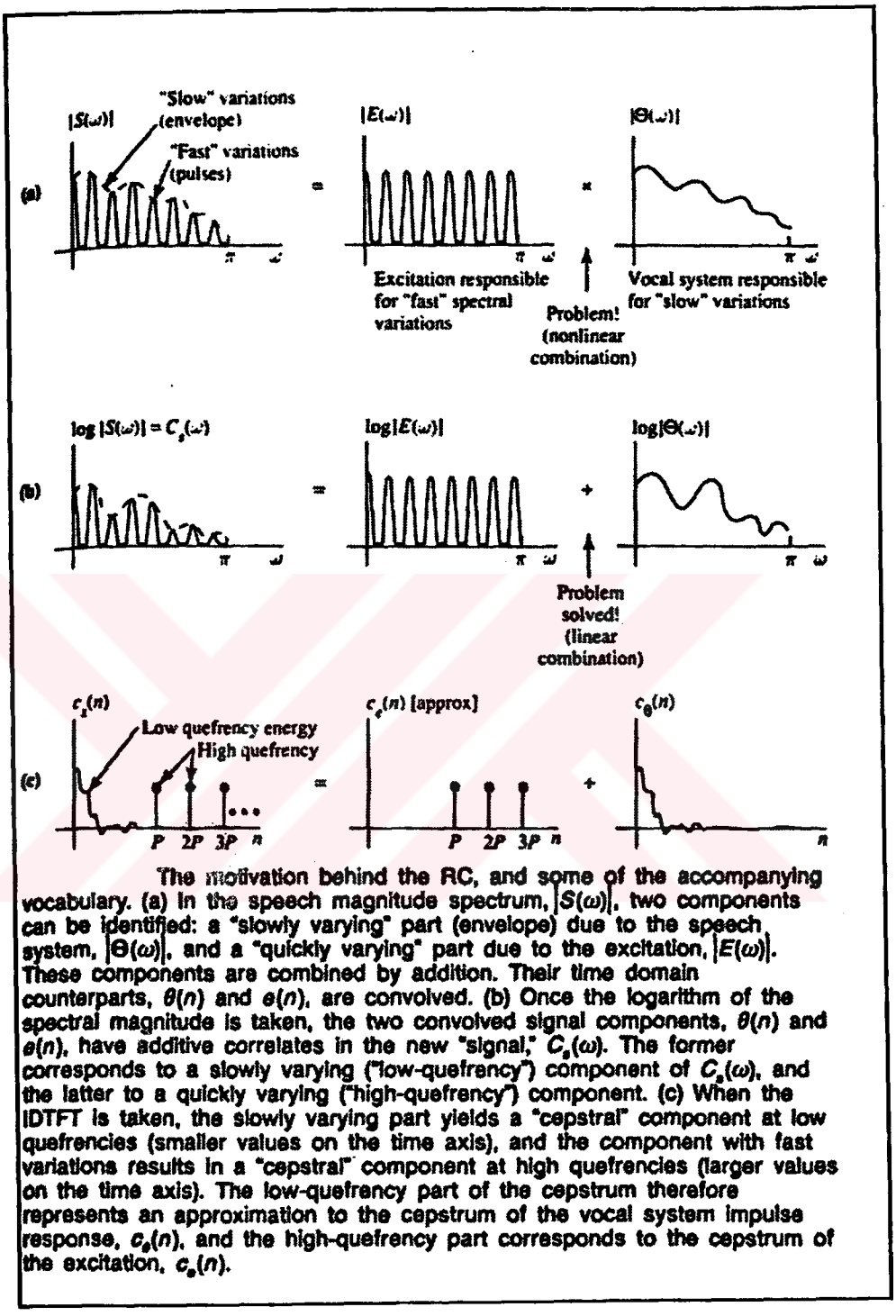
In Figure 4.4.3.1 explains very well why cepstral coefficients should be used for speaker independent speech recognition. Spectrum due to vocal track and spectrum due to excitation can be separated by logarithm operation. So if spectrum due to excitation is eliminated, the spectrum due the vocal tract will be essential for comparing spectrums[16].

A cepstral weighting, or liftering procedure , w(n), can be therefore be designed to control the noninformation-bearing cepstral variabilities for reliable discrimination of sounds. The index weighting, as used in Eq. (4.4.3.1) , is one of the simple form of the cepstral weighting. Another, more sophisticated weighting is raised sine function of the form

$$w(n)=\begin{cases}1+h\sin(\frac{n\pi}{L}) & \text{for } n=1,2,\ldots,L \\ 0 & \text{for } n\le 0,\ n\ge L\end{cases} \qquad (4.4.3.2)$$

Where h is usually chosen as $L/2$ and L is typically 10~16 for speech of 4 kHz bandwidth. The weighted sequence ,$w(n)c_n$, correspond to a smoothed of power spectrum. Figure 4.4.3.2 shows the effect of the liftering process through a sequence of spectral plots. Figure 4.4.3.2 (a) is a hidden line plot of a series of 30 consecutive LPC log spectra, corresponding to a wovel-like sound. The randomness of the spectral components and the sharp spectral peaks that lead to excessive spectral sensitivity are clearly shown, Figure 4.4.3.2 (b) is the same log spectral plot after the lifter defined in Eq (4.4.3.2) with L = 12 is applied. The undesirable (noiselike) components of the LPC spectrum are reduced or removed, while the essential characteristics of the "formant" structure are retained[20].

Based on the above, a useful form of weighted cepstral distance is thus where $w(n)$ is any reasonable lifter function. This kind of distance measures are robust against environmental noises or adverse conditions.



The motivation behind the RC, and some of the accompanying vocabulary. (a) In the speech magnitude spectrum, $|S(\omega)|$, two components can be identified: a "slowly varying" part (envelope) due to the speech system, $|\Theta(\omega)|$, and a "quickly varying" part due to the excitation, $|E(\omega)|$. These components are combined by addition. Their time domain counterparts, $\theta(n)$ and $e(n)$, are convolved. (b) Once the logarithm of the spectral magnitude is taken, the two convolved signal components, $\theta(n)$ and $e(n)$, have additive correlates in the new "signal," $C_s(\omega)$. The former corresponds to a slowly varying ("low-quefrency") component of $C_s(\omega)$, and the latter to a quickly varying ("high-quefrency") component. (c) When the IDTFT is taken, the slowly varying part yields a "cepstral" component at low quefrencies (smaller values on the time axis), and the component with fast variations results in a "cepstral" component at high quefrencies (larger values on the time axis). The low-quefrency part of the cepstrum therefore represents an approximation to the cepstrum of the vocal system impulse response, $c_\theta(n)$, and the high-quefrency part corresponds to the cepstrum of the excitation, $c_e(n)$.

Figure 4.4. 3.1. Illustration of cepstral processing.

Another distance measure is defined by Juang[21]. According to Juang, it has been shown that additive white noise reduces the norm of the cepstral coefficients. The reduction of vector norm is a function of the SNR, the lower the SNR, the more

the reduction. It is further observed that, compared to the traditional Euclidean distance,

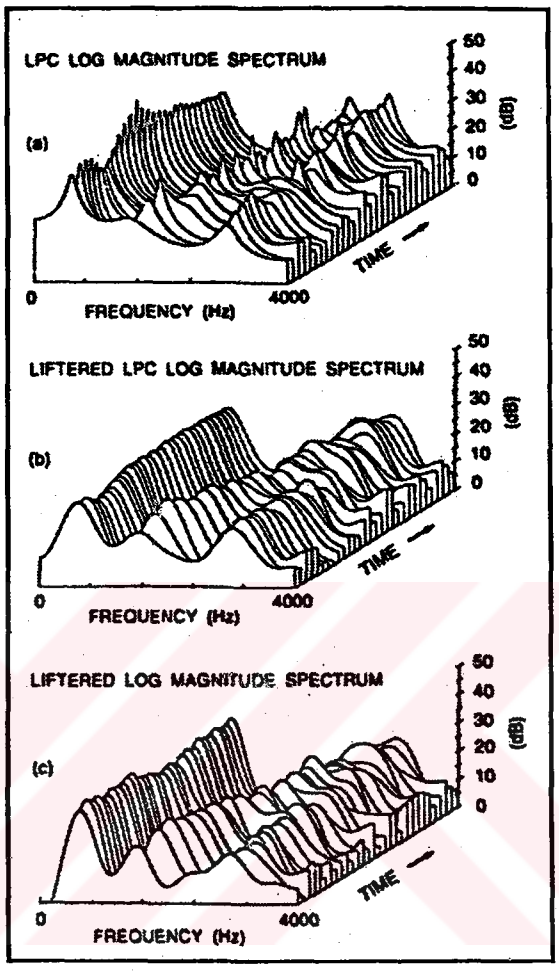


Figure 4.4.3.2. Comparison of (a) original sequence of LPC magnitude spectra; (b) liftered LPC log magnitude spectra; and (c) liftered log magnitude spectra.[21]

the angle between two cepstral vectors is less sensitive to additive white noise. Therefore this measure is a robust distortion measure. In [21] a family of distortion measures have been given to compensate the effect of the noise in the speech. One of the distance measure-will be explained in the next chapter - is used in this thesis to see performance of the distortion measure compared to other measures such as Itakura LPC distance measure, cepstral distance measure , cepstral plus lifter distance measure.

4.5 TIME ALIGNMENT AND NORMALIZATION

An utterance, which is to be recognized, involves a sequence of short time acoustic features. The pattern comparison technique for speech recognition has to be able to compare sequences of acoustic features.

The problem with spectral sequence comparison for speech comes from the fact that different acoustic renditions, or tokens, of the same speech utterance (e.g., word, phrase, sentence) are seldom realized at the same speed across the entire utterance. When comparing different tokens of the same utterance, speaking rate variations as well as duration variation not contribute to the dissimilarity score. Thus there is a need to normalize out speaking rate variations in order for the utterance comparison to be meaningful before a recognition decision can be made [15].

Time registration of a test and a reference pattern is one of the fundamental problems in the area of the automatic isolated word recognition. This problem is important because the time scales of a test and reference patterns are generally not perfectly aligned. In some cases the time scales can be registered by a simple linear compression and expansion, however in most cases , a nonlinear time warping is required to compensate for local compression of the time scale. For such cases, the class of the algorithms known as dynamic time warping methods have been developed. These algorithms have been shown to be applicable to the isolated speech recognition problem and greatly improve the accuracy of such systems[5]. Time alignment procedure is defined in the following text.

Consider two speech pattern X and Y represented by the spectral sequences $(x_1, x_2, \ldots, x_{T_x})$ and $(y_1, y_2, \ldots, y_{T_y})$,respectively, the dissimilarity between X and Y is defined by considering some function of short-time spectral distortions $d(i_x, i_y)$ where i_x =1,2,...,T_x and i_y=1,2,...,T_y. The interaction between these sequential constraints and the natural speaking rate variation constitutes one of the central problems in speech recognition, that is, that of time alignment and normalization.

The simplest solution to time alignment problem and normalization, is a linear time normalization technique. In linear time normalization, the dissimilarity between X and Y *is* defined

$$d(X,Y)=\sum_{i_x=1}^{T_x} d(i_x,i_y) \qquad (4.5.1)$$

where i_x, i_y satisfy

$$i_y = \frac{T_y}{T_x} i_y \qquad (4.5.2)$$

Figure 4.5.1 shows linear time alignment of two sequences of different duration. Linear time normalization alignment implicity assumes that the speaking rate variation is proportional to the duration of the utterance and is independent of the sound being spoken. Therefore, evaluation of the distortion measure takes places along the diagonal in the figure. Obviously, this rigid constraint of invariant speaking rate fluctuation does not adequately model the true situation for real speech utterances.

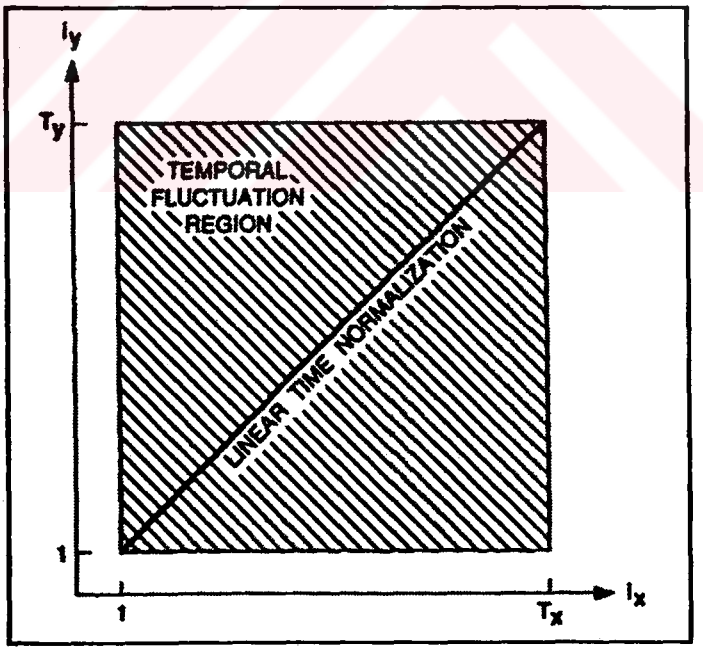


Figure 4.5.1. Linear time alignment for two sequences of different duration.

A more general time alignment and normalization scheme involves the use of two warping functions Φ_x and Φ_y, which relate the indices of the two speech patterns.

$$i_x = \Phi_x(k) \qquad k=1,2,...,T \qquad (4.5.3)$$

$$i_y = \Phi_y(k) \qquad k=1,2,...,T \qquad (4.5.4)$$

A global pattern dissimilarity measure $d_\Phi(X,Y)$ is defined based on the warping function pair $\Phi=(\Phi_x,\Phi_y)$ as the accumulated distortion over the entire utterance, that is,

$$d_\Phi(X,Y)=\sum_{k=1}^{T} d(\Phi_x(k),\Phi_y(k))m(k)/M_\Phi \qquad (4.5.5)$$

where $d(\Phi_x(k),\Phi_y(k))$ is a short time spectral distortion measure (kinds of distortion measures was explained before), $m(k)$ is a nonnegative path weighting coefficient and M_Φ is a path normalizing factor. Figure 4.5.2 shows an example of general time normalization scheme. The solid line in the lower grid represents the path along which $d_\Phi(X,Y)$ is evaluated. The indices i_x and i_y as functions of the normal time scale k, have been shown at the top of the figure.

To specify the path $\Phi=(\Phi_x,\Phi_y)$ is to define dissimilarity $d(X,Y)$ as the minimum of $d_\Phi(X,Y)$, over all possible paths, such that

$$d(X,Y)=\min_{\Phi} d_\Phi(X,Y) \qquad (4.5.6)$$

where Φ must satisfy a set of requirements. Finding the "best" alignment between a pair of patterns is functionally equivalent to finding the "best" path that minimize the distortion between compared utterances. The specific form of the accumulated distortion of Eq (4.5.5), namely, the problem of finding the optimal path between compared utterances ,can be found by using dynamic programming techniques [4][13][14][15]. The dynamic programming technique must be able to choose the

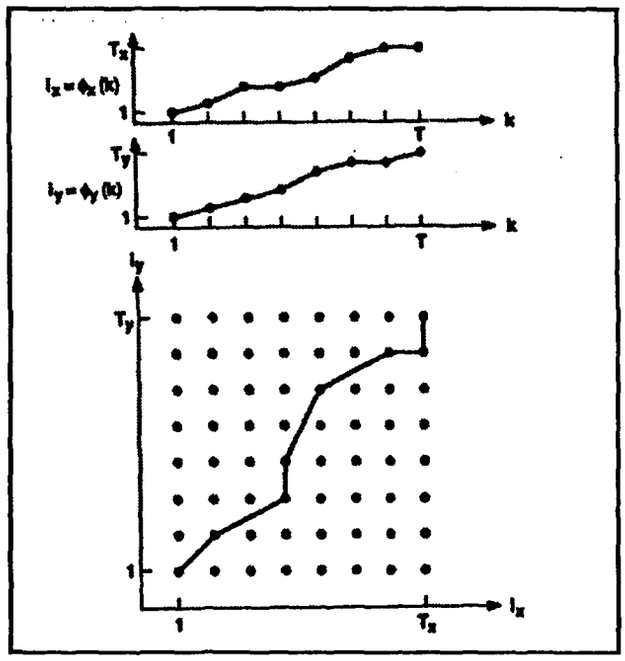


Figure 4.5.2. An example of time normalization of two sequential pattern to a common time index ; time warping functions.

correct path (the correct path minimizes the accumulated distortion measure) by discarding a lot of paths. In the following text dynamic time warping procedure with dynamic programming will be described. For alignment process to be meaningful in terms of time normalization for different utterances, some constraints on the warping functions are necessary. These constraints are described below.

In order to find the best path in the Φ_x and Φ_y plane, based on the parametric formulation, several factors of the DTW algorithm must be specified, including[5]. Figure 4.5.3 shows an example for possible paths with constraints. Here is constraints.

1) endpoint constraints on the path,
2) monoticity conditions,
3) local continuity constraints,
4) global path constraints,

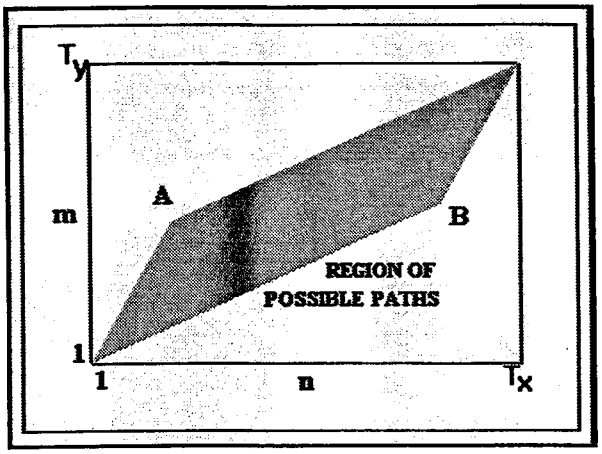


Figure 4.5.3. Allowable region of time warping paths for a constraint on the maximum ratio of duration's of test and reference utterances.

A. Endpoint Constraints

For the case of isolated words with precisely determined endpoints for the reference and test patterns, It is denoted that T_x and T_y are endpoints of the of the compared utterances. It is assumed that initial points and final points of the test and reference pattern are the same and formulated below[13].

beginning point $\Phi_x(1)=1, \quad \Phi_y(1)=1$ (4.5.7)

ending point $\Phi_x(T)=T_x \quad \Phi_y(T)=T_y$ (4.5.8)

B. Monoticity Conditions.

The temporal order of the spectral sequence is a speech pattern has a crucial importance to linguistic meaning. To maintain temporal order while performing the time normalization, it is reasonable to impose a monoticity constraint of the form.

$$\Phi_x(k+1)\geq\Phi_x(k) \tag{4.5.9}$$

$$\Phi_y(k+1)\geq\Phi_y(k) \tag{4.5.10}$$

C. Local Continuity Constraints.

To ensure proper time alignment while keeping any potential loss of information to a minimum, generally , a set of local continuity constraints on the warping function are added. One example, proposed by Skoe and Chiba [14] is

$$\Phi_x(k+1)-\Phi_x(k)\leq 1 \qquad (4.5.11)$$

$$\Phi_y(k+1)-\Phi_y(k)\leq 1 \qquad (4.5.12)$$

Such constraints specifications are often quite complicated and therefore it is convenient to express them in terms of incremental path changes.

D. Global Path Constraints.

Equations require that $\Phi(x, y)$ monotically increase with a maximum slop of 2, and minimum slop of 0 except when the slop at the preceding frame was 0, in which case the slope is 1. The boundary conditions together with the continuity conditions constraints the warping function $\Phi(x, y)$ to lie within parallelogram in the (n,m), as shown in the figure 3.3.5.2. The vertices (labeled A and B in fig) are obtained as the intersections of the lines[13]

$m-1 = 2(n-1)$ point A

$m-T_y = (n-T_x)/2$ (4.5.13)

and

$m-1 = 1/2(n-1)$ point B

$m-T_y = 2(n-T_x)$ (4.5.14)

The DTW function $\Phi(x, y)$ is therefore constrained to fallow a path inside the shaded region in Figure 4.5.3.

These are the sample constraints when warping area is selected as in Figure 4.5.3. They can be generalized for different assumptions.

4.6 DYNAMIC TIME-WARPING SOLUTION.

To minimize pattern similarity measure of Eq. (4.5.6), constraints explained before and dynamic programming are used together. The minimum partial accumulated distortion along a path connecting $(1,1)$ and (i_x, i_y) is

$$D(i_x, i_y) = \min_{\Phi_x, \Phi_y, T} \sum_{k=1}^{T} d(\Phi_x(k), \Phi_y(k)) m(k) \tag{4.6.1}$$

where

$$\Phi_x(T) = i_x \quad \text{and} \quad \Phi_y(T) = i_y \tag{4.6.2}$$

are implied. The dynamic programming recursion with constraints thus becomes

$$D(i_x, i_y) = \min_{(i_x', i_y')} \left[D(i_x', i_y') + \zeta((i_x', i_y'), (i_x, i_y)) \right] \tag{4.6.3}$$

where ζ is weighted the local distance between point (i_x, i_y) and (i_x', i_y')

$$\zeta((i_x', i_y'), (i_x, i_y)) = \sum_{l=0}^{L_s} d(\Phi_x, \Phi_y) m(l) \tag{4.6.4}$$

where L_s is the number of moves in the path from (i_x', i_y') to (i_x, i_y).

The incremental distortion ζ is evaluated only along the allowable paths as defined by the chosen local continuity constraints for efficient implementation of the DP. Table 4.6.1 summarizes the key recursion formula for several type of local continuity constraints with typical slope weighting.

An algorithmic approach given with its steps to evaluate dynamic time warping procedure is below. The beginning point is at (1,1) and ending point is at (T_x, T_y).

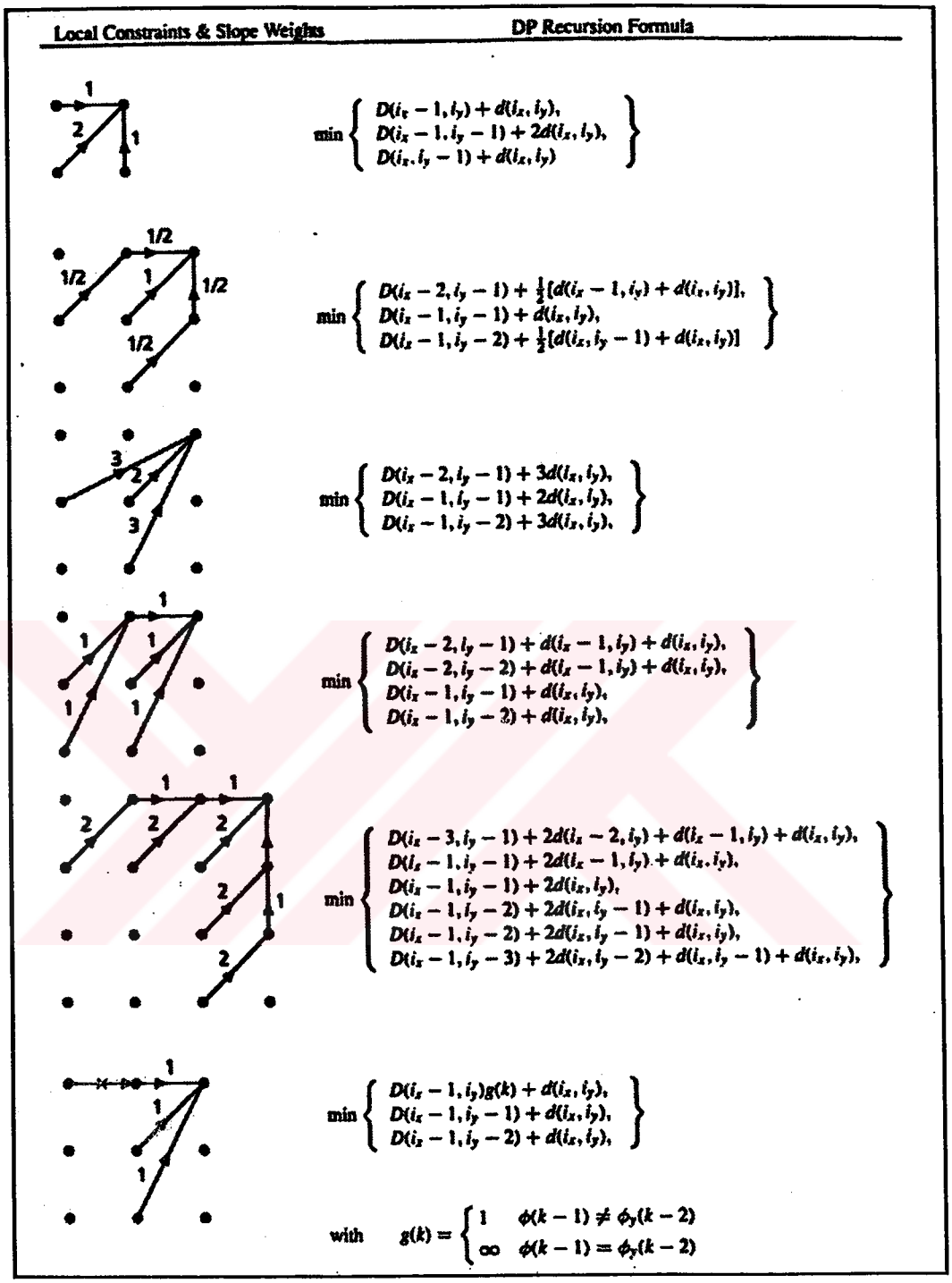


Table 4.6.1. Local constraints.

1. Initialization

$$D_A(1,1)=d(1,1)m(1)$$

2. Recursion

For $1\le i_x\le T_x$, $1\le i\le T_y$ such that i_x and i_y stay within the allowable grid, compute

$$D_A(i_x,i_y)=\min_{(i_x',i_y')}\left[D(i_x',i_y')+\zeta((i_x',i_y'),(i_x,i_y))\right]$$

where $\zeta((i_x',i_y'),(i_x,i_y))$ is defined in Eq. (4.6.4).

3. Termination

$$d(X,Y)=\frac{D_A(T_x,T_y)}{M_\Phi}$$

This is the basic recursion for DTW (dynamic time warping).

CHAPTER 5. IMPLEMENTATION AND RESULTS

5.1 INTRODUCTION

In the previous chapters, every stage of a speech recognizer was explained. A brief summary has been given before implementation. The first step in the recognizer is the analysis stage. In this stage, features about speech signal is obtained. In the speech recognition, spectral information of the speech signal is mostly used for comparing utterances. So, generally, speech features represent spectral information of speech signal. One of the used popular analysis method is the LP(linear predictive) analysis. In this analysis, a set of predictive coefficients are obtained. By using LP coefficients, some other features can be obtained such as cepstral coefficients.

After analysis stage, the test and reference utterances will be compared how much they are similar to each other in spectral domain. There is two main problem for comparison; local distance measure for feature(spectral vector) comparison and time alignment problem due to variation of speech rate. What kind of distance measure should be used in the framework of the linear prediction technique. In order to discuss this question, let us consider a more simplified problem; what is the optimal distance measure to find similarity of the spoken words. An approach to this problem can be described from the statistical point of view, and it is shown that the log likelihood ratio, which is one of the best criterion to test patterns, is reduced to the logarithm of the ratio of the prediction residuals, and can be used as a powerful distance measure. Other local distance(distortion measures) measure are cepstral distances as explained previous text.

The time alignment problem is solved by using dynamic time warping process. Time registration of a test and a reference pattern is one of the fundamental problems in the area of the automatic isolated word recognition. This problem is important

because the time scales of a test and reference patterns are generally not perfectly aligned. In some cases the time scales can be registered by a simple linear compression and expansion, however in most cases , a nonlinear time warping is required to compensate for local compression of the time scale. For such cases, the class of the algorithms known as dynamic time warping methods have been developed.

After DTW, an ordered accumulative distance measure between test utterance and reference utterances are used for decision. Generally, the reference utterance with minimum accumulative distance is accepted as recognized utterance after setting a meaningful distance threshold. In the following text, implemented speech recognizer block diagram and explanations and results will be given.

5.2 IMPLEMENTED ISOLATED WORD SPEECH RECOGNIZER.

Implemented DTW based isolated speech recognition block diagram shown in Figure 5.2.1.

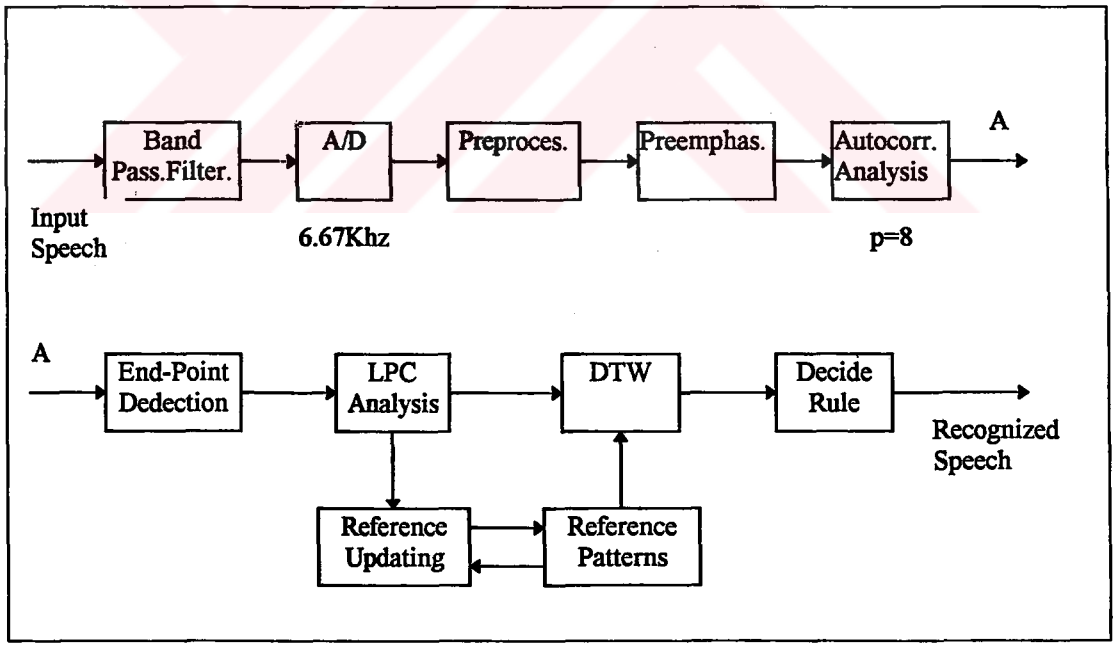


Figure 5.2.1. implemented DTW based speech recognition block diagram.

The speech utterances (in this implementation utterances mean isolated words) have been obtained by using a sound card on PC(Personal Computer). The spoken word are sampled at 11.025 kHz and a bandpass filter is used to limit signal spectrum between 100-3300 Hz. Then bandpassed speech is sampled at 6.67 Khz. Sampling at

6.67 Khz is a decimating process. Preprocessing such as noise suppressing is not used, then preemphasis and autocorrelation process is applied to speech signal. The analysis frame is 300 sample overlapped 200 sample namely 45 ms analysis frame with 15 ms intervals. The first autocorrelation coefficient (namely energy for that frame) is used for endpoint detection. After endpoint detection LPC analysis process begins and during LPC analysis LP coefficients are obtained for prediction order p equal 8. Obtained features are stored in files as a reference files for reference utterances. A linear warping procedure is used to time normalization of the utterances. The normalize/warp approach to dynamic warping has significant implementation advantages because of the fixed length of the reference and test patterns. Since all reference and test patterns are a fixed length of , no normalization of distances is required. On the other hand , it has implementation advantage when some constraints are set ,especially, when as memory requirements and in a real-time implementation are considered. More details on linear warping procedure can be found in paper[22]. Frame length after normalize warp procedure is selected 40. It means 600 ms speech duration. Speech features vector number below 40 frame is fixed to 40 frame with spectral (feature) interpolation and speech feature vectors more than 40 frames are decimated and fixed to 40 frame. All analysis stages are implemented in MATLAB for Windows application program.

After analysis phase, recognition phase begins. Test utterances are compared with reference utterances by using feature vectors for utterances(feature vectors were saved to files after analysis). Recognition steps such as local distance calculation, dynamic time warping and decision algorithms are implemented in ANSI C.

Spoken utterances are isolated words and number of words is 27. These words are numbers between zero and nine , some commands, some names in Turkish etc. List of these words have been given in Table 5.2.1.

aç	altı	beş	bir	böl	bülent	çarp	çıkart	dokuz
dört	dur	iki	kod	liste	oku	on	osman	sakla
sekiz	sıfır	şifre	topla	üç	yardım	yaz	yedi	yükle

These words are repeated 5 times during speech library construction. Each different utterances for the same words are named with increasing number extensions such as osman1, osman2,osman3, osman4, osman5. So, there are 135 words in the library. They are different spoken words for the same word. They are logically grouped, and group name is "osman" for these words. Speech recognition algorithms are applied to this library.

Experiments on speech recognition have been made in the following ways. Four types of local distance measures have been used. These are

1. LPC distance measure (Itakura's log likelihood distortion measure) . Equation for this distance measure[5] is

$$D(a_R, a_T) = \log \frac{\left[a_r V_T a_R'\right]}{\left[a_T V_T a_T'\right]} \qquad (5.2.1)$$

2. Cepstral distance measure in Eq. (4.4.2.2)

3. Weighted cepstral distance (cepstral distance plus liftering operation) in Eq . (4.4.3.2)

4. Projection distortion measure with liftering. Equation for these distance measure (local distance measure) is

$$D(a_R, a_T) = |C_T| - \frac{C_T C_R}{|C_R|} \qquad (5.2.2)$$

One of the word in 27 main words in listed table 5.2.1 is compared with 135 word in the library. Comparison results are ordered according to accumulative distortion measure (accumulative distortion is result of DTW). The most similar 10 word is listed in the Table 5.2.2 with its distortion measures by using above distortion measures. The other constraint in this experiment is that no symmetric comparison is done. In single comparisons ,the accumulated distortion measure between test and reference is $D = D(a_T, a_R)$. In symmetric comparison the accumulated distortion measure $D = (D(a_T, a_R) + D(a_R, a_T))/2$. Some distortion measures, for example, Itakura log likelihood distance measure is not symmetric distortion measure. Cepstral distance measures are symmetric. Another asymmetry may come from warping process. Because of these reasons this symmetric distortion measure is defined. In the first

experiment single comparison has been used to see the effect of asymmetry especially for LPC distance measure. In Table 5.2.2 comparison results have been given.

Some speech signals for the same utterances have been given in figures. Figure 5.2.2 shows speech signal for word 'osman'.

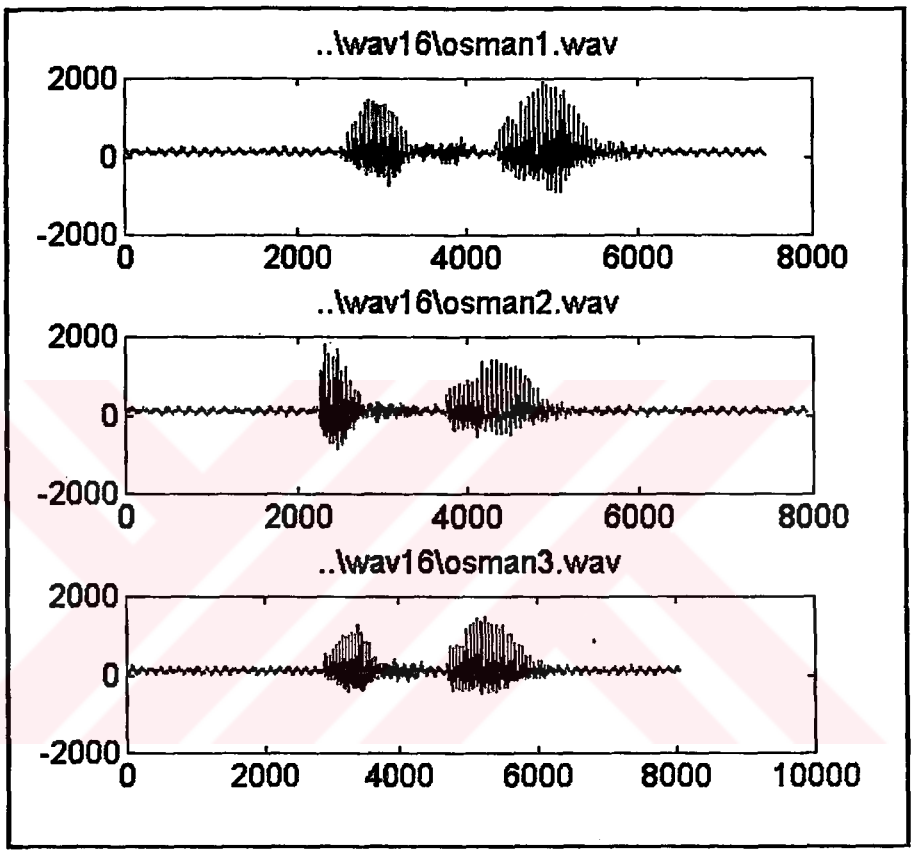


Figure 5.2.2. Speech signal representations for word 'osman'.

It seems that all three graphic nearly similar to each other. It begins a vowel 'o' which has higher energy than unvoiced sounds. An unvoiced speech sector comes. These unvoiced sector corresponds to 's' and 'm' sounds. These sector of voice has less energy than voiced speech sectors. Then a vowel 'a' comes again. It has high energy. Then unvoiced speech 'n' comes just after vowel. It is seeing that duration of unvoiced and voiced speech sectors are not same but there is no big differences. Total energy of the words are not equal as seen. The speech signal has two vowel with high energy. Spectral shape of the vowels are more characteristic than spectral shape of the

accumulated distance between osman1 osman2 by using lpc distance measure

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
0	0.1	0.2	0.4	0.4	0.4	0.4	0.4	0.4	0.3	0.3	0.2	0.1	0.1	0	0.1	0.1	0.1	0.1	0	0	0.1	0.1	0	0.1	0.1	0.2	0.3	0.4	0.4	0.5	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.2	0.1
1	1	0.5	0.4	0.7	0.9	0.9	1	1	0.7	0.5	0.3	0.2	0.3	0.4	0.4	0.4	0.4	0.3	0.3	0.3	0.4	0.5	0.5	0.4	0.4	0.4	0.4	0.3	0.5	0.7	0.8	0.8	0.8	0.7	0.7	0.7	0.8	0.7	0.2	0.3
2	2	1.3	0.8	0.6	0.9	1	1.2	1.3	1.1	0.9	0.7	0.6	0.4	0.5	0.6	0.8	0.8	0.7	0.6	0.5	0.4	0.6	1	0.9	0.9	0.9	0.9	0.9	0.8	1	1.2	1.3	1.3	1.3	1.3	1.3	1.3	1.2	0.7	0.7
3	3.2	2.2	1.4	0.9	0.7	0.8	1.1	1.4	1.5	1.3	1.1	0.9	0.4	0.4	0.8	0.9	1.1	1	0.7	0.5	0.6	0.8	1.2	1.5	1.5	1.6	1.6	1.6	1.5	1.5	1.7	1.9	2	2.1	2	2	2	1.8	1.4	1.4
4	4.5	3.2	2.2	1.4	1	0.8	0.9	1.2	1.6	1.6	1.3	1	0.8	0.5	0.8	0.6	1	1	0.8	0.7	0.9	1.1	1.6	2	2.3	2.4	2.4	2.4	2.4	2.4	2.4	2.6	2.8	3	3	2.9	2.9	2.6	2.1	2.1
5	5.7	4.2	3	2	1.4	1.1	1.1	1.1	1.2	1.6	1.4	1.1	0.9	0.8	0.8	0.6	0.7	1	0.9	1	1.2	1.6	2	2.5	3	3.3	3.3	3.4	3.4	3.4	3.4	3.5	3.7	4	4.2	4.1	4	3.6	2.9	2.9
6	7	5.3	3.9	2.7	2	1.6	1.4	1.5	1.3	1.4	1.5	1.2	1.2	1	1	0.6	0.7	0.7	1.1	1.3	1.6	2	2.5	3	3.5	4	4.3	4.4	4.4	4.5	4.6	4.7	4.8	5.1	5.4	5.6	5.4	4.7	3.8	3.7
7	8.2	6.3	4.7	3.3	2.4	1.9	1.7	1.6	1.5	1.5	1.5	1.4	1.5	1.2	1.1	0.7	0.8	0.7	0.9	1.4	1.9	2.4	2.8	3.4	4	4.5	5	5.3	5.3	5.4	5.6	5.8	5.9	6.1	6.4	6.7	6.6	5.8	4.7	4.5
8	9.3	7.2	5.4	3.9	2.8	2.2	2	2	1.8	1.6	1.6	1.6	1.7	1.5	1.3	0.7	0.8	0.8	1	1.4	2.3	2.8	3.3	3.8	4.4	4.9	5.4	5.9	6.1	6.2	6.3	6.5	6.7	6.9	7.1	7.4	7.6	6.6	5.4	5.2
9	10	8	6.1	4.4	3.2	2.5	2.2	2.3	2	1.9	1.9	1.8	1.9	1.7	1.4	0.9	0.9	0.8	1	1.5	2.2	3.2	3.7	4.1	4.7	5.3	5.7	6.2	6.6	6.8	6.9	7.1	7.3	7.5	7.8	8	8.2	7.3	6	5.8
10	11	8.6	6.7	4.9	3.7	2.9	2.7	2.7	2.5	2.3	2.2	2.1	2	1.7	1.5	1.2	1.3	1	0.9	1.3	2.2	3	3.9	4.4	4.9	5.4	6	6.5	6.9	7.3	7.6	7.7	7.8	8	8.3	8.6	8.7	8	6.6	6.3
11	12	9.3	7.5	5.7	4.5	3.8	3.6	3.6	3.3	3.1	2.9	2.7	2.3	1.8	1.9	1.9	1.9	1.5	1	0.9	1.8	2.7	3.5	4.6	5.1	5.6	6.2	6.8	7.4	7.8	8.3	8.6	8.8	8.8	9	9.3	9.6	8.9	7.3	6.8
12	12	10	8.8	7	5.8	5.1	4.9	4.9	4.6	4.3	4	3.6	3	1.8	2.1	2.6	2.7	2.2	1.5	1.2	0.8	1.9	3.2	4.4	5.5	6.1	6.7	7.4	8.2	8.8	9.3	9.7	10	10	10	10	11	10	8.4	7.6
13	13	11	10	8.4	7.1	6.4	6.3	6.2	5.9	5.6	5.3	4.8	4	2.2	1.8	2.6	3.2	2.7	1.8	1.3	1.2	1.3	2.8	4.3	5.5	6.7	7.3	8	8.7	9.4	10	11	11	11	11	11	12	12	9.7	8.9
14	14	13	12	9.8	8.6	7.9	7.7	7.7	7.4	7	6.7	6.1	5.2	2.8	2.1	2.1	3.1	3	1.9	1.3	1.5	1.6	2.4	4.1	5.6	6.8	8	8.7	9.4	10	11	11	12	12	13	13	13	13	11	10
15	16	14	13	11	10	9.4	9.2	9.2	8.9	8.5	8.1	7.6	6.8	3.7	2.7	2.2	2.1	3.1	2.2	1.8	2.2	2.4	2.9	3.8	5.4	7	8.2	9.5	10	11	12	12	13	13	14	14	14	14	13	12
16	17	16	15	13	12	11	11	11	10	9.9	9.5	9	7.9	4.7	3.4	2.5	2.3	2.4	2.7	2.5	2.8	3.2	3.6	4.2	5.1	6.6	8.4	9.6	11	12	12	13	14	14	15	15	15	15	14	13
17	18	17	16	14	13	12	12	12	12	11	11	10	9.3	5.6	4	2.8	2.5	2.6	2.8	3	3.3	3.8	4.4	4.9	5.5	6.4	8.1	9.7	11	12	13	14	14	15	15	16	16	17	15	15
18	19	18	17	16	14	14	14	13	13	13	12	12	11	6.4	4.6	3.1	2.7	2.7	2.8	3.2	3.6	4	5	5.7	6.2	6.8	7.8	9.5	11	12	14	14	15	15	16	17	17	18	17	16
19	20	20	19	17	16	15	15	15	14	14	14	13	12	7.1	5.2	3.3	2.7	2.8	2.9	3.1	3.6	4.2	5.2	6.2	7	7.5	8.2	9.1	11	12	14	15	15	16	17	17	18	19	18	17
20	21	21	20	18	17	16	16	16	15	15	15	14	13	7.4	5.3	3.4	2.9	2.9	2.8	2.9	3.2	3.9	5	6.1	7.1	7.9	8.5	9.2	10	12	13	15	16	16	17	18	18	19	19	18
21	22	21	21	19	18	17	17	16	16	16	15	14	13	7.5	5.3	3.4	3	2.9	2.9	2.9	2.9	3.4	4.4	5.5	6.6	7.7	8.5	9.1	9.8	11	13	14	15	16	17	18	18	19	20	19
22	22	22	21	19	18	17	17	17	16	16	15	15	13	7.6	5.4	3.5	3	3	2.9	2.9	3	3	3.5	4.5	5.7	6.9	8.1	9.1	9.7	10	11	13	14	16	17	17	18	19	20	19
23	22	22	22	20	19	18	18	17	17	16	16	15	13	7.8	5.6	3.7	3.3	3.2	3.1	3	3	3.1	3.1	3.7	4.7	5.9	7.2	8.5	9.6	10	11	12	14	15	16	17	18	19	19	19
24	23	22	22	20	19	18	18	18	17	16	16	15	14	8.4	6.3	4.5	4.1	4	3.8	3.7	3.5	3.4	3.4	3.2	3.7	4.8	6	7.4	8.8	9.8	10	11	12	14	15	16	17	18	19	19
25	23	23	22	21	19	18	18	18	17	17	16	15	14	9.1	7.1	5.3	4.9	4.8	4.5	4.3	4	3.8	3.8	3.4	3.3	3.8	4.9	6.3	7.7	9	10	11	11	12	14	15	16	18	18	19
26	24	23	23	21	20	19	18	18	17	17	16	15	15	9.8	7.9	6.2	5.8	5.6	5.3	5	4.4	4.2	4.2	3.5	3.3	3.3	4	5.3	6.7	8.1	9.5	10	11	11	12	14	15	17	18	18
27	24	23	23	22	20	19	19	18	18	17	18	16	15	11	8.7	7.1	6.7	6.4	6.1	5.6	4.8	4.6	4.5	3.5	3.4	3.3	3.5	4.5	6	7.5	8.8	10	11	11	12	13	14	16	17	18
28	25	24	24	22	21	20	19	19	18	17	16	16	15	11	9.4	7.9	7.5	7.2	6.8	6.2	5.1	4.8	4.8	3.6	3.5	3.4	3.4	4	5.2	6.8	8.3	9.6	11	12	12	13	13	15	16	17
29	25	24	24	23	21	20	20	19	18	17	16	16	15	12	10	8.7	8.3	8	7.4	6.6	5.4	5	5	3.7	3.8	3.6	3.4	3.9	4.7	6	7.6	9	10	11	12	13	13	14	15	16
30	25	24	24	23	21	20	20	19	18	17	17	16	15	12	11	9.4	9	8.6	7.9	7	5.5	5.1	5.1	3.8	4	3.8	3.5	3.7	4.3	5.2	6.6	8.1	9.5	11	12	13	13	13	14	15
31	25	24	24	23	21	20	20	19	18	16	17	16	15	12	11	9.9	9.6	9.1	8.4	7.3	5.6	5.2	5.3	3.9	4	4	3.6	3.7	4	4.7	5.6	7	8.5	9.9	11	12	13	13	14	14
32	26	25	24	23	21	20	20	19	18	18	17	18	15	13	12	11	10	9.8	8.8	7.7	5.8	5.4	5.4	3.9	3.9	4	3.6	3.7	3.8	4.1	4.9	5.9	7.2	8.8	10	11	12	13	14	14
33	26	25	25	23	21	20	20	19	18	18	17	17	16	13	12	11	11	10	9.2	8.1	6.2	5.7	5.7	4.2	4.2	4.2	3.8	3.7	3.7	3.8	4.2	4.9	5.9	7.3	8.9	10	11	13	13	14
34	27	25	25	23	22	21	20	19	19	18	18	17	16	13	12	12	11	11	9.8	8.7	6.8	6.3	6.3	4.8	4.7	4.7	4.2	3.9	3.7	3.8	3.9	4.3	4.9	5.9	7.4	8.9	10	12	13	14
35	27	26	25	24	22	21	20	20	19	19	19	18	17	14	13	12	12	11	11	9.4	7.6	7.1	7	6.5	5.4	5.3	4.7	4.3	3.9	3.8	3.9	4	4.3	5	6	7.4	8.9	11	12	14
36	28	26	25	24	22	21	21	20	20	20	19	19	18	15	14	13	13	12	11	10	8.4	7.9	7.8	6.3	6.2	6.1	5.4	4.8	4.2	3.9	3.8	4	4.2	4.5	5.2	6.2	7.5	9.2	11	13
37	29	27	25	25	23	22	22	21	21	21	21	20	19	16	15	14	15	13	13	11	9.6	9	8.8	7.4	7.3	7.1	6.3	5.4	4.6	4	3.9	3.9	4.1	4.4	4.8	5.4	6.2	7.7	9.7	12
38	30	29	26	25	24	23	23	22	22	22	22	22	21	18	17	16	16	15	14	13	11	10	10	8.8	8.7	8.4	7.4	6.1	4.8	4.1	4.1	4.1	4	4.1	4.4	4.9	5.7	6.2	8.3	11
39	32	30	26	26	25	24	23	23	23	24	24	24	23	20	19	18	17	17	16	15	13	12	12	10	10	9.8	8.5	6.6	4.9	4.5	4.8	4.7	4.6	4.5	4.6	5.1	5.8	6.2	6.7	9.4

Figure 5.2.3 Frame by frame accumulated distance measure for ‘osman1’ and ‘osman2’ when Itakura’s log likelihood distance (LPC distance) measure has been used.

accumulated distance between osman1 osman2 by using ceps project measure

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
0	0.6	1.7	2.9	2.8	2.7	2.7	2.6	2.6	2.6	2.5	2.4	2.2	1.9	1.2	1.1	1.1	1.2	1.4	1.5	1.6	1.8	1.7	1.4	1.5	1.7	2	2.4	2.9	3.1	3.1	3	2.9	2.6	2.8	2.9	3	2.9	2.7	2.1	0.7
1	6.3	2.7	2.2	3.2	3.1	3.1	3	3.1	3.2	3.3	3.4	3.7	4.1	5.3	5.5	5.9	6.1	6.2	6.5	6.9	7.2	7.8	6.7	4.4	4.2	3.8	3.2	3	3.7	4.3	4.4	4.1	4	3.9	3.9	4	4.1	3.6	3.3	5.2
2	12	5.3	3.3	2.9	3.5	3.5	3.4	3.4	3.5	3.6	3.8	4.3	5.4	7.6	9.4	10	11	11	11	12	13	13	12	7.4	7.1	6.2	4.7	4	4.4	5.5	6.2	6	5.7	5.7	5.5	5.6	5.6	5	4.9	6
3	18	8.2	4.8	3.7	3.4	3.9	3.6	3.7	3.7	3.7	4	4.5	5.8	8.7	12	14	15	15	16	16	17	18	17	10	9.8	8.5	6.3	5.3	5.7	6.6	7.9	8.2	7.8	7.8	7.6	7.5	7.8	6.8	6.8	9.7
4	24	11	6.8	5.1	4.2	4	4.2	4.1	3.8	3.8	3.8	4.4	5.7	8.9	13	16	18	19	20	21	22	23	22	13	12	10	7.8	6.8	7.6	9.7	9.7	11	11	11	10	10	11	9.1	9.3	12
5	30	15	8.9	6.6	5.1	4.6	4.5	4.4	4	3.8	3.9	4.1	5.4	8.6	13	17	20	23	24	25	26	28	27	15	14	12	9.3	8.5	9.3	11	12	13	14	14	13	13	13	12	12	14
6	36	18	11	8.2	6.1	5.4	5.1	4.8	4.2	3.9	3.9	4.1	5.1	8.2	12	17	21	25	27	29	30	32	33	17	16	14	11	10	11	13	14	15	16	17	16	16	16	14	14	17
7	42	21	14	10	7.3	6.2	5.6	5.3	4.4	4	3.9	4.1	5	7.8	12	16	21	26	29	32	34	36	37	19	18	16	13	12	13	15	16	18	19	19	19	19	19	17	17	19
8	47	24	18	12	8.5	7.1	6.5	5.8	4.7	4.1	4	4.1	4.6	7.3	11	16	20	25	30	34	37	39	41	21	20	18	14	14	15	17	18	20	21	22	22	22	22	20	20	21
9	52	27	19	14	9.7	8	7.4	6.6	5.2	4.5	4.2	4.2	4.7	6.7	10	14	19	24	29	33	39	42	44	23	22	20	17	17	17	18	20	21	23	24	25	25	25	22	22	23
10	56	30	22	16	11	9.5	8.7	7.9	6.2	5.3	4.8	4.7	4.7	6	9.5	12	16	21	26	31	37	42	45	25	25	23	20	19	20	21	22	23	25	26	27	28	28	25	24	25
11	60	34	25	18	14	12	11	10	8.4	7.3	6.6	6.3	5.9	5.9	7.2	9.9	13	18	23	28	34	40	45	29	29	27	24	23	23	24	25	26	27	29	30	32	32	29	28	27
12	63	38	30	22	18	16	15	14	12	11	9.7	9.1	7.9	6.9	6.6	7.9	11	14	19	24	31	37	43	33	34	33	30	28	28	28	29	30	31	33	34	36	37	35	32	31
13	66	41	35	27	22	19	19	18	16	14	13	12	10	7.7	6.9	6.9	8.2	11	15	20	26	33	40	37	38	38	35	33	33	33	33	34	36	36	38	40	42	40	37	35
14	68	45	40	31	26	24	23	23	20	18	17	16	13	8.6	7.3	7.1	7.1	8.5	11	16	21	28	36	42	42	43	41	39	38	38	38	39	39	40	42	44	46	46	43	39
15	71	49	45	36	31	28	27	27	24	22	21	19	16	9.7	7.8	7.4	7.3	7.3	8.9	12	17	23	30	39	46	47	46	44	44	44	44	44	44	45	46	48	50	52	49	43
16	73	53	51	41	36	33	32	32	28	26	24	22	18	11	8.8	8.1	7.9	7.7	7.6	9.1	13	17	24	33	43	50	50	49	49	49	49	49	49	49	50	51	54	56	55	47
17	76	57	56	46	40	37	36	36	32	30	28	26	22	13	10	9.3	9	8.4	8.1	8	9.5	13	19	27	36	46	54	54	54	54	56	56	56	54	56	56	57	60	61	52
18	79	62	62	51	45	42	41	40	36	34	32	30	25	16	12	11	11	9.6	8.9	8.6	8.5	10	14	21	30	39	50	58	59	60	60	60	60	60	60	60	62	64	66	56
19	83	67	69	57	50	47	46	45	41	38	36	33	29	18	15	13	13	11	9.9	9.4	9	9	12	17	24	33	43	54	64	66	66	66	66	65	65	65	68	69	70	61
20	87	72	72	62	54	51	50	49	45	42	40	37	32	21	17	15	15	13	11	10	9.6	9.6	10	14	19	27	36	47	69	69	70	71	70	70	69	70	71	72	73	65
21	91	78	78	66	58	54	54	52	48	45	42	39	36	23	19	17	17	14	12	11	10	10	11	12	15	21	29	39	51	64	73	74	75	74	73	74	74	75	76	69
22	94	79	80	69	61	57	57	56	51	48	45	42	37	26	22	20	19	16	14	13	11	11	11	12	13	17	23	32	43	56	66	77	78	77	77	77	78	78	79	71
23	97	82	83	73	64	60	59	56	53	50	47	44	39	28	24	22	21	18	16	14	13	13	12	12	13	14	18	26	36	47	59	72	81	81	80	80	81	81	81	73
24	100	86	86	76	67	63	62	60	55	52	49	46	42	32	28	26	25	22	19	18	16	16	14	12	12	13	15	21	30	40	51	63	75	84	83	83	84	84	84	75
25	104	89	89	79	69	65	64	62	57	54	51	48	44	36	32	30	29	26	23	21	19	19	16	13	12	12	14	17	24	33	44	54	66	78	86	86	86	86	86	78
26	108	92	92	82	72	67	66	64	59	56	53	49	46	38	36	34	33	30	26	24	22	22	19	13	12	12	13	16	21	28	37	47	58	69	80	89	89	89	89	80
27	111	95	95	85	74	70	69	68	61	58	54	51	48	41	39	38	37	33	30	28	25	25	21	13	12	12	13	15	19	24	32	41	50	61	72	83	92	92	92	83
28	115	98	98	88	77	72	71	68	63	60	56	53	51	45	43	42	41	37	33	31	28	28	24	14	13	13	13	15	19	23	29	35	44	53	63	74	86	95	95	86
29	118	101	101	91	79	74	73	69	64	61	57	54	53	48	46	46	45	41	37	35	31	31	27	14	13	13	13	15	18	22	26	31	38	46	56	66	77	89	97	88
30	121	103	102	92	80	75	74	70	66	62	59	56	56	51	50	49	49	45	40	38	35	35	30	15	13	13	13	14	17	20	24	28	33	40	48	57	68	79	91	92
31	125	104	103	93	81	76	75	71	66	63	60	57	57	54	54	53	53	49	44	42	39	39	33	17	14	13	13	13	15	18	22	26	30	34	41	49	59	69	81	96
32	129	106	104	94	82	77	76	72	68	64	61	59	60	58	58	58	57	53	49	47	43	43	37	19	16	15	13	13	14	16	19	23	26	31	36	42	50	60	71	86
33	134	108	106	95	83	78	77	73	69	66	64	62	63	63	63	63	63	58	54	52	49	49	43	22	19	17	14	13	13	14	16	20	23	27	31	35	42	51	61	76
34	139	111	106	96	86	80	79	75	72	69	67	68	67	69	69	69	69	66	61	59	56	56	49	27	23	20	16	14	14	14	14	17	20	23	27	31	36	43	53	67
35	145	114	108	97	87	82	81	78	74	72	70	70	71	73	75	75	76	71	67	66	62	63	56	32	28	24	19	16	14	14	14	15	17	20	24	27	31	36	45	59
36	152	117	107	99	89	86	84	80	77	75	74	74	75	77	79	82	82	79	75	73	70	70	64	37	32	27	21	16	16	14	14	14	15	17	20	24	27	32	38	51
37	158	121	109	101	91	87	86	83	80	78	77	78	79	82	84	87	89	86	82	81	77	77	71	42	37	31	24	17	16	15	14	14	14	16	17	20	24	28	34	44
38	165	125	110	103	94	90	89	86	83	82	81	82	84	86	89	92	94	94	90	89	85	85	78	47	41	36	27	19	16	15	15	15	15	15	15	17	21	24	30	40
39	171	128	112	105	96	93	92	89	87	86	86	86	87	90	93	96	99	101	97	98	92	92	85	52	45	39	30	21	17	16	15	15	15	15	15	16	18	21	26	39

Figure 5.2.4. Frame by frame accumulated distance measure for ‘osman1’ and ‘osman2’ when cepstral projection measure has been used.

unvoiced speech signal. It is a benefit for distortion measures, so, speech recognition process.

Comparison results of DTW for 'osman1' and 'osman2' has been given in the Figure 5.2.3 and Figure 5.2.4. These figures shows that minimum accumulated distance values are occuring in the diagonal from top-left to bottom-right. Difference between frames numbers for test and reference vectors are small (e.g. 3-5) then it is expected that accumulated distance value will be relatively smaller because it is possible that compared frames have the same spectral shape. Figure shows that this expectation is true.

To see speech signal similarities of some of the characteristic speeches have been given in figures. Words 'carp' and 'dort' are mono syllable and has a lot of unvoiced sounds. Words 'iki' and 'oku' have wovel at the beginning and end of speech. Word 'bes' has unvoiced sound at the beginning and end. Word 'osman' has two syllable and different vowels and unvoiced sounds. Words 'sakla' and 'topla' acoustically looks like very similar. Comparison result shows this similarity is true or not.

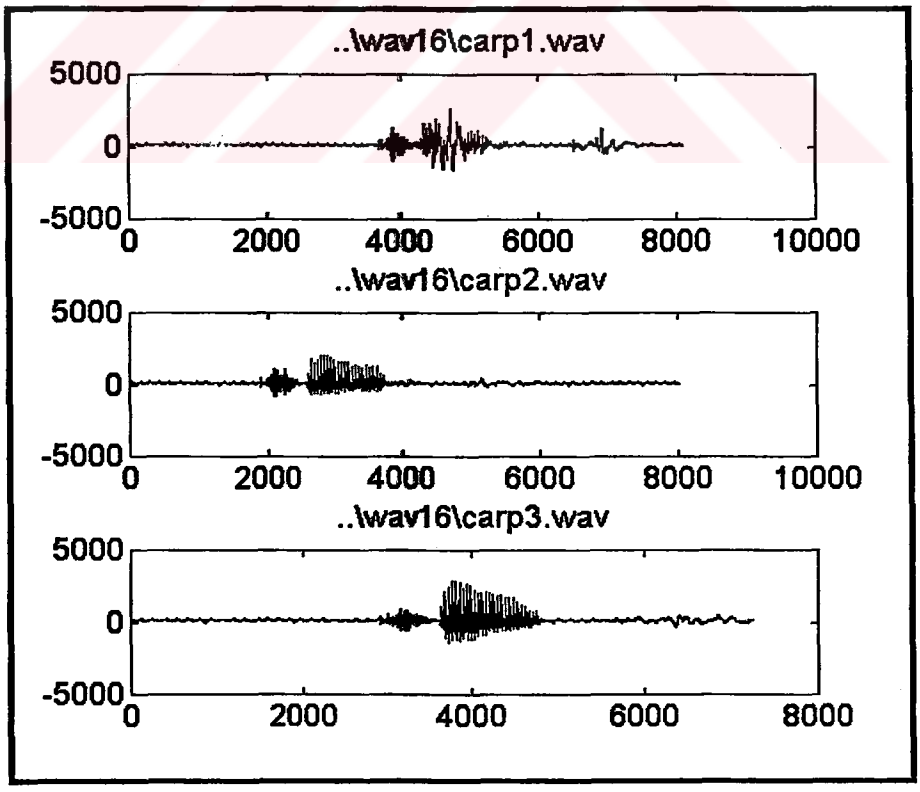


Figure 5.2.5. Speech signal representation for word 'carp'.

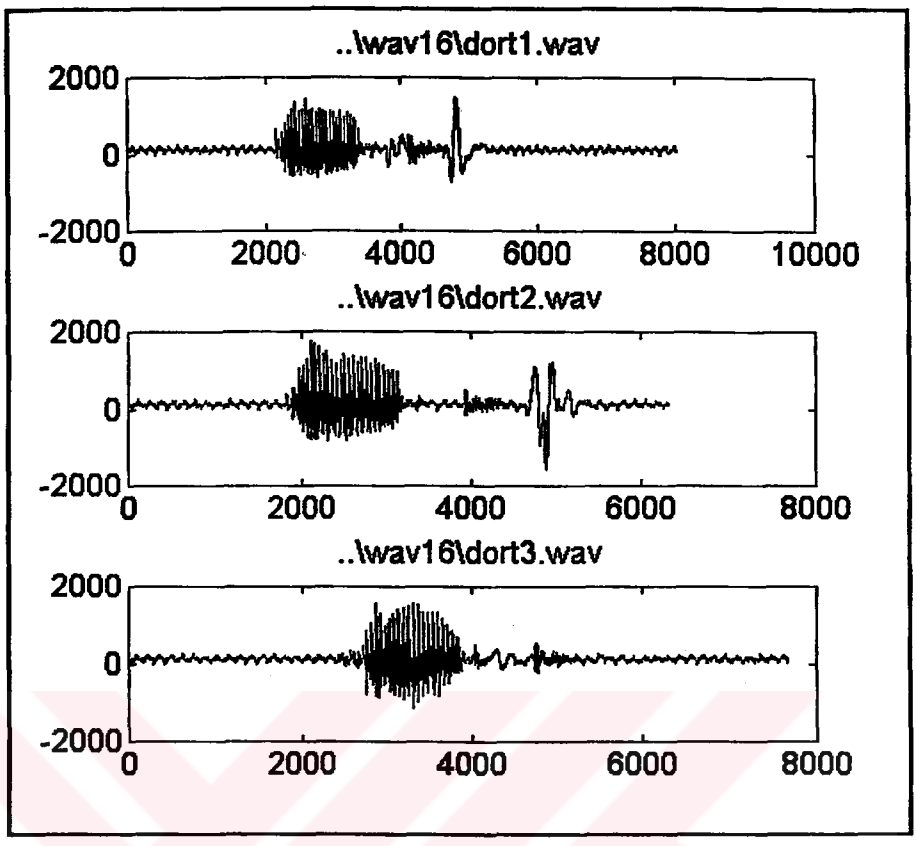


Figure 5.2.5. Speech signal representation for word 'dort'

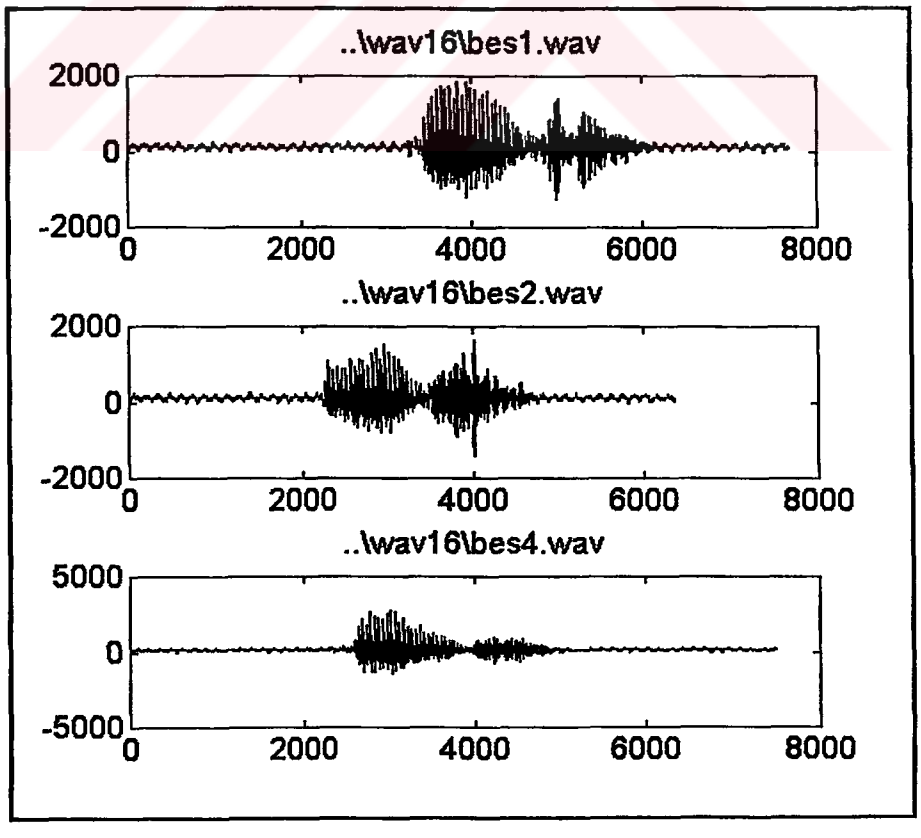


Figure 5.2.6 Speech signal representation for word 'bes'

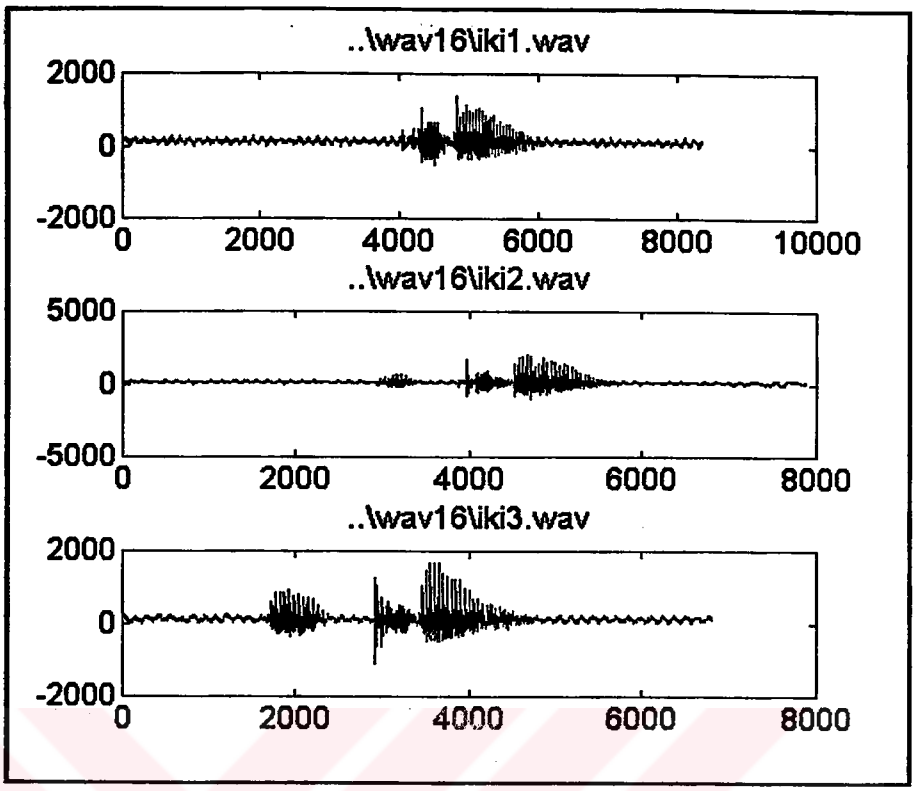


Figure 5.2.7. Speech signal representation for word 'iki.

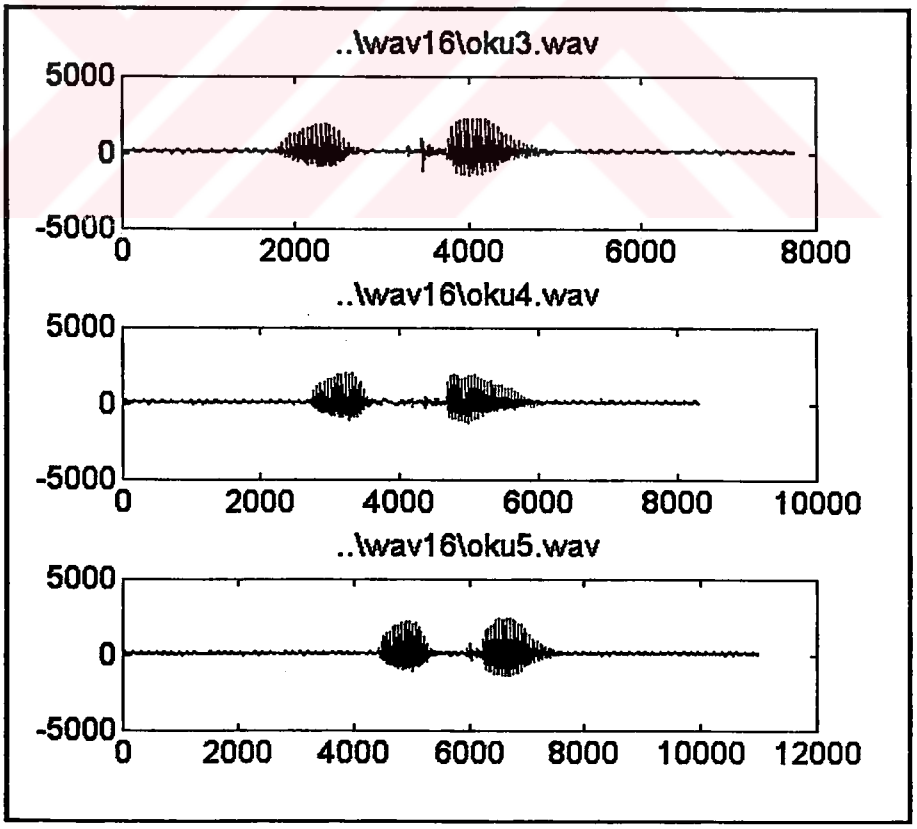


Figure 5.2.8. Speech signal representation for word 'oku'

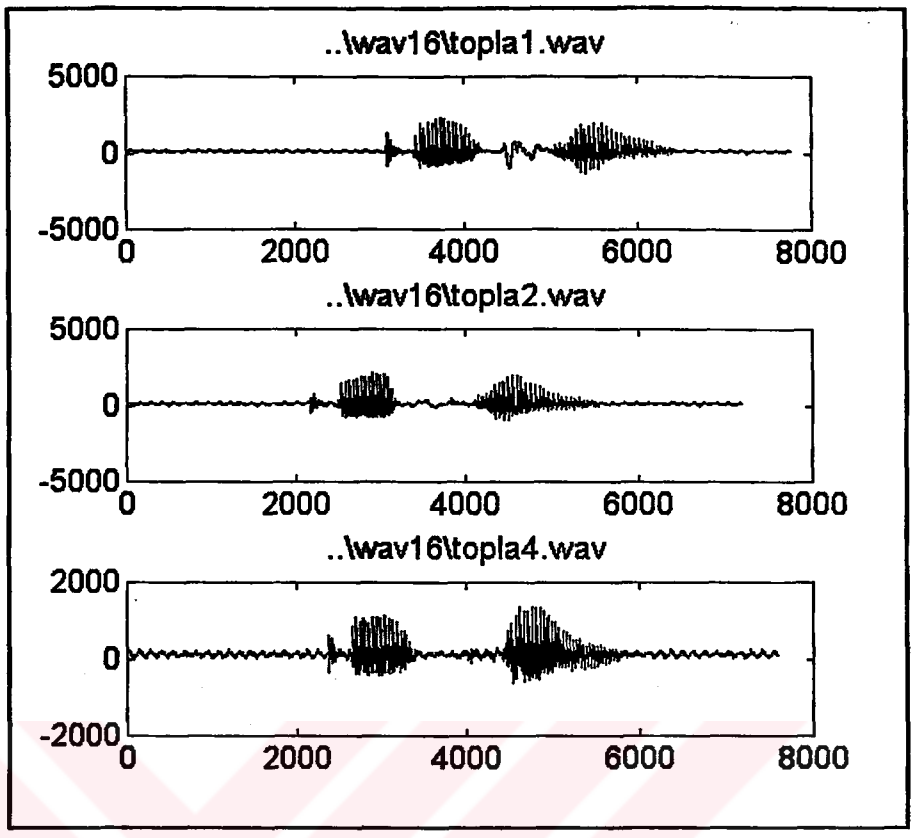


Figure 5.2.8. Speech signal representation for word 'topla'.

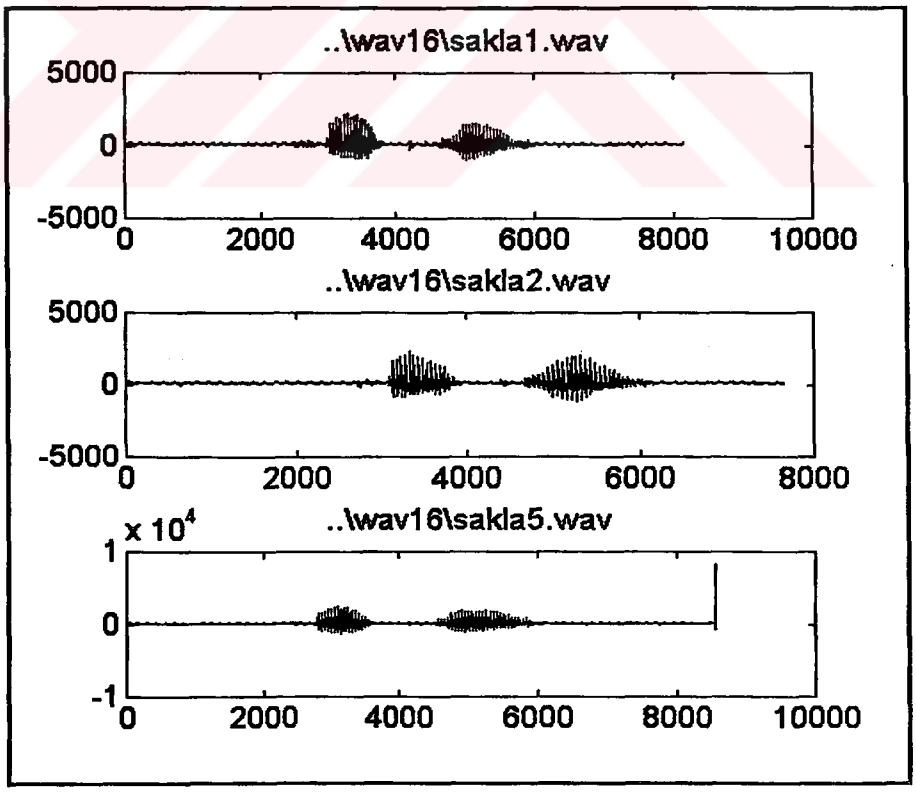


Figure 5.2.9. Speech signal representation for word 'sakla'.

To see a word is how much similar to each other, this word is compared with all the words in library (27*5 words are available in the library). Table 5.2.2 shows ordered ten word with its distance measure which are most similar to the test word. For local distance measure (previously defined) four distance measure is used. Single comparison method has been used for this experiment. Table 5.2.3 is the statistical result of Table 5.2.2. For 27 test words comparisons results have been shown with number of wrong recognized words. It has been shown that LPC distance measure is the worst distance measure because 1st, 2nd, 3rd and 4th similar words are mostly wrong recognized and it's percentage is more than %50.

RECOG. ERROR	LPC distance	LPC ceps.	LPC cep_lift.	LPC_cep_pro
1st most similar	15	1	1	1
2nd most similar	>15	2	2	1
3rd most similar	>15	6	6	6
4th most similar	>15	10	9	8

Table 5.2.3. Statistical representation of Table 5.2.2. It shows recognition error for different distance measures(single comparison).

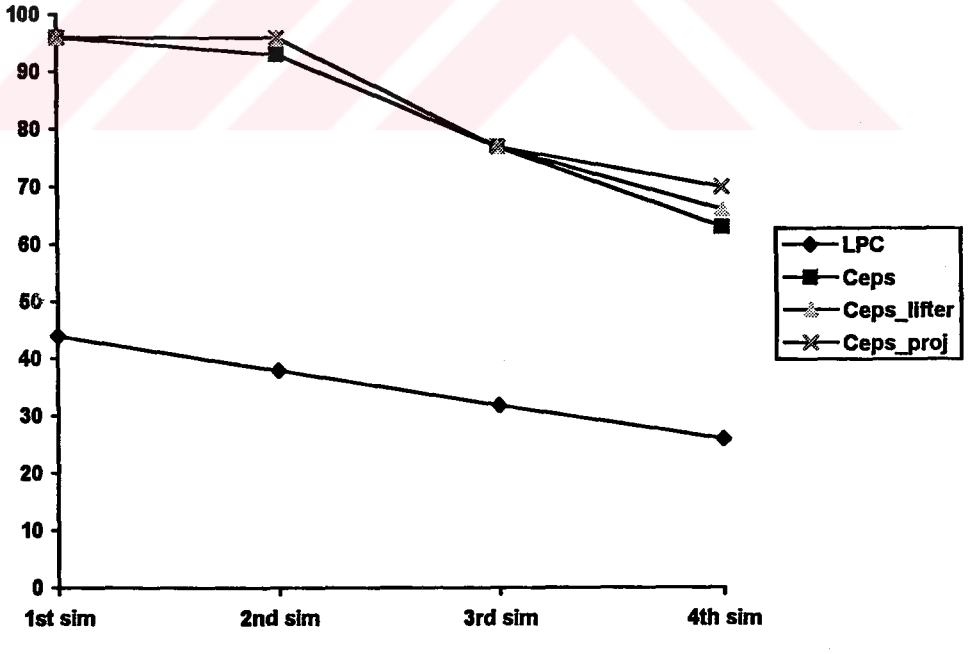


Figure 5.2.10. Recognition percentage according to distance measure type

These is due to LPC distance measure is asymmetric distance measure as explained before. When single comparison is made with LPC distance measure, then wrong recognized word percentage is very high. On the other side, the other distance

Table 5.2.2

ac2 (lpc)		ac2 (lpc_cep)		ac2 (lpc_cep_lif)		ac2 (lpc_cep_pro)	
ac2	0	ac2	0	ac2	0	ac2	0
yaz2	3.9	ac3	3.9	ac4	81.8	ac4	8.4
yaz3	4	ac4	4.7	ac3	83.2	ac3	9.8
ac3	4.3	ac5	4.9	ac1	96.3	ac5	10.6
kod5	4.6	ac1	5	ac5	112.2	ac1	10.9
cikart4	4.6	kod1	6.6	kod1	161.3	kod1	15.7
cikart1	4.9	kod2	7.6	kod5	163.1	kod5	18.4
yaz5	4.9	kod5	8	kod2	184.3	kod4	19.7
ac1	4.9	kod3	8.5	kod4	185.9	kod2	20.2
ac4	5.1	yaz3	9.6	yaz3	226.7	yaz4	24.4

alti2 (lpc)		alti2 (lpc_cep)		alti2 (lpc_cep_lif)		alti2 (lpc_cep_pro)	
alti2	0	alti2	0	alti2	0	alti2	0
topla2	9.1	alti1	9.4	alti4	224.7	alti4	23.5
alti1	9.9	alti4	10.7	alti1	224.8	alti3	23.9
alti4	10.6	topla2	11.8	alti3	288	alti1	26.9
cikart4	11.2	alti3	12.4	topla2	325.2	topla2	38.2
carp4	11.3	kod4	16.6	topla1	397.6	topla1	46.2
carp1	12.3	topla1	17.8	kod4	413.5	kod4	48.9
osman4	12.9	topla5	20.4	osman4	457.2	topla4	55.8
kod4	13	topla4	20.9	kod3	466.3	dokuz4	56.4
yardim3	13	kod2	21.9	on2	467.9	on2	56.5

bes2 (lpc)		bes2 (lpc_cep)		bes2 (lpc_cep_lif)		bes2 (lpc_cep_pro)	
bes2	0	bes2	0	bes2	0	bes2	0
bol1	3	bes1	4.5	bes1	106.1	bes4	9.6
bes1	3.9	bes3	5.7	bes4	116.8	bes1	10.7
sekiz4	4.2	bes4	6	bes5	126	bes5	11.3
yaz2	4.6	bes5	6.3	bes3	138.6	bes3	11.3
ac2	4.7	sekiz5	7.9	sekiz5	195	sekiz5	20
cikart2	4.8	sekiz1	9.2	sekiz4	225.2	bir3	22.5
bes3	4.8	sekiz4	10.2	sekiz1	227.9	sekiz1	23.7
yaz3	5	sekiz3	10.7	sekiz3	233.5	sekiz4	23.7
ac3	5.1	ac2	11.4	sekiz2	248.9	sekiz3	24.6

bir2 (lpc)		bir2 (lpc_cep)		bir2 (lpc_cep_lif)		bir2 (lpc_cep_pro)	
bir2	0	bir2	0	bir2	0	bir2	0
bulent4	5.3	bir1	12.8	bir1	335.2	bir1	32.7
cikart2	5.5	bir5	13.8	bir5	349.5	bir5	33
dur1	5.7	uc2	16.2	yedi3	355.5	yedi3	38.1
uc4	6.1	uc3	16.4	sekiz2	363	sekiz4	38.4
bulent2	6.3	sekiz3	16.7	sekiz1	365.5	yedi1	38.5
uc2	6.7	sekiz4	16.9	yedi1	365.6	sekiz1	39.3

kod5	6.7	sekiz1	17.2	sekiz4	373.9	yedi5	39.5
sekiz4	7	bulent2	17.6	sekiz3	375.2	sekiz2	40.3
yaz2	7	uc1	17.7	yedi5	390.1	sekiz3	40.4
bol2		bol2		bol2		bol2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
bol2	0	bol2	0	bol2	0	bol2	0
bol3	3.9	bol3	4.9	bol5	127	bol3	13.9
yardim2	4.6	bol5	5.8	bol3	129.1	bol5	14.2
bol5	4.9	bol4	5.9	bol4	170.4	bol4	15.7
yardim1	4.9	yardim5	9.6	sifir4	231.4	sifir4	22
carp4	5.1	yardim3	10.2	sakla1	239.5	sifir1	24.5
yardim4	5.5	yardim2	10.7	sifir1	250.9	sifir2	25.5
yardim5	6.2	yardim1	10.8	sifir2	255.4	yardim3	29.5
bol4	6.4	carp5	11.7	yardim5	275.2	yardim1	31
yardim3	6.6	cikart4	11.7	yardim1	295.4	sakla1	31.2
bulent2		bulent2		bulent2		bulent2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
bulent2	0	bulent2	0	bulent2	0	bulent2	0
kod2	5	bulent3	8.7	bulent5	270.9	bulent3	31.2
kod5	5.8	bulent5	9.5	bulent3	282.4	bulent5	34.7
cikart2	5.8	bulent4	11.8	cikart2	339.5	bulent4	44.1
kod1	6.6	cikart1	13.6	cikart1	366.6	bulent1	46.3
bulent5	6.6	bulent1	14.3	bulent1	368.8	cikart2	49.3
bulent4	6.6	cikart2	14.3	bulent4	373.9	cikart1	50.3
dort4	6.8	kod1	15.8	kod5	412.3	sifre3	52.1
kod3	6.8	kod5	16	yaz3	434.2	bol1	54.2
cikart1	6.8	cikart3	16	kod3	434.3	sifre5	59
carp2		carp2		carp2		carp2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
carp2	0	carp2	0	carp2	0	carp2	0
topla1	4.6	topla4	7.3	topla4	202.7	topla4	25.1
topla5	6.3	topla5	8.3	topla2	210.2	topla2	26.6
alti2	7.2	topla1	10.2	topla5	224	carp4	28.3
topla3	7.6	topla2	10.7	topla1	240.3	topla5	28.8
topla4	7.9	alti1	11.2	carp4	253.6	topla1	31.7
topla2	8.5	carp4	11.7	alti1	310	alti1	36.2
alti1	8.7	alti2	12.4	alti2	318.6	alti2	37.6
alti4	8.9	alti4	14.7	sakla3	357.3	alti4	41.1
osman3	11.2	osman1	14.8	alti4	358.9	alti3	43.9
cikart2		cikart2		cikart2		cikart2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
cikart2	0	cikart2	0	cikart2	0	cikart2	0
cikart5	5.8	cikart5	5.5	cikart5	131.1	cikart5	16.5
yaz3	6.5	cikart1	6.5	cikart1	147.1	cikart1	18.7
kod4	7.3	cikart3	8.1	cikart3	174.9	cikart3	21.9
kod3	7.3	cikart4	11.1	cikart4	223.8	cikart4	25.7

cikart1	7.4	carp5	17.4	carp5	375.5	carp5	45.4
kod5	7.5	carp3	19.9	carp3	411.2	carp3	48.6
yaz5	7.5	carp4	21.4	dort5	484.7	sifir2	55.6
kod2	7.7	carp1	24	carp4	496.5	alti4	57.3
bulent2	8.4	bulent2	25	bulent2	507.5	sifir3	57.6
dokuz2 (lpc)		dokuz2 (lpc_cep)		dokuz2 (lpc_cep_lif)		dokuz2 (lpc_cep_pro)	
dokuz2	0	dokuz2	0	dokuz2	0	dokuz2	0
kod5	5	dokuz3	9.2	dokuz1	233	dokuz1	25.7
cikart4	5.3	dokuz1	9.9	dokuz3	258.3	dokuz3	29.2
yaz3	5.5	dokuz4	11.5	dokuz4	339.4	dokuz5	35.4
dur1	6.2	dokuz5	13.7	dokuz5	359.1	dokuz4	36.6
bulent4	6.3	kod3	15.7	sifir2	425.9	sifir2	56.4
dokuz3	6.6	sifir4	15.9	carp5	481.2	sifir4	61.8
sifir2	6.6	cikart2	16.5	dur1	502.6	kod1	62.9
bol1	7	dur1	16.6	sifir3	509.9	sifir3	64.7
kod1	7.1	cikart1	17	cikart1	510.7	kod2	65.2
dort2 (lpc)		dort2 (lpc_cep)		dort2 (lpc_cep_lif)		dort2 (lpc_cep_pro)	
dort2	0	dort2	0	dort2	0	dort2	0
carp5	6.4	dort1	6.1	dort1	139.5	dort1	18.2
dort1	6.5	cikart4	13.1	cikart4	242.5	dort4	32.2
cikart4	6.5	dort3	13.2	kod3	249.9	cikart4	32.5
kod4	6.6	kod3	13.3	dort3	260.6	dort5	32.9
yaz5	7.1	dort4	13.5	dort4	264.3	yaz4	33.4
carp1	7.2	dort5	13.6	kod2	269.6	kod2	33.7
carp4	7.3	carp3	13.6	cikart1	276.9	cikart1	34.3
bol1	7.5	cikart2	13.9	carp3	283.6	kod3	34.6
carp3	7.6	kod2	14.2	dort5	284.7	dort3	36.1
dur2 (lpc)		dur2 (lpc_cep)		dur2 (lpc_cep_lif)		dur2 (lpc_cep_pro)	
dur2	0	dur2	0	dur2	0	dur2	0
sifir3	5.3	dur5	7	dur5	208.1	dur5	21.6
dur1	7.9	dur4	8.8	dur4	244.4	dur4	25.6
sifir2	8.2	dur1	9.3	dur1	248.8	dur3	32.1
sifir5	8.3	dur3	12.8	dur3	300.4	dur1	32.2
dur5	8.9	sifir1	13	sifir1	328.5	sifir1	33.7
cikart4	8.9	sifir2	14.5	sifir2	383.8	sifir2	40.2
cikart3	9.1	sifir5	15.2	sifir5	420.6	sifir5	43.3
kod5	9.3	sifir4	19	carp5	434.2	dokuz5	45.1
yaz3	9.7	cikart2	19	sifir4	438.2	dokuz1	46
iki2 (lpc)		iki2 (lpc_cep)		iki2 (lpc_cep_lif)		iki2 (lpc_cep_pro)	
iki2	0	iki2	0	iki2	0	iki2	0
bol1	10.6	iki3	8.2	iki3	222.4	iki3	20.4
bulent3	10.8	iki5	13.7	iki5	290.4	iki5	30.8
sekiz4	10.9	sekiz2	14.6	sekiz2	342.3	iki1	31.8

kod5	11	yedi1	15.9	sekiz4	361.4	iki4	33.1
bulent1	11.2	yedi3	16.4	iki1	373.6	sekiz2	39.5
bulent4	11.2	sekiz4	16.4	sekiz3	379.5	sekiz4	40.9
iki3	11.6	sekiz3	17.9	yedi1	395.4	yedi1	41.9
yukle2	11.8	yedi4	17.9	yedi3	407.2	sekiz3	42.1
kod2	11.8	sekiz5	18.1	sekiz5	411.7	yedi3	44.8
kod2		kod2		kod2		kod2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
kod2	0	kod2	0	kod2	0	kod2	0
cikart4	3.9	kod4	4.6	kod4	109	kod4	12.7
kod4	4.7	kod1	5.1	kod1	130.9	kod1	13.6
kod1	5	kod3	5.3	kod3	134.2	kod3	17.1
kod3	5.4	kod5	6.5	kod5	169.7	kod5	20.9
kod5	7.7	dokuz4	12.8	oku1	311.8	dokuz4	28.2
yaz5	8.9	dokuz5	15.7	dokuz4	313.1	dokuz5	34.3
bulent3	9.8	dokuz3	16.9	oku3	402.2	dokuz3	42.3
bulent4	10.1	cikart4	17	on3	416.1	dokuz1	48
yaz3	10.1	dur1	19.1	on1	431.8	on1	50.3
liste2		liste2		liste2		liste2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
liste2	0	liste2	0	liste2	0	liste2	0
liste3	5.5	liste1	4.9	liste1	99.9	liste1	10.3
liste1	6.9	liste3	5.1	liste3	135.7	liste3	12.4
liste4	7.2	liste4	7.2	liste4	170.8	liste4	16.8
yukle3	7.8	liste5	9.6	liste5	187.7	liste5	22.5
bulent3	9.1	yedi4	13.7	yedi4	342.1	yedi4	35.4
liste5	9.2	yukle1	15.1	yedi1	386.3	yedi1	38.8
yukle4	9.8	yedi1	16	sifre1	403.9	yukle1	42.6
dort1	9.9	yukle5	16.4	yukle1	405	sifre3	43
dort2	10	sifre3	17.8	yukle5	427.7	sifre4	44.1
oku2		oku2		oku2		oku2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
oku2	0	oku2	0	oku2	0	oku2	0
osman4	9.5	oku4	13.3	oku3	323.9	oku3	37.4
osman1	9.8	oku3	13.9	oku4	333.2	oku4	38.9
yardim2	10.8	osman1	14.3	osman1	340.6	alti5	42.1
osman3	11	osman3	15.9	osman3	383.4	oku1	42.4
yardim1	11.2	osman5	16.5	osman5	389.2	osman5	45
osman2	11.2	oku1	17	oku1	410.4	osman1	45.5
dort2	11.4	oku5	17.6	oku5	434.4	osman3	46.9
osman5	12.5	yaz1	17.6	yaz1	434.4	oku5	47.5
yardim4	12.7	alti5	21.8	osman4	501.4	yaz1	47.5
on2		on2		on2		on2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
on2	0	on2	0	on2	0	on2	0
yardim2	4.5	on1	5.3	on1	125.8	on1	12.5

on5	5.6	on5	5.8	on5	139	on5	13.9
bir4	6.2	on4	7.1	on3	180.9	on4	17.5
on1	7	on3	7.4	on4	186.6	on3	18
yardim5	7.1	osman1	14.1	osman4	287.9	osman4	31.3
yardim1	8	oku1	15.1	osman1	345.5	osman3	43.5
osman4	8	osman4	15.8	osman3	350.3	osman1	43.7
yardim3	8.6	osman3	16.3	osman5	442.8	osman5	44.5
on4	9.1	oku2	17.9	oku1	454.8	alti5	45.2
osman2 (lpc)		osman2 (lpc_cep)		osman2 (lpc_cep_lif)		osman2 (lpc_cep_pro)	
osman2	0	osman2	0	osman2	0	osman2	0
osman1	8.3	osman3	8.4	osman1	241.9	osman3	28.2
osman3	8.3	osman1	8.5	osman3	243.6	osman5	29.4
dort1	11.2	osman5	10.4	osman5	277.9	osman1	29.7
topla1	11.5	osman4	13.4	osman4	344.4	osman4	42.2
alti2	11.7	topla2	20.3	topla5	477.1	topla5	62.8
osman4	12.8	topla5	20.8	topla2	553.7	topla3	63.7
alti1	13.4	alti2	23.7	topla3	557.6	topla4	69.1
topla5	14	topla4	23.9	topla1	562.2	topla2	69.8
carp3	14	topla1	24.3	kod2	573.8	topla1	70.7
sakla2 (lpc)		sakla2 (lpc_cep)		sakla2 (lpc_cep_lif)		sakla2 (lpc_cep_pro)	
sakla2	0	sakla2	0	sakla2	0	sakla2	0
topla5	5.4	sakla3	6.6	sakla5	166.8	sakla5	24
sakla3	6.9	sakla5	7	sakla3	187.3	sakla4	24.1
sakla5	7.2	sakla4	7.4	sakla4	191.3	sakla1	27.1
topla3	7.9	sakla1	7.8	sakla1	205.7	sakla3	27.3
sakla4	8	topla5	8.9	topla5	261.7	topla5	32.4
sakla1	8	topla1	10.4	topla1	316.6	carp2	37.2
topla1	8.1	topla2	11.8	topla2	330.9	topla1	39.5
topla4	8.2	alti2	13	yardim4	339.1	carp4	39.6
yardim4	8.5	yardim3	13.9	yardim3	342.3	topla2	42.9
sekiz2 (lpc)		sekiz2 (lpc_cep)		sekiz2 (lpc_cep_lif)		sekiz2 (lpc_cep_pro)	
sekiz2	0	sekiz2	0	sekiz2	0	sekiz2	0
bulent4	4.7	sekiz3	4.5	sekiz3	116.9	sekiz3	13.1
sekiz1	4.9	sekiz4	5.9	sekiz4	141.6	sekiz4	15.3
yaz2	5.2	sekiz5	6.7	sekiz5	169.7	sekiz5	18.3
sekiz3	5.3	sekiz1	12	sekiz1	237.4	sekiz1	23.8
yaz3	5.5	yaz3	15.3	yaz3	437.9	yedi5	46.4
sekiz5	5.5	bol1	16	bol1	438.2	sifre1	46.9
kod2	5.8	yaz2	16.5	kod5	440.1	bir3	47.7
yaz5	5.9	kod5	16.6	sifre1	445.3	bir2	47.7
dort3	6.1	yaz5	17	yedi3	448.5	yedi1	48.8
sifir2 (lpc)		sifir2 (lpc_cep)		sifir2 (lpc_cep_lif)		sifir2 (lpc_cep_pro)	

sifir2	0	sifir2	0	sifir2	0	sifir2	0
kod5	6	sifir1	5.8	sifir1	147.7	sifir1	16.2
kod2	6.4	sifir4	7.6	sifir4	173.9	sifir4	19
dur1	6.9	sifir3	7.7	sifir3	181.8	sifir3	20.9
sifir4	7.2	sifir5	10.1	sifir5	260.6	sifir5	28.9
sifir1	7.4	cikart1	19.9	carp5	492	carp5	61.3
dort5	8.3	cikart3	20.7	cikart1	515.4	dokuz1	64.8
cikart4	8.4	carp5	20.7	yardim3	522	bol2	65
bulent4	8.6	cikart4	21.1	yardim2	534.1	cikart1	66.6
dokuz5	8.6	cikart2	21.4	dort5	538.1	dokuz2	66.6
sifre2		sifre2		sifre2		sifre2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
sifre2	0	sifre2	0	sifre2	0	sifre2	0
yukle3	6.5	sifre1	5.5	sifre1	106	sifre1	10.3
sifre5	6.5	sifre5	5.8	sifre5	127.2	sifre5	13.5
sifre3	8	sifre3	8.2	sifre3	155.5	sifre3	17
yukle4	9.3	sifre4	10.9	sifre4	199	sifre4	20.7
yukle5	10.2	yukle4	13.2	yukle4	309.2	yukle4	35.4
sifre1	10.8	yukle3	14	yukle5	317.6	yedi5	37.5
kod4	10.9	yukle1	14	yukle1	345.7	yukle1	38.4
sifre4	10.9	yukle5	14.3	sekiz2	351.3	yukle5	38.5
dort2	12.1	iki3	15.8	yukle3	354	bir2	40.1
topla2		topla2		topla2		topla2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
topla2	0	topla2	0	topla2	0	topla2	0
topla1	10.7	topla5	16.6	topla1	359.6	topla1	39.9
topla5	11.3	topla4	16.8	topla5	373.6	alti4	45.1
topla4	11.6	topla1	17.4	sakla1	382.5	sakla1	47
alti2	11.9	sakla1	17.5	topla4	389.9	topla5	48.7
alti1	12.2	sakla2	19	alti4	394.1	topla4	49.4
dort1	12.3	sakla3	20.2	sakla3	440.4	topla3	50.4
alti4	12.3	alti2	20.4	alti2	449.7	carp2	53.8
dort2	13.3	alti1	20.8	sakla2	452.3	sakla4	57.5
alti3	13.5	alti4	22.7	sakla4	459.4	sakla3	57.9
uc2		uc2		uc2		uc2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
uc2	0	uc2	0	uc2	0	uc2	0
uc4	4.2	uc1	5.7	uc1	122.9	uc1	12.3
uc5	4.4	uc4	5.7	uc3	128.9	uc3	15.1
kod5	4.6	uc3	6.1	uc4	144	uc4	17.4
uc3	4.8	uc5	7.5	uc5	196.5	uc5	22.2
uc1	5.2	sekiz3	18.1	bir2	392	bir2	34.8
bulent4	5.4	bir2	18.2	bir1	423.4	bir1	35.4
yaz3	5.8	yedi5	18.2	yukle5	444	bir5	40
cikart2	6.3	sekiz1	19	sifre1	467.1	bir3	51.1
bol1	6.6	sifir1	19.1	yukle4	471.4	sifre1	51.2

yardim2		yardim2		yardim2		yardim2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
yardim2	0	yardim2	0	yardim2	0	yardim2	0
yardim1	4.1	yardim3	5.5	yardim3	136.4	yardim3	15.5
yardim3	5.1	yardim1	6.5	yardim1	147.7	yardim1	16.2
alti2	5.7	alti2	10.4	alti2	264.2	alti2	33.1
alti4	7	yardim4	11	yardim4	321.3	yardim4	33.3
carp4	7.5	topla2	13.6	topla2	351.7	bol3	47.8
kod4	7.7	yardim5	14	alti1	403	topla2	48.2
bol2	8	carp4	15.3	sakla1	404.7	alti4	50.2
dort2	8.1	sakla1	15.8	carp4	405.8	bol4	51
alti1	8.2	carp2	16	cikart1	414.7	bol2	51.2
yaz2		yaz2		yaz2		yaz2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
yaz2	0	yaz2	0	yaz2	0	yaz2	0
yaz3	3.5	yaz3	5.8	yaz3	118.6	yaz3	13.9
kod4	4.7	yaz4	7.5	yaz5	176.3	yaz5	20.9
cikart4	4.7	yaz5	8.1	yaz4	197.8	yaz4	21.3
kod5	5.6	cikart1	10.4	cikart1	289.3	cikart1	38.7
carp5	6.3	bulent2	12.8	cikart2	329.3	cikart2	38.8
kod1	6.3	cikart4	13.4	bulent2	362.6	bulent2	42
yaz5	6.4	cikart2	14.1	cikart4	367.2	cikart4	44.3
kod2	6.5	bulent5	14.7	cikart5	397.7	cikart5	48
kod3	7	cikart5	15.2	cikart3	404.8	bulent5	48.4
yedi2		yedi2		yedi2		yedi2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
yedi2	0	yedi2	0	yedi2	0	yedi2	0
yedi4	5.1	yedi5	4.3	yedi5	94.4	yedi5	8.8
yedi1	5.3	yedi4	4.4	yedi4	121.8	yedi4	12.5
yedi3	6.7	yedi1	4.9	yedi1	125.1	yedi1	12.6
yedi5	7.3	yedi3	6.4	yedi3	137.4	yedi3	12.9
iki5	7.5	iki4	8	iki4	173.3	iki4	15.3
kod5	8.7	iki2	8.3	iki2	206.9	iki2	19.9
iki3	8.8	iki5	9.4	iki5	222.9	iki5	21.3
bulent4	9.3	iki3	9.6	iki3	229.8	iki3	22.4
sekiz4	9.3	iki1	11.8	iki1	249.6	iki1	23.6
yukle2		yukle2		yukle2		yukle2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
yukle2	0	yukle2	0	yukle2	0	yukle2	0
yukle5	6.8	yukle5	6.3	yukle5	139.1	yukle5	18.6
yukle3	8.7	yukle3	10.3	yukle3	233.1	yukle3	20.4
yukle4	9	yedi3	13	yukle4	284.4	yukle4	26.5
kod4	10.1	sifre2	13.4	sifre5	324.8	yukle1	31.1
bol1	10.2	yedi2	13.5	sifre2	333.5	sifre2	33
sifre2	10.4	yedi1	14	yukle1	336.3	sifre3	35
carp1	10.7	yukle4	14.6	sifre3	351.9	sifre1	35.3
kod5	10.9	sifre5	14.7	sifre1	354.8	sifre5	36
yukle1	11.1	yedi4	14.9	yedi3	364.1	yedi5	38.5

measures are acceptable recognition ratio. If 1st most similar word is assumed the recognized word, then recognition ratio is 96%. It seems that 2nd most similar word is the tested word with percentage 92 for LPC ceps. and LPC ceps_lifter.

Figure 5.2.11 shows that LPC distance measure with single comparison is not very efficient. The other three distance measure has very high recognition ratio for 1st most similar and 2nd most similar word. Word 'dort' could not be recognized for all distance measures. If this word is excluded, then the recognition ratio for 26 word is 100%.

The same test has been made for the symmetric comparison. Table 5.2.4 shows ordered ten words with its distance measure which are most similar to the test word. For local distance measure previously defined four distance measure is used. Symmetric comparison method has been used for this experiment. Table 5.2.5 is the statistical result of Table 5.2.4.

RECOG. ERROR	LPC distance	LPC ceps.	LPC cep_lift.	LPC_cep_pro
1st most similar	1	1	1	1
2nd most similar	8	2	1	1
3rd most similar	13	6	3	2
4th most similar	18	10	7	5

Table 5.2.5. Statistical representation of Table 5.2.2. It shows recognition error for different distance measures(single comparison).

Figure 5.2.11 shows that although LPC distance measure with symmetric comparison is better than single comparison, this distance measure is not very efficient especially when other 2nd ,3rd and 4th most similar word recognition ratio is considered. The other three distance measure has very high recognition ratio for 1st most similar and 2nd most similar word. Word 'dort' could not be recognized for all distance measures. If this word is excluded, then the recognition ratio for 26 word is 100%. The highest recognition rate has been obtained with LPC cepstrum projection distance measure as shown in Figure 5.2.11. The positive effect of lifter is seen in

Table 5.2.4							
ac2		ac2		ac2		ac2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
ac2	0	ac2	0	ac2	0	ac2	0
ac3	4.1	ac3	3.9	ac4	81.8	ac4	8.6
ac4	5	ac4	4.7	ac3	83.2	ac3	10.3
ac5	5.3	ac5	4.9	ac1	99.9	ac5	11.6
ac1	5.5	ac1	5.1	ac5	112.2	ac1	11.7
yaz2	7.2	yaz4	17.6	dort2	373.8	dort2	43.9
yaz4	8.5	dort3	18.4	cikart1	382.5	cikart1	47
bes4	9.7	kod3	18.9	dort5	406.6	dort5	48.5
dokuz4	9.8	dort5	19.3	dort1	425.4	yaz4	49.7
bes2	10	dort4	19.9	dort3	426.3	dort1	50.3
alti2		alti2		alti2		alti2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
alti2	0	alti2	0	alti2	0	alti2	0
alti1	7.9	alti1	8.2	alti1	205.9	alti3	23.1
yardim2	10.1	alti3	11.8	alti3	254.7	alti1	24.7
yardim3	10.3	alti4	13.2	alti4	263.5	alti4	29.4
topla2	10.5	topla2	16.1	topla2	387.4	topla2	50.2
carp4	10.9	yardim2	16.9	topla1	451.1	carp2	51.8
osman4	11	carp2	17.8	carp2	453.6	topla4	52.8
cikart4	11.5	sakla1	18.8	yardim2	455.2	topla1	53.8
sakla2	11.9	yardim3	19.4	topla5	460.2	topla5	55.6
alti4	12	topla1	19.5	carp4	469.2	yardim2	57.5
bes2		bes2		bes2		bes2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
bes2	0	bes2	0	bes2	0	bes2	0
bes1	4.3	bes1	4.5	bes1	109.6	bes1	11.1
bes3	5.1	bes3	6.2	bes3	153.6	bes3	13.7
bes4	6.6	bes5	8.4	bes4	170.8	bes4	15.6
bes5	7.4	bes4	8.9	bes5	173.6	bes5	16.6
ac1	8.6	bir3	14.8	bir3	339.7	bir3	32
ac2	10	sekiz1	16.3	sekiz1	413.3	sekiz1	41.6
ac3	10.7	sekiz2	17.6	sekiz2	421	sekiz2	43.2
sekiz1	11.2	sekiz3	18.5	sekiz3	440.6	sekiz3	46.6
sekiz2	11.5	sekiz5	18.5	sekiz5	476.7	sekiz5	48
bir2		bir2		bir2		bir2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
bir2	0	bir2	0	bir2	0	bir2	0
bir1	6.9	bir1	10.6	bir1	268.9	bir1	27
uc2	7	bir5	13.8	bir5	341.7	bir5	34.9
uc4	7.1	bir3	16.5	yedi3	382.6	yedi5	36.4
bir5	7.3	uc2	17.2	bir3	388.3	yedi3	37.8
uc5	7.6	uc3	17.3	yedi5	390.8	yedi1	39.2
sekiz4	7.7	uc5	18.2	yedi1	393.3	bir3	39.3

uc3	8	uc1	18.2	yedi2	414.3	uc2	39.7
bir3	8.4	yedi5	18.6	uc2	414.7	uc3	41.3
sekiz5	8.6	uc4	18.8	yedi4	429	yedi2	42.4
bol2		bol2		bol2		bol2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
bol2	0	bol2	0	bol2	0	bol2	0
bol3	4.8	bol3	5	bol5	127	bol5	14
bol5	5	bol4	5.6	bol3	130.4	bol3	14.3
bol4	5.4	bol5	5.7	bol4	152	bol4	14.9
yardim2	6.3	bol1	12.8	bol1	341.7	bol1	36.9
yardim1	8.2	yardim2	14	yardim1	392.9	yardim1	40.4
yardim4	9.4	yardim1	14.2	yardim2	406.6	sifir4	43
yardim3	10.2	yardim5	14.6	yardim5	417.9	yardim3	43.1
yardim5	11	yardim3	15.6	sifir4	419	yardim2	44.4
bol1	12	yardim4	17.2	yardim3	421.2	sifir2	45.2
bulent2		bulent2		bulent2		bulent2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
bulent2	0	bulent2	0	bulent2	0	bulent2	0
bulent5	5.5	bulent5	7.1	bulent5	194.4	bulent5	24.8
bulent4	5.8	bulent3	7.8	bulent3	258.2	bulent3	29.1
bulent1	5.9	bulent4	8.6	bulent1	266.8	bulent4	32.1
bulent3	6.9	bulent1	10	bulent4	272.7	bulent1	33.8
cikart2	7.1	yaz2	17	yaz2	406.8	yaz2	51.3
cikart1	7.1	yaz3	17.2	yaz3	410.6	yaz3	53.2
dort5	7.7	dort5	17.3	cikart2	423.5	cikart2	55.4
bol1	8	dort4	17.9	cikart1	450.5	dort5	56.1
kod5	8.1	dort3	18.5	dort5	463.6	cikart1	57.5
carp2		carp2		carp2		carp2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
carp2	0	carp2	0	carp2	0	carp2	0
topla1	8.8	carp4	11.3	carp4	240.2	carp4	26.5
topla5	10.5	topla1	14	topla5	322.1	topla4	37.4
carp4	11	topla5	14.4	topla4	339.9	topla5	37.5
topla3	11	topla4	16.1	sakla4	341.8	topla2	40.2
topla4	12	sakla1	16.7	topla1	343.9	topla1	41.9
sakla1	12.1	topla2	17.1	sakla3	350.5	sakla3	43.8
sakla4	12.3	alti2	17.8	topla2	361	sakla4	45.2
yardim2	12.6	yardim2	18.5	sakla1	365	sakla1	48.3
alti2	12.8	sakla4	19.7	sakla5	424.2	bol5	49.9
cikart2		cikart2		cikart2		cikart2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
cikart2	0	cikart2	0	cikart2	0	cikart2	0
cikart5	6.1	cikart5	5.6	cikart5	136.1	cikart5	16.8
yaz3	6.5	cikart1	7.4	cikart1	165.9	cikart1	19.4
bulent1	7	cikart3	9.1	cikart3	192.8	cikart3	23.4
bulent2	7.1	cikart4	14.6	cikart4	303.5	cikart4	35.9

bulent4	7.1	carp5	17.6	dort5	362.3	carp5	44.9
kod3	7.4	dort3	19	carp5	375.5	carp3	48.5
dort5	7.5	bulent4	19.3	yaz3	392.9	yaz2	48.7
dort4	7.6	bulent3	19.6	dort3	395.7	ac4	52.1
bulent3	7.7	dort4	19.7	carp3	415.8	dort5	52.2
dokuz2		dokuz2		dokuz2		dokuz2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
dokuz2	0	dokuz2	0	dokuz2	0	dokuz2	0
dokuz3	7.7	dokuz1	9.6	dokuz1	234.6	dokuz1	28.8
dokuz1	8	dokuz3	10.7	dokuz3	290.8	dokuz3	37.1
dokuz4	8.5	dokuz4	11.3	dokuz4	330.6	dokuz5	38.7
dokuz5	9.2	dokuz5	13.7	dokuz5	359.1	dokuz4	41.2
kod1	9.7	kod3	16.9	kod3	501.6	kod3	57.3
kod3	10.1	dort3	19.3	sifir2	513.6	dur2	60
bol1	10.2	dort5	19.8	dort5	513.9	bol2	60.9
kod5	10.2	dur2	20.4	dort3	514.5	kod1	61
yaz3	10.3	kod2	20.5	kod4	526.3	sifir2	61.5
dort2		dort2		dort2		dort2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
dort2	0	dort2	0	dort2	0	dort2	0
dort1	6.8	dort1	7.3	dort1	156.8	dort1	20.7
cikart4	8.7	dort5	16.5	dort3	298.1	dort3	42.7
bol1	9.3	dort3	16.7	dort5	329.5	ac2	43.9
carp3	9.3	carp3	17.1	dort4	336.7	ac5	44
carp5	9.9	cikart4	17.5	ac1	361	ac1	44.6
carp4	10.8	dort4	18.1	ac2	373.8	dort5	45.7
carp1	10.9	carp1	18.7	ac5	378.7	dort4	48.1
cikart3	13.1	ac1	19	carp3	395.5	ac3	49.7
dur3	13.1	ac2	20	ac3	406.5	ac4	49.7
dur2		dur2		dur2		dur2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
dur2	0	dur2	0	dur2	0	dur2	0
dur1	7.8	dur4	8.5	dur4	231.1	dur4	23.2
dur4	9.2	dur1	9.3	dur1	248.8	dur5	26.3
sifir3	9.4	dur5	9.5	dur5	255.9	dur1	29.6
sifir2	9.8	dur3	13.7	dur3	303.1	dur3	30.5
dur5	10.2	dort4	18.1	dort4	439.1	dort4	52
sifir5	11.8	sifir2	19.2	dokuz1	483.8	dokuz1	53.1
dur3	12.8	dokuz1	19.2	dort5	499.3	carp5	56.6
dokuz2	12.8	dort5	20	carp5	499.7	sifir1	56.6
dokuz5	12.9	dokuz4	20.4	sifir1	521.8	sifir2	57.6
iki2		iki2		iki2		iki2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
iki2	0	iki2	0	iki2	0	iki2	0
iki4	9.7	yedi1	12.2	iki1	242.7	iki1	21.2
iki5	9.9	iki5	12.4	iki5	256.2	iki4	27

yedi1	12.1	yedi3	12.7	yedi1	282	iki5	29.2
iki3	12.9	iki3	12.8	iki4	306.6	yedi1	29.4
yukle2	13.1	yedi4	13.3	yedi3	311.3	iki3	30.9
yedi3	14.2	iki4	13.3	iki3	323	yedi4	33
iki1	14.2	iki1	13.6	yedi4	326.3	yedi3	33.3
yedi4	14.5	yedi2	14.4	yedi2	366	yedi2	35
sekiz4	15.4	yedi5	17.1	sekiz2	416.2	sekiz2	46.2
kod2		kod2		kod2		kod2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
kod2	0	kod2	0	kod2	0	kod2	0
kod4	4.7	kod4	4.6	kod4	109	kod4	12.5
kod1	5.4	kod1	5.3	kod3	134.2	kod1	14.3
kod3	5.7	kod3	5.3	kod1	134.6	kod3	15.8
yaz5	7.3	kod5	6.4	kod5	159.8	kod5	18.8
kod5	7.5	dokuz4	15.2	dokuz4	403.3	dokuz4	40.5
bulent3	7.8	dokuz3	17	dort4	412.2	dokuz3	47.6
bulent2	8.2	dort5	18.2	dort5	425.3	ac5	50
bulent4	8.2	dort4	18.6	ac5	455.4	dort4	52.7
cikart4	8.2	dokuz2	20.5	dokuz3	466	dort5	53.1
liste2		liste2		liste2		liste2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
liste2	0	liste2	0	liste2	0	liste2	0
liste3	7.1	liste3	6.6	liste3	159.4	liste3	16.7
liste4	8.8	liste4	8.6	liste5	193.9	liste4	21.1
liste5	9.7	liste1	9.5	liste4	202.9	liste5	22.3
liste1	10.2	liste5	9.7	liste1	226.3	liste1	22.5
bol2	13.1	yukle1	15.8	yedi4	387.6	yedi4	38.8
bol1	13.4	yedi4	16	yedi1	408.6	sifre3	42.2
sifre2	14	yedi1	17	sifre2	412.8	yedi5	42.8
yukle3	14	sifre3	17.5	sifre3	413.4	yedi1	43
bol3	14.2	sifre2	17.7	yedi5	414.6	sifre2	44.5
oku2		oku2		oku2		oku2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
oku2	0	oku2	0	oku2	0	oku2	0
oku4	10	oku3	9.6	oku3	225	oku3	26.7
oku3	11.6	oku1	11.6	oku4	274.5	oku1	31.2
oku5	14.7	oku4	11.6	oku1	288.5	oku4	33.4
yaz1	14.7	oku5	12.8	oku5	315.7	oku5	36
osman4	15.6	yaz1	12.8	yaz1	315.7	yaz1	36
oku1	15.7	osman1	20.7	osman1	447	alti5	46.1
sakla2	16.1	on3	21.4	osman3	472.2	osman1	55.1
yardim2	16.4	osman3	22.4	osman5	515	osman3	58
osman1	16.6	on5	22.7	alti5	529	osman5	60.5
on2		on2		on2		on2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
on2	0	on2	0	on2	0	on2	0

on1	6.1	on1	5.7	on1	135.4	on1	13.4
on5	6.3	on5	6.4	on5	171.7	on5	18.8
on4	9.4	on4	10.4	on4	264.5	on4	26.7
yardim2	9.6	on3	11.5	on3	271.2	on3	28.8
on3	11	oku1	21.4	kod1	491.8	kod1	56.7
yardim1	11.7	yardim2	21.8	kod2	507.3	kod2	62.3
yardim5	13.4	kod4	22.7	osman4	516.2	osman4	62.4
bol2	13.9	oku3	22.9	kod3	527.6	oku1	62.8
yardim3	14.4	yardim1	23.1	alti2	538.7	kod4	64.6
osman2 (lpc)		osman2 (lpc_cep)		osman2 (lpc_cep_lif)		osman2 (lpc_cep_pro)	
osman2	0	osman2	0	osman2	0	osman2	0
osman1	10.6	osman1	10.2	osman1	273	osman1	32.7
osman3	11.5	osman4	15.2	osman3	369.1	osman3	41.8
topla1	12.5	osman3	15.8	osman4	373.9	osman4	43.9
sakla3	13.7	osman5	22.4	osman5	475.2	osman5	55.3
osman4	14.1	topla5	23.1	topla5	567.7	oku4	68.3
topla5	14.4	topla2	24.6	oku4	578.2	alti5	69.4
sakla2	14.7	topla1	25.8	oku1	593.3	topla5	70.7
sakla1	15.1	topla4	26.8	on5	595.2	topla3	73.4
osman5	15.3	oku4	27	oku2	612.8	on4	74.1
sakla2 (lpc)		sakla2 (lpc_cep)		sakla2 (lpc_cep_lif)		sakla2 (lpc_cep_pro)	
sakla2	0	sakla2	0	sakla2	0	sakla2	0
sakla5	6.5	sakla5	5.9	sakla5	147.6	sakla5	20.8
sakla3	6.8	sakla4	6.4	sakla4	170.7	sakla4	22
sakla4	6.9	sakla3	6.6	sakla3	184.8	sakla3	25.8
sakla1	7.5	sakla1	7.5	sakla1	202.7	sakla1	26.8
topla5	8.2	topla5	11.6	topla5	294.6	topla5	39.2
topla1	9	topla1	13.4	topla1	373.6	topla3	47.7
topla3	9.5	topla2	15.4	topla2	391.6	topla1	48.5
yardim4	10.2	topla3	16.5	topla3	424.5	topla2	50.4
topla4	11.7	yardim4	17.3	yardim4	451.2	carp4	56.6
sekiz2 (lpc)		sekiz2 (lpc_cep)		sekiz2 (lpc_cep_lif)		sekiz2 (lpc_cep_pro)	
sekiz2	0	sekiz2	0	sekiz2	0	sekiz2	0
sekiz3	4.8	sekiz3	4.3	sekiz3	108.8	sekiz3	12.4
sekiz1	5.7	sekiz4	5.7	sekiz4	133.6	sekiz4	14.6
sekiz4	5.8	sekiz5	6.3	sekiz5	161	sekiz5	17.9
sekiz5	6	sekiz1	8.9	sekiz1	179.9	sekiz1	18.9
yaz2	10.5	yedi5	14.6	yedi5	377.3	yedi5	38.8
bir2	11.2	yedi3	15.7	yedi4	379.3	bir3	40
bulent4	11.4	iki3	15.9	yedi3	384.4	yedi4	40.4
bes2	11.5	yedi4	16.4	sifre1	400.7	bes5	40.6
bir5	11.6	bes1	17.4	bes5	403.3	yedi3	42.3
sifir2		sifir2		sifir2		sifir2	

(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
sifir2	0	sifir2	0	sifir2	0	sifir2	0
sifir4	7.1	sifir3	7.7	sifir3	180.3	sifir3	20.8
bulent4	7.7	sifir1	8.2	sifir1	188.5	sifir1	20.9
dur1	7.7	sifir5	8.9	sifir4	215.3	sifir4	24.7
sifir3	8.6	sifir4	9.8	sifir5	237.3	sifir5	26.3
yaz3	8.6	dur2	19.2	bol2	440.9	bol2	45.2
sifir1	9	dort5	19.7	bol1	481.8	bol4	51.6
sifir5	9.5	dort4	20.3	carp5	496.3	bol3	53.2
bulent5	9.6	dort3	20.9	uc5	510.9	bol5	53.3
dur2	9.8	uc5	21.4	dokuz2	513.6	bol1	53.8
sifre2		sifre2		sifre2		sifre2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
sifre2	0	sifre2	0	sifre2	0	sifre2	0
sifre5	6.6	sifre5	5.9	sifre5	132.1	sifre5	13.6
yukle3	7	sifre3	7.6	sifre3	140.3	sifre3	14.8
sifre3	8.1	sifre1	8.8	sifre1	165.6	sifre1	16.3
yukle4	8.3	sifre4	10.3	sifre4	185.9	sifre4	19.1
sifre4	10.2	yukle4	11.5	yukle4	268.1	yukle4	29.8
yukle5	10.8	yukle3	11.7	yukle5	276.8	yukle5	29.8
sifre1	11.1	yukle5	11.9	yukle3	300	yukle3	33.1
liste5	11.7	yukle1	13.1	yedi3	325.5	yedi5	34.9
yukle1	12.6	yukle2	16.5	yukle1	345.9	yedi3	36
topla2		topla2		topla2		topla2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
topla2	0	topla2	0	topla2	0	topla2	0
topla4	9.9	topla5	11.4	topla5	264	topla1	32
topla1	9.9	topla1	12.3	topla1	265.4	topla5	35.1
topla5	9.9	topla4	12.4	topla4	299.7	topla4	37.6
alti1	10.5	sakla1	14.5	sakla1	338	carp2	40.2
alti2	10.5	sakla2	15.4	topla3	355.2	topla3	41.1
alti4	13.1	alti2	16.1	carp2	361	alti4	41.4
topla3	13.5	sakla3	16.4	alti4	361.1	sakla1	42.3
alti3	13.6	alti1	16.5	sakla3	374.3	sakla4	47.7
carp2	14.2	sakla4	17	alti2	387.4	sakla3	48
uc2		uc2		uc2		uc2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
uc2	0	uc2	0	uc2	0	uc2	0
uc1	4.5	uc1	5.5	uc1	122.6	uc1	12.5
uc3	5.1	uc4	6.4	uc3	134	uc3	15.4
uc4	5.7	uc3	6.7	uc4	160.9	uc4	18.9
uc5	5.9	uc5	7.3	uc5	194.7	uc5	22.2
bir2	7	bir2	17.2	bir2	414.7	bir2	39.7
bir1	8.4	yedi5	17.3	bir1	427	bir1	40.7
bir5	9.2	bir1	17.7	sifre5	440.5	bir5	49.2
dokuz1	9.3	sekiz3	18.7	sifre2	448.4	sifre5	49.5
sekiz5	10.7	sekiz1	18.8	sifre1	456.2	sifre1	50

yardim2		yardim2		yardim2		yardim2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
yardim2	0	yardim2	0	yardim2	0	yardim2	0
yardim1	3.9	yardim1	5.6	yardim1	124.3	yardim1	14
yardim3	5.6	yardim3	6.6	yardim3	148.6	yardim3	17.8
yardim4	6.1	yardim4	8	yardim4	221.2	yardim4	23.3
bol2	6.3	bol3	12.1	bol3	355	bol3	38.7
bol4	6.6	yardim5	12.3	yardim5	381.6	bol2	44.4
yardim5	7.2	bol5	12.7	yaz5	395.4	bol5	44.7
bol3	8	bol2	14	bol2	406.6	bol4	46.1
on5	8.8	bol4	14.8	sakla1	415.8	yaz5	49.1
bol5	9.3	sakla1	16.4	bol5	427.8	yardim5	49.3
yaz2		yaz2		yaz2		yaz2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
yaz2	0	yaz2	0	yaz2	0	yaz2	0
yaz3	4.7	yaz3	5.6	yaz3	121.9	yaz3	13
ac2	7.2	yaz4	7.2	yaz5	186.6	yaz4	20.8
yaz4	7.5	yaz5	8.6	yaz4	192.8	yaz5	22.1
yaz5	7.9	bulent5	15.6	bulent5	406.2	dort5	47.8
dort3	8.2	bulent4	15.9	bulent2	406.8	bulent5	48.7
cikart2	8.9	dort5	16.9	cikart1	430.6	cikart2	48.7
ac1	9.2	bulent2	17	bulent4	434.9	dort4	50.1
kod5	9.5	dort3	17.6	cikart2	444.5	bulent4	50.6
bol1	9.8	dort4	17.9	dort5	450.7	bulent2	51.3
yedi2		yedi2		yedi2		yedi2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
yedi2	0	yedi2	0	yedi2	0	yedi2	0
yedi4	5.1	yedi4	4.2	yedi3	111.3	yedi3	9.8
yedi1	5.4	yedi1	4.8	yedi4	113	yedi4	10.9
yedi3	6.5	yedi3	5.2	yedi1	120.4	yedi1	11.4
yedi5	7.8	yedi5	6.4	yedi5	140	yedi5	12.6
yukle1	14.1	yukle2	14.1	iki2	366	iki2	35
yukle3	15.2	iki2	14.4	yukle2	370.3	iki3	36.2
yukle2	15.2	iki3	14.5	iki3	377.7	yukle2	37.4
yukle5	16	yukle5	16.1	sifre2	381.9	sifre2	39.6
iki3	16.3	sifre2	16.8	yukle1	387.7	iki4	39.9
yukle2		yukle2		yukle2		yukle2	
(lpc)		(lpc_cep)		(lpc_cep_lif)		(lpc_cep_pro)	
yukle2	0	yukle2	0	yukle2	0	yukle2	0
yukle5	7.6	yukle5	6.7	yukle5	146.5	yukle3	18
yukle3	8.4	yukle3	8.2	yukle3	179.3	yukle5	19
yukle1	10.7	yukle1	12.2	yukle1	250.6	yukle1	25.6
yukle4	10.7	yedi3	12.5	yukle4	273.1	yukle4	31.1
iki2	13.1	yedi5	13.9	yedi5	318.4	yedi5	32.2
sifre2	13.1	yedi2	14.1	yedi3	328.8	yedi3	34.1
liste5	13.2	yedi4	14.2	sifre2	369.8	yedi2	37.4

figure. Recognition rate of LPC cepstrum with lifter distance measure is higher than recognition rate of LPC cepstrum distance measure.

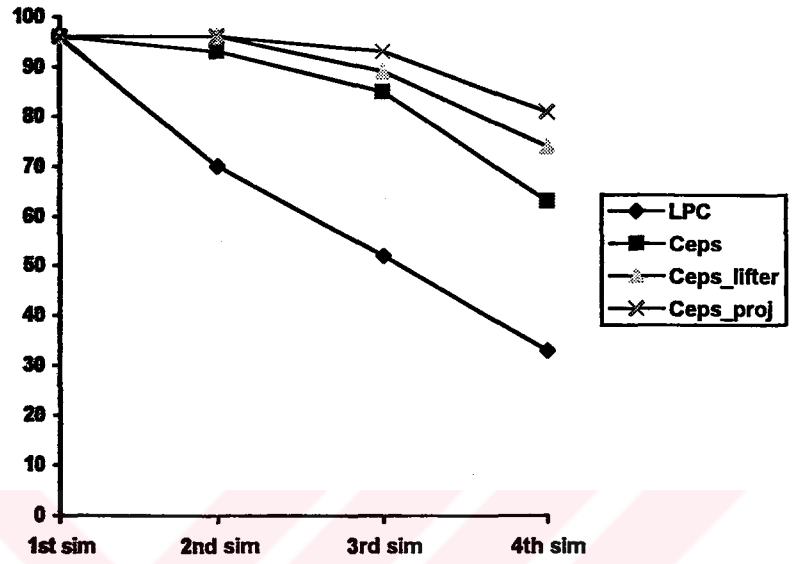


Figure 5.2.10. Recognition percentage according to distance measure type.

CHAPTER 6. CONCLUSION

During experiments some objective and subjective results have been obtained. Objective results are based on mathematical and statistical results. Subjective results are based on only observation. Subjective results are named as subjective observations.

6.1 OBJECTIVE RESULTS

1. Single comparison is not adequate for LPC distance measure.
2. Single comparison does not very dominant effect on recognition with LPC cepstrum distance measures. This is the expected result because cepstrum distance measures are symmetric.
3. Recognition with LPC distance measure has the least recognition rate with single and symmetric comparisons.
4. Recognition rate with LPC cepstrum with lifter distance measure has higher ratio than recognition rate with LPC cepstrum. This is good effect of liftering operation as explained previously.
5. Recognition ratio with LPC cepstrum projection measure has the highest recognition rate. This distortion measure is the most robust one.

6.2 SUBJECTIVE OBSERVATIONS

These observations have been obtained from Table 5.2.5.

1. Recognition rate for the words that have more than one syllable is high. These words also have more vowel than the other words. It is known that characters of vowels are dominant in a word due to its characteristic or unique spectral shape. Unvoiced sounds don't have very characteristic spectral shape. It is known that unvoiced sounds are modeled like white noise (white noise has flat spectrum).

Very good recognized words are 'alti', 'bulent' , 'cikart', 'dokuz', 'liste', 'osman', 'sakla', 'sekiz', 'sifir','sifre', 'yardim', 'yedi', 'yukle'. 1st, 2nd, 3rd and 4th most similar words are all in group of test word. So the recognition rate for these words is 100%.

2. Recognition rate for mono syllable words that has a vowel with a long duration is high. These words are 'ac', 'bes', 'bol','dur', 'on', 'uc', 'yaz'. For these words, 2nd, 3rd, and 4th most similar words are mostly in the group of test words.

3. Recognition rate for mono syllable words that have more than one unvoiced sounds is not very high. 2nd, 3rd, and 4th most similar words are not in the group of test word. These words are 'carp' , 'dort'.

4. Acoustically similar words for human perception are also mathematically similar. For example, the most similar word for 'topla' is word 'sakla'. These similarity can be seen from table and speech signal representations.

5. If two or more stage is used for recognition decision algorithm by considering some other constraints, recognition rate can be 100%.

6.3 FEATURE STUDY ON SPEECH RECOGNITION.

In this thesis, a very limited speech database has been used. To see the real performance of the speech recognizer a large speech database should be used. On the other hand, speeches are recorded in noise free environment nearly in ideal conditions(SNR > 20 dB). What is the performance of the recognizer in noisy environments are not known clearly. In the literature [7][8][9][23][24][25][26] [27][28] ,to increase performance of the speech recognizer in noisy environments different techniques are proposed. Three main category is available for robust speech recognition. First category is speech enhancement methods. In this method, noise suppression algorithms are used. These are filtering, adaptive noise cancellation algorithms, constraint iterative methods, etc. Second method is to find robust distortion measures to suppress effect of noise in recognition. LPC cepstral projection

distance measure used against noise, a kind of projection measures are used for these methods. The third method is training in simulated adverse conditions (like noisy environment simulating) or real adverse conditions. It has been seen that simulated adverse condition training or training in the real adverse condition is the most robust. To increase robustness of the recognizer, training in very different conditions is used. Training set have different training templates (reference template) in different conditions. In the recognition phase, recognition rate increases because recognizer has trained references in adverse conditions, so, it can find that word in the set.

For feature study, a robust pattern comparison based speech recognizer can be implemented. For this purpose, one or more robust local distance measure should be used and a strategy for training should be chosen. Mel cepstral features and subband cepstral measures are most robust features. A robust speech recognizer should use mel or subband cepstral features included delta (derivative) cepstral features, projection measure method, and training in simulating or real adverse condition method. If recognizer has two or more stage decision algorithm by eliminating bad candidate words in each stage, recognition ratio might be 95% even at 5dB (SNR).

REFERENCES

[1] RABINER, L.R. and R.W.SCHAFER, Digital Processing of Speech Signals, Prentice-Hall,1978

[2] RABINER, L.R and , A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proc. of IEEE, Vol 77, No 2, February, 1989

[3] LEA,W.A. Trends in Speech Recognition, Prentice-Hall, 1980

[4] ITAKURA, F. Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Transactions on A.S.S.P, Vol. ASSP-23, NO.1, February 1975

[5] MYERS,C. and RABINER, L.R and ROSENBERG,A.E. Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition. IEEE Trans. on ASSP, Vol 28, No 6, December 1980

[6] ACKENSEN, J.G and OH, Y.H Single Chip Implementation of Feature Measurement for LPC-Based Speech Recognition. AT&T Tech. Journal. Vol 64, No 8, October 1985.

[7] YANK,R. and Haavisto, P. Noise Compensation for Speech Recognition in Car Noise Environments. ICASSP 95, Vol 1, page 433-436

[8] WHIPPLE, G. Low Residual Noise Speech Enhancement Utilizing Time-Frequency Filtering. Volume1,page 5

[9] ARSLAN L, and McCREE,A. and VISWANATHAN, V. New Methods for Noise Suppression ,ICASSP 95, Volume 1,Page 812-815

[10] HIRCH,H.G and ERLICHER,C. Noise Estimation Techniques for Robust Speech Recognition, ICASSP 95, Volume 1, Page 153-156

[11] HERMANSKY,H. WAN,E. Speech Enhancement Based on Temporal Processing ,ICASSP 95, Volume 1, Page 405.

[12] PAPAMICHALIS, P.E Practical Approaches to Speech Coding, Prentice-Hall, 1987

[13] RABINER,L.R and ROSENBERG,A.E, LEVINSON, S.E. Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition, IEEE Transactions on A.S.S.P, Vol ASSP-26, NO.6 ,December 1978

[14] SAKOE,H. and CHIBA,S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Trans. on ASSP, vol ASSP-26, pp.575-582 , Dec,1978

[15] RABINER,L.R and JUANG, B.H. Fundamentals of Speech Recognition ,Prentice Hall, 1993

[16] DELLER, J.H and PROAKIS, J.G and HANSEN, J.H.L. Discrete-Time Processing of Speech Signals, Prentice Hall,1995

[17] MARKEL, and J.D. GRAY, A.H. Linear Prediction of Speech, Sringer-Verlag,1976

[18] RABINER, L.R, "On the performance of isolated word speech recognizers vector quantization and temporal energy contours", AT&T Tech. J, 63 1245-1260 ,1984

[19] LAMEL, L.F. and RABINER, L.R, "An improved endpoint detector for Isolated Speech Recognition", IEEE Trans. ASSP, ASPP-29, No 4, (August 1981)

[20] JUANG, B.H and WILPON, J.G, ..."On the use of bandpass liftering in speech recognition", IEEE Trans ASSP, 35(7), 947-954, July 1987,

[21] MONSOUR D. and JUANG, B.H, "A family of distortion measures based upon projection operation for robust speech recognition". IEEE Trans ASSP , Vol 37, No 11 , November 1989

[22] MYERS C. and RABINER, L.R "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition", IEEE Trans. ASSP, Vol ASSP-28, NO.6. December 1980.

[23] MILNER, B.P and VESEGHI, S.V. "Speceh recognition in impulsive noise", IEEE ICASSP-95, Vol 1.

[24] ERZIN, E. and CETIN, A.E and YARDIMCI , Y. "Subband analysis of robust speech recognition in the presence of car noise", IEEE ICASSP-95, Vol 1.

[25] HANSEN, J.H.L and ARSLAN ,M.A "Robust feature-estimation and objective quality assessment for noisy speech...", IEEE Trans. on Speech and Audio Processing. Vol 3, No 3, May 1995

[26] HANSEN, J.H.L and NANDKUMAR ,S. "Dual channel iterative speech enhancement with constraints on an auditory-based spectrum". IEEE Trans. on Speech and Audio Processing. Vol 3. No 1. January 1995

[27] JUNQUA, J and MAK, B. "A robust algorithm for word boundary detection in the presence of noise", IEEE Trans.on Speech and Audio Processing. Vol 2, No 3, July 1994

[28] ITAKURA, F and KAJITA, S. "Robust speech feature extraction using SBCOR analysis", IEEE ICASSP-95 Vol 1.

[29] AKANSU, A.N. and HADDAD, R.A "Multiresulation Signal Decomposition" Academic press ,1995

Curriculum Vitae

Osman MERAL was born in Giresun, in 1970. He received B.S. degree in Computer and Control Engineering from Istanbul Technical University in 1992. His B.S. thesis subject is telephone responding device and real-time implementation of DTMF tones using a DSP.

In the same year, he began to work in Hardware Design Engineering Department of Research & Development Center in NETAS. He dealt with system software design for telecommunication systems, digital signal processing, speech coding, speech recognition, and image coding. His new subject is now wireless digital communication, and he is dealing with DECT(Digital European Cordless Telephone) project in NETAS.