Similarity Measuring Algorithm

Ms. Munazah Gul[1], Ms. Sukhvinder Kaur[2], Muheet Ahmed Butt[3], Majid Zaman[4]

[1]M. Tech Student, Department of Electronics and Communication, Swami Devi Dyal Inst. of Engg. & Technology, Kurukshetra University, Kurukshetra
[2]Assistant Professor and Head of Department of Electronics and Communication, Swami Devi Dyal Inst. of Engg. & Technology, Kurukshetra University, Kurukshetra
[3]Scientist, PG Department of Computer Sciences, University of Kashmir, Srinagar
[4]Scientist, Directorate of IT&SS. University of Kashmir, Srinagar

Abstract- Speech recognition has been an integral part of human life acting as one of the five senses of human body, because of which applications developed on the basis of speech recognition, have high degree of acceptance. This thesis has tried to analyze different steps involved inartificial speech recognition by man-machine interface. The various steps we followed in speech recognition are feature extraction, distance calculation, dynamic time wrapping. The most generic objective of the thesis was to analyze the similarity measuring algorithms in ASR systems. We at first calculated the different type of feature vectors such as LPC, RASTA and MFCC. After performing such operations we analyzed MFCC in particular, and selected it as the preferred mode of feature vector coding because they follow the human ear's response to the sound signals. We also found different methods of distance measurement and compared them and concluded that Euclidean distance measure is a preferred one when the template database of sound is very low. We also performed a quick analysis of dynamic time wrapping algorithm and found the least path between two sounds. Then we designed a small model by writing a simple code which was able to recognize small set of isolated words.

I. INTRODUCTION

Great progress in ASR has yielded many practical applications in recent years, such as user-friendly speech interfaces in control consoles of cars, credit card number recognition and the verbal selection of menus over the telephone. However, after 50 yearlong research efforts and considerable advances in ASR notwithstanding, robust speech recognition for human-machine interface still remains a challenging problem today. The performance of the modern speech recognizers may turn out to be poor under adverse conditions, especially when classifiers are trained under high signal-to-noise ratio (SNR) environments like noise-free chambers (typically where SNR $\geq$ 30 dB) and operated in real-world surroundings of relatively lower SNR [1]. In contrast, a healthy human listener's performance is usually far more stable on average under similar training and operating conditions. Unfortunately many researchers agree that human-quality; adaptively-learning and noise-robust machines that recognize and interpret human speech will not be achieved in the near future [1,2].However, even incremental improvements leading toward this ultimate goal in ASR are of great importance.

A brief introduction to how the speech signal is produced and perceived by the human system can be regarded as a starting point in order to go into the field of speech recognition. The process from human speech production to human speech perception, between the speaker and the listener, is shown in Figure 1 [3].

Fig: 1 Human speech communication Speech recognition

In electronic engineering, speech recognition (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or just "speech to text" (STT). Speech recognition systems try to establish a similarity to the human speech communication system. The aim of human speech communication is to transfer ideas. They are made within the speaker's brain and then, the source word sequence is performed to be delivered through her/his text generator. The human vocal system, which is modeled by the speech generator component, turns the source into the speech signal waveform that is transferred via air (a noisy communication channel) to the listener, being able to be affected by external noise sources. When the acoustical signal is perceived by the human auditory system, the listener's brain starts processing this waveform to understand its content and then, the communication has been completed.
Speech Recognition also known as automatic speech recognition is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program or we can say it is the ability of the

computer to accept speech in audio format and then generate its content in text format. Speech recognition in computer domain includes various steps with issues attached with them. The general model begins with a user creating a speech signal which is amplitude versus time waveform. The digitized speech signal is used to extract various spectral and temporal features like zero crossing rate, short time energy, fundamental frequency, mfcc etc. Some of these features are used for word boundary detection, silence detection etc. which are done during the preprocessing of the speech signal and many along with these are used for recognition in subsequent phases by making a feature vector. These feature vectors are compared against stored and trained knowledge model to categorize phonemes which are further combined to form the target words. These words depending upon their probabilistic confidence either are accepted or rejected. Despite of years of research speech recognition system is still challenging field. There are multiple factors which affects the performance of the speech recognition accuracy. Such as, background noise effect, speaker variability, same word spoken differently by different region of people within same county like India, types of words i.e isolated, continuous, dictation type. So, various researcher works take into consideration during literature survey. Speech is the most natural way for communication between different people. The aim of speech recognition system is to make interaction between human and machine possible [18]. It seems to be a straight forward problem, but from research it has been revealed that it's difficult to achieve accurate results. The speech recognition system faces multidimensional problems such as non-stationary nature of speech, large vocabulary size, confusable words, speaker dependency, and large processing time

II. RELATED WORK

Speech is the natural and the fundamental way of communication for most humans. Technically speaking, Automatic Speech Recognition (ASR) refers to mechanism (hardware and software combined) that stores some representations of distinguishing characteristics of speech with a source of input equipment, such as a microphone and further processes these representations to match them to incoming speech in an effort to interact with machines, computers and/or human users. The first primitive recognizer was developed at Bell Labs during the early 1950s. However, it was the 1960s that brought many major breakthroughs to the field of ASR. Some of these achievements are noteworthy to mention herein because they did not only develop significant tools for speech recognition but also established the very basic concepts on which most of the research work is mainly based. Specifically, the development of the Fast Fourier Transform (FFT) by Cooley and Tukey in 1965 decreased the computational load of Discrete Fourier Transform (DFT) with a faster algorithm, thereby enabling the practical implementations of Digital Signal Processing (DSP) custom chips [6,2]. Oppenheim, Schafer, and Stockham introduced Cepstral Analysis which performs deconvolution of the speech signal to separate an excitation sequence from an impulse response convolved with it [7]. Cepstral coefficients and many derivatives have been widely used to represent the short-term spectral envelope of speech signals thus far. It was also the late 1960s and early 1970s that saw another useful method for speech analysis, known as Linear Predictive Coding (LPC). One of the earliest and complete papers on the application of linear prediction to speech analysis was published by Atal and Schroeder [8,1]. Basically, LPC uses a pole-only (autoregressive) filter to model the speech signal. LPC coefficients and its derivatives are extensively used for transmitting speech spectral envelope information [2].Most notably, the foundations for the statistical technique of Hidden Markov Modeling, which models an observed sequence as produced by a sequence of hidden states, dates back to the 1960s as well. However, the first successful applications of Hidden Markov Modeling to speech recognition were realized in the 1970s [2].

Baum and his colleagues developed a popular expectation-maximization (EM) algorithm, known as the Baum-Welch Re-estimation Algorithm (or Forward-Backward Algorithm),to estimate the parameters of a Hidden Markov Model (HMM) iteratively [9,10]. Hidden Markov Models (HMM) and the Baum-Welch Re-estimation Algorithm are widely used today in contemporary state-of-the-art speech recognition systems. Dynamic Time Warping (DTW), a deterministic alternative approach to the statistical HMM was also introduced in the 1970s. DWT normalizes the different-length utterances of the same word and applies template-based classification to speech recognition. Many different approaches incorporating DTW, HMM and Artificial Neural Networks (ANN) were developed for speech recognition in the 1970s. Among these studies, the project of the Advanced Research Projects Agency (ARPA), was are markable achievement in that it performed a 1000-word ASR system by using connected speech from a few speakers with a word error rate of less than 10% [2]. In the 1980s, the project of the Defense Advanced Research Projects Agency(DARPA Project) and the major other programs conducted by Texas Instruments and the Massachusetts Institute of Technology (TIMIT Project) and the National Institute of Standards and Technology (NIST) primarily concentrated on the collection of large corpora used for training and testing speech recognizers. These large corpora were subsequently used by the ASR research community at large for performance comparison of different approaches applied to speech recognition. The ASR community witnessed some other important developments in the 1980sas well. Among those, the Mel-Cepstrum Analysis introduced by Davis [11],and the Dynamic Cepstral Coefficients proposed by Furui [12] can be considered remarkable techniques for speech feature-extraction due to significant improvement in recognition accuracy. As for the speech recognizers of the 1980s, many researchers were experimenting with frame-based HMM recognizers, ANN recognizers or hybrid schemes combining HMM and ANN in isolated and/or continuous contexts of speech [2]. Most importantly, the contemporary speech recognition systems of today still use these methods predominantly. Waibel and Lee addressed the question of complexity involved in ASR in as "dimensions of difficulty."

These are the factors that determine the complexity and the specifications of an ASR system. Deller et al. summarizes some of these factors that render speech recognition methods complicated.

III. PROPOSED METHODOLOGY

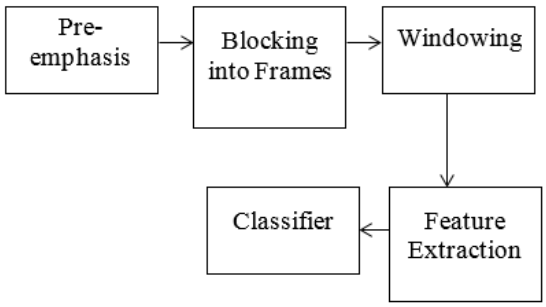

Fig: 2. Proposed method

A. Pre emphasis

The digitized speech signal is processed by a first order digital network in order to spectrally flatten the signal. This pre emphasis is easily implemented in the time domain by taking difference.

$$\tilde{A}(n) = A(n) - a^* A(n-1)$$

a= scaling factor = 0.95, A(n)= Digitized Speech Sample, A(n-1) = Previous digitized Speech Sample, $\tilde{A}(n)$ = Pre emphasised Speech Sample, n = No. of Samples in the whole frame.

B. Blocking into Frames

Section of N (e.g. 300) consecutive speech samples are used as a single frame. Consecutive frames are spaced M (e.g. 100) samples apart.

$$X_l(n) = \tilde{A}(M^*l + n) \quad , \; 0 <= n <= N-1 \text{ and } 0 <= l <= L-1$$

N = Total No. of samples in a frame, M = Total No. of sample spacing between the frames. [Measure of overlap], L = Total number of frames.

C. Frame Windowing

Each frame is multiplied by an N sample window W (n). Here we use a hamming window. This hamming window is used to minimize the adverse effects of chopping an N sample section out of the running speech signal. While creating the frames the frames the chopping of N sample from the running signal may have a bad effect on the signal parameters. To minimize this effect windowing is done.

$\hat{U}_l(n) = X_l(n) * W(n)$, $0 <= n <= N-1$

W(n) = Scale factor i.e. (0.54 - 0.46*Cos(2*pie*n/ N)) , $0 <= n <= N-1$

N = Total No. of samples in a frame.

The multiplicative scaling factor ensures appropriate overall signal amplitude

IV. RESULT AND DISCUSSION

As every journey begins with a small step here we are trying to achieve that small step in the field of speech recognition. Here we have presented at first the analysis of different feature extraction procedures. Then we have tried to present an analysis of MFCC as how it is a good approach of feature extraction. Then we have tried to analyze different methods of distance measure used to calculate to the distance between the feature vectors extracted by us. Then we try to do a small analysis of dynamic time warping using dynamic programming approach. At the last but not the least we try to present a small program for small speaker dependent recognition system to recognize isolated words.Here we want to state that as we were motivated by the application of speech recognition in mobile phones we here are trying to recognize the English numerical digits from 'zero' to nine'. It should be also noted that this applications are not restricted by this and can be used to recognize any isolated words with appropriate changes. All the programming used here is done in Matlab due to obvious reasons of it being the most efficient tool for mathematical and signal analysis.

MFCCs are commonly derived as follows:

1. Take the Fourier transform of (a windowed excerpt of) a signal
2. Map the log amplitudes of the spectrum obtained above onto the Mel scale, using triangularoverlapping windows.
3. Take the Discrete Cosine Transform of the list of Mel log-amplitudes, as if it were a signal.
4. The MFCCs are the amplitudes of the resulting spectrum.

A set of Matlab modules were written to find the above mentioned coefficients and the corresponding raps for letters 'zero' to 'nine' are given below.

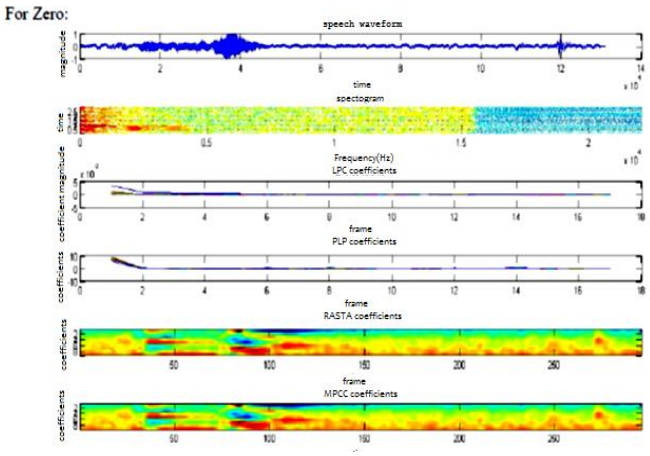

Fig: 3 Coefficients for zero

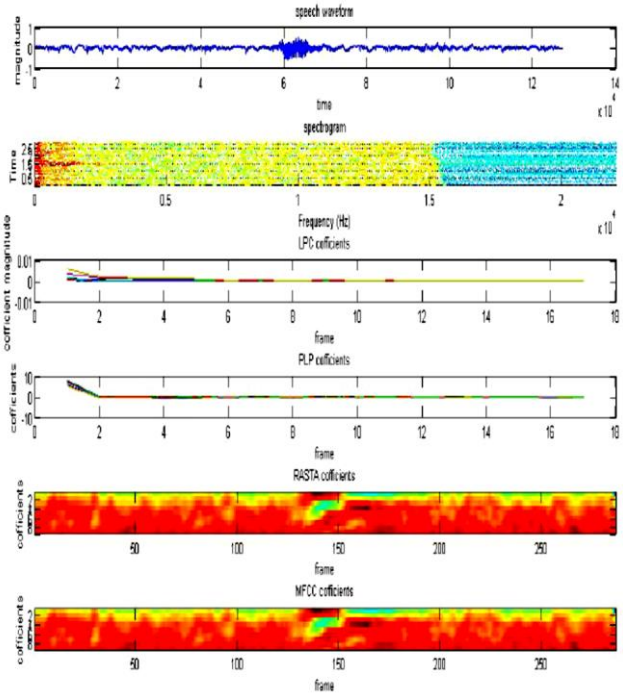

Fig: 4 Coefficients for One

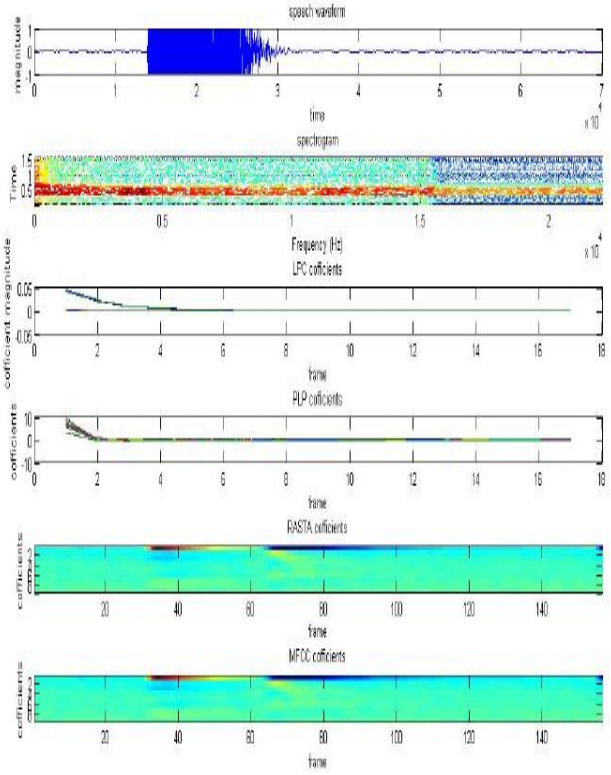

Fig: 5 Coefficients for Two

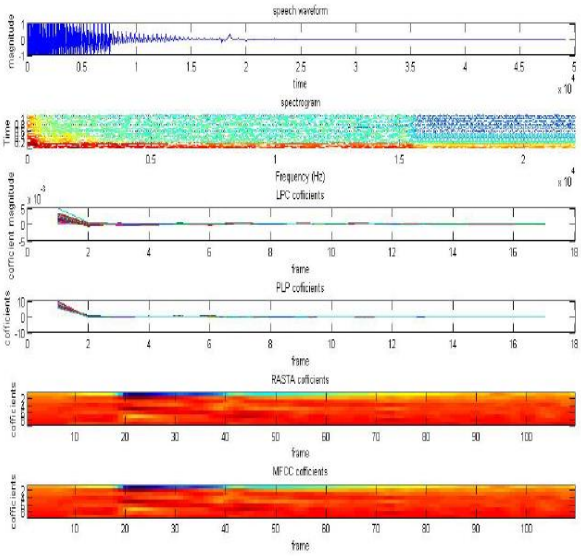

Fig: 6 Coefficients for Nine

An important conclusion that we can make from the last set of experiments is that one of themain reasons for the need of large training databases for LPC based analysis (without filtering) ,is the large difference between the different telephone lines, which is reflected in a difference inspectral distortion.

Out of all the different options available for feature extraction we selected the MFCCoefficients as in the MFC, the frequency bands are positioned logarithmically (on the mel scale) which approximates the human auditory system's response more closely than the linearly spacedfrequency bands obtained directly from the FFT (Fast Fourier Transform) or DCT (DiscreteCosine Transform). This can allow for better data processing. This feature of MFCC can be analyzed by a Matlab programme which takes in a speech waveform converts it into the MFCCcoefficients and then reconstructs the waveform from the MFCC and thus compare the powerspectra of the original sound and the reconstructed sound.

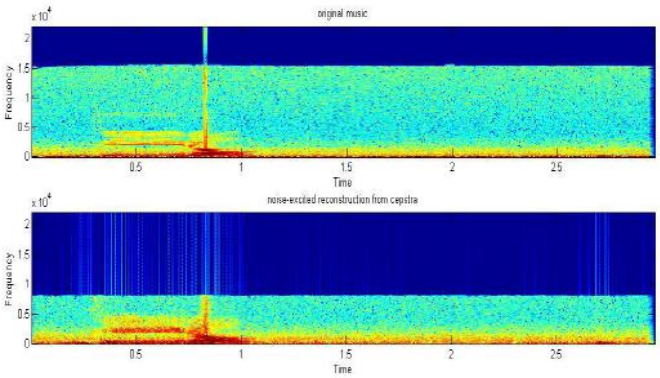

Fig: 7 MFCC co-efficient analysis

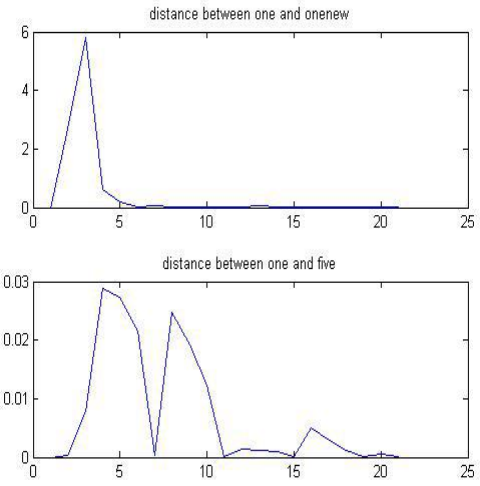

Fig: 8 Euclidian distance

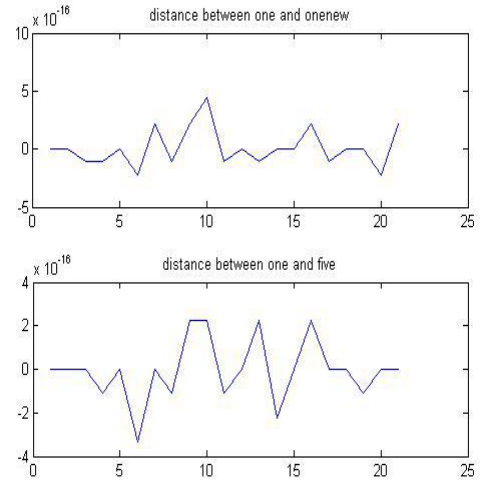

Fig: 9 Itakura-saito distance

Thus it can be easily seen that even though Itakura-saito distance is a very good form of distancemeasure its performance for the case of isolated word recognition with very little database isvery poor. Thus we have decided to use Euclidean distance for our purpose.

A. Dynamic Time Warping

One of the difficulties in speech recognition is that although different recordings of the samewords may include more or less the same sounds in the same order, the precise timing – thedurations of each sub word within the word - will not match. As a result, efforts to recognizewords by matching them to templates will give inaccurate results if there is no temporalalignment.

Although it has been largely superseded by hidden Markov models, early speech recognizersused a dynamic-programming technique called Dynamic Time Warping (DTW) to accommodatedifferences in timing between sample words and templates. The basic principle is to allow arrange of 'steps' in the space of (time frames in sample, time frames in template) and to find thepath through that space that maximizes the local match between the aligned time frames, subject to the constraints implicit in the allowable steps. As the duration of speaking for differentpersons are different DTW is highly unavoidable. The most common algorithm used for thispurpose is dynamic programming. Here we bring aMatlab program to calculate the DTW fortwo given signal, the input signal is two different versions of word ‘one’.

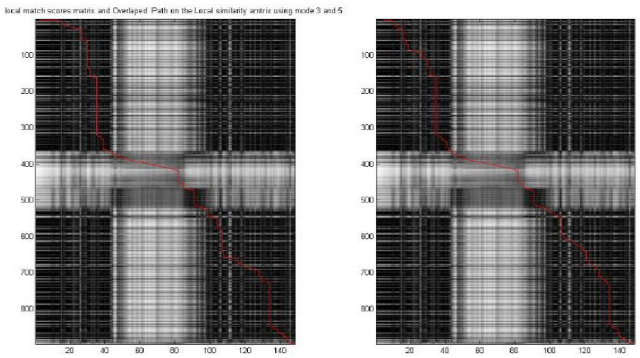

Fig 10 Dynamic Time Warpping

After analyzing the different parts of the speech recognition analysis here we try to present asmall program which does two tasks. The first task is to produce a data base of templates foronce spoken wards for example ‘zero’ to ‘nine’. This is known as the training of the recognizer.The next task is to recognize. The MFCC feature coefficient is used here for reasons statedearlier Euclidean distance is used to measure the distance between the feature vectors. Here wefirst give the process of training.

V. CONCLUSION

In this project we at first calculated the different type of feature vectors such asLPC,RASTA and MFCC. After performing such operations we analyzed MFCC in particular, and selected it as the preferred mode of feature vector coding because they followthe human ear’s response to the sound signals. We also found different methods of distanceMeasurement and compared them and concluded that equiledean distance measure is a preferred one when the template database of sound is very low.We also performed a quick analysis ofdynamic time wrapping algorithm and found the least path between two sounds. Then we designed a small model by writing a simple code which was able to recognize small set ofisolated words.

The performance of this model is limited by a single template generated by the trainingprogrammers, as it does not incorporate training algorithm of any sort. The performance factorcan be optimized busing high quality audio devices in a noise free environment. There is apossibility that thespeech can be recorded and can be used in place of the original speaker.Thiswould not be a problem in our case because the

MFCCs of the original speech signal and therecorded signal are different. Finally I conclude that although the project has certain limitations,its performance and efficiency have outshined these limitations at large.

VI. REFERENCES

[1]. J. R. Deller, J. H. L. Hansen and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, New York, 2000.

[2]. B. Gold and N. Morgan, *Speech and Audio Signal Processing*, John Wiley & Sons, 2000.

[3]. Meseguer, NoeliaAlcaraz, "Speech analysis for automatic speech recognition, "Norwegian University of Science and Technology, Master's Thesis 109 (2009).

[4]. Fujimoto, M. and Ariki, Y., 2000. Noisy speech recognition using noise reduction method based on Kalman filter. IEEE transactions on acoustic, speech and signal processing, vol. 3, pp. 1727-1730.

[5]. Yoon, T.J., Zhuang, X., Cole and Jhonsen, M.H., 2009. Voice Quality Dependent Speech Recognition. Linguistic patterns in spontaneous speech, Academia Sinica.

[6]. J. W. Cooley and J. W. Tukey, "An algorithm for the machinecomputation of complex Fourier series," *Mathematical Computations*, Vol. 19, pp.297-301, 1965.

[7]. A. V. Oppenheim, R. W. Schafer and T. G. Jr. Stockham,"Nonlinear filtering of multiplied and convolved signals," *Proceedings of the IEEE*,Vol. 56, No. 8, pp. 1264-1291, 1968.

[8]. B. S. Atal and L. S. Hanauer, "Speech analysis and synthesis bylinear prediction of the speech wave," *Journal of the Acoustic Society of America*,Vol. 50, pp. 637-655, 1971.

[9]. L. E. Baum and T. Petrie, "Statistical inference for probabilisticfunctions of finite state Markov chains," *Annals of Mathematical Statistics*, Vol. 37,pp. 1554-1563, 1966.

[10]. L. E. Baum, T. Petrie and G. Soules, "A maximization technique inthe statistical analysis of probabilistic functions of Markov chains," *Annals ofMathematical Statistics*, Vol. 41, pp. 164-171, 1970.

[11]. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 28,No. 4, pp. 357-366, 1980.

[12]. S. Furui, "Speaker-independent isolated word recognition usingdynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech andSignal Processing*, Vol. 34, No. 1, pp. 52-59,1986.

[13]. A. Waibel and K. F. Lee, *Readings in Speech Recognition*,Morgan-Kauffmann, Palo Alto, California, 1990.

[14]. S. Moon and J. N. Hwang, "Robust speech recognition based on joint model and feature space optimization of hidden Markov models," *IEEE Transactions on Neural Networks*, Vol .8, No .2, pp. 194-204, March 1997.

[15]. M. Shozakai, S. Nakamura and K. Shikano, "Robust speechrecognition in carenvironments," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP `98)*, Vol. 1, pp.269-272, May 1998.