International Journal of Advanced Trends in Computer Applications
www.ijatca.com

KASHMIRI SPEECH RECOGNITION SYSTEM USING LINEAR PREDICTIVE CODING AND ARTIFICIAL NEURAL NETWORKS

[1]Aabid Rashid Wani
Department of Electronics & Communication Engineering,
Islamic University of Science & Technology, Awantipora, India.
[2]Er. Tabish Gulzar
Department of Electronics & Communication Engineering,
Islamic University of Science & Technology, Awantipora, India.
[3]Er. Mamoon Rashid
Department of Computer Science & Engineering,
Chandigarh University, Mohali, Punjab,India.

Abstract: *This paper suggests an approach to recognize Kashmiri words corresponding to digits zero (safer) to Nine (Nov) spoken in an isolated way by different male and female speakers. The study performs feature extraction for isolated word recognition using Linear Predictive Coding (LPC) and Artificial Neural Network (ANN). The dataset comprising of males and females voices were trained and tested where each word has been repeated 5 times by the speakers. An accuracy of 92% is acquired by the combination of features, when the suggested approach is verified using a dataset of 350 speech samples, which is excess than those acquired by using the features singly.*

Keywords- *LPC, ANN, Extraction, Kashmiri, Database.*

I. INTRODUCTION

The main aim of this research is to provide an easy and natural way for human beings to communicate with the computer or mobile device. Human beings mostly communicate with each other through the natural way called speech. But in most cases, where speech is not way of communicating there we found other ways of .The objective of this research work is to make computer capable to understand the natural speech through the Automatic Recognition System(ARS)[1].Present computers can understand the human natural speech in very limited capacity. As present interfaces are usually mouse and keyboard, and it is necessary for a person to learn certain skills before communicating with the computer. But it would be easy for a person to communicate with the computer when computer could understand the natural language of the person. Speech recognition is very helpful for those handicapped persons who are unable to use keyboard and mouse. Computers also need a way to be able to recognize who is trying to excess them. In the present technology this is done by the use of passwords which gives user identification. But the problem with the passwords is that they are not effective because of several reasons. The first reason is that the computer identifies the user through the sequence of characters input by the user. So it is easy to excess the computer if the sequence of characters is known to anyone. Another drawback of the passwords is that they can be easily guessed or broken. These problems can be overcome by using person's voice features which is unique to the individual. Because of this individual's uniqueness, a person's voice could be very accurate to authenticate the user. So voice recognition has the properties of user friendly and secure. We want to design such a speech recognition system which can understand a few simple commands and identify them clearly. Linear predictive parameters and their derived parameters related to speaker's vocal tract have been used for speaker identification system [2, 3]. The linear predictive coding (LPC) derived parameter with hidden Markov models has been used in both speech and speaker identification system [4, 6]. Automatic speech recognition and speaker identification using artificial neural network (ANN) is described in [5]. Not so much work has been done on speech recognition for Kashmiri Language. The other feature extraction methods like zero-crossing rate,

short-time energy, pitch-extraction, formant frequencies have been studied [7].

The remainder of the paper is organized as follows. Section II explains the Speech Recognition System. In Section III, the authors describe the theoretical background regarding the work to be presented in the paper. Section IV gives database development. In Section V, the authors provide the end point detection of the present architecture. In Section VII the authors discuss results and comparison study of implemented architecture with other models. Finally, Section VIII discusses future extensions and concludes the paper.

II. SPEECH RECOGNITION SYSTEM

This can be explained in two steps, in which the first step consists of speech processing and the second is pattern recognition. The speech processing depends on the various number of speech recognition stages. which include the speech end point detection, windowing of the speech signal, filtering the samples of speech so that The resulting speech would be free from the noise, computing the linear predictive coding and cepstral coefficient from the LPC coefficients and then implement the vector quantization on the signal to obtain the code book, which helps us in the pattern recognition step.

The codebook acts as an input to the recognizer in pattern recognition step .The recognizer using here is ANN (Artificial Neural Network) and hence the codebook acts as input feature vectors to the ANN. The back propagation algorithm of the ANN is used for training purposes and then the training parameter is stored for future to test the same sample speech.

III. THEORETICAL BACKGROUND

Speech signal is a slowly time varying signal, when seen over a relatively short interval of time, and its characteristics are quite stationary [8]. In order to meet the requirements that include computational accuracy, complexity, response time etc different applications make use of different algorithms. These applications include those which are based on energy threshold, pitch detection, ZCR .The end points of speech are usually obscured by speaker generated artifacts such as clicks or by dial tone. Long-distance telephone transmission channels may also introduce these types of artifacts and some background noise [9]. Conventional short-time or spectral energy or ZCR based endpoint detection algorithms are usually susceptible to speech artifacts such as breath, mouth and lip clicks and break down easily in the presence of noise. These classical energy threshold methods i.e. energy and ZCR methods, the threshold value need to be recalculated at each and every silence (voice-inactive) segment [10]. And in case, the noise is non-stationary, these methods fail to track the exact value of thresholds, resulting in falsely detected endpoints. In some of the applications which may include speaker verification, name dialing, speech control etc, where the speech (voice-active) part of the signal is sometimes less (e.g. less than 2 s) and the recognition process can be done within 1 s or even less, that are usually provided by embedded systems, such as wireless phones or portable devices; or in multi-user systems, such as speaker verification system for several users, usually a low computational complexity for low cost or for faster response of the system is required [11]. One solution for the abovementioned cases is to make use of an accurate endpoint detection algorithm to remove all silence (voice-inactive) part. Intrusion of the proposed Bit-wise method uniquely defines the threshold first instance.

IV. DATABASE DEVELOPMENT

Fifty dissimilar words are used for database. Three female and four male speakers are used for tape session, thus assembly a sum of three hundred fifty utterances. Stereo headset H 250 with frequency response of 20 Hz-20 KHz was used. A sampling rate of 16000 Hz was used for all data. For different of the recordings, there is some surroundings or a few lip and mouth clicks or breath noise.

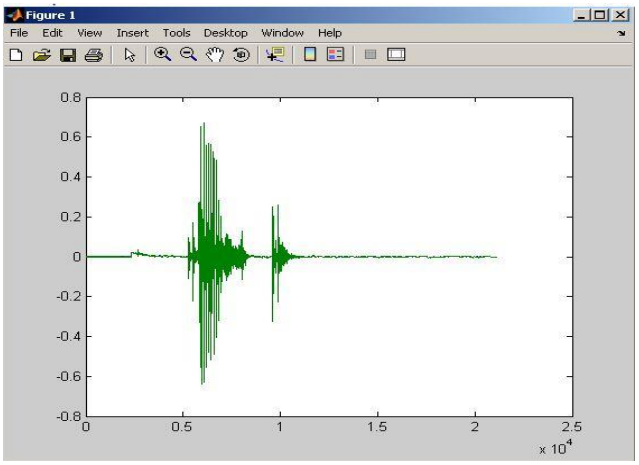
V. END POINT DETECTION

End point detection is to find the utterance and separates the digit spoken word from the background noise or silence. The techniques of End point detection that we used in this paper is based on the short time energy as well as zero crossing rate of the speech signal[12]. For detection of voiced part from the signal we use the short time energy. However if the digit spoken starts or ends with unvoiced phoneme then there is a chance of it not getting detected. Because of this we take the help of zero crossing rate to find the unvoiced part without going detected. Initially the start and end points S1 and E2 respectively are find by assigning lower and upper threshold values determined by formula given by [13]:

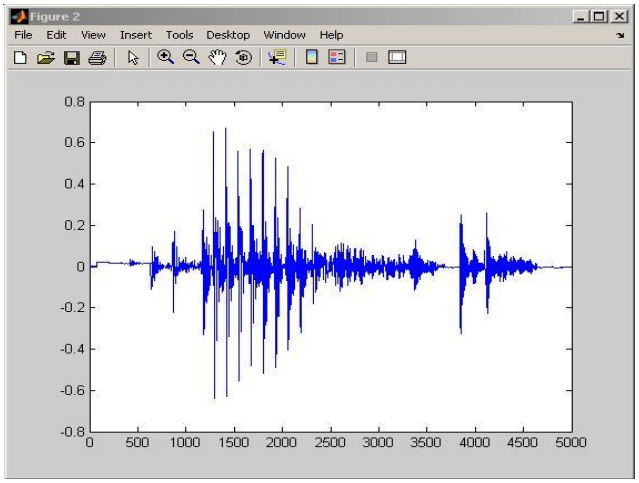
$$K1=0.03 \times (Lmax-Lmin)+Lmin$$
$$K2=4 \times Lmin$$
$$LTH=min(K1,K2)$$
$$UTH=5 \times LTH$$

Where LTH is the lower threshold and UTH is the upper threshold. The point at which the signal energy

exceeds the lower threshold and then the upper threshold before falling below lower threshold is the start point S1 and the point at which it falls again below LTH is the end point E2.Next the ZCR can be used before the start and the end points to detect any unvoiced part that might have been missed. Fig shows the plot of signal before and after endpoint detection in MATLAB tool.



(a) Speech signal before end point detection



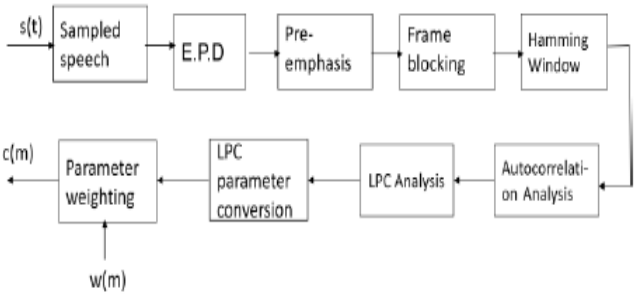
(b) Speech signal after end point detection

VI. LINEAR PREDICTIVE CEPSTRAL COEFFICIENTS

Linear Predition can be used for predicting the present speech sample using the past samples. The predicted vale can be given by [14]:

$$S_n = \quad {}_{n-k}\, a_k$$

Here a_k are the prediction coefficients and S_{n-k} are the previous values used to determine the present value S_n. Actually the prediction coefficients are determined for minimizing the prediction error in the least squares sense using autocorrelation function. The total number of predictions or the order is defined by m. The following Fig shows the block diagram computation of LPCC computation.



Here the input speech signal is sampled and endpoint detection and pre-emphasis are implemented to compensate for the loss in higher frequencies. It is then divided into the frames of 10ms duration and then multiplied by overlapping Hamming windows. After the LPC coefficients are determined with the help of using autocorrelation functions and then these coefficients are converted into cepstral coefficients C_m using the equations (A),(B),(C).the quantity of cepstral coefficients Q should be 1.5 times the quantity of LPC parameters m[12].The coefficient C_0 indicates the average energy determined from the gain G in each spoken digit and hence is discarded that is amplitude is normalization is done

$$C_0 = \ln(G) \qquad (A)$$

$$C_m = -a_m + 1/m \qquad \sum_{k=1}^{m} {}_kC_{(m-k)}]\, 1 \le m \le p \qquad (B)$$

$$C_m = {}_kC_{(m-k)}], p < m < Q \qquad (C)$$

VII. RESULTS

The parameters to obtain the accuracy of the recognition system is the Recognition Rate [15] in which $R_{correct}$ is the number of words that are recognized accurately and R_{total} is the number of words in the vocabulary.

$$\text{Recognition Rate} = R_{total} \times 100\%$$

An experiment was performed where the Recognition Rate was estimated for each of the feature extraction technique for 4 males and 3 female speakers each. This is a speaker dependent recognition system that we implemented here that is the digits spoken by the same speaker are recognized. The results obtained by this experiment are shown in table 1 and table 2

Feature Extraction Technique	Speaker 1	Speaker 2	Speaker 3	Speaker 4
LPCC	87.08%	91.93%	93.34%	92.51%

TECHNIQUE FOR MALE SPEAKERS

Feature Extraction Technique	Speaker 1	Speaker 2	Speaker 3
LPCC	90.73%	94.87%	92.23%

TECHNIQUE FOR FEMALE SPEAKERS

VIII. CONCLUSION

In this paper, the authors tried to make the recognition rate of female speakers better than the male speakers by the application of techniques. However the performance can be further improved with the help of other techniques like MFCC and so many.

REFERENCES

[1] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*,Prentice-Hall, Englewood Cliffs, 1978.

[2] L. Rabiner and G. Juang, "Fundamentals of Speech Recognition," Prentice-Hall, 1993.

[3] M. Eunus Ali, "An Approach to Implementation of Bangla Speech Recognition using Hidden Markov Model," A Thesis Paper, Dept. of Computer Science and Engineering, BUET,Dhaka.

[4] J. Tebelskis, "Speech Recognition Using Neural Networks," PhD Dissertation, Carnegie Mellon University, 1995.

[5] H. Hasegawa, M. Inazumi, "Speech Recognition by Dynamic Recurrent Neural Networks,"Proceedings of 1993 International Joint Conference on Neural Networks.

[6] David J. DeFatta, Joseph G. Lucas, William S.Hodgkiss, "Digital Signal Processing: A System Design Approach," John Willy & Sons, Inc.,1998

[7] M.N. Minhaz, M.S. Rahamn and S.M. Rahamn,"Feature Extraction for Speaker Identification,"Int. Conf. on Comp. and Info. Tech., Dhaka,December 18-20, 1998

[8] Rabiner, L.R. and Juang, B.H 1993, Fundamentals of speech recognition, 1st Indian Reprint, Pearson Education.

[9] Qi Li, Zheng, J., Tsai, A.and Zhou, Q. 2002, Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition, IEEE Transactions on speech and audio processing, Vol.10, NO.3.

[10] Tanyer, S.G. and Özer, H. 2000, Voice Activity Detection in Non Stationary Noise, IEEE Transactions on speech and audio processing, Vol. 8, NO. 4.

[11] Qi.Li and Tsai, A. 1999, A language- independent personal voice controller with embedded speaker verification, in Proc. Eurospeech'99, Budapest, Hungary.

[12] L.R. Rabiner and R.W.Schafer, ,"Digital Signal Processing for Man-Machine Communication by Voice" in Digital processing of Speech Signals, 3rd ed.Pearson Education,2009,pp. 505-516

[13] L.R.Rabiner and M.R.Sambur, ,"An Algorithm for Determining the Endpoints of Isolated Utterances"*Bell Syst.Tech.J.,vol.24,no.2,pp.297-315,1975*

[14] B.S.Atal, ,"Effectiveness of linear prediction characteristics of the speech wave fpr automatic speaker identification and verification,"J.Acoust.Soc.Am.,vol.55,no.6,pp.1304-1312,1974

[15] Ahmad A.M.Abushariah,Teddy S. Gunawan,et. al., ,"English Digits Speech Recognition Based on Hidden Markov Models,"*ICCCE,* Kuala Lampur,Malaysia,May 2010.