Proceedings of the 18th International Conference on Automation & Computing, Loughborough University, Leicestershire, UK, 8 September 2012

TESPAR Feature Based Isolated Word Speaker Recognition System

Munaza Sher, Nasir Ahmad, Madiha Sher
Department of Computer Systems Engineering
University of Engineering and Technology
Peshawar, Pakistan
engrmunaza@yahoo.com, n.ahmad@nwfpuet.edu.pk, madiha@nwfpuet.edu.pk

***Abstract*—This paper presents a time domain feature extraction method of speaker identification using Time Encoded Signal Processing and Recognition (TESPAR) approach. TESPAR matrices are not only generated for English words but also for the *Urdu* and *Pashto* words. For classification, the standard Artificial Neural Network (ANN) classifier and its variant have been used. The recognition results obtained show that when the user spoke a word from the vocabulary in an isolated fashion, 99% of the time it is correctly recognized. The results of TESPAR based feature are compared with features extracted using Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coefficients (LPC). The MFCC and LPC features are obtained using the Hidden Markov Model toolkit, HTK. Feed forward neural network with back propagation has been used for the recognition. The results show that the speaker recognition systems with TESPAR features gives better performance with a high recognition rate and low computational complexity as compared with MFCC and LPC based features.**

Keywords-component; speaker recognition; TESPAR features; word recognition; Artificial Neural Network.

I. INTRODUCTION

There are numerous applications of speaker recognition and the use of the appliances performing speaker's recognition is persistently mounting. These techniques make it possible to use the speaker's voice for the verification of their identity and therefore can be used in the security applications to permit the controlled access to the services through voice. Speaker recognition process typically contains a feature extraction stage and a classification stage.

Various techniques for the extraction of useful voice features have been reported in literature. Generally the features are extracted using frequency domain analysis of the speech signal, such as the extraction of MFCC [1] and LPC [2] features. Along with other frequency based feature the pitch of the speaker has also been used to enhance the recognition performance [3]. Unlike the usual frequency based approaches, TESPAR based feature extraction is a purely time domain based extraction of features. The key feature of using TESPAR coding for speech signal is its capability to separate signals that cannot be separated in the frequency domain. Moreover, it has the ability to code time varying speech waveform into optimum configurations for processing with neural network in a parallel architecture way with a low computational complexity [4].

TESPAR and Dynamic time Warping (DTW), have also been used for speaker verification system. DTW is used for the successive alignments of the epochs in order to generate the verification decision. All calculations are performed in time domain and a data reduction of 15 to 20 times is achieved [5].

Average F-Ratio score of the TESPAR feature is another efficient optimization technique used in Automatic Speech Recognition (ASR) in order to achieve the reduced size of the speaker models. Such reduced speaker model result in faster convergence of the Artificial Neural Network (ANN). However, it has been observed that the TESPAR features have greater redundancy as compared with the MFCC [6].

MFCC based features have shown consistently superior performance for speaker recognition as it captures significant information from the audio signal. MFCC are positioned logarithmically in the frequency bands thus approximating the human auditory response system more closely than the frequency bands spaced linearly using Fast Fourier Transform or Discrete Cosine Transform [7].

For classification purposes a number of techniques have been proposed. These can be mainly divided into three models: Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and Artificial Neural Network (ANN). Some hybrid approaches such as DTW/GHMM classifier have also been implemented showing an improvement of 2% to 10% in the recognition [8] [9].

The accuracy of the speaker recognition system also depend on many other factors such as, background noise, the lengths of the voice samples used for the training and testing, whether the recognition is text-dependent or text-independent, and also to a lesser extent on the sampling frequency of the speech waveform [7] .

II. SPEAKER RECOGNITION SYSTEM

The typical model of a speaker recognition system is shown in Fig. 1. The Pre-processing of speech signal involves unraveling the voiced region from the silence/unvoiced segment of the signal. This is because most of the speech or speaker's precise attributes are present in the voiced part of the speech.

After the pre-processing, the next step is the feature extraction, where the speech signal is transformed into some type of parametric representation. Larger number of features requires more computations for classification and more memory space for storage. Thus, the purpose of the feature extraction is to capture the salient aspects of the speaker in a compact set of reduced number of dimensions. In this work, the TESPAR based feature extraction approach as proposed in [4], [10], has been used.

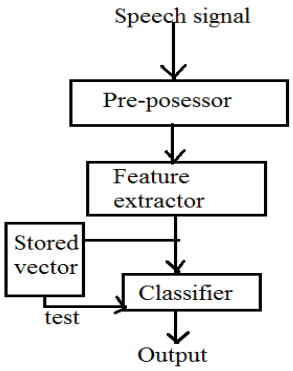

Figure 1. Speaker Recognition Model

The TESPAR approach is based on the zeros theory where the speech waveform is divided into periods identified by the successive zero crossings of the signal. So the time information along with the simple approximation of the waveform between two successive zero crossing is upheld.

The simplest implementation of TESPAR coder uses two descriptors associated to each epoch. An epoch symbolizes a waveform between two successive zero crossings. The duration between two successive zeros in a given sample of the signal is represented by "D" while the shape of the signal between the zero crossings is represented by "S".

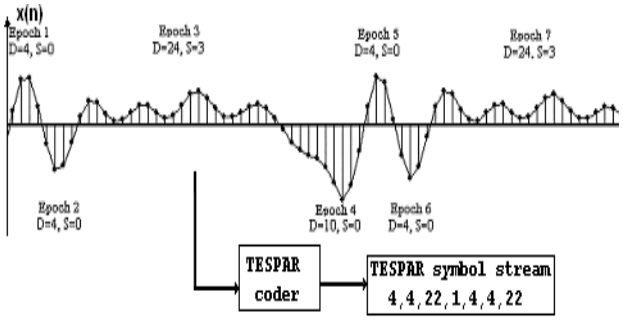

Figure 2: TESPAR Coding

For classification purpose Artificial Neural Network and its variants have been used. Feed-forward back-propagation network is implied. Feed-forward networks used have one hidden layers of sigmoid neurons followed by an output layer of linear neurons. Multiple layers of neurons with nonlinear transfer functions allow the network to learn nonlinear and linear relationships between input and output vectors. The purpose is to constrain the outputs of a network (such as between 0 and 1), so the output layer is using a sigmoid transfer function (such as logsig). For a detailed study on transfer function and neural network refer to [11].

III. SYSTEM OPERATION

A. *Database Used:*

A speaker database was first developed by recording the speech samples from different speakers uttering the selected words several times. On the average 10 speech samples from each of the speaker have been collected. All these speech samples were then loaded into the TESPAR program: coded and converted to TESPAR Alphabets and stored into the database.

B. *Quality Processing:*

The speech samples for each speaker are averaged to generate the mean reference matrix. This averaging procedure is useful in reducing the inconsistency among the speech samples from the same speaker. The next task is to suppress the noise present in the speech waveform. The area of interest in the speech signal, which lies between the first rise and final drop point of the speech waveform, is then found.

C. *Coding process:*

The duration/shape (D/S) pairing of each epoch is used to produce the TESPAR alphabet symbols. A TESPAR codebook comprises of a symbol table of 28 different symbols and is used to map the D/S parameters of each epoch into a single symbol. In most applications, a TESPAR alphabet of 28 different symbols is sufficient enough to represent the original waveform.

D. *Classification process:*

The TESPAR alphabets are used as input to the Neural Network. The TESPAR matrices are ideally matched to the processing requirements of Neural Networks which needs that the training data samples must be of the same sizes. The network receives the 10 samples as a 28-element input vector. It then identifies the speakers by responding with a 10-element output vector. Each of the 10 elements of the output vector represents a speaker. The neural network responds with a 1 in the position of the speaker being presented to the network. All other values in the output vector remains to be 0.

E. *Test setup:*

HTK tool kit has been used for the extraction of LPC and MFCC coefficients while Matlab has been used to extract the TESPAR features. Neural Network toolbox of the Matlab was used for the verification of the results.

Histogram for every spoken word is attained after passing through the TESPAR feature extraction

procedure. Figure 3 shows frequency of occurrence of TESPAR alphabets for the spoken word "CLOCK". Histograms of all the spoken samples are given to neural network as an input for recognition purpose.

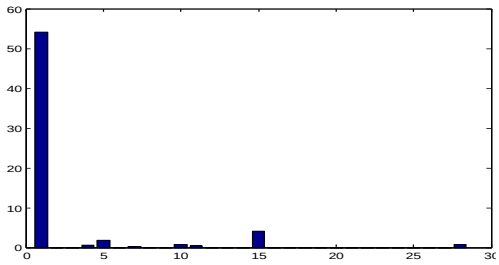

Figure 3: Histogram of speech signal

The Neural Network (NN) used in the model is multilayer perception (MLP) with two layers of neurons, an input layer and a hidden layer. The reiteration of the NN can be stopped in two ways; either by specifying a certain number of iterations or until the time when no sample is left unclassified. The number of epochs in the training phase differs from one example to another. If the number of epochs is set to be high, the NN will saturate or there will be an over fitting of the NN. This case is avoided by setting an appropriate number of epochs. The performance of the neural network is significantly affected by the number of neurons in the hidden layer. The number of neurons in the hidden layer of the NN has been varied from 20 to 150. For 20 hidden layers the result obtained is only 60%. By increasing the number of neurons in hidden layer the performance is also improved. The performance achieved with 100 hidden layers is shown in Figure 4, where the goal is met after only 146 epochs. It is evident from the Figure 5 that when the number of hidden neurons is reduced to 50, the training is achieved after 282 epochs. Moreover the training time is also increased from 4 seconds to 8 seconds.

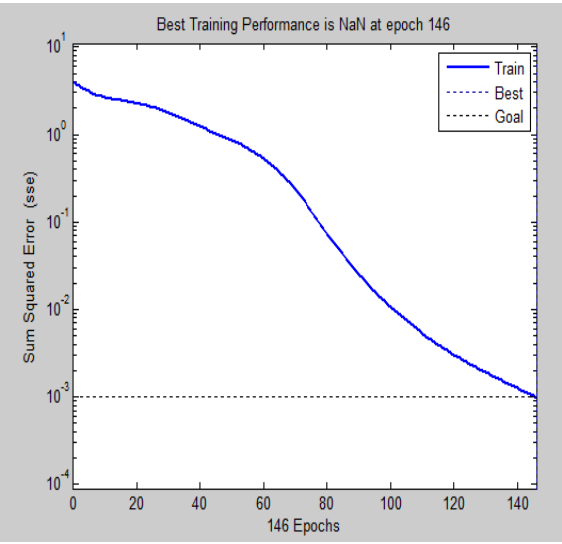

Figure 4: Training performance (example 1)

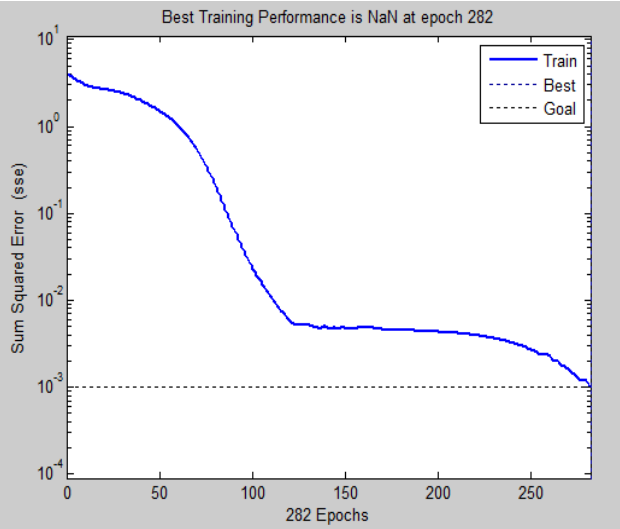

Figure 5: Training performance (example 2)

IV. ASSESSMENT

The Performance of TESPAR based features is compared with that of the LPC based features and MFCC and their first and second derivatives. For acquiring those features, Cambridge University Hidden Markov Model Toolkit (HTK) [12] has been used. HTK tools provide sophisticated facilities for such speech analysis.

MFCC and LPC features are extracted using the procedure mentioned in [1], [2] and the steps performed are shown in Figure 6 and Figure 7, respectively.

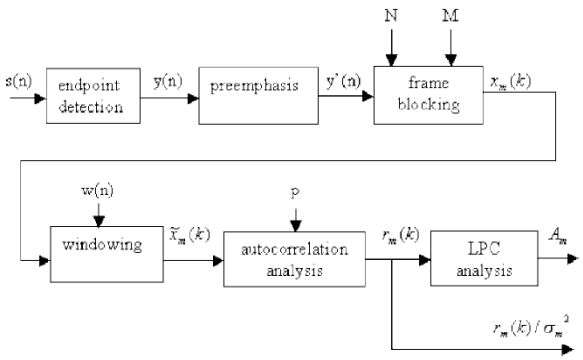

Figure 6: Block diagram of LPC based Feature Extraction

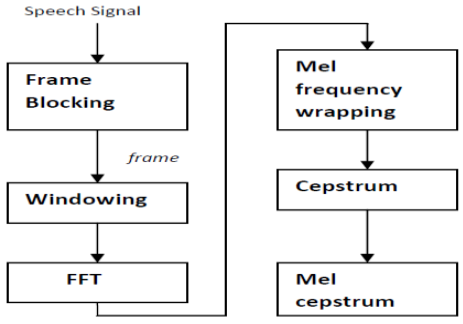

Figure 7: Block diagram of MFCC based Feature Extraction

V. RESULTS AND DISCUSSION

Experiment # 1:

TESPAR matrices are generated for different speakers uttering different words. First the features for English words were extracted but later on the features from the **Urdu** and **Pashto** words were also tested. The recognition performance for all the three languages was almost the same.

For training purpose, the ANN with Feed-forward network was used. TESPAR features from English, Pashto and Urdu words were provided as an input to the neural network. The number of neurons in the hidden layer was varied and the recognition results were obtained. The percentage accuracy abstained for the 10 speaker‘s is shown in Table1. It is evident that the recognition rate improves with the increase in the number of neurons in the hidden layer.

TABLE 1 TESPAR FEATURE BASED RESULTS

Number of hidden layer neurons	Recognition Rate
20	60%
30	70%
50	80%
100	95%
150	**99%**

Experiment # 2:

The same speech signals of the three different languages used for the TESPAR feature were used for attaining the LPC features. The results obtained are shown in Table 2. As can be seen from Table 1 and Table 2, the recognition rate for the LPC features is lower than that of TESPAR features.

TABLE 2 LPC FEATURE BASED RESULTS

Number of hidden layer neurons	Recognition Rate
20	50%
30	70%
50	80%
100	90%
150	**97%**

Experiment # 3

In the third experiment, MFCC features along with delta and delta-delta coefficients are extracted using HTK. The results obtained using MFCC based features are shown in Table 3.

TABLE 3 MFCC FEATURE BASED RESULTS

Number of hidden layer neurons	Recognition Rate
20	40%
30	60%
50	70%
100	85%
150	**95%**

These results shows that for speaker recognition purpose LPC features performed better than MFCC while TESPAR based features gave the best performance among all the three.

VI. CONCLUSION

This approach extracts features from the speech signal in the time domain, so, lower computational requirements are employed as compared with the traditional frequency domain analysis. Successfully testing for URDU and PASHTO speakers, it is shown that the TEAPAR features are language independent. LPC and MFCC based features exhibit low performance especially when the number of speakers increases.

The advantage of using the TESPAR feature is that the size of matrices remains the same irrespective of the duration of the speech signal. From the above results it can be inferred that the smaller size TESPAR metric based features are the most efficient one. Moreover the training time is also the lowest while using the TESPAR based Feature for recognition using ANN. Although the training time of MFCC feature is small compared to LPC, its recognition performance is comparatively low.

REFERENCES

[1] G. S. Kumar, K. A. P. Raju, M. R. Cpvnj, P. Satheesh, (2010), “Speaker Recognition using GMM”, *International Journal of Engineering Science and Technology*, vol. 2, no. 6, pp. 2428-2436

[2] Thiang, S. Wijoy, (2011), “Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot” *International Conference on Information*

[3] S. K. Gaikwad, B. W. Gawali, P. Yannawar (2010), “A Review on Speech Recognition Technique”, *International Journal of Computer Applications*, (0975 – 8887), vol. 10, no.3.

[4] E. Lupu, Z. Feher, P. G. Pop (2003), “On The speaker Verification Using The TESPAR coding Method”, *International Symposium on signals, Circuits and Systems*, July 10-11, 2003, Romania, vol. 1, pp. 173-176.

[5] PETRE G. POP, EUGEN LUPU (2006), “Speaker Verification With Combined Time And Spectral Domain Methods” Proc. of the 8th WSEAS Int. Conf. on Mathematical Methods and Computational Techniques in Electrical Engineering, Bucharest,

[6] K. Anitha Sheela, K. Satya Prasad (2007), “Linear Discriminant Analysis F-Ratio for Optimization of TESPAR & MFCC Features for Speaker Recognition“ Journel of Multimedia, vol. 2, No. 6

[7] S. Singh, E. G. Rajan (2011), “MFCC VQ based Speaker Recognition and its Accuracy Affecting Factors”, *International Journal of Computer Applications*,vol. 21, no.6, pp. 1-6.

[8] E. H. Bourouba, M. Bedda and R. Djemili (2006), “Isolated Words Recognition System Based on Hybrid Approach DTW/GHMM”, Informatica, vol. **30, pp. 373-384.**

[9] A. Lipeika, J. Lipeikien, L Telksnys (2002), “Development of Isolated Word Speech recognition system”, *INFORMATICA*, vol. 13, no. 1, pp. 37–46.

[10] R. A. King, T. C. Phipps (1998), “Shannon,TESPAR and Approximation Strategies”, *ICSPAT 98*, Toronto, Canada, vol. 2, pp. 1204-1212.

[11] S. Hykin. “Neural Networks-a comprehesive foundation” 2nd Edition, Printice Hall. *and Electronics Engineering*, vol. 6, pp. 179-183, IACSIT Press, Singapore.

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland (2006), *The HTK Book V3.4*.