Speaker Identification for Biometric Access Control Using Hybrid Features

Avnish Bora
Associate Prof. Department of ECE, JIET Jodhpur, India
Dr.Jayashri Vajpai
Prof. Department of EE,M.B.M.M Engg. College Jodhpur, India
Gaur Sanjay B.C ,
Associate Prof. Department of ECE, JIET CoE Jodhpur, India

***Abstract*— This paper presents the application based Neural Network for speaker recognition in a voice authenticated access control system in high security applications. The Neural Network designed for this purpose employs hybrid feature extraction techniques for speaker identification. These features include the time domain as well as frequency domain features and are hence named as hybrid features. Speaker dependent features obtained by Linear Predictive Coding (LPC) and Mel Frequency Cepstrum Coefficients (MFCC) have been used in this work. The hybrid features are used for testing the voice authenticated access control. The performance of different features individually and in combination has been analysed by finding recognition efficiency of speaker identification system. The proposed system uses Feed Forward Neural Network classifier. The results obtained from the hybrid features show the improvement in recognition efficiency. The work has been implemented in MATLAB environment. The experiments have been conducted with English language database and Rajasthani Language.**

***Keywords*-** LPC; MFCC; MFLPC; PCA

I. INTRODUCTION

Speaker identification is among the fastest growing smart technologies. Human-machine communication using voice enabled services, with special reference to applications in the areas of medical, industrial robotics, forensic, defence, aviation and e-learning requires the identification of speaker. The speech signal conveys many levels of information of speaker which is encoded in a complex form. Speaker specific information can be extracted by use of the different feature extraction techniques.

Speaker identification is the process of recognizing automatically the particular speaker out of the group of the enrolled speakers, on the basis of the individual's specific information ingrained in speech signal. No two individuals sound identical because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different [1].The process of speaker identification involves two phases training phase and testing phase. During both the phases the input speech signal undergoes the front end processing that include feature extraction. Figure.1 shows the components of speaker identification system. Training phase is the enrollment phase of the system. Speaker's identity is registered with features extracted during training phase and compared with the enrolled identity during testing phase.

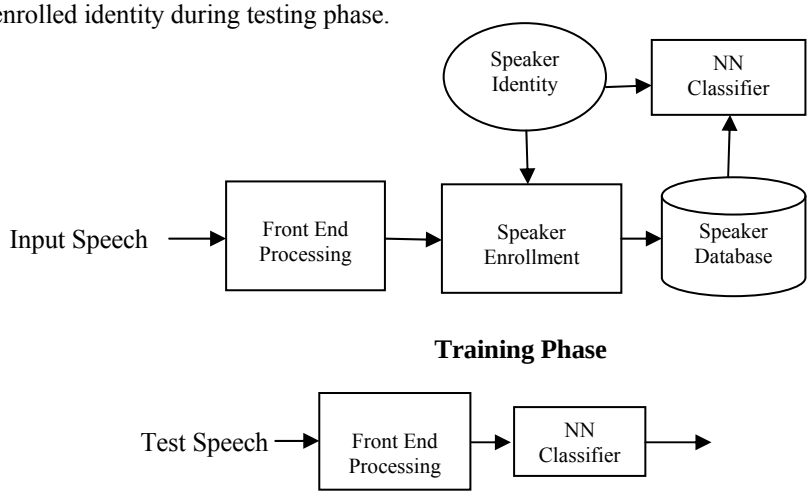

Figure 1 Components of voice authenticated access control system: Training and Testing Phases

A feed forward neural network is used as the classifier in this work. The classifier compares the speaker's identity with the enrolled database, in testing phase and the identified speaker is displayed. Encouraging results have been obtained, which shows the suitability of the proposed technique for access control in voice based security systems.

The paper is organized as follows, section II describes the voice based access control, section III discusses the important feature extraction techniques, section IV presents the methodology used in the work, section V gives the experimental results and discussion , in the final section VI conclusions from the work done are drawn.

II. VOICE BASED ACCESS CONTROL

Voice based access control may be categorised as one of the biometric application of speech processing. Speaker identification under the high security application areas demands high recognition efficiency. A speaker identification system uses the physiological information embedded in speech signal to identify the speaker specific characteristics. Voice based access control may be designed as non contact control mechanism, wherein speaker is not required to be physically present at the location of the security system. This system may be used for highly secured medical database access and may be adapted to provide multilevel multi user access control. These data may be secured by providing access to the authorised speaker only, even if he is located remotely.

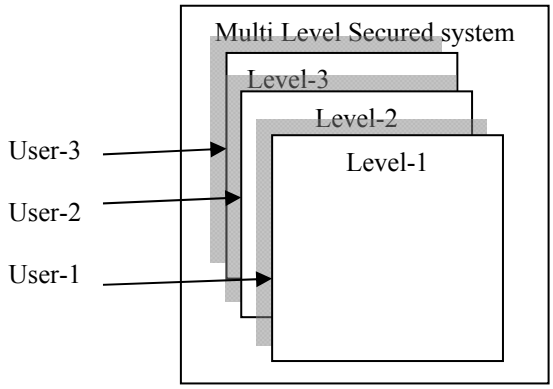

Figure 2 Multiple User Multiple level Access Control

Figure 2 shows the example of multiple level of security in an access control. Here the multiple users are assigned with the voice authenticated access control at multiple levels. This is a great advantage as compared to other biometric applications where sensing device is required to be placed near the user. User which is not enrolled prior will not be able to access the control of levels. This speaker identification system uses the physiological information in speech to identify the speaker specific characteristics.

III. STATE OF THE ART

Speaker identification technology has evolved in past four decades. For speaker recognition various parameters, statistical and predictive were extracted by researchers. That included: average autocorrelation, instantaneous spectra covariance matrix, fundamental frequency histograms, LPC coefficients and long term average spectras. Texas instruments attempted for automated speaker verification system [9].

Due to the mobile nature of speaker identification systems for accessing the different security levels important feature characteristics are required for proper identification and access control [10]. H. Misra et. al. have presented the concept of speaker-specific mapping for the task of speaker recognition. The speaker specific mapping was realized using a multilayer feed forward neural network[11].

O. O. Khalifa et.al. have developed text independent speaker recognition system. Mel Frequency Cepstrum Coefficient (MFCC) feature extraction method was used to extract a speaker‟s discriminative feature from the mathematical representation of the speech signal. Vector quantization was implemented as feature matching method using the LBG Algorithm. Feature matching was carried out in order to cluster the speech features into groups of specific sound classes. Final analysis was carried out to identify parameter values that could be used to increase the accuracy of the system [12]. However 20th century witnessed the advances in feature extraction and pattern matching techniques.

IV. FEATURE EXTRACTION TECHNIQUES

Feature extraction is the fundamental front end processing for identification and meaningful representation of speech signal. The speaker specific speech information is represented in compact form by extracting the

important time and frequency characteristics and information from the speech signals. Feature extraction transforms the input speech signal into a set of features, which provides the relevant information for performing a desired task without the need of a large data set. Speech signal is analysed for a short window of time frame, due to its highly variable nature. A compact representation of the input speech signal is obtained by the various feature extraction techniques, proposed by the researchers. The popular and well established feature extraction techniques used in speaker identification are LPC, MFCC, and their combination as hybrid features.

A. Linear Prediction Coefficients

Linear Prediction Coefficients (LPC's) [2] are used as time domain features for representing the speaker characteristics. These features are derived from the speech production model. Linear prediction provides an estimation of the current sample of a discrete speech signal as a linear combination of several previous samples. By minimizing the sum of the squared differences (over a finite interval) between the actual speech samples and the linearly predicted ones, a unique set of predictor coefficients can be determined.

Speech sample $s(n)$ at time n can be approximated as linear combination of past p speech samples [3].

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p) \quad (1)$$

Here a_p represents the coefficients over the analysis frame.

Speech is modelled as the output of linear, time-varying system excited by either quasi-periodic pulses (during voiced speech), or random noise (during unvoiced speech). The linear prediction method provides a robust, reliable, and accurate method for estimating the parameters that characterize the linear time-varying system representing vocal tract[4],[5].

The major limitations of LPC model is that in many instances, a speech frame cannot be classified as strictly voiced or strictly unvoiced. Furthermore, the use of strictly random noise or a strictly periodic impulse train as excitation does not match practical observations using real speech signals.

B. Mel-Frequency Cepstral Coefficients

The Mel Frequency Cepstral Coefficients (MFCC) [6] have been proposed by S.B. Davis and P. Mermelstein for speech recognition systems, and is based on the fundamental concept of perception of human hearing. MFCC represents the short term power spectrum of speech signal to create voiceprints of the input speech signal.

The MFCC are based on frequency domain signal decomposition with the help of filter bank, which uses the Mel scale. The MFCC are obtained by discrete cosine transform of the real logarithm of the short-term energy expressed on a Mel frequency scale. The Mel Scale uses the relationship with linear frequency f as

$$\text{mel frequency} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (2)$$

The speech signal is broken down into a sequence of frames and each frame undergoes a sinusoidal transform (Fast Fourier Transform) in order to obtain certain parameters which then undergo Mel-scale perceptual weighting and de-correlation. The result is a sequence of feature vectors describing useful logarithmically compressed amplitude and simplified frequency information.

V. METHODOLOGY

The proposed methodology is categorized in two different phases training phase and testing phase. The features obtained are tested by evaluating the recognition efficiency. Recognition Efficiency is defined as

The major steps in the process include preprocessing, feature extraction, classification and Identification.

- Pre Processing

  The speech signal is acquired from the microphone of the system. In the first step of pre-processing, the input speech signal is pre-emphasized to boost the higher frequency components. It was shown in [8] that the absence of high frequency components of speech signal resulted in the loss of performance of recognition system. Figure 3 shows the input speech signal and it's pre-emphasized output. The baseline level is also adjusted by evaluating the mean of the signal samples and subtracting it from the input speech.

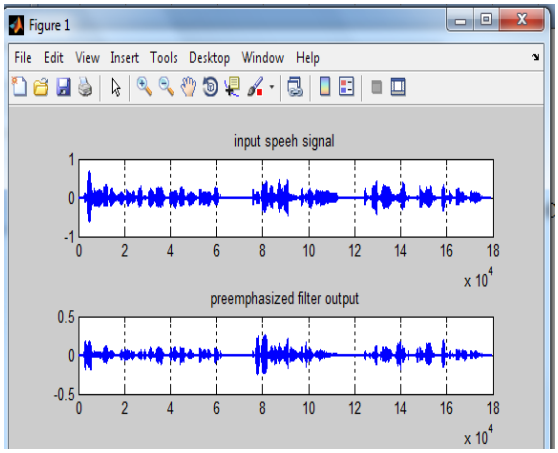

Figure 3 Input Speech Sample and Pre-Emphasized output

- Feature Extraction

  In the next step, spectral and time domain features are extracted from the acquired speech signal. The features extracted in this step are MFCC, LPC. Different combinations and hybridization of these features have been investigated, with an aim to get better recognition efficiency, even when there are multiple speakers. This gives large number of features that form the raw feature vector.

- Dimensionality Reduction

  After being extracted the raw feature vectors are applied for dimension reduction with Principal Component Analysis (PCA) method. Thus converting the large number of feature set in to more compact representation. The reduction gives the principal components of the features. It allows analyzing data and detecting patterns more clearly.

- Classification and Identification

  Speaker and speech is then recognized by employing neural network classifier. Different structures of feed forward neural network were designed and different training functions were explored with an aim to improve the recognition efficiency. In the training phase, a dataset of speakers is prepared along with feature vectors that represent their distinct characteristic features. The speech samples of speakers are collected from the different databases to extract the same sentences from speakers. These samples are used to train the Feed Forward Neural Network model. The sets of training features and corresponding speaker identifiers form the models of different speakers and their collection forms the speaker database. Thus, the speaker database is a repository of the important speaker dependent characteristics, which vary from speaker to speaker and serve as means to discriminate the different speakers.

  In the next phase, which is the identification phase, a test sample from an unknown speaker is compared against the speaker database. Feature Extraction gives the speaker specific information from the given speech signal by performing complex transformations. The acoustic features contain the characteristic information of the speech signal and are useful for recognizing the speaker. The neural network is trained for pattern recognition to match the extracted features with the speaker identity, stored in the database during training. The claimed speaker is identified at the output in test phase for access control.

  Due to the time varying nature of the speech signal it is analysed over a short period of time. It is considered to be quasi stationary signal during 20ms time period. A segmented speech signal for 20ms is shown in Figure 4. This corresponds to the pre-emphasized signal shown in Figure 3.

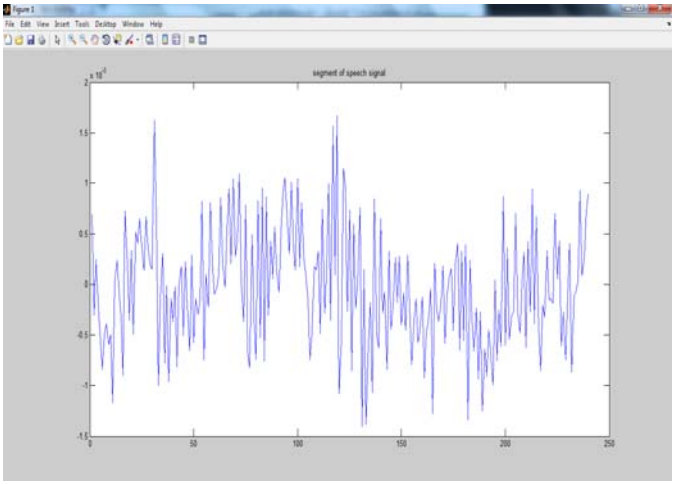

Figure 4: Segmented Speech Signal

In this work different features such as LPC, MFCC and extended hybrid have been used to train the classifier and identify the recognition efficiency for two different languages. Figure 5 and figure 6 shows the LPC and MFCC features of the input speech sample under test.

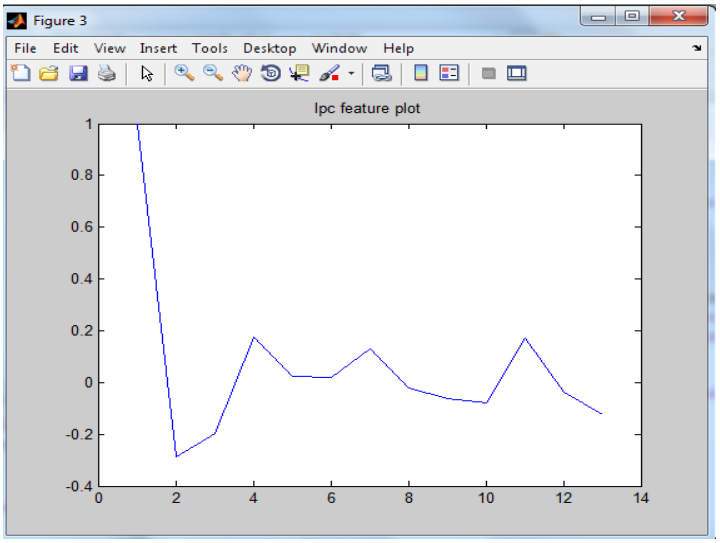

Figure 5: LPC features of input Speech Signal

In the first attempt for the identification of speaker, the neural network classifier was trained with the LPC and MFCC features. The recognition efficiency obtained by these features was quite low. Attempts were made to improve the recognition efficiency of network by experimenting with the different parameters of the network and by using extended hybrid features. The different combinations of feature vectors used for this analysis are Mel Frequency Cepstrum Coefficient- Linear Predictive Coefficient (MFLPC).

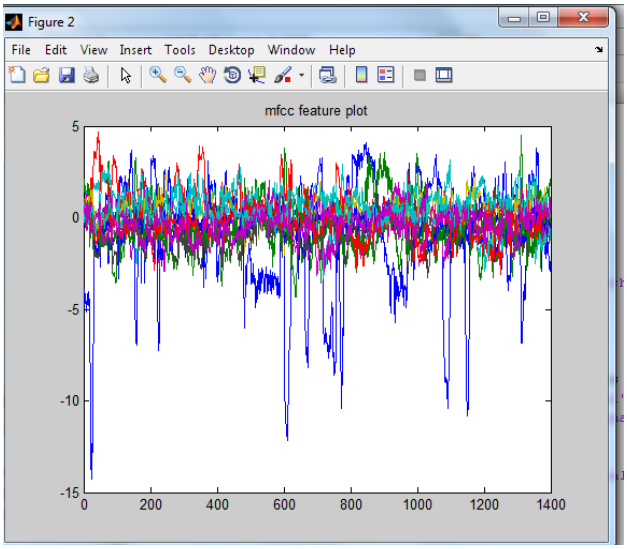

Figure 6: MFCC features of input Speech Signal

The neural network used and regression plot are shown in the figure 7 and figure 8 respectively.

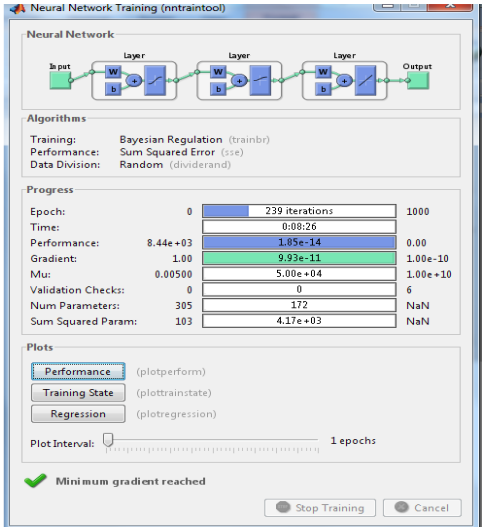

Figure 7 Neural Network with two Hidden Layers

Section VI gives the results of the experiment conducted with the different features and discussion. The recognition efficiency is compared.

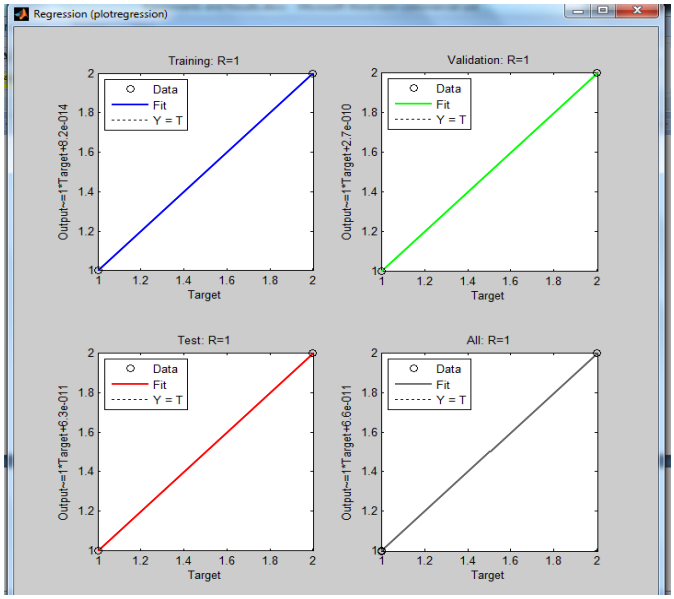

Figure 8 Regression Plot

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The neural network classifier was trained for 154 utterances from ELSDSR database. The effect of extended hybrid features can be observed from the results obtained in table 1. The percentage recognition efficiency has been obtained by using the relation:

$$\frac{\text{Total no. of correctly recognised speakers}}{\text{Total no. of speakers}} \times 100$$

Comparative results of Recognition Efficiency for different features, for two languages are shown in Table 1 and Table2.

Table 1 Recognition Efficiency with Features Used (English)

S. No.	*Features Used*	*Recognition Efficiency*
1	*LPC*	*85%*
2	*MFCC*	*78%*
3	*MFLPC*	*92%*

Table 2 Recognition Efficiency with Features Used (Rajasthani)

S. No.	*Features Used*	*Recognition Efficiency*
1	*LPC*	*80%*
2	*MFCC*	*76%*
3	*MFLPC*	*89%*

Feature extraction is the influential component of pre-processing. It extracts the important speaker specific information from the speech signal. Speaker identification largely depends on the selection of features. Independent features have shown poor efficiency in the experiment conducted, however the hybrid features have improved the recognition efficiency. The results obtained suggest that the use of LPC with MFCC results highest recognition efficiency.

VII. CONCLUSION

It may be concluded that the voice authenticated access control may provide an added security level to the existing biometric technologies. However the practical implementation of these systems may encounter many challenges. As shown in the results in table 2, when implemented for Rajasthani language the recognition efficiency has reduced. This may be due the higher prosody in the language compared to the English language used. Noise robustness of the access control system may be tested with extended hybrid features as a part of future work. One of these is environmental noise.

VIII. REFERENCES

[1] T. Kinnunen, H Li, "An overview of text-independent speaker recognition: From features to supervectors", Elsevier, Speech Communication Vol. 52 pp: 12‑40, 2010

[2] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identifi cation and verifi cation," J. Acoustic. Soc. Amer., vol. 55, p. 1304, 1974

[3] L.Rabiner, B.H.Juang and B.Yegnanarayana, Fundamentals of Speech Recognition Pearson Education, 2009

[4] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals. Englewood Cliffs, New Jersey: Prentice-Hall, 1978.

[5] W. C. CHU, "Speech Coding Algorithms Foundation and Evolution of Standardized Coders", Wiley-Interscience, A John Wiley & Sons, Inc., Publication, 2003

[6] S.B.Davis and P.Mermelstein,"Comparison of Parametric Representations for Monosyllabic Word Recognation" IEEE Transactions on Acoustics, Speech, And Signal Processing, ASSP Vol.-28, No. 4, August 1980

[7] M. H. Farouk , "Application of Wavelets in Speech Processing", Springer Briefs in Electrical and Computer Engineering Speech Technology, 2014

[8] H. Misra, S. Ikbal , B. Yegnanarayana , "Speaker-specific mapping for text-independent speaker recognition", Elsevier Speech Communication 39 (2003)

[9] S.Furui, "50Years of Progress in Speech and Speaker Recognition Research", ECTI Transactions on Computer and Information Technology Vol.1, No. 2 Nov. 2005

[10] Ji Ming_, Timothy J. Hazen, James R. Glass, and Douglas A. Reynolds, "Robust speaker recognition in unknown noisy conditions",

[11] H. Misra, S. Ikbal , B. Yegnanarayana , "Speaker-specific mapping for text-independent speaker recognition", Elsevier Speech Communication 39(2003)

[12] O. O. Khalifa, S. Khan, Md. R. Islam, M. Faizal and D. Dol, "Text Independent Automatic Speaker Recognition", 3rd International Conference on Electrical & Computer Engineering ICECE 2004, 28-30 December2004