Text Dependent Speech based Biometric for Mobile Security

L. Thulasimani
Asst. Professor, Dept. of ECE, PSG College of Technology
Coimbatore-4,Tamilnadu,INDIA

ABSTRACT

Mobile is a powerful data communication media through which confidential information can be exchanged. To communicate to the authorized person and network, biometric is used. In this paper, an efficient speaker recognition technique is proposed to solve the authenticity and security problem for the mobile in noisy environment. An effective feature extraction technique and two different speaker verification technique is used and compared to improve the recognition rate of the speaker in the noisy system for effective communication.

Keywords

Feature extraction, wavelet transform, MFCC, Speaker verification, GMM, FFBNN

1. INTRODUCTION

Speech is natural mode for communication and is highly efficient. Everyone in the society including old citizens and illiterate people can very conveniently use the voice-based interface (unlike keypad based passwords) for any electronic equipment or system. Thus, Speaker Recognition has been an active topic of research for many years. Wireless networks have always been vulnerable to security threats from intruders encroaching upon user data. Biometric verification as such is not foolproof but when used in conjunction with the existing security arrangements, they can provide powerful means of security. Varying levels of security can also be provided according to the need of the application. The choice of Speaker Recognition is due to its inherent robustness, easy adaptability, and dynamic nature [2]. Good use of Speaker Recognition as a biometric can help reduce the vulnerability of wireless communication and can help to protect sensitive information.

2. Literature Survey

Speaker recognition is basically a pattern classification task preceded by feature extraction stage. Given a sequence of feature vectors representing a given test utterance, it is the job of the classifier to find out which speaker has produced this utterance. Several classifiers have been applied to the task of speaker recognition, many of which originate in speech recognition. Early methods of speaker recognition consisted of classifiers like Vector quantization(VQ) [2], Dynamic Time warping[3],hidden Markov Modeling [HMM[4] and Gaussian Mixture Models(GMM)[5]. Vector quantization classifier had success for speaker recognition but ,it suffered from drawback that search for closet centroid within the codebook for large speaker population can be cumbersome. One way to alleviate this problemalleviate this problem is to use tree structured VQ. However, the tradeoff in using tree structured VQ is that the search time will be reduced at the expense of optimal cluster assignment. The technique of DTW performed better for text dependent speaker recognition but since the method of template construction is inherently text dependent, is a drawback of DTW. In [4] D.M. Weber et al. showed that the performance of HMM for speaker recognition was considerably better than VQ. However, if the nature of the speech is not the same as the given sample or the next phoneme in the word depends on more than just the previous state, then the recognition rate drops dramatically. Further training stage is time consuming. Recognition accuracies attained by GMM were quite promising, but for noisy environment, the performance of GMM degrades considerably. Further GMM require several minutes of training, which is not practical for real world applications. Other classifiers are neural networks such as multi layer perceptron, radial basis function network, decision trees such as C4, ID3 and CART. Neural networks learn complex mappings between input and output and are capable of solving much more complicated recognition tasks. They can handle low quality, noisy data. Another advantage of neural network is that they require smaller number of parameters than independent speaker models. Further we need not to train individual model to represent an individual speaker. Rather neural networks are trained to model decision functions which best discriminates the speakers with in a known set [5]. However, optimal neural architecture to solve a particular problem must be selected by trial, which is a drawback. Decision trees have an advantage over neural networks that they have self organizing architectures that do not have to be specified a priori as with neural networks. BayesNet are less sensitive to small data set size and are therefore more suited for environments that change rapidly.

Chularat Tanprasert et al. [6] presented a neural network based text-dependent speaker identification system for Thai language. Linear prediction coefficients were extracted from the speech signal and feature vector was formed. Performance of MLP was compared with Euclidean distance. MLP gave better identification rate. Hassen Seddik et al. in [7] proposed a method of speaker recognition based on formant frequencies position in first voiced speech frame. MLP was used for training and classification. Two classifications methods were used: serial classification and cascade classification. Cascade classification produced better results than serial classification. Mohammad M.Tanabian, Bahram Zahir Azami [8] proposed a method of speaker recognition based on tracking Formant frequency trajectories. They used neural network and CART as classifier and neural network out performed CART with significantly less misclassification rate. R.V Pawar et al.[9] proposed a text dependent speaker identification using neural networks. Linear prediction coefficients were extracted to form feature vector and results showed that system worked fine for identifying a speaker from number of different speakers. D. Gharavian et al. in [10] evaluated the effect of prosodic parameters such as pitch, formants on gender dependent speech recognition. Authors have shown that speech parameters as fundamental frequency, formants and their slopes are gender- dependent and appending these parameters to the feature vector can lead to improved recognition results. T. Lalith Kumar et al. [11] proposed a speaker dependent speech recognition system using MLP and RNN. Linear prediction coefficients were used as feature vector. Accuracy levels obtained during testing reveals that MLP's recognition accuracies are better than RNN. We found that different features have been used to characterize the speaker's voice which includes features like mel frequency cepstral coefficients, linear predictive coefficients,linear predictive cepstral coefficients, linear spectral frequency, formants and pitch [7].

.

3. Our Proposal

In this paper, speaker recognition program is implemented using Matlab [2]. Speaker recognition systems can be characterized as text-dependent or text-independent. The system we have developed is the latter, text-independent, meaning the system can identify the speaker regardless of what is being said.The program will contain two functionalities: A training mode, a recognition mode. The training mode will allow the user to record voice and make a feature model of that voice. The recognition mode will use the information that the user has provided in the training mode and attempt to isolate and identify the speaker.

A robust feature extraction algorithm for speech signals is proposed. This algorithm depends on combining both the wavelet transform and the MFCCs for the feature extraction stage. First, the wavelet transform is applied to decompose the speech signal into two different frequency channels. The components of the low frequency channel are the approximations, while the high frequency channel components are the details. The decomposition process can be iterated with successive approximations being decomposed. Second, for capturing the characteristics of the individual speakers, the MFCCs of the approximations and detail channels are calculated. Based on this mechanism, the multi-resolution features of the speech signal can easily be extracted using the wavelet decomposition and calculating the related coefficients. The proposed technique is used in the feature extraction stage of a text-dependent speaker identification system. GMM and FFBNN are used for the identification or verification stage and compared to know the better recognition rate in noisy environment.

The rest of this paper is organized as follows. Section 3 gives a description of the proposed feature extraction technique and provides detailed description of each constituting part. Section 4 introduces the recognition techniques used. The experiment and the results obtained are given in section 5. Concluding remarks are given in section 6.

4. Description of Implemented Technique

The general architecture of speaker recognition system is given in Fig.1. It is based on feature extraction, speaker modelling and classification. Feature extraction is the first phase carried out to obtain compact and speaker dependant information. After extracting features, we transform these features to create a model for each speaker and store it. In Patten matching or classification, for an unknown speaker, we match the model for the unknown speaker to stored templates. Decision is based on how closely model for an unknown peaker matches with the stored ones.

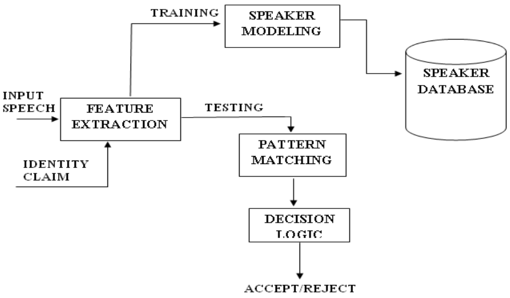

Fig. 1. General Structure of Speaker Recognition

Feature Extraction

The feature extraction technique consists of two phases i.e. use of discrete wavelet transform (DWT) and extraction of Mel-coefficients. Approximation coefficients from DWT of a signal are obtained in order to denoise the speech signal and Mel filters are later used due to their perception of speech which is similar to that of human ear.

4.1 Discrete wavelet Transform

First step is to compute discrete wavelet transform of a signal. The DWT of a signal x is calculated by passing it through a series of filters. First the samples are passed through a low pass filter with impulse response g resulting in a convolution of the two given in (1).

$$y[n] = (x * g)[n] = \sum_{k=-\infty}^{\infty} x[k]g[n-k] \qquad (1)$$

The signal is also decomposed simultaneously using a high pass filter h. The output gives the detail coefficients (from the high-pass filter h) and approximation coefficients (from the low-pass g). These two filters are related to each other and are known as Quadrature Mirror Filters. Approximation coefficients give characteristics of lower frequencies in the signal, whereas details give information about higher frequency characteristics. The Approximation coefficients at each level can be used for another level decomposition (after a down-sampling of 2) and this can be extended to multiple levels in order to get more frequency resolution. This is shown in Fig 2.

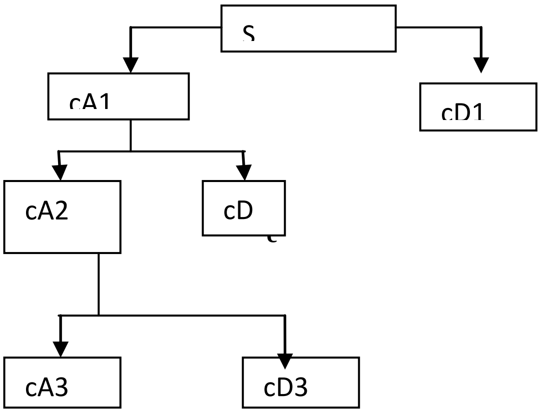

Fig. 2. A 2-level Decomposition

After getting approximation coefficients, we use these coefficients to model the speech signal. Details coefficients contain high frequency signal data which is affected by noise and contains little information about the identity of the speaker as it varies greatly with change in the text spoken and recording/acquisition conditions. Therefore details coefficients are not used in speech signal modelling. An analysis of different wavelets for speaker authentication has been performed as a part of this work. The wavelet level gives optimized results with symlet-7 wavelet or even in some other wavelets in which we can use a simple filtering stage analogous to low pass filtering with a filter based on the symlet-7. This will reduce the computational load of the technique further.

Fig. 3 shows results for increasing number of decomposition levels which were computed up to three levels and experiments show that highest performance is achieved on level 2. This is possibly due to the reason that speech signal loses its characteristics as the decomposition levels are increased.

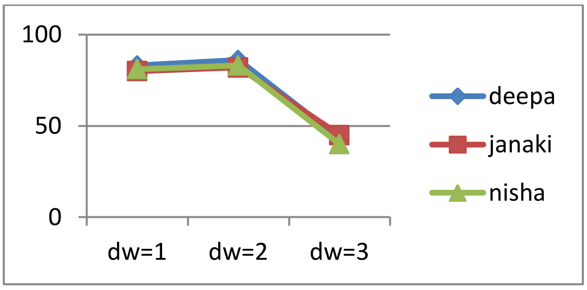

Fig 3. Effect of Decomposition Levels

The purpose of the Mel-bank is to simulate the critical band filters of the hearing mechanism. Mel-Filters emphasize on low frequencies and ignore higher frequencies just like human ear behaves. Fifth step is to take log of the spectrum and compressing it by discrete cosine transform, DCT. The resultant matrices are referred to as Mel-Frequency Cepstrum Coefficients. This spectrum provides a fairly simple but unique representation of the spectral properties of the voice signal. Important property of cepstral coefficients is that they are fairly uncorrelated with each other.Fig. 5 shows that as the

4.2 Mel-Frequency Cepstral Coefficients

Mel-cepstrum is one of the most commonly used feature extraction technique used in both speech and speaker recognition. MFCC technique is based on the known variation of the human ear's critical bandwidth frequencies with filters that are spaced linearly at low frequencies and logarithmically at high frequencies to capture the important characteristics of speech. MFCC is composed of five phases as shown in Fig. 4. First phase is of framing speech signal in order to analyse speech signal in shorter frames due to its non stationary nature. Frame size is 256 in this case. The next step involves windowing of each frame which minimizes the discontinuities at start and end of each frame. Then windowed speech signal is converted from time domain to frequency domain by taking Fast Fourier transform (FFT) which gives insight to frequencies present in that speech signal. Once converted to frequency domain, the signal is passed through Mel-frequency wrapping block.

The purpose of the Mel-bank is to simulate the critical band filters of the hearing mechanism. Mel-Filters emphasize on low frequencies and ignore higher frequencies just like human ear behaves. Fifth step is to take log of the spectrum and compressing it by discrete cosine transform, DCT. The resultant matrices are referred to as Mel-Frequency Cepstrum Coefficients. This spectrum provides a fairly simple but unique representation of the spectral properties of the voice signal. Important property of cepstral coefficients is that they are fairly uncorrelated with each other.Fig. 5 shows that as the number of MFCCs are increased, recognition rate increases rapidly in start and then varies gradually.

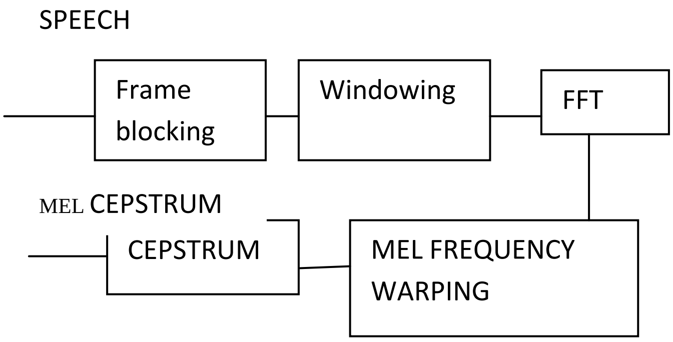

Fig. 4. Steps for Computing MFCCs

number of MFCCs are increased, recognition rate increases rapidly in start and then varies gradually.

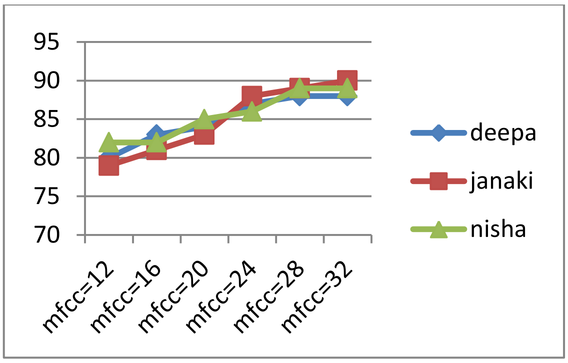

Fig 5. Number of Mel-Filter Cepstral Coefficients

4.3 Wavelet-Based MFCCs Feature Extraction Technique

Speech signals contain two types of information, time and frequency. In time space, sharp variations in signal amplitude are generally the most meaningful features. In the frequency domain, although the dominant frequency channels of speech signals are located in the middle frequency region, different speakers may have different responses in all frequency regions [6]. Thus, the traditional methods which just consider fixed frequency channels may lose some useful information in the feature extraction process. In this paper, the multi resolution decomposing technique using wavelet transform is used. Based on this technique, one can decompose the speech signal into different resolution levels.

The characteristic of multiple frequency channels and any change in the smoothness of the signal can then be detected to perfectly represent the signals. Then, the MFCCs are applied to the wavelet channels to extract features characteristics. MFCCs have the advantage that they can represent sound signals in an efficient way because of the frequency warping property. In this way, the advantages of both techniques are combined in the proposed technique.

5. Recognition Techniques

In speaker identification, the goal is to design a system that minimizes the probability of identification errors. Thus, the objective is to discriminate between the given speaker and all other speakers. This is done by computing a match score. This score is a measure of the similarity between the input feature vectors and some model.

Here two different models are used. Gaussian Mixture Model (GMM) and Feed Forward Back Propagation Neural Network (FFBNN).

5.1 Gaussian Mixture Model (GMM)

GMM as described in this paper is referred from [7]. A Gaussian mixture density is a weighted sum of M component densities given by equation (2)

$$p\left(\frac{\vec{x}}{\lambda}\right) = \sum_{i=1}^{M} P_i\, b_i(\vec{x}) \qquad (2)$$

Where $\vec{x}$ is a D-dimensional random vector,

$b_i(\vec{x})$ are the component densities and P_i are the mixture weights. Component density is a D-variate Gaussian function of the form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} \lfloor\Sigma_i\rfloor^{1/2}} exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^t {\Sigma_i}^{-1}(\vec{x}-\vec{\mu}_i)\right\} \quad (3)$$

with mean vector $\vec{\mu_i}$ and covariance matrix Σ_i . The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities as shown in (4).

$$\lambda = \{P_i, \vec{\mu_i}, \Sigma_i\}, i = 1, \ldots M \qquad (4)$$

λ refers to each speaker in the GMM. In this paper, the GMM model has one covariance matrix per Gaussian component which is called the “nodal covariance”.

speaker	1	2	3	4	5
1	**2.579**	- 328.544	- 120.139	- 302.883	- 501.963
2	- 279.523	**1.2415**	- 132.812	- 247.714	-83.208
3	-51.997	-23.699	**2.3190**	-46.279	-37.064
4	-80.729	- 755.607	- 499.695	**2.3726**	- 245.351
5	-64.843	-55.313	-34.257	-86.724	**1.5638**

Fig 6. Covariance matrix of 5 speakers

5.2.FFBNN

Feed Forward Back Propagation Neural Network (FFBNN) is the most widely used architecture. It is very popular technique that is relatively easy to implement. It requires large amount of training data for conditioning the network before using it for predicting the outcome. A back-propagation network includes at-least one hidden layer. The approach is considered as “feed-forward/ back propagation” approach. The network is created by fixing the number of input layers, hidden layers, and output layers. By giving the input vectors and fixing the target vector,network is created and trained.

After that network simulates the network outputs (the weights and the biases) with each model stored in the system and then error rate and recognition rate is calculated between imposter and model.

6. EXPERIMENTAL RESULTS

Our experiments were conducted with 10 speakers. For each speaker 5 signals of English word "congratulations" are recorded. 3 of the signals are used for training the system and the other 2 signals used for testing the system. Recognition rate is calculated for each speaker.

Two recognition techniques are compared with different noise level and with their recognition rate is calculated for each speakers and from that we can conclude best recognition technique for mobile security. Fig 7 shows the recognition rate using neural network for different sample of same speaker with different decomposition level. Fig 8 shows the recognition rate using neural network compared between two different speakers Fig 9 shows the recognition rate using neural network where different level of noise is added to the speakers. Fig 10 shows the recognition rate using neural network compared with different coefficients level Fig 11 shows the recognition rate using neural network compared with different decomposition level of same SNR

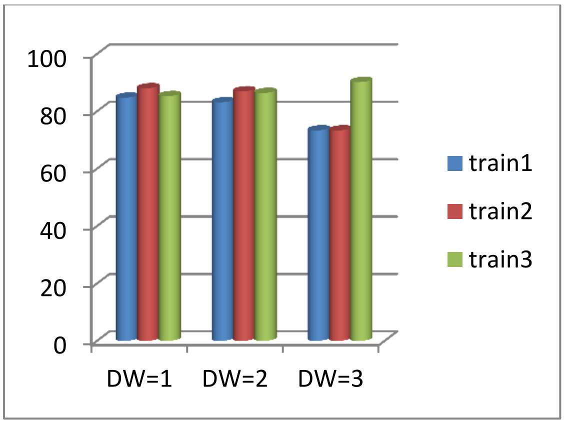

Fig 7. Recognition rate compared with different decomposition level of same speaker

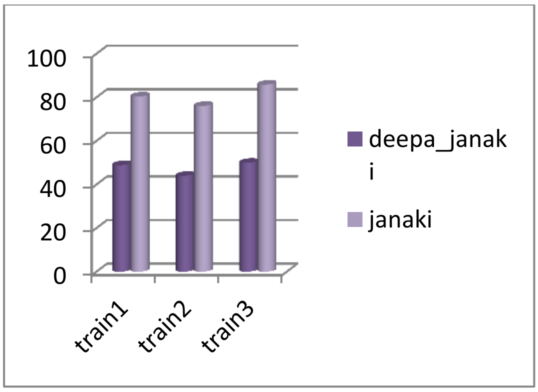

Fig 8. Recognition rate compared between different speakers

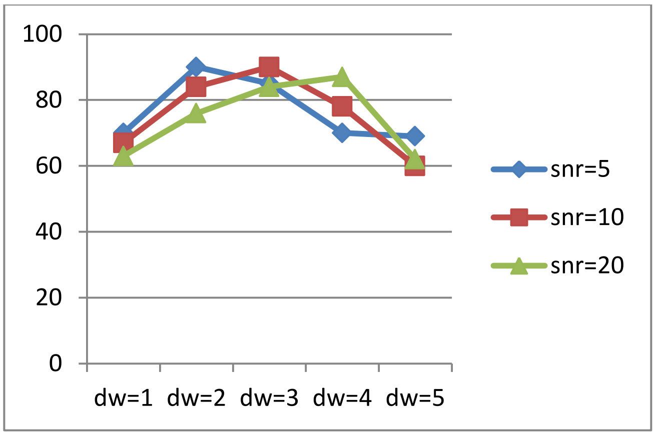

Fig 9. Recognition rate with different SNR and decomposition level

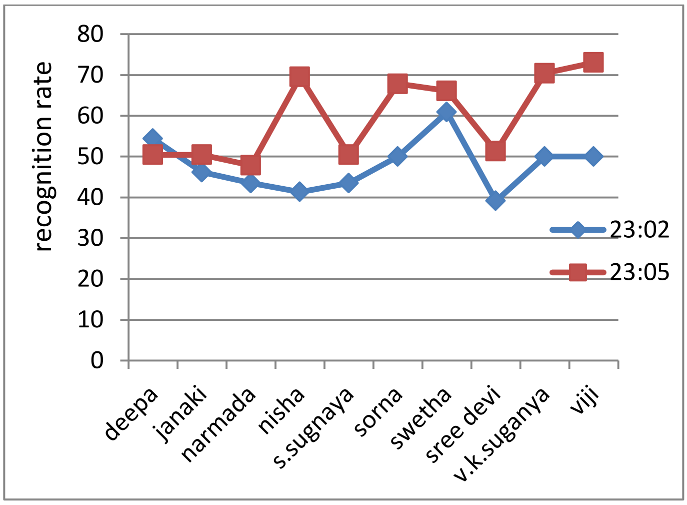

Fig 10. Recognition rate with 5 db SNR and with different coeffecients level

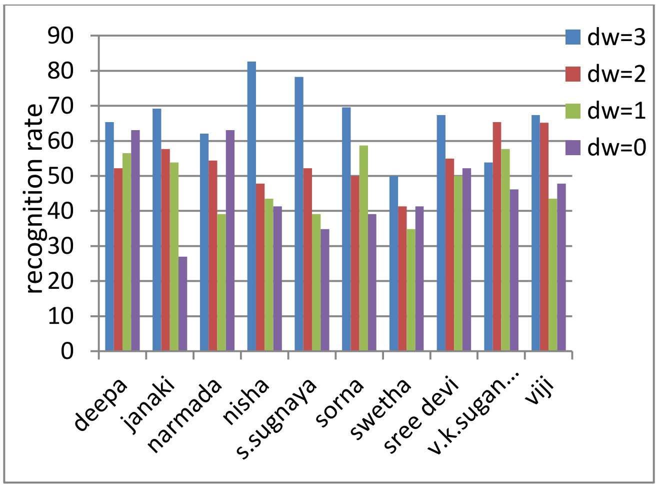

Fig 11. Recognition rate with 5 db SNR and with different decomposition level

7. CONCLUSION

In this paper efficient speaker recognition system is proposed for mobile security to solve the authentication problem. The advantage of using speech biometric is that it does not require any extra hardware to be implemented. The effective feature extraction technique is implemented successfully which gains better performance in the system where the noise level is high. Feed Forward Back Propagation Neural Network (FFBNN) verifies better than Gaussian Mixture Model (GMM). GMM is more complex compare to FFBNN if there is increase in feature size more. Time taken to calculate is also high compare to FFBNN. By applying DWT, GMM recognize better but increase in covariance value cannot be limited moreover, the system cannot tolerate where the noise is high. Thus, FFBNN performs better and faster than GMM.

8. REFERENCES

[1] Dennis C. Tanner, Matthew E. Tanner, “Forensic aspects of speech patterns: voice prints, speaker profiling, lie and intoxication detection”, Learn How publishing, 2004.

[2] Kevin R. Farrell et al., “Speaker recognition using neural networks and conventional classifiers”, IEEE Transactions on Speech and AudioProcessing, vol. 2, Jan 1994, pp. 194-204.

[3] K. Yu et al., “Speaker recognition using hidden Markov models, dynamic time warping, and vector quantization”, in Proc. Image Signal Process, vol. 142, Oct 1995, pp. 313-318.

[4] D.M. Weber et al., “A comparison between hidden Markov models and vector quantization for speech independent speaker recognition”,IEEE Comsig, 1993, pp. 139-144.

[5] Douglas A. Reynolds et al., “Robust text-independent speaker identification using Gaussian mixture speaker models”, IEEE Transactions on Speech and Audio Processing, vol.3, Jan 1995, pp. 72-83.

[6] Chularat Tanprasert et al., “Text-dependent speaker identification using neural network on distinctive Thai tone marks”, Technical Journal, vol.1, Jan-Feb 2000, pp. 249-253.

[7] Hassen Seddik et al., “Text independent speaker recognition based on attack state formants and neural network classification”, IEEEInternational Conference on Industrial Technology, 2004, pp. 1649-1653.

[8] Mohammad M. Tanabian et al., “Automatic speaker recognition with formant trajectory using CART and neural networks”, IEEE, May 2005, pp. 1225-1228.

[9] R.V Pawar et al., “Speaker identification using neural networks”, World Academy of Science, Engineering and Technology, 2005, pp. 31-35.

[10] D. Gharavian et al., “Statistical evaluation of the effect of gender on prosodic parameters and their influence on gender dependent speech recognition”, IEEE ICICS, 2007.

[11] T. Lalith et al., “Speech recognition using neural networks”, IEEE International Conference on Signal Processing Systems, 2009, pp. 248-252.