International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-3 Issue-4, April 2014

Comparison of Two Speaker Recognition Systems

A. Vaishnavi, B.Chanakya Raju, G. Prathiksha, L. Harshitha Reddy, C. Santhosh Kumar

Abstract – This paper presents a comparison between two speaker recognition systems. One system uses 30 Shannon entropy values extracted from a four level wavelet packet decomposition method in addition to the first three formant frequencies as features and a cascaded feed forward back propagation neural network is used as classifier. The second system uses Mel frequency cepstral coefficients (MFCC) as features and a support vector machine (SVM) as classifier. Results suggest that wavelet based system has better performance than the classic MFCCs with an efficiency of 89.56%.

Index terms – Shannon entropy, Formant frequencies, cascaded neural network, MFCC, SVM.

I. INTRODUCTION

Speaker recognition has gained more importance in recent years due to the wide range of applications in the fields of banking, forensics etc. Speaker recognition refers to the process of recognizing a person's identity through his/her spoken utterance [1]. The process of speaker recognition comprises three steps: feature extraction, speaker modeling, and decision making using pattern classification methods. The feature extraction involves extracting certain characteristics from the speech utterance of each speaker. Speaker modelling refers to the training of speaker models for the target speakers and finally in the recognition phase, the features extracted from the test speech segments are compared with the target speaker models using pattern recognition methods. There are many factors that determine the efficiency of the system. Important among them is the type of features selected. The features selected must be robust to noise and distortion, must occur frequently and naturally in the speech, and must be difficult to mimic or impersonate. Choosing a feature like this prevents the chance of false acceptance and false rejection [1].

Manuscript published on 30 April 2014.
***** Correspondence Author (s)
A.Vaishnavi, Department of Electronics and Communication, Amrita Vishwa Vidyapeetham, Coimbatore, India.
B.Chanakya Raju, Department of Electronics and Communication, Amrita Vishwa Vidyapeetham, Coimbatore, India.
G.Prathiksha, Department of Electronics and Communication, Amrita Vishwa Vidyapeetham, Coimbatore, India.
L.Harshitha Reddy, Department of Electronics and Communication, Amrita Vishwa Vidyapeetham, Coimbatore, India.
C.Santosh Kumar, Department of Electronics and Communication, Amrita Vishwa Vidyapeetham, Coimbatore, India.

Short-term spectral features are the most popular and easily obtained features. In this method, the speech is divided into several short duration frames of 20-30 milliseconds duration and each frame is analyzed. Adjacent frames overlap usually by 10 milliseconds to ensure that there is no abrupt change in the feature characteristics, and also ensure continuity in the characteristics of the features in adjacent frames. Other high level features like pitch and prosody can give better results, but their extraction requires highly complex front end system [1]. An ideal feature doesn't exist and the type of feature chosen depends on the purpose. For example, in military applications, there cannot be any mistaken identities. So a complex system should be used there.

If different feature types are to be discussed, the basic and well-known feature would be Fast Fourier transform (FFT) features. This transform helps us to analyze the signal in the frequency domain. Only magnitude spectrum is considered, as the phase plot will not hold much perceptual information. It is also possible to use time domain features, though it is not very popular for speaker recognition applications. Both time and frequency information of a signal at a particular instant cannot be known. To overcome this, Mel frequency cepstral coefficients are introduced. MFCCs are proven to be the best features in speech recognition. Their extraction involves the decomposition of the signal using filter banks and extracting the DCT coefficients of the FFT spectrum. But the resolution to be used during the time to frequency domain decomposition is fixed and follow the Mel scale [7]. This may lead to loss of information in the high frequency regions of the signal, as in the case of human perception. An alternative to the MFFC is wavelet features. Wavelet features give information regarding both time and frequency simultaneously. Wavelet packet transform helps us in analyzing the signal in different bands of frequencies with flexible resolution, and the wavelet decomposition also can reflect the human perception of different frequency bands [2].

II. SYSTEM DESCRIPTION:

In this paper there are two speaker recognition systems of which one system is neural network based system with wavelet packet features, where 30 shannon entropy values and three formant frequencies are taken as features and the other is UBM-SVM based system using MFCCs as features.

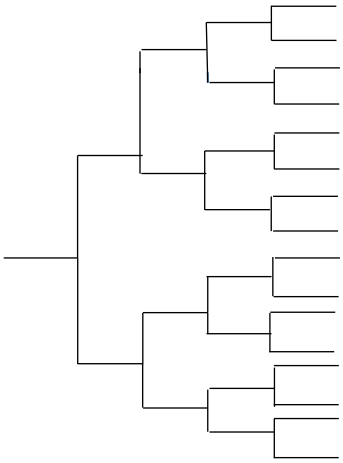

Fig.1. 30 sub-band Wavelet Packet Tree

A. Neural network system using wavelet packet features:

Wavelet Packet Transform can be viewed as a tree shown in Fig.1.The root of the tree is the original signal. The next level of the tree is the resultant of one step of wavelet transform(i.e., multiplying the signal by mother wavelet).Subsequent levels in the tree are constructed recursively by applying the wavelet transform set to low and high pass filter results of the previous wavelet transform done[5]. The advantage of using wavelet transform is that the signal can be analyzed at multiple levels of frequency resolution.

The pre-processed signal is decomposed to depth 4 by wavelet packet transform by using Daubechies mother wavelet (dB1). This results in 30 frequency bands. Shannon entropy is then calculated from each of these 30 bands by "(1)".

$$E_i(s) = -\sum s_i^2 \log s_i^2 \quad (1)$$

Where s_i is the signal at each level.

These 30 entropy values are taken as features for our work.

Production of sound involves the vocal folds vibrating in a periodic manner. The signal produced by this oscillation modulates the resonator system of the vocal tract. While harmonics near the resonant frequency are boosted, the other harmonics are attenuated. These vocal tract resonances are referred to as formants. Since formant features are distinguishable for each person they can be effectively used for speaker recognition [1].

The formant frequencies are found out by calculating Power Spectral Density (PSD) by Yule-Walker Auto-Regressive (AR) method. This method fits the AR linear prediction filter model to the signal by minimizing the forward prediction error in the least squares [2]. The spectral estimate returned by this method has magnitude square of the AR estimate. The speaker's PSD is represented in logarithmic scale (F) as given in "(2)".

$$F_{P_{XX}}(i) = \sum_{i=1}^{M} 10 \log_{10}(P_{XX}(i)) \quad (2)$$

Table i. Optimized parameters of Neural Network

Functions	Description
Neural Network type	Cascade Feed Forward Back Propagation
No. of layers	1 input, 2 hidden and 1 output
No of neurons per layer	33 in input, 40 in each hidden and 1 in output
Weight function	DOTPROD
Training function	Levenberg Marquardtback-propagation (trainlm)
Activation function	Log-sigmoid
Performance function	MSE (10e-06)
No. of epochs	300

The first three local maxima frequencies of the spectral estimate are taken as the formant frequency values and given as features to the neural network.

Recently there has been a great deal of interest in the application of artificial neural nets (ANNs) to the problems of pattern and speech recognition. In these particular areas, ANNs have outperformed well and this motivated to investigate the application of ANNs to the speaker recognition task.

Initially during training, feature vectors are given to the ANN system which assumes some random weights, computes the output based on this weights, compares it with the target values and tries to minimize the mean square error by adjusting the weights through back propagation [2].The Back propagation algorithm allows multilayer feed forward neural networks to learn input/output mappings from training samples. Back propagation network adapts itself to learn the relationship between the set of training speech, by means of error propagation and could be able to apply the same relationship to the test input. The output is obtained by the dot product of both input and weights. The output may take a large range of values. So to normalize the data we use activation function.

All the optimized parameters tabulated in Table i., were selected after several experiments, such as using different number of hidden layers, the activation functions, and the size of hidden layers (number of neurons) for best performance.

B. UBM-SVM system using MFCC:

After the extraction of Mel frequency cepstral coefficients (MFCC), we use UBM-SVM system for the classification of the speaker recognition system. Fig.2 shows the steps involved in UBM-SVM speaker recognition system.

A universal background model (UBM) is a speaker-independent model. It represents speaker independent distribution of the feature vectors used to form the model. It is trained with a huge amount of speech data from the set of speakers. When a speaker enrols into the system, the UBM is updated with speaker-dependent features from the new speaker [6]. UBM is a GMM-based model; it acts as a large GMM, composed of large number of mixtures. The method is to first select a speaker-specific trained model, then determining a likelihood ratio of the match score of a test speech sample with the trained model and the universal background model. The mean parameters of this speaker independent UBM model are then Maximum a Posteriori (MAP) adapted with the extracted MFCC features of each utterance individually. MAP is a model adaptation technique that maximizes the posterior probability of the adaptation data given the model parameter [7]. MAP can be used to determine the parameters of the speaker model that maximize the likelihood according to "(3)".

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

$$\lambda_{ML} = argmax\ (p(x/\lambda)) \quad (3)$$

Where x is the utterance from the speaker and λ denotes the parameter for the speaker model.

The speaker independent UBM model is trained on a large set of speech signals from different speakers. The MAP algorithm is then used to determine a better estimate of λ, from the old estimate λ and the process is repeated until a convergence criterion is satisfied.

From the MAP adapted models of each utterance, the modified mean values of all the Gaussians are concatenated to obtain the GMM supervectors. These supervectors and their corresponding labels of the training data are used to train the SVM model.

SVM uses hyper planes to classify the data into two classes. The hyper plane is represented as in "(4)".

$$x.w + b = 0 \quad (4)$$

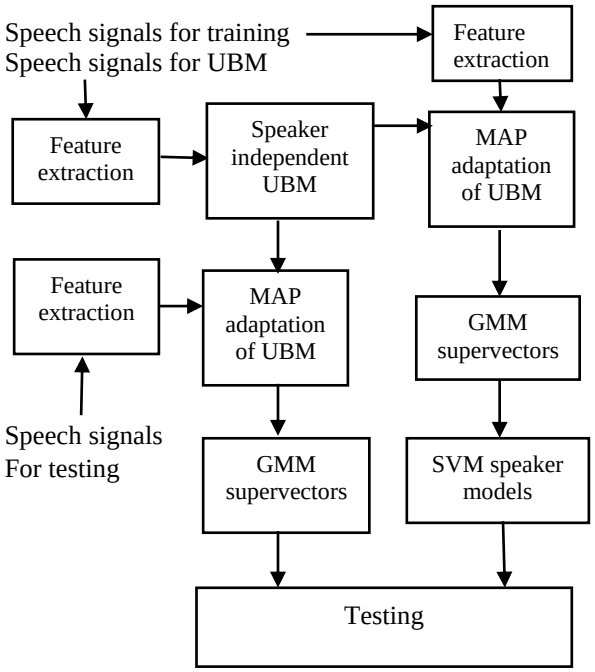

Fig.2. Development of UBM-SVM system

Type of Feature	Performance (%)
MFCC	81.33
WPT	89.56

Table ii. Performance measures

Where x is the training data, w is normal to the hyper plane. The criterion to obtain an optimum hyper plane is to maximize the normal distance between two classes. SVM can be classified as linear and non-linear based on the data. If the data is linearly separable linear SVM will be used non-linear kernels can be used when the data cannot be linearly separable. Later the GMM super vectors are extracted from the MAP adapted models of each utterance by concatenating the modified mean values of all the Gaussians. Now, the supervectors and their corresponding labels of the training data are used to train the Support Vector Machine (SVM) model. This SVM model is then used to classify the speakers using the test super vectors.

III. EXPERIMENTS AND RESULTS:

Speech signals from 20 different speakers are taken, where each speaker may have two or more utterances. The number of speech utterances available for both training and testing is 128 in total, Out of which 42 speech signals are used for training and 86 for testing.

The speech signals are pre-processed for silence removal and amplitude normalization in order to make them comparable with the others.

In wavelet based system, pre-processed signals are used to obtain 33 wavelet features. 128*33 feature matrix is obtained from feature extraction stage and is given to the feed forward neural net with all the specifications given in Table i. for training and testing. While extracting MFCC features the signals are segmented to frames of size 20-30 milliseconds with an overlap period of 10ms. The universal background model (UBM) is created using the MFCC features. UBM is then trained using the train speech signals, creating models for each speaker. Here 64 Gaussians are used in UBM. These speaker independent UBM models are MAP adapted. GMM supervectors are obtained by the mean vectors of the MAP adapted GMM. These GMM supervectors are given as the input to the SVM classifier for training and for classification.

From Table ii. we observe that Neural network system using wavelet packet features has outperformed UBM-SVM system using MFCC.

IV. CONCLUSION:

In MFCC based system, generally the frequency resolution is high in the lower frequency bands and gets considerably coarser in the higher frequency bands. This structure has worked very well for speech recognition but the need of speaker recognition might be different. In wavelet packet decomposition method, information from the frequency bands which are more important for speaker discrimination can be obtained, resulting in better performance over MFCC method.

REFERENCES:

1. T. Kinnunen and H. Li, "An overview of text-independent speaker recognition from features to super vectors," Speech communication.vol. 52, pp. 12-40. 2010.
2. K. Daqrouq, "Wavelet entropy and neural network for text-independent speaker identification," Engineering Applications of Artificial Intelligence., vol. 24, pp. 796-802. 2011.
3. K. Daqrouq, T. Abu Hilal, M. Sherif, S. El-Hajjar and A. Al-Quawasmi, "Speaker identification system using wavelet transform and neural network," Advances in Computational Tools for Engineering Applications (ACTEA)., ZoukMosbeh, Lebanon, pp. 559-564. Jul. 2009.
4. R. Sarikaya, J.H.L. Hansen and L. Bryan, "Wavelet transform features with application to speaker identification," in Proc. of IEEE Nordic Signal Processing Symp., Visgo, pp. 81-84. 1998.
5. M. Siafarikas, T. Ganchev and N. Fakotakis, "Objective wavelet packet features for speaker verification," in Proc. Of the InterSpeech-2004-ICSLP., Jeju, Korea, pp. 2365-2368. Oct. 2004.
6. Kuruvachan K. George, Arunraj K and Sreekumar K.T, "Towards Improving the Performance of Text/Language Independent Speaker Recognition Systems."
7. D. A. Reynolds, Thomas F. Quatieri and Robert B. Dunn,"Speaker verification using Adaptive Gaussian Mixture Models," in Digital Signal Processing Vol. 10.Nos. 1-3,January 2000.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)