WAVELETS FOR SPEAKER RECOGNITION USING GMM CLASSIFIER

Keerthi Anand V D,
Department of Electronics and Communication Engineering,
SNS College of Technology,
Coimbatore, India
keerthianandkm11@gmail.com

Abstract: Speaker recognition plays an important role in a biometric based identification of the person using the information available in their speech signals. In any speaker recognition system, feature extraction using signal processing approaches is an important stage. In this paper, an efficient speaker recognition system is presented by extracting the energy features of the speech signals using Discrete Wavelet Transform (DWT). Then, the extracted DWT energy features are modeled using Gaussian mixture model (GMM) classifier for the recognition of the speaker. Results prove the efficiency of the speaker recognition system with an accuracy of 96.31% at 4th level DWT features with 16 Gaussian densities.

Keywords: Speaker Recognition, Speech Signal, DWT, GMM Classifier.

I. INTRODUCTION

Speaker recognition is widely used in telephone based applications. There are many methods developed in the last two decades. A speaker recognition system that uses the modified Mel-Frequency Cepstral Coefficients (MFCC) method is discussed in [1]. It includes Blackman windowing which is based on different classifiers like back-propagation neural network, Euclidean distance, and self-organizing map. The keyword choice effect that includes and excludes the plosive sounds are investigated in [2] on isolated speaker recognition system. The system employs MFCC and Dynamic Time Warping (DTW) for feature extraction and time equalization respectively.

A method of recently developed Deep Neural Network (DNN) model called as time delay DNN that is used for the large-vocabulary continuous speech recognition tasks is discussed in [3]. It investigates a lightweight change in which a supervised GMM is obtained using time delay DNN posteriors. A deep learning process to derive speaker identifies d-vector by a DNN is discussed in [4]. It discusses two schemes for the deep learning process: a scoring method based on DTW and a phone dependent DNN structure to normalize phone variation.

An enhanced vector quantization algorithm is discussed in [5] for speaker recognition where the classification is attained by the results of two set of codebooks. They are constructed by the use of a portion of words and the whole words respectively. Two divergent speaker recognition systems such as i-vector system and GMM with a universal background model are compared using the performance utterance duration variability and are discussed in [6].

A method to recognize the Bangla words is discussed in [7]. It also recognizes the speaker that uses a semantic modular time delay DNN. The acoustic fuzziness of human utterance disturbances is reduced by using the Fuzzy C Means clustering method. The Bangla words and speaker are recognized by MFCC. The utilization of MFCC and Shifted MFCC features are discussed in [8]

for speaker recognition. To enhance the execution at a high recurrence area, fuzzy demonstrating strategies and vector quantization are used.

Two core objectives of the speaker recognition system are discussed in [9]. (i) By the back-end algorithm and ii) By modifying the intrinsic and extrinsic back-end algorithm. The former one is more robust back-ends, and multi-session enrollment data is used to address the noisy environment for speaker recognition. The later one is a highly discriminative speaker verification system. Kernel matching pursuit algorithm and a particle swarm optimization based on the chaotic scheme of speaker recognition method is discussed in [10]. Streamlining MFCC feature parameters is used to transform the MFCC parameters at first, and then the kernel matching pursuit algorithm is used for the reduction of the streamlining MFCC features. GMM is used for recognition.

In [11], for the observation of the effect of the type of text based on the speaker recognition performance, the training, and testing of various types of texts are done by the developed automated speaker recognition system. The various types of text used are words, digits, sentences, and paragraphs. Genetic algorithms based feature selection method for speaker recognition is explained in [12]. It is compared to two well-known reduction scheme of dimensionality called as linear discriminant and principal component analysis.

II. METHODS AND MATERIALS

Figure 1 illustrates the proposed speaker recognition system using DWT+GMM. It consists of three modules; preprocessing, feature extraction, and recognition. All these modules are explained in the following sub-sections.

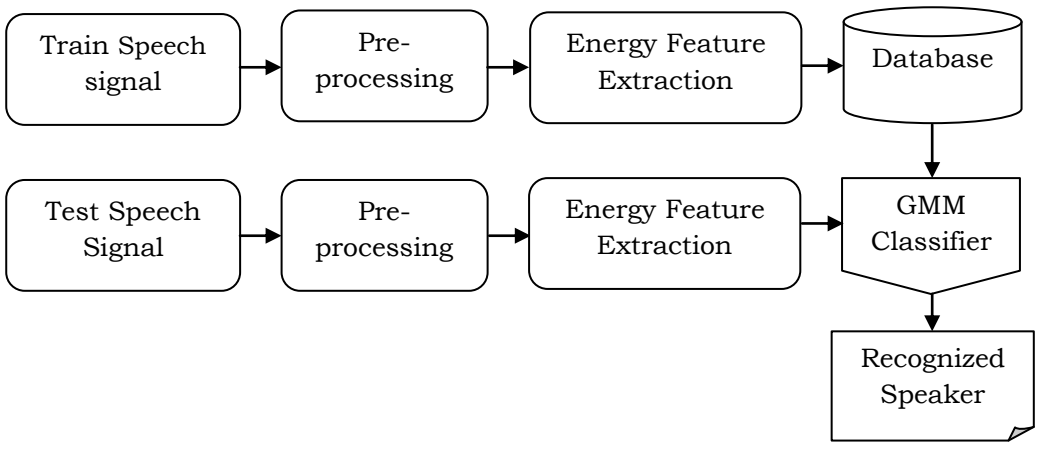

Fig. 1 Speaker recognition system using DWT+GMM

A. Pre-Processing

At first, pre emphasis filter is applied to the input speech signal. In this study, pre emphasis coefficient is set to 0.97. Then, a band pass filter is applied to extract the information between two frequencies. In this study, the speech signal of 4000-8000Hz is filtered using 6th order Butterworth band pass filter. Finally, all the speech signals are sampled with analysis frame duration of 25 milliseconds and analysis frame shift of 10 milliseconds. Features are extracted from all the

frames of a speech signal. Figure 2 shows the preprocessing steps of speaker recognition system.

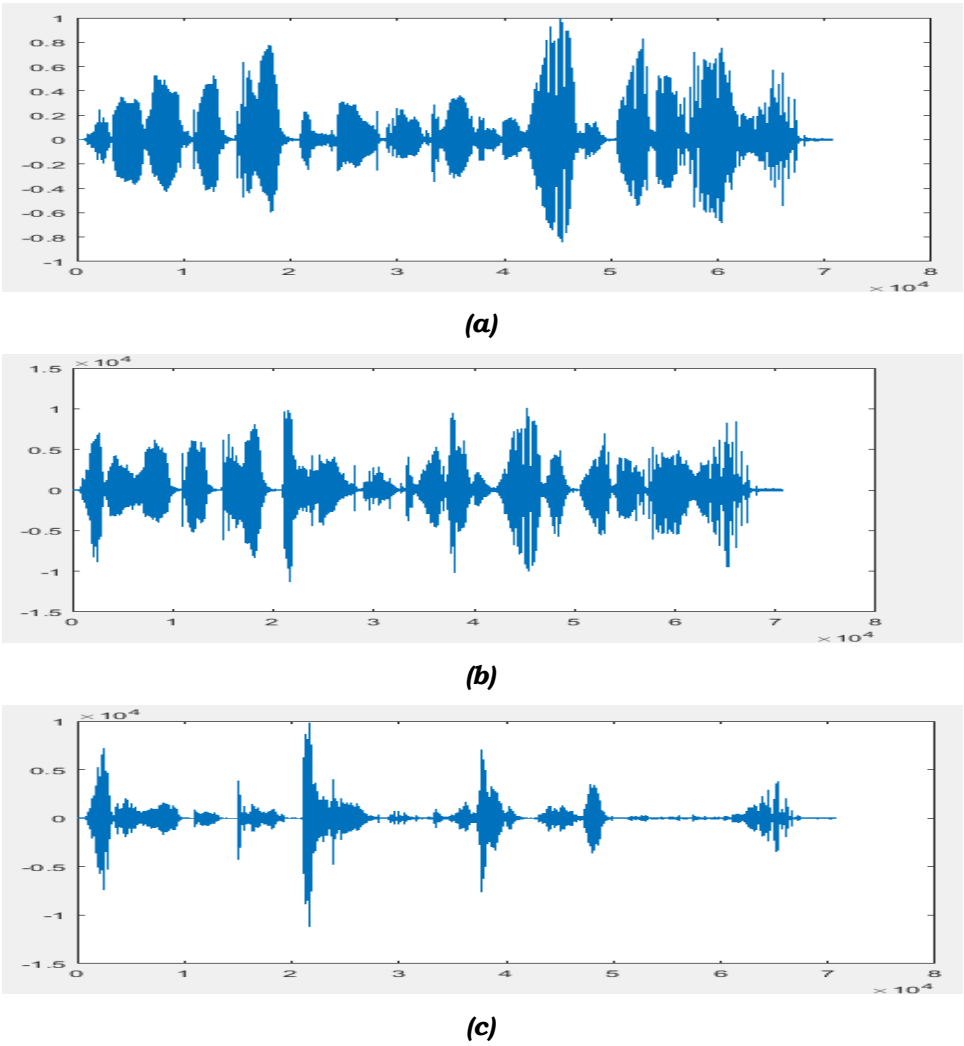

Fig. 2 Pre-processing steps of speaker recognition system (a) Input speech signal (b) pre emphasis filtered signal (c) Butterworth filtered signal

B. Feature Extraction

Feature extraction is applied after preprocessing speech signals. In this study, DWT based energy features are extracted from each frame of a speech signal. To obtain the features, the each frame of a speech signal is decomposed using DWT which represent the speech signal in the time-frequency domain. The main advantage of DWT is that it localizes time and frequency very well so that the local characteristics of the input signal can be revealed very well. After the computation of the signals into wavelet domain, the energy features of the signals are extracted. Then the wavelet energy coefficient s(t) is defined as:

$$E(s(t)) = \sum_{j=1}^{N} c_j^2 \tag{1}$$

where c_j is the wavelet coefficient value at location j in a sub-bands of a frame.

C. GMM based Classification

GMM is used for the recognition of speaker in this paper. It is an efficient density estimator based on the summation of different Gaussian distributions. Features of each class in the given training samples are modeled by GMM independently. In general, the classifier needs two inputs for its operation. So the extracted features of the trained database which are stored before and the testing features of the speech signals that are obtained are given as the input to the classifier.

The computation of Gaussian mixtures utilizes the data or features of each class in the training set. Eqn. 2 gives the likelihood associated with GMM classifier. The superscript in Eqn. 2 indicates the corresponding mixture class, and the subscript indicates the Gaussian within the mixture of Gaussians.

$$likelihood(Y|\beta) = \log P(Y|\beta) = \log \sum_{k=1}^{c} \pi \sum_{k}^{(\beta)} N(Y^{b} \mu \sum_{k}^{(\beta)}, \sum_{k}^{(\beta)}) \quad (2)$$

where N and c represent the Gaussian distribution and number of components in the mixture of Gaussians. $\sum_{k}^{(\beta)}$,and $\mu_{k}^{(\beta)}$ are the covariance and mean of each Gaussian distribution. The mixing factor is denoted by $\pi_{k}^{(\beta)}$. More information can be found in [13].

III. RESULTS AND DISCUSSION

The speaker recognition system is evaluated using chain corpus database [14]. Among 36 speakers, only eight speakers are randomly chosen for the evaluation. A total of 33 speech signals per speaker is available. They are recorded under different conditions. As the system is a recognition system, the dataset is split into two sets; training (24 speech signals) and testing (9 speech signals). These signals are decomposed at different decomposition level using wavelets. The recognition accuracy is obtained at each decomposition level with various Gaussian levels which are shown in Table 1.

TABLE 1 Performances of speaker recognition system

No. of Levels	**Classification Accuracy (%)**			
	No. of Gaussian Levels for 8 Speaker			
	G4	**G8**	**G16**	**G32**
1	57.75	59.14	67.23	64.56
2	66.06	74.45	77.12	71.50
3	74.23	81.13	88.38	82.61
4	86.12	91.02	**96.31**	85.01
5	50.65	63.89	78.45	68.72
6	39.45	49.16	72.50	67.33

From Table 1, it is clear that the accuracy level varies at each decomposition level and different Gaussian levels. The speaker recognition system

provides a maximum average accuracy of 96.31% at the 4th level of decomposition with the Gaussian level of 16.

IV. CONCLUSION

In this paper, speaker recognition system is presented using wavelets and GMM classifier. The features are extracted by wavelets, and then GMM is modeled for the classification. Chain corpus database is used to evaluate the speaker recognition system. DWT based energy features of 1-6 decomposition levels are obtained in the feature extraction stage for the recognition of individual speakers. Results show that GMM based identification process is more useful in speaker recognition system.

REFERENCES

[1]. Sukhwal, and M. Kumar, "Comparative study of different classifiers based speaker recognition system using modified MFCC for noisy environment", IEEE International Conference on Green Computing and Internet of Things, 2015, pp. 976-980.

[2]. Z. Senturk, and O. Salor, "Effect of plosives on isolated speaker recognition system performance", IEEE 9th International Conference on Electrical and Electronics Engineering, 2015, pp. 1263-1265.

[3]. D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition", IEEE Workshop on Automatic Speech Recognition and Understanding, 2015, pp. 92-97.

[4]. L. Li, Y. Lin, Z. Zhang, and D. Wang, "Improved deep speaker feature learning for text-dependent speaker recognition", IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2015, pp. 426-429.

[5]. M. Jamali, V. Ghafarinia, and M.A. Montazeri, "Recognition of speaker-independent isolated Persian digits using an enhanced vector quantization algorithm", IEEE Signal Processing and Intelligent Systems Conference, 2015, pp. 164-168.

[6]. A. Poddar, M. Sahidullah, and G. Saha, "Performance comparison of speaker recognition systems in presence of duration variability", IEEE Annual India Conference, 2015, pp. 1-6.

[7]. M.Y.A. Khan, S.M. Hossain, and M.M. Hoque, "Isolated Bangla word recognition and speaker detection by semantic modular time delay neural network", IEEE 18th International Conference on Computer and Information Technology, 2015, pp. 560-565.

[8]. P. Bansal, S.A. Imam, and R. Bharti, "Speaker recognition using MFCC, shifted MFCC with vector quantization and fuzzy", IEEE International Conference on Soft Computing Techniques and Implementations, 2015, pp. 41-44.

[9]. G. Liu, and J.H. Hansen, "An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 22, No. 12, 2014, pp. 1978-1992.

[10]. D. An, M. Shao, Z. Yuan, H. Shi, and Q. Pan, "Speaker Recognition Method Based on CPSO Clustering and KMP Algorithm", IEEE 7th International

Symposium on Computational Intelligence and Design, Vol. 1, 2014, pp. 556-559.

[11]. M. Alsulaiman, "Effect of Spoken Text on Text-Independent Speaker Recognition", IEEE 5th International Conference on Intelligent Systems, Modelling and Simulation, 2014, pp. 279-284.

[12]. M. Zamalloa, L.J Rodríguez-Fuentes, M. Penagarikano, G. Bordel, and J.P. Uribe, "Feature dimensionality reduction through genetic algorithms for faster speaker recognition", IEEE 16th European Signal Processing Conference, 2008, pp. 1-5.

[13]. C.M Bishop, "Pattern recognition and machine learning", Springer, Chapter 9, Vol. 1, 2006, pp.435.

[14]. F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The CHAINS corpus: CHAracterizing INdividual Speakers", In Proceedings of SPECOM & RQUO; 2006, pp. 431-435