Speaker Recognition System Based on Wavelet Features and Gaussian Mixture Models

K. Sajeer, Paul Rodrigues

Abstract:. Identification of a person's voice from the different voices is known as speaker recognition. The speech signals of individuals are selected by means of speaker recognition or identification. In this work, an efficient method for speaker recognition is made by using Discrete Wavelet Transform (DWT) features and Gaussian Mixture Models (GMM) for classification is presented. The input speech signal features are decomposed by DWT into subband coefficients. The DWT subband coefficient features are the input for the classification. Classification is made by GMM classifier at 4, 8, 16 and 32 Gaussian component levels. Results show a better accuracy of 96.18% speaker signals using DWT features and GMM classifier.

Index Terms: Speaker recognition, DWT transform, subband coefficient features, GMM classifier.

I. INTRODUCTION

Speaker recognition is that the systems are created to identify, verify and differentiate the individual speaker. It was recognized by the frequency and flow of their voice in normal pronunciation. GMM and Convolutional Neural Network (CNN) hybrid method for speaker recognition in short utterance are discussed in [1]. The signal features are extracted by Mel-Frequency Cepstral Coefficients (MFCC) and classification is made by hybrid classifiers like CNN and GMM.

I-vector based speaker recognition for Discriminatively Learned Network (DLN) is discussed in [2]. The input signal features are extracted by MFCC with mean and variance. DLN with i-vector is used for classification. Speaker recognition using wavelet analysis and multimodal neural networks is discussed in [3]. The input speech signal features are extracted by DWT, wavelet packet transform, wavelet subband frequency and MFCC. The classification is made by radial basis function neural network models, probabilistic neural network and general regressive neural network.

Speaker verification with a combination of DWT and MFCC feature wrapping is discussed in [4]. Feature extraction and classification are made by DWT based MFCC features for enhanced forensic speaker verification. Automatic speaker recognition based on wavelet transform is discussed in [5]. The speech signal features are extracted by DWT based MFCC and linear predictive coding features are used for automatic speaker recognition.

Speaker recognition based on Stationary Wavelet Transform (SWT) and Principal Component Analysis (PCA) is discussed in [6]. The input speech signal features are decomposed by SWT and PCA. An artificial neural network is used for classification. Comparative analysis of MFCC and Bark Frequency Cepstral Coefficient (BFCC) for speaker recognition system is discussed in [7]. The feature extraction for speech signals features is made by MFCC and BFCC for speaker recognition system with vector quantization method.

Forensic speaker recognition in noise for mitigating effects is discussed in [8]. In preprocessing stage voice activity detector is used to achieve the robustness. The speech signal features are extracted by gammatone frequency cepstral coefficients. Classification is made by universal background model. Natural voice disguise technique for automatic speaker recognition is discussed in [9]. MFCC is used to extract the speech signal features and GMM is used for classification.

Speaker recognition based on MFCC and Back Propagation Neural Network (BPNN) is discussed in [10]. The input speech signal features are extracted by MFCC and the classification is made by BPNN. Speaker recognition using Butterworth Filter (BF) and wavelet cepstral coefficient is discussed in [11]. In the preprocessing stage, BF is applied to remove noise. Then the speech signals are decomposed into different frequency channels. The wavelet cepstral coefficient is applied for the individual speakers.

Automatic speaker recognition based on a machine learning algorithm is discussed in [12]. Initially, voice activity is detected. The speech signal features are extracted by time, frequency and cepstral domain. Finally, classification is made by support vector machine, k- nearest neighbor, multilayer perceptrons and random forest classifier. Speaker recognition using MFCC and Locality Sensitive Hashing (LSH) is discussed in [13]. At first, the speech signal features are extracted by MFCC. Then LSH classifier is used for classification.

Speaker recognition with MFCC application using Matlab is discussed in [14]. The speech signal features are extracted by MFCC. At last, the paper compares the rectangular window and hamming window technique based on the filters. In this paper, an efficient method for speaker recognition is presented based on DWT features and GMM classifier. The organization of this paper as follows: In section 2 the methods and materials of the speaker recognition are described. Section 3 gives the results of the

Revised Manuscript Received on 12 October, 2019.

K. Sajeer, Research Scholar, Department of Computer Science, Research and Development Centre, Bharathiar University, Coimbatore-641 046, Tamil Nadu, India
(Email: sajeerkarattil@gmail.com)
Paul Rodrigues, DMI College of Engineering, Palanchur, Chennai-600123, Tamil Nadu, India
(Email: drpaulprof@gmail.com)

speaker recognition at different Gaussian levels. The last section concludes the speaker recognition system..

II. METHODS AND MATERIALS

Figure 1 explains the workflow speaker recognition system. Implementation of this system is based on DWT features and GMM classifier.

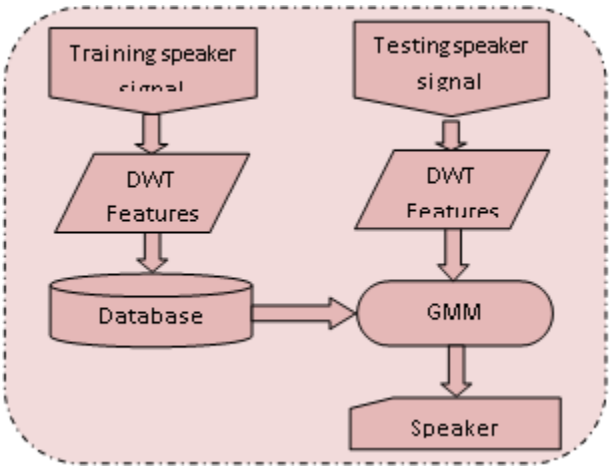


Figure 1 Speaker recognition system

A. *Wavelet Decomposition*

The speaker signals features are decomposed by DWT features. It produces higher and lower frequency subbands and also it is familiar. The discrete set of rules and translation scales are implemented by wavelet transform. In the set of input signal features the discrete set of shrinking signal features are decomposed. The DWT decomposition is defined as,

$$\Phi(K)=\sum_{v=-\infty}^{\infty}(-1)^{v}L_{T-1-v}\Phi(2K-v) \tag{1}$$

where T is even integer and it decomposes the set of wavelets and forms the decomposition signal. The input speech signal features are extracted by DWT. DWT is also widely used in other fields like epileptic seizure detection [15] and video steganography method for secure video transaction [16].

B. *GMM Classifier*

GMM is a supervised learning classifier algorithm and it classifies the wide variety of signals. GMM is a superior algorithm to use for classification in pattern recognition. It is a probabilistic model and generates the data point from the overall finite Gaussian distributions with their parameters. Parameters in the GMM are derived from the well trained earlier model. GMM is used for modelling data from several groups.

Let us consider N clusters, and is estimated for each n. It has only one estimation and it is estimated by the maximum-likelihood method. The n clusters have the probability density and it is defined by a linear function of densities of all the n distributions,

$$q(Y)=\sum_{n=1}^{N}\gamma_n H(Y|\delta_n,\lambda_n) \tag{2}$$

where γ_n is the coefficient of n.

The parameter estimation of the maximum log-likelihood method is $q(Y|\delta,\lambda,\gamma)$.

$$\ln q(Y|\delta,\lambda,\gamma)=\sum_{j=1}^{K}q(Y_j) \tag{3}$$

$$=\sum_{j=1}^{K}\ln\sum_{n=1}^{N}\gamma_n H(Y_j|\delta_n,\lambda_k) \tag{4}$$

The random variable is defined as $\psi_n(Y)$ such that $\psi_n(Y)=q(n|Y)$.

Baye's theorem, $\psi_n(Y)$

$$=\frac{q(Y|n)q(n)}{\sum_{n}^{N}q(n)q(Y|n)} \tag{5}$$

$$=\frac{q(Y|n)\gamma_n}{\sum_{n=1}^{N}\gamma_n q(Y|n)} \tag{6}$$

The maximum derivative function is $q(Y|\delta,\lambda,\gamma)$ with respect to δ,λ and δ should be zero. The derivative of q is equalled by $q(Y|\delta,\lambda,\gamma)$ with respect δ to zero and rearranged by the terms is,

$$\delta_n=\frac{\sum_{j=1}^{J}\psi_j(y_j)y_j}{\sum_{j=1}^{J}\psi_j(y_j)} \tag{7}$$

The derivative is taken with respect to λ and γ respectively, the obtained expressions are,

$$\lambda_n=\frac{\sum_{j=1}^{J}\psi_j(y_j)(y_j-\delta_j)(y_j-\delta_j)^{M}}{\sum j=1^{J}\psi_n(y_j)} \tag{8}$$

$$\gamma_n=\frac{1}{J}\sum_{j=1}^{J}\psi_n(y_j) \tag{9}$$

where $\sum_{j=1}^{J}\psi_n(y_j)$ denotes the sample points in the nth cluster. The total K number of samples and each sample contains g features are denoted by y_j. The classification is made by GMM classifier at 4, 8, 16 and 32 Gaussian component levels. GMM classifier is also used for texture image classification [17], single image super-resolution method [18] and in online signature verification method [19].

III. RESULTS AND DISCUSSION

A set of 36 speakers in the Chain corpus [20] database is used to estimate the performance of DWT-GMM based speaker recognition system. Performance of DWT-GMM is implemented by the DWT features with Gaussian

components like 4,8,16 and 32. The speech signal features are extracted by DWT at 5 levels for all speech signals of the speaker. GMM classifier is used for the classification of 8, 16 and 32 speakers set respectively. Figure 2 shows some of the speech signals in the database. Figure 3 shows the DWT feature extraction at 4 levels and approximate subband. Table 1, 2 and 3 shows the accuracies obtained by DWT-GMM system for 8, 16 and 32 speaker.

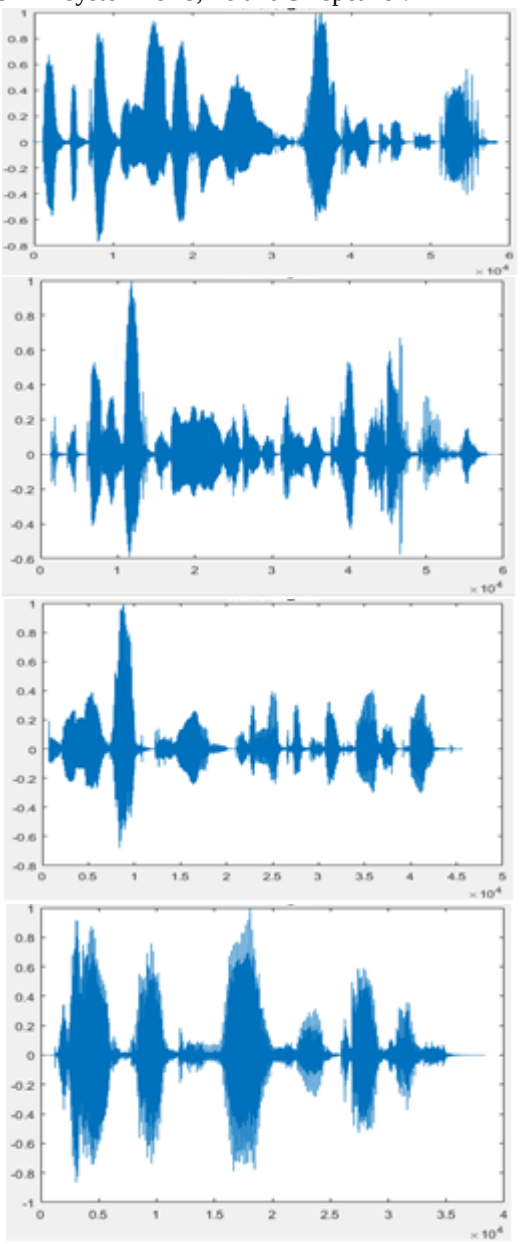
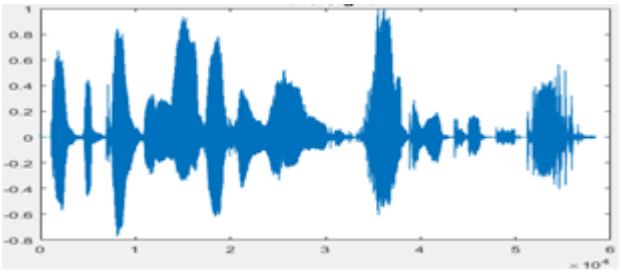
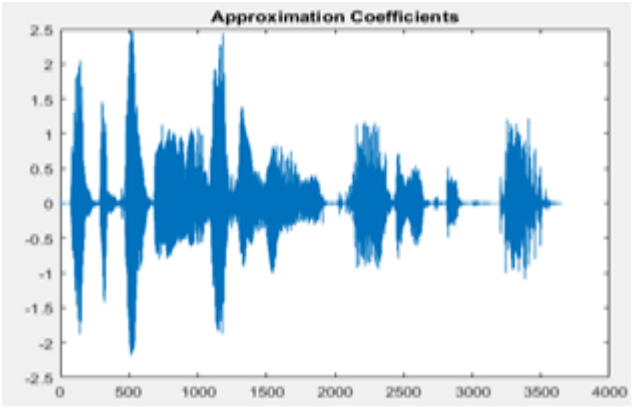


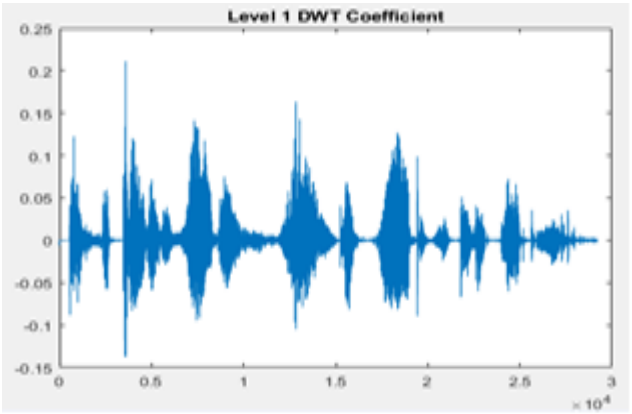
Figure 2 Speech signals in the chain corpus database



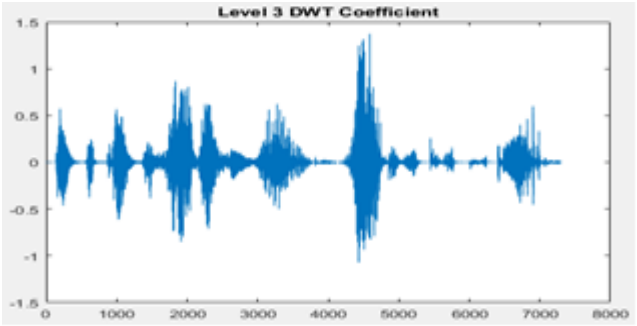
(a) Original speech signal



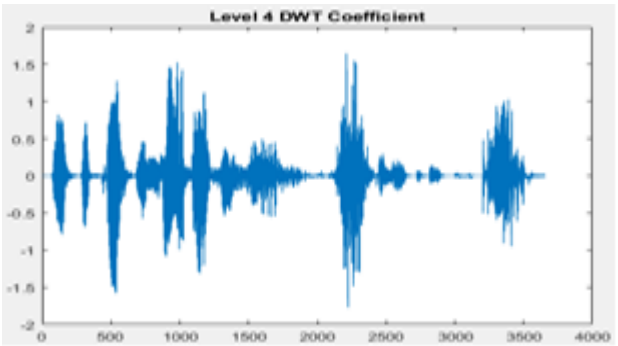
(b) Approximate coefficient



(c) DWT decomposition at 1st level



(e) DWT decomposition at 3rd level



(f) DWT decomposition at 4th level

Figure 3 DWT feature extraction stage at 4 levels with approximate subband

DWT Levels	Speaker recognition accuracy (%)			
	Gaussian levels for 8 Speaker			
	G4	G8	G16	G32
1	56.94	58.33	66.66	63.88
2	65.27	73.61	76.38	70.83
3	86.11	90.27	95.83	83.33
4	73.61	80.55	87.5	81.94

Table 1 Accuracies of DWT based GMM system on 8-speaker set

DWT Levels	Speaker recognition accuracy (%)			
	Gaussian levels for 16 Speaker			
	G4	G8	G16	G32
1	56.94	59.02	67.36	64.58
2	65.97	77.01	77.08	71.52
3	86.11	90.97	95.83	84.02
4	74.30	79.86	86.80	81.94

Table 2 Accuracies of DWT based GMM system on 16-speaker set

DWT Levels	Speaker recognition accuracy (%)			
	Gaussian levels for 32 Speaker			
	G4	G8	G16	G32
1	57.29	57.63	65.27	62.84
2	64.23	75.34	76.04	70.13
3	84.72	88.88	96.18	82.98
4	72.56	77.77	84.02	79.16

Table 3 Accuracies of DWT based GMM system on 32-speaker set

From table 1 to 3 it is observed that DWT-GMM system provides 95.83% accuracy at 8-speaker set, 95.83% accuracy at 16-speaker set and 96.18% for 32-speaker set and these accuracies are obtained at 3rd level of DWT-GMM system for 8, 16 and 32 sets of the speaker. Also, it is observed that the 32-speaker set produces a higher accuracy when compared to the 8 and 16 speaker set.

IV. CONCLUSION

An efficient method for speaker recognition based on DWT-GMM is presented. The speech signal features are extracted by DWT. The extracted signal features are the input for the classification of GMM with different Gaussian components for 8, 16 and 32 speaker sets. The higher classification accuracy is produced by 3rd level of DWT decomposition in 32 speaker set. CHAINS corpus speech signal database is used for implementation. Results show better classification accuracy.

REFERENCES

1. Liu, Z., Wu, Z., Li, T., Li, J. and Shen, C., 2018. GMM and CNN hybrid method for short utterance speaker recognition. IEEE Transactions on Industrial Informa`tics, 14(7), pp.3244-3252.
2. Yao, S., Zhou, R., Zhang, P. and Yan, Y., 2018. Discriminatively learned network for i-vector based speaker recognition. Electronics Letters, 54(22), pp.1302-1304.
3. Almaadeed, N., Aggoun, A. and Amira, A., 2015. Speaker identification using multimodal neural networks and wavelet analysis. IET Biometrics, 4(1), pp.18-28.
4. Al-Ali, A.K.H., Dean, D., Senadji, B., Chandran, V. and Naik, G.R., 2017. Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions. IEEE Access, 5, pp.15400-15413.
5. Malik, S. and Afsar, F.A., 2009, December. Wavelet transform based automatic speaker recognition. In 2009 IEEE 13th International Multitopic Conference (pp. 1-4). IEEE.
6. Jayakumar, A., Vimal, K.V. and Babu, A.P., 2009, December. Text dependent speaker recognition using discrete stationary wavelet transform and PCA. In 2009 International Conference on the Current Trends in Information Technology (CTIT) (pp. 1-4). IEEE.
7. ur Rehman, F., Kumar, C., Kumar, S., Mehmood, A. and Zafar, U., 2017, December. VQ based comparative analysis of MFCC and BFCC speaker recognition system. In 2017 International Conference on Information and Communication Technologies (ICICT) (pp. 28-32). IEEE.
8. Athulya, M.S. and Sathidevi, P.S., 2017, March. Mitigating effects of noise in Forensic Speaker Recognition. In 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 1602-1606). IEEE.
9. Staroniewicz, P., 2018, September. Influence of Natural Voice Disguise Techniques on Automatic Speaker Recognition. In 2018 Joint Conference-Acoustics (pp. 1-9). IEEE.
10. Wang, Y. and Lawlor, B., 2017, June. Speaker recognition based on MFCC and BP neural networks. In 2017 28th Irish Signals and Systems Conference (ISSC) (pp. 1-4). IEEE.
11. Rathor, S. and Jadon, R.S., 2017, July. Text indpendent speaker recognition using wavelet cepstral coefficient and butter worth filter. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.
12. Mokgonyane, T.B., Sefara, T.J., Modipa, T.I., Mogale, M.M., Manamela, M.J. and Manamela, P.J., 2019, January. Automatic Speaker Recognition System based on Machine Learning Algorithms. In 2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA) (pp. 141-146). IEEE.
13. Awais, A., Kun, S., Yu, Y., Hayat, S., Ahmed, A. and Tu, T., 2018, May. Speaker recognition using mel frequency cepstral coefficient and locality sensitive hashing. In 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 271-276). IEEE.
14. Bhattarai, K., Prasad, P.W.C., Alsadoon, A., Pham, L. and Elchouemi, A., 2017, April. Experiments on the MFCC application in speaker recognition using Matlab. In 2017 Seventh International Conference on Information Science and Technology (ICIST) (pp. 32-37). IEEE.
15. Sharmila, A. and Geethanjali, P., 2016. DWT based detection of epileptic seizure from EEG signals using naive Bayes and k-NN classifiers. Ieee Access, 4, pp.7716-7727.
16. Mstafa, R.J., Elleithy, K.M. and Abdelfattah, E., 2017. A robust and secure video steganography method in DWT-DCT domains based on multiple object tracking and ECC. IEEE Access, 5, pp.5354-5365.
17. Boulemnadjel, A., Hachouf, F. and Kharfouchi, S., 2015. GMM estimation of 2D-RCA models with applications to texture image classification. IEEE Transactions on Image Processing, 25(2), pp.528-539.

AUTHORS PROFILE

Sajeer Karattil is currently working as Assistant Professor in the department of Computer Science and Engineering, MES College of Engineering Kuttipuram, Kerala, India. Previously he worked as Assistant Professor in Information and Technology Department, Hindustan University, Chennai, Tamil Nadu, India(www.hindustanuniv.ac.in). He is currently pursuing PhD in Speech processing from Bharathiar University. He finished B.Tech in computer science and engineering from School of Engineering Cusat, Kochi and M.Tech in Software Engineering from SRM University. He has more than 12 years of industry and academic experience and his area of expertise include software engineering and speech processing.

Paul Rodrigues is a professor in Computer Engineering department of DMI Engineering College, Chennai which is registered under Anna University. He is the CTO of WisdomTree Software Solutions, Chennai, India. He has received his B.Tech. from Karnataka University, India and M.Tech from NIT-Allahabad, India andPh.D. from Pondicherry University, India. He has more than 20 years of Teaching and Industry experience in Delivery Management, Software Engineering, Budget Management and Business Development. He has published more than 50 (refereed) papers in International Conferences/Journals which include Extreme Programming, Software Architecture, Databases and Object Oriented Analysis and Design. Also, he was an author of a chapter in the book-Recent Trends in Network Security and Applications? during July 2010 published by Springer Berlin Heidelberg publications, New York, USA. He is first in the world to apply Vastu to Software Architecture. He has worked in various domains that include Insurance, Retail, Digital Forensic, Content Management and Application Migrations. He is an active member of many professional-bodies like Identity Research Group, PMP and CISSP.