Optimised Features for Speaker Identification using Daubechies Wavelet based Variance Spectral Flux

Chander Prabha
Chandigarh University
Sukhvinder Kaur
Swami Deva Dyal Institute of Engineering and Technology
Meenu Gupta (gupta.meenu5@gmail.com)
Chandigarh University https://orcid.org/0000-0001-7366-0841
Fadi Al-Turjman
Cyprus Near East University: Yakin Dogu Universitesi

Research Article

Keywords: Bayesian Information Criterion, Discrete Wavelet Transform, Mel Frequency Spectral Coefficients, optimized Variance spectral flux

Posted Date: February 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-178374/v1

Optimised Features for Speaker Identification using Daubechies Wavelet based Variance Spectral Flux

Chander Prabha[1], Sukhvinder Kaur[2], Meenu Gupta[3*] and Fadi-Al Turjman[4]
[1, 3]Department of CSE, Chandigarh University, Punjab
[2]Department of Electronics and Computer Science Engineering, SDDIET, Barwala, Haryana
[4]Department of Artificial Intelligence Engineering and Research Centre for AI and IoT, Near East University, Nicosia, Cyprus, Mersin 10 Turkey
chander.e9251@cumail.in[1], er1971sukhvinderkaur@rediffmail.com[2], meenu.e9406@cumail.in[3], fadi.altujman@neu.edu.tr[4]

Abstract: An important application of speech processing is speaker recognition, which automatically recognizes the person speaking in an audio recording, basis of which is speaker-specific information included in its speech features. It involves speaker verification and speaker identification. This paper presents an efficient method based on discrete wavelet transform and optimized variance spectral flux to enhance the enactment of speaker identification system. An effective feature extraction technique uses Daubechies 40 (db40) wavelet to compress and de-noised the speech signal by its decomposition into approximations and details coefficients at level 1. The approximation coefficients contain 99.9% of speech information as compared to detailed coefficients. So, the optimized variance spectral flux is applied on wavelet approximation coefficients which efficiently extract the frequency contents of the speech signal and gives unique features. The distance between extracted features has been obtained by applying traditional Bayesian information criteria. Experimental results were computed on recording data of 33 speakers (23 female and 10 males) for text independent identification of speaker. Evaluation of effectiveness of the proposed system is done by applying detection error trade-off curves, receiver operating characteristic, and area under curve. It shows 94.38% of speaker identification results when compared with traditional method using Mel frequency spectral coefficients which is 90.70%.

Keywords: Bayesian Information Criterion; Discrete Wavelet Transform; Mel Frequency Spectral Coefficients; optimized Variance spectral flux.

1. INTRODUCTION

Speaker recognition is a process to determining identity of speaker's from the utterance present in the database. Two basic steps of speaker recognition system are there i.e. speaker identification (SI) and another is speaker verification (SV). These systems are generally used

for person authentication as in biometric, in forensic labs for voice identification. Other applications related to security are transactions over Telephone, computer access control and banking access. In speaker identification system, features of a person is compared with all the speakers feature stored in a database and for speaker verification, the features of the speaker is compared only with its stored voice in a database. Over the past years, the performance of speaker recognition systems and speech and has improved efficaciously. The conventional approaches, like Gaussian mixture models and Hidden Markov model have attain high accuracies results on refine speech as compared to real innate speeches [1][2], resulting in degradation of their performance. To improve on and to minimize this drop in performance, this paper proposes to inflate the robustness of the speaker recognition system by feature matching techniques and extracting more robust features. Mel-frequency spectral coefficients (MFCC) [3] technique is used for speaker recognition tasks.

The paper is organised as follows. Second section defines the algorithms of feature extraction for detecting the unique features of speakers. Feature matching using Bayesian information criteria is presented in third section and fourth section illustrates the performance evaluation criteria. The proposed speaker identification system is discussed in section five. Section six illustrates and evaluates various experimental results including recognition tests. The final remarks are concluded in last section based on our findings.

2. FEATURE EXTRACTION ALGORITHMS

The objective of feature extraction is to reduce data and save memory space, transmission bandwidth and power by capturing the essential characterstics of speaker. Various algorithms for extracting features of speech signal are discrete wavelet transform(DWT), Variance spectral flux(VSF) and Mel Frequency Cepstral Cofficients (MFCCs).

Artificial Intelligence and machine learning can be employed in different domains like drug discovery [11-12], fraud prediction [13-14], cancer prediction [15-16], etc. Authors in [17-19] describe the security and privacy aspects of the information especially the sensitive attributes like location and user identification present in the datasets used for empirical studies, while some good works discusses the same issue for discrete point dataset used for publishing the user data publicly [20-21].

2.1 Discrete Wavelet Transform (DWT) for doing Speech Compression

Since 1990's DWT has been extensively used to solve engineering problems due to its highfrequency and time resolution property. It can examine a signal simultenously in time-frequency domain. It also denoised the speech signal and improves its strength [4]. In the process of tansforming wavelet, the signal (speech) is decomposed into successive levels of low and high frequency compnents. The low frequency components are known as approximations and high frequency componenets are details. When DWT is applied on speech signal, about 98% of its information lies in approximation cofficients as shown in Fig. 1.

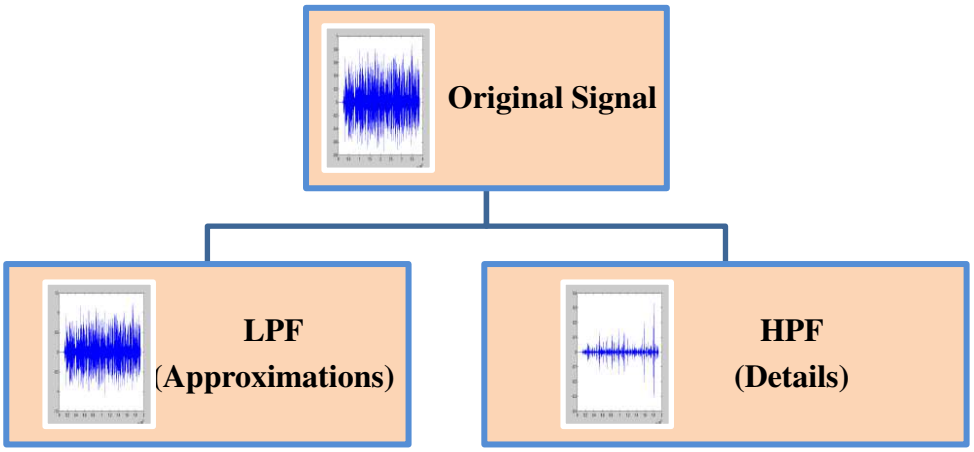

Fig. 1 Decomposition of speech signal using DWT to get noise free compressed signal

So, it is used as a best compression technique in speech processing. The definition of wavelet transform is the inner product of a input signal x(t) and mother wavelet ψ(t) represented as:

$$W_{\psi}x(m,n)= \frac{1}{\sqrt{m}} \int_{\infty}^{-\infty} x(t)\psi * \left(\frac{t-n}{m}\right) dt, \quad (1)$$

In above, mother wavelet is :

$$\Psi_{m,n}(t)=\psi\left(\frac{t-n}{m}\right) \quad (2)$$

Where, n and m are shift and scale parameters respectively. The DWT functions at level N and time location t_N can beexpressed as:

$$D_N(t_N)= x(t)\ \psi_m \left(\frac{t-t_N}{2^N}\right) \quad (3)$$

Where, ψ_N is known as decomposition filter at frequency level N that scaled the output by a factor 2^N.

2.2 Variance spectral flux (VSF)

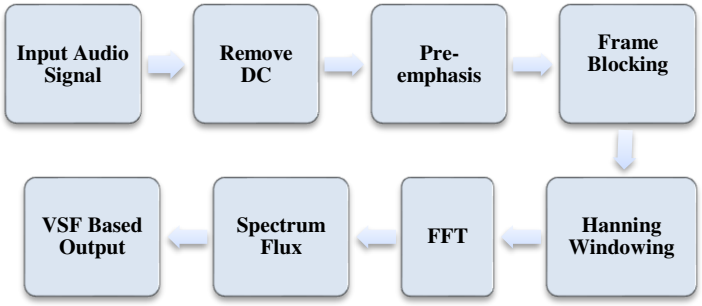

Fig. 2 Steps to extract VSF [5]

VSF feature extraction process flow diagram is shown in Fig. 2. The spectrum flux (SF) is the ordinary Euclidean norm of the Δ spectrum magnitude, and is as follows:

$$SF = \|S_i - S_{i-1}\|_2 = \frac{1}{N}\left(\sum_{k=0}^{N-1}(S_i(k) - S_{i-1}(k))^2\right)^{\frac{1}{2}} \quad (4)$$

Where S_i is the spectrum magnitude vector of frame and is defined as:

$$S_i(k) = \left|\sum_{n=0}^{N-1} s\left(n + \frac{N_i}{2}\right)\omega(n)\exp\frac{-2\pi kn}{N}\right| \qquad k \in [0, N-1] \quad (5)$$

Where $s\left(n + \frac{N_i}{2}\right)$ is audio data, N is the window size, $\omega(n)$ and is the window function. In this case hanning window is used [5].

When equation (5) is applied on the frames of apprximation coefficients obtained from DWT, it detects the variance in the frequency of speech signal.

2.3 Mel Frequency Spectral Coefficients (MFCC)

The most important part of speech processing is feature extraction to reduce the data size. Mel frequency cepstral coefficient is a common technique to extract features of speech signal in speaker identification system. Its performance degrades in noisy environment. The term Mel is a unit of pitch and an abbreviation of the word melody. The relation between linear frequency scale and mel scale is expressed as: f_{mel} = 2595log(1+f/700). Coefficient calculation steps are as follows [6]:

1. Take the discrete fourier transform of(a windowed signal.

$$h(n) = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & \text{else} \end{cases} \quad (6)$$

$$X(K) = \sum_{n=0}^{N-1} x(n)e^{-j2nK/N}, \qquad (0 \le n, K \le N-1) \quad (7)$$

2. After converting the powers of the spectrum into mel scale,take its log as follows:

$$L_M = \ln \sum |X(K)^2| H_M(K) , \qquad (0 \le m \le M) \tag{8}$$

Finally cofficients of MFCC are obtained by taking discrete cosine transform of equation (8) and is shown in Fig. 3.

$$c(n) = \sum L_M \cos\left(\frac{\pi n\left(m+\frac{1}{2}\right)}{M}\right), \tag{9}$$

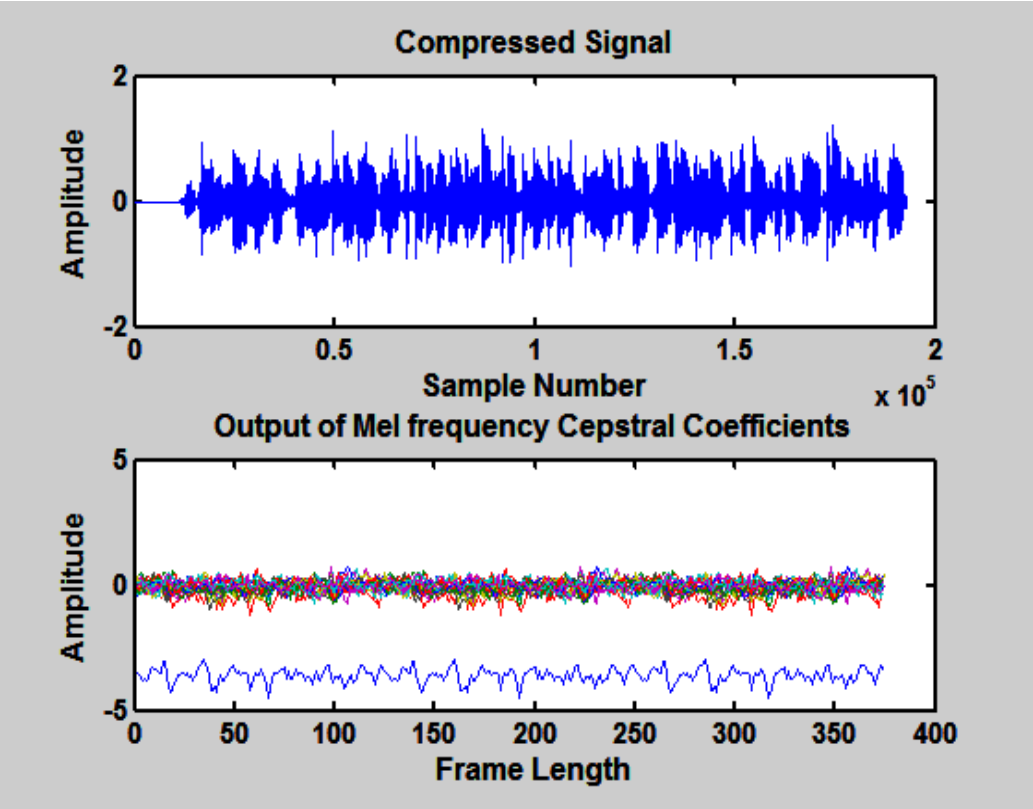

Fig. 3 Compressed speech signal and its coefficient of MFCC

3. SPEAKER CLASSIFICATION USING BAYESIAN INFORMATION CRITERION

Many speaker classification algorithms were proposed in past for speaker identification. Widely used techniques are Bayesian information criteria, generalized likelihood ratio and cross likelihood ratio. In this research work, for doing speaker identification, delta Bayesian Information Criterion (BIC) is used to find distance between two speakers, which expand the log-likelihood penalized by the intricacy of the model [7]. We considered two speakers i and j of parameterized acoustic vectors of Xi and Xj of frame lengths Ni and Nj respectively, and with mean and standard deviation values μi, σi and μj, σj. On fusing the speaker's features into

X, their mean and variance is μ, σ respectively with frame length N. The distance between two speakers is given as follows:

$$\Delta BIC = \frac{N}{2}\log|\sum X| - \frac{Ni}{2}\log|\sum Xi| - \frac{Nj}{2}\log|\sum Xj| - \lambda P \quad (10)$$

Where, λ is a free design parameter and it depends on the data being modelled, its value is 10, P is the penalty term, and is a function of the number of free parameters in the model.

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right) \quad (11)$$

4. PERFORMANCE EVALUATION CRITERIA

In this research work the speaker identification system performance is evaluated by two techniques to check whether a given speaker belongs to the specified database or not. During evaluation two types of errors were detected: missed detections and false alarms [8].

- Missed detection: Speaker is not attributed when speaker's speech exists in the database.
- False alarms: Speaker is attributed when there is no speaker's speech in the database.

4.1 Receiver operating characteristic (ROC)

The ROC is a frequently used methodology to compare the performance of classifier in speaker recognition system. It is based on hit and error detection probabilities. The maximum value of ROC curve is 1 and minimum value is 0 on both axes. The horizontal axis represents false positive rate and perpendicular axis is for true positive rates. It can be calculated as [9]:

$$\text{True positive rate (TPR)} = \frac{\text{No. of outputs greater than or equal to threshold}}{\text{No. of one targets}} \quad (12)$$

$$\text{False positive rate (FPR)} = \frac{\text{No. of outputs greater than threshold}}{\text{No. of zero targets}} \quad (13)$$

The area under ROC curve is calculated as:

$$AUC = \left\{ \begin{matrix} 0.5 * [TPR(2; end) + TPR(1: end-1)] * \\ [FPR(2: end) - FPR(1: end-1)] \end{matrix} \right\} \quad (14)$$

The value of AUC will always lies in between 0 and 1.

4.2 Detection Error Trade-off (DET)

In speaker recognition system, the performance of detection task is represented by DET curves. It involves the trade-off between two errors: missed speech and false alarm. The operating point at which two errors rates are equal is called equal error rate (EER). The performance of system

is determined by the value of EER. When the DET curve is closed to the origin, EER will be low, and then the quality of the system is improved [10].

5. Proposed Methodology

Speaker Identification system flow chart follows the same procedure of conventional identification system but with some alteration. The flow chart is shown in Fig.4. Based on the discrete wavelet transform, the audio signals were first enhanced and compressed in the ratio of 1:2 at level 1 using Daubechies 40 (db40) wavelet with energy of 99.9% (approx.). DWT based MFCC and VSF are used for extracting the features of compressed signal. Then BIC is used for speaker classification. The performance of proposed method was evaluated by ROC, DET and area under curve.

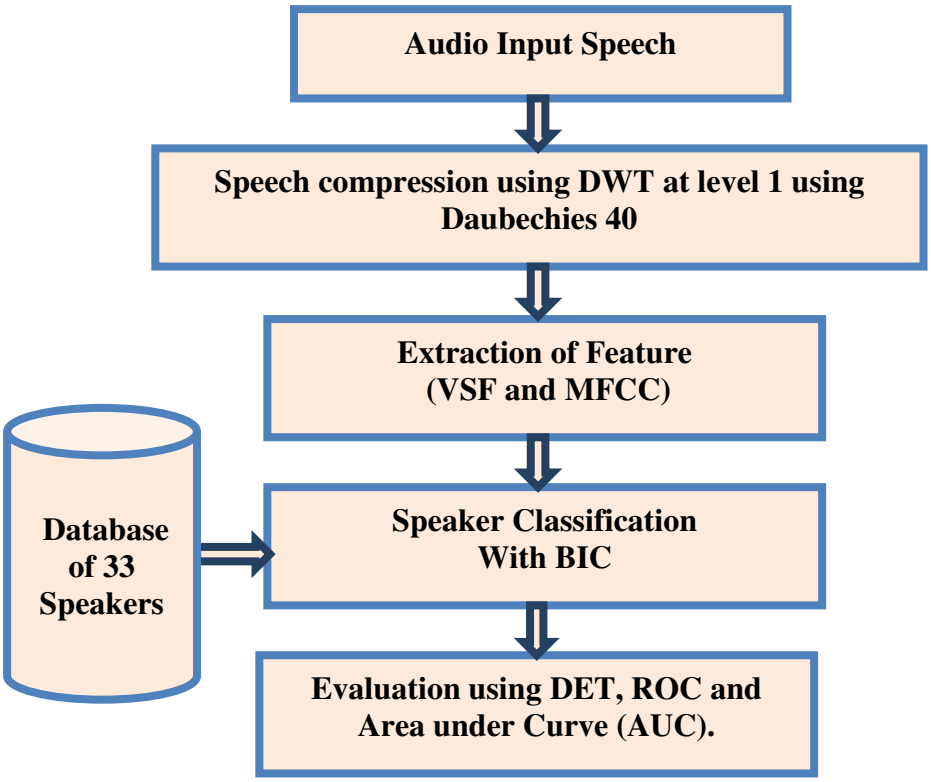

Fig. 4 Proposed diagram of speaker identification system

6. EXPERIMENTS AND RESULTS

6.1 Database

The recordings of utterances of 33 speakers (23 females and 10 males) of 15-20 seconds were used in this research work. Recordings of 11 speakers were taken from Personal Digital

Assistant (PDA) speech dataset. In this dataset, the speech of various speakers was recorded by four small microphones mounted around a PDA. Remaining 22 recordings were taken by using mobile phone in MP3 format. Further these recordings were converted into .wav form to use it in MATLAB software. Sampling frequency of each recording is 44100Hz.

6.2 Results and Discussion

After doing the audio signal compression and framing, their features extraction is done by using MFCC and DWT based VSF with distance metrics delta BIC. For testing, the distance between speaker number 5 and all other 33 speakers is calculated using MFCC and BIC and shown in Fig. 5. It shows that when speaker 5 is compared with itself its value is negative otherwise its value is positive. Similar test is applied on our proposed method using DWT based VSF and BIC and its output is shown in Fig. 6. It also shows that the value of distance between two same speakers is negative and for different speakers is positive. The performance of proposed system shows that the dissimilarity measure is improved as compared to existing system.

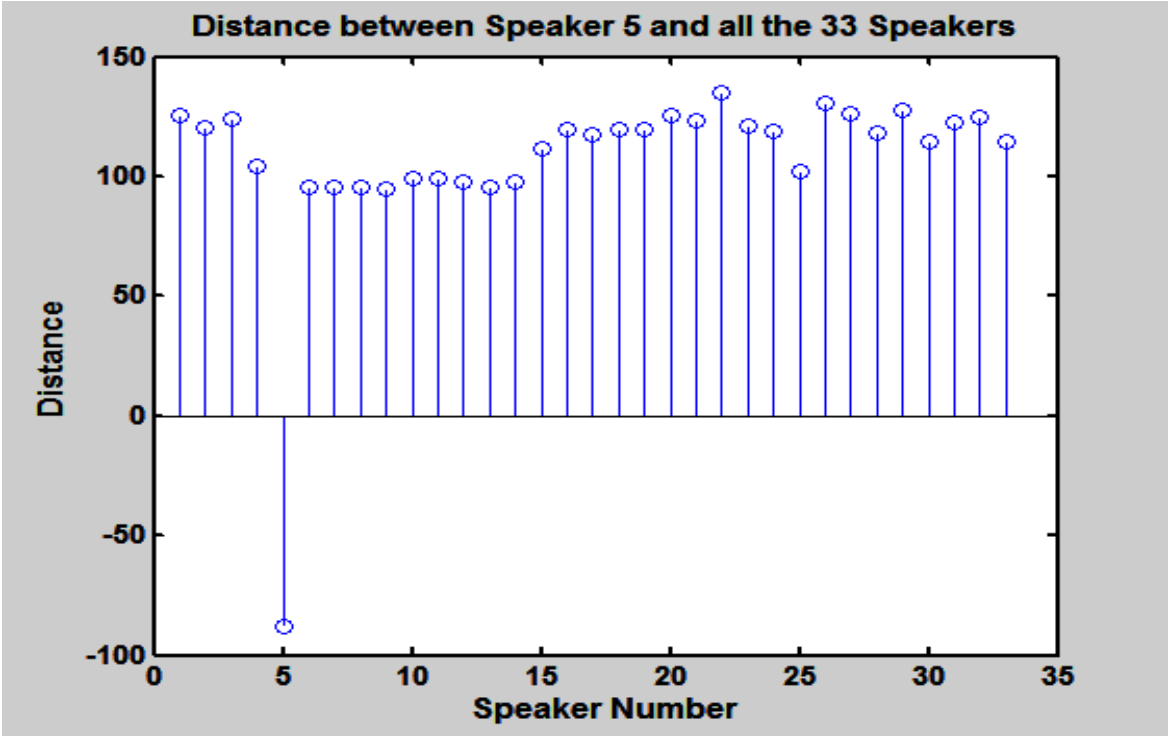

Fig. 5 Output (MFCC vs Distance Metric Delta BIC)

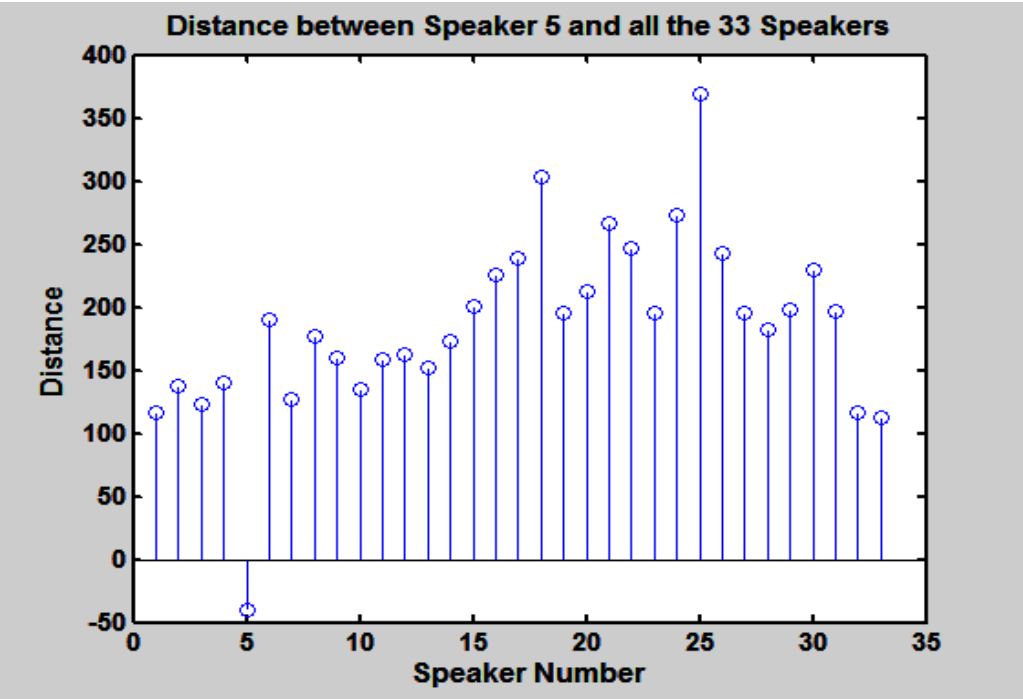

Fig. 6 Output (Distance Metric Delta BIC vs proposed algorithm)

Proposed algorithm performance for speaker identification system is weigh by traditional ROC curve. In this graph, true positive rate (miss speech rate) is plotted in function of the false positive rate (false alarm rate) for different cut-off points. The ROC curves for two techniques are shown in Fig. 7 and AUC for these curves are calculated using equation (14) and given in Table 1.

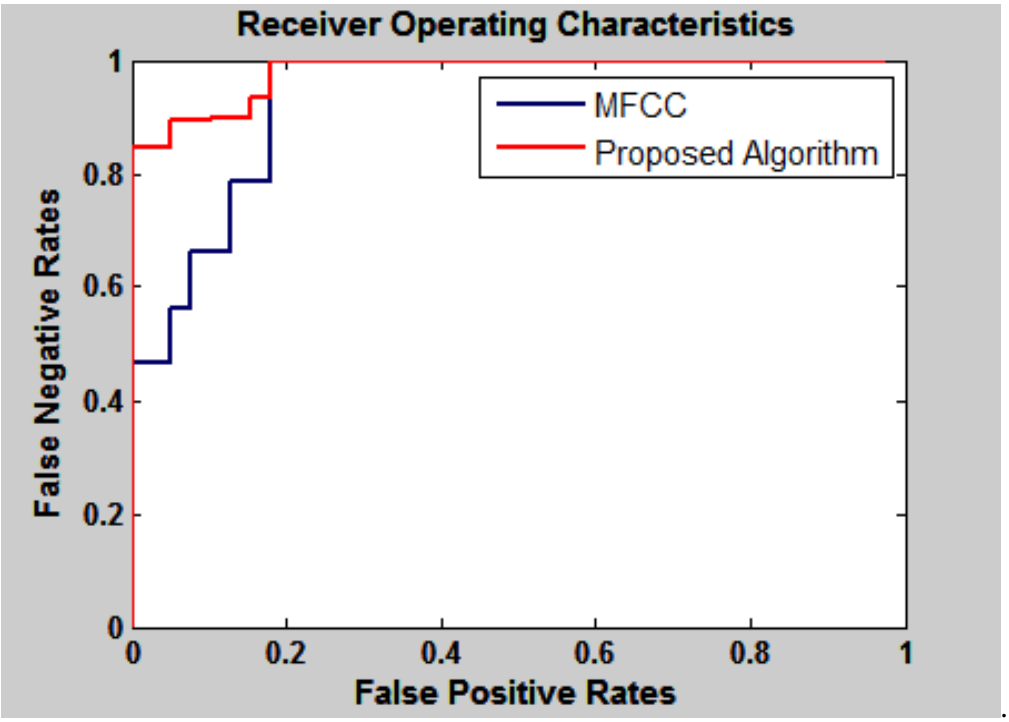

Fig. 7 ROC curves for MFCC and proposed method using VSF

Proposed method performances can also be weigh by Detection Error Trade-off (DET) curves as shown in Fig. 8. It is a graph of two error rates: false alarm rate and miss rate, drawn on the x and y axis respectively.

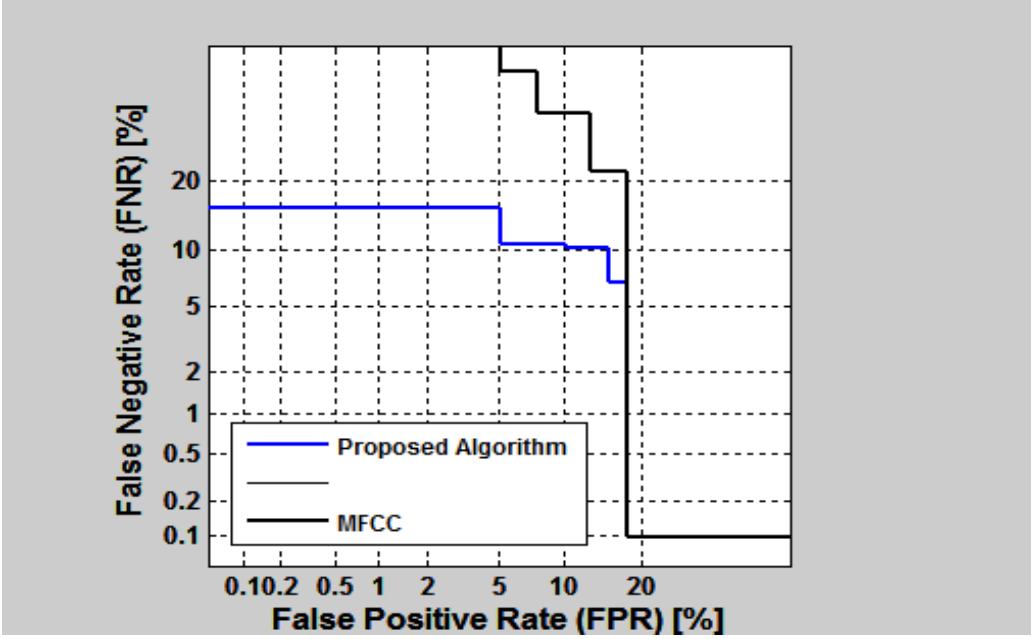

Fig. 8 DET curves for BIC with MFCC and VSF algorithm.

The curve for BIC with DWT based VSF is close to the origin, so, it performs better. The equal error rate by using BIC with MFCC is 17.9487 and that of proposed method based on DWT is 10.3564. Table 1 compares the results attained by MFCC and proposed algorithm using ROC and DET. After the estimation of the area under the ROC curve, it is found that BIC with proposed method cover max. area of 94.38%.

Table 1. Comparison of results using ROC and DET

Algorithm	Area under curve (%)	Equal Error Rate
MFCC	90.70	17.9487
Proposed Method	94.38	10.3564

7. CONCLUSIONS AND FUTURE SCOPE

This research work, presents an efficient algorithm for the speaker identification system which is performed on recordings of independent speech of 33 speakers. For extracting the features

of speech signal of different speakers, proposed algorithm based on DWT and optimized VSF is applied. Initially Daubechies 40 (db 40) wavelet is used to compress and denoised the speech signal. Its approximation coefficient carries 99.9 % of speech information on which optimized VSF is applied to extract its unique features. Moreover, for the classification of speakers, feature matching technique using traditional BIC classifier is applied. The system performance is evaluated by ROC curves, DET graphs and AUC. Their results are compared with traditional method using MFCC and observed that, AUC is increased and EER is reduced by using proposed method. Further research can be done on the classification of small utterances of length less than five seconds.

Declaration: On the behalf of all the authors, I mentioned there is no conflict of interest.

REFERENCES

1. Beigi H, *Fundamentals of Speaker Recognition*. (2011).
2. Kinnunen T & Li H , "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication* 52 (1) (2010) 12–40.
3. Hossain M M, Ahmed B, & Asrafi M, "A Real Time Speaker Identification Using Artificial Neural Network," in *proceedings of ICCIT*, 2007.
4. Wu J D, & Lin B F, "Speaker identification using discrete wavelet packet transform technique with irregular decomposition," *Expert Syst. Appl.*, 36(2) (2009) 3136–3143.
5. Rongqing Huang, , and John H. L. Hansen, "Advances in Unsupervised Audio Classification and Segmentation for the Broadcast News and NGSW Corpora", IEEE Transactions on Audio, Speech, and Language Processing, 14(3) (2006).
6. Anguera Miró X, Bozonnet S, Evans N, Fredouille C, Friedland G, & Vinyals O, "Speaker Diarization: A Review of Recent Research," *IEEE Trans. Audio Speech Lang. Process*, 20(2) (2012) 356–370.
7. Almpanidis G & Kotropoulos C, "Automatic Phonemic Segmentation Using the Bayesian Information Criterion", *in proc. of EUSIPCO*, (2007) 2055–2059.
8. Reynolds D, "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication*, 17(1) (1995) 91-108.
9. Kaur S & Sohal J S, "Speech Activity Detection and its Evaluation in Speaker Diarization System", *International Journal of Computers & Technology*, 16(1) (2017) 7567-7572.

10. Garcia Perera L P, Raj B, & Nolazco Flores J A, "Optimization of the DET curve in speaker verification under noisy conditions," *in proc. of ICASSP*, (2013).

11. B. Duhan and N. Dhankhar, "Hybrid Approach of SVM and Feature Selection Based Optimization Algorithm for Big Data Security," Lect. Notes Electr. Eng., vol. 605, pp. 694–706,
2020.

12. Hooda, N., Bawa, S., & Rana, P. S. (2018). B2FSE framework for high dimensional imbalanced data: A case study for drug toxicity prediction. *Neurocomputing*, *276*, 31-41.

13. Hooda, N., Bawa, S., & Rana, P. S. (2018). Fraudulent firm classification: a case study of an external audit. *Applied Artificial Intelligence*, *32*(1), 48-64.

14. Hooda, N., Bawa, S., & Rana, P. S. (2020). Optimizing Fraudulent Firm Prediction Using Ensemble Machine Learning: A Case Study of an External Audit. *Applied Artificial Intelligence*, *34*(1), 20-30.

15. Bhardwaj, R., & Hooda, N. (2019). Prediction of Pathological Complete Response after Neoadjuvant Chemotherapy for breast cancer using ensemble machine learning. *Informatics in Medicine Unlocked*, *16*, 100219.

16. Gupta, R., Rao, U.P.: Achieving location privacy through CAST in location based services. Journal of Communication Network 19(3), 227–238 (2017)

17. Gupta, R., Rao, U.P.: VIC-PRO: Vicinity protection by concealing location coordinates using geometrical transformations in location based services. Wireless Personal Communication 107(2), 1041–1059 (2019)

18. Gupta, R., Rao, U.P.: A hybrid location privacy solution for mobile lbs. Mobile Information System, 2017 (2017)

19. Gupta, R., Rao, U.P.: An exploration to location based service and its privacy preserving techniques: A survey. Wireless Personal Communication 96(2), 1973–2007 (2017)

20. Gupta, R. and Rao, U.P.: Preserving location privacy using three layer RDV masking in geocoded published discrete point data. *World Wide Web*, *23*(1), pp.175-206 (2020)

21. S. Singh, M. Singh, C. Prakash, M. K. Gupta, M. Mia, and R. Singh, "Optimization and reliability analysis to improve surface quality and mechanical characteristics of heat-treated fused filament fabricated parts," Int. J. Adv. Manuf. Technol., vol. 102, no. 5–8, pp. 1521–1536, 2019.

Figures

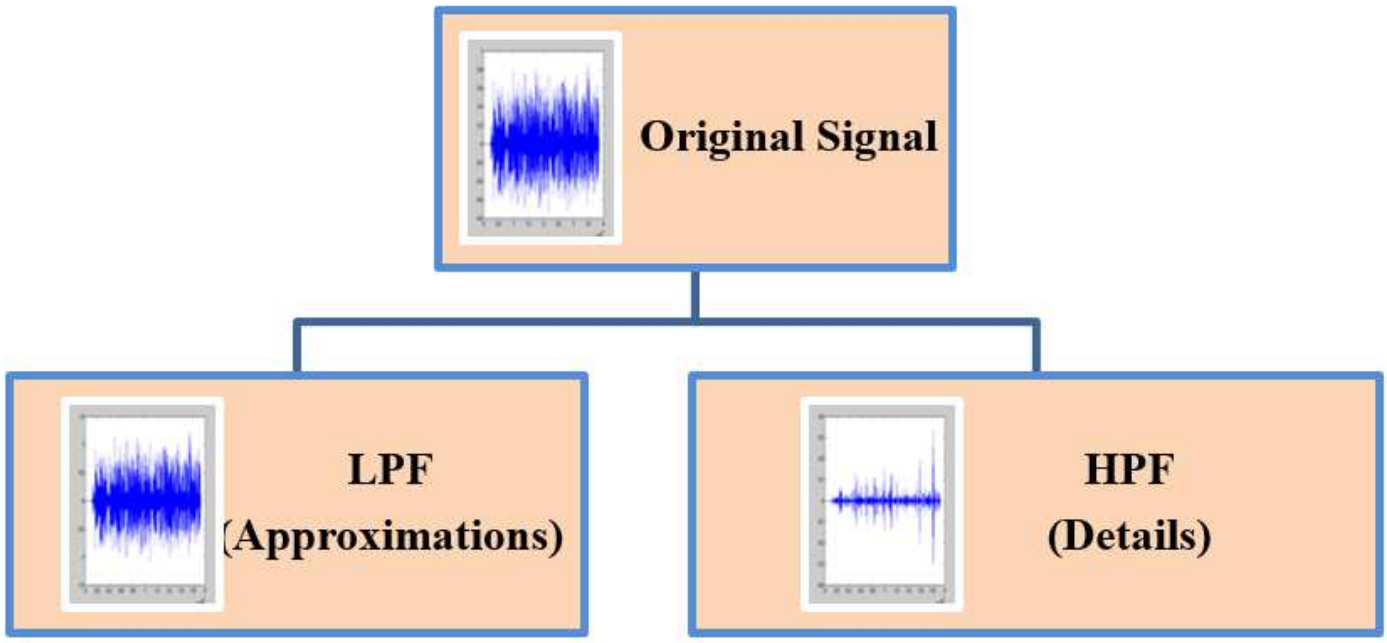

Figure 1

Decomposition of speech signal using DWT to get noise free compressed signal

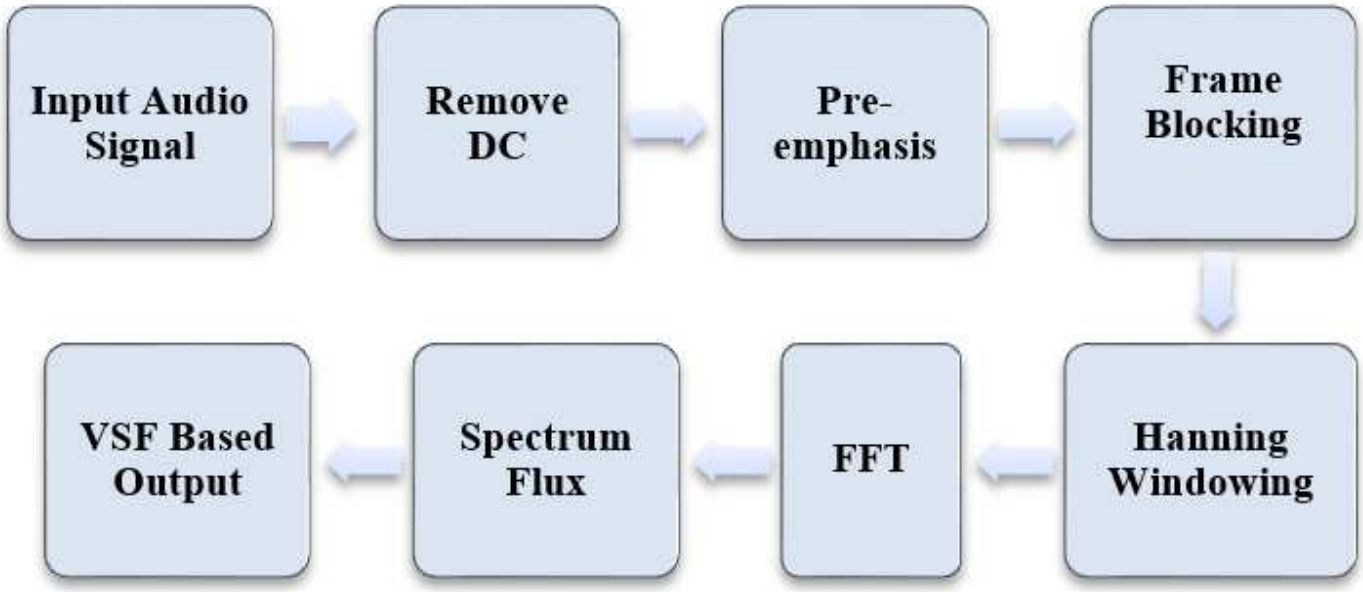

Figure 2

Steps to extract VSF [5]

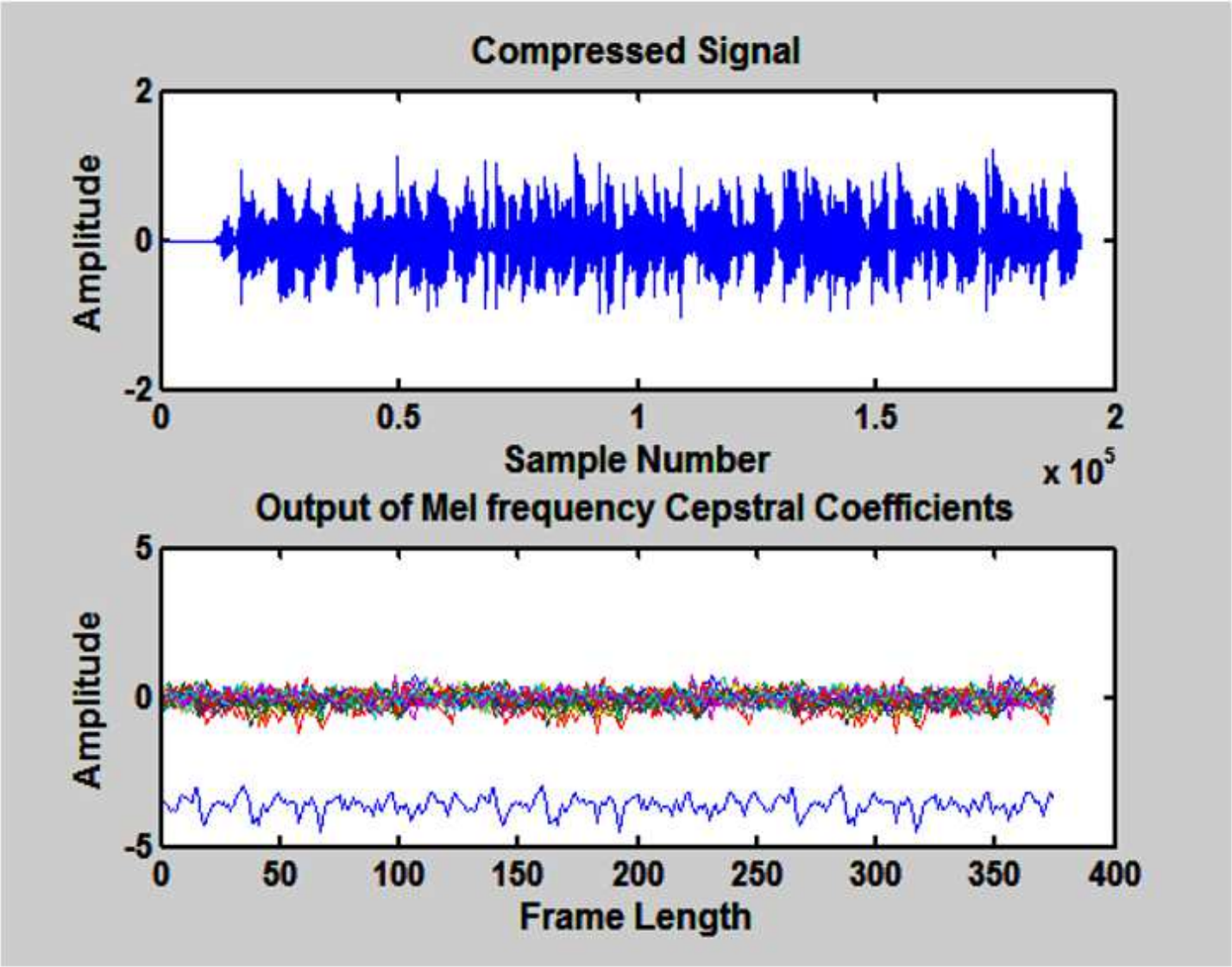

Figure 3

Compressed speech signal and its coefficient of MFCC

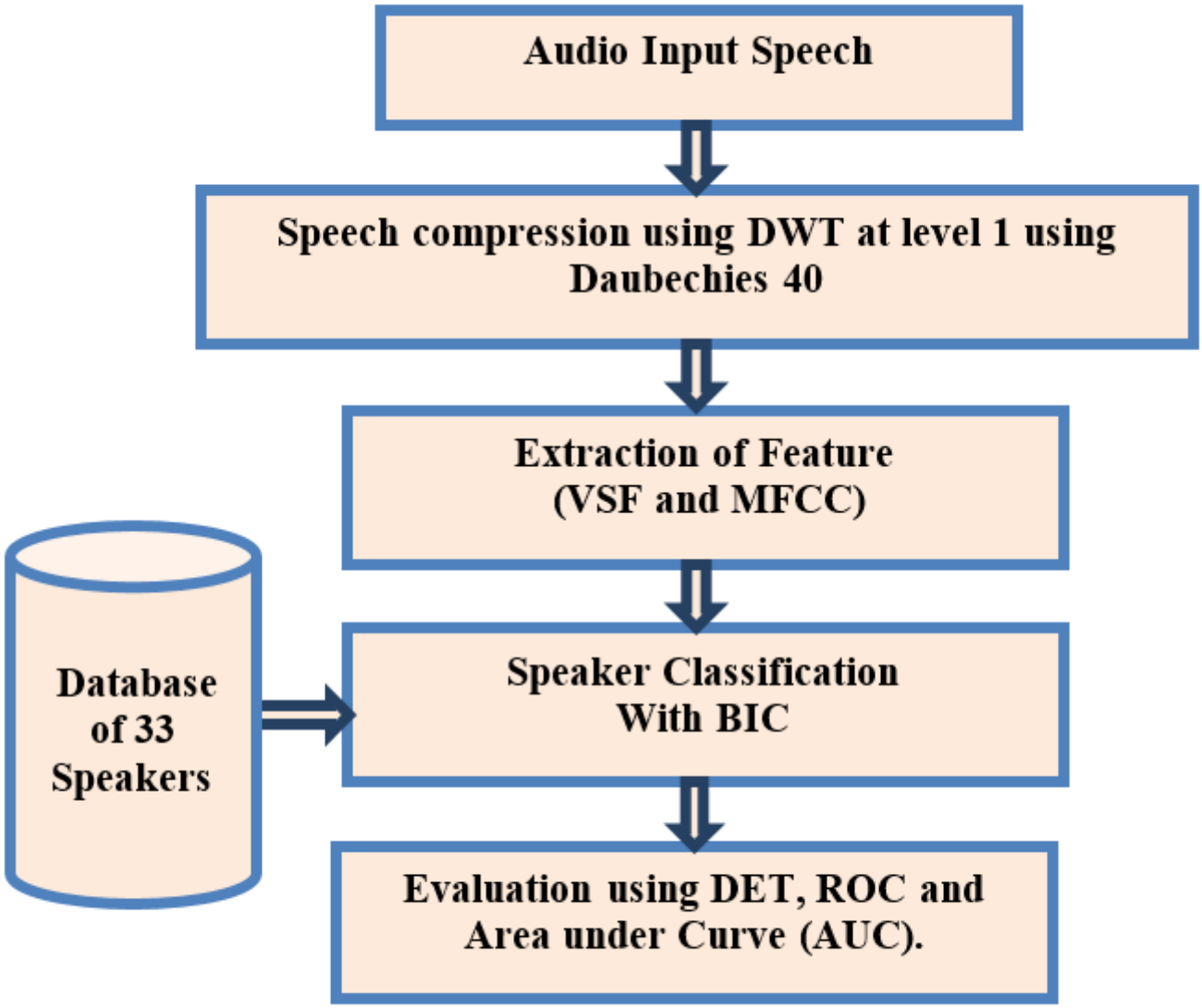

Figure 4

Proposed diagram of speaker identification system

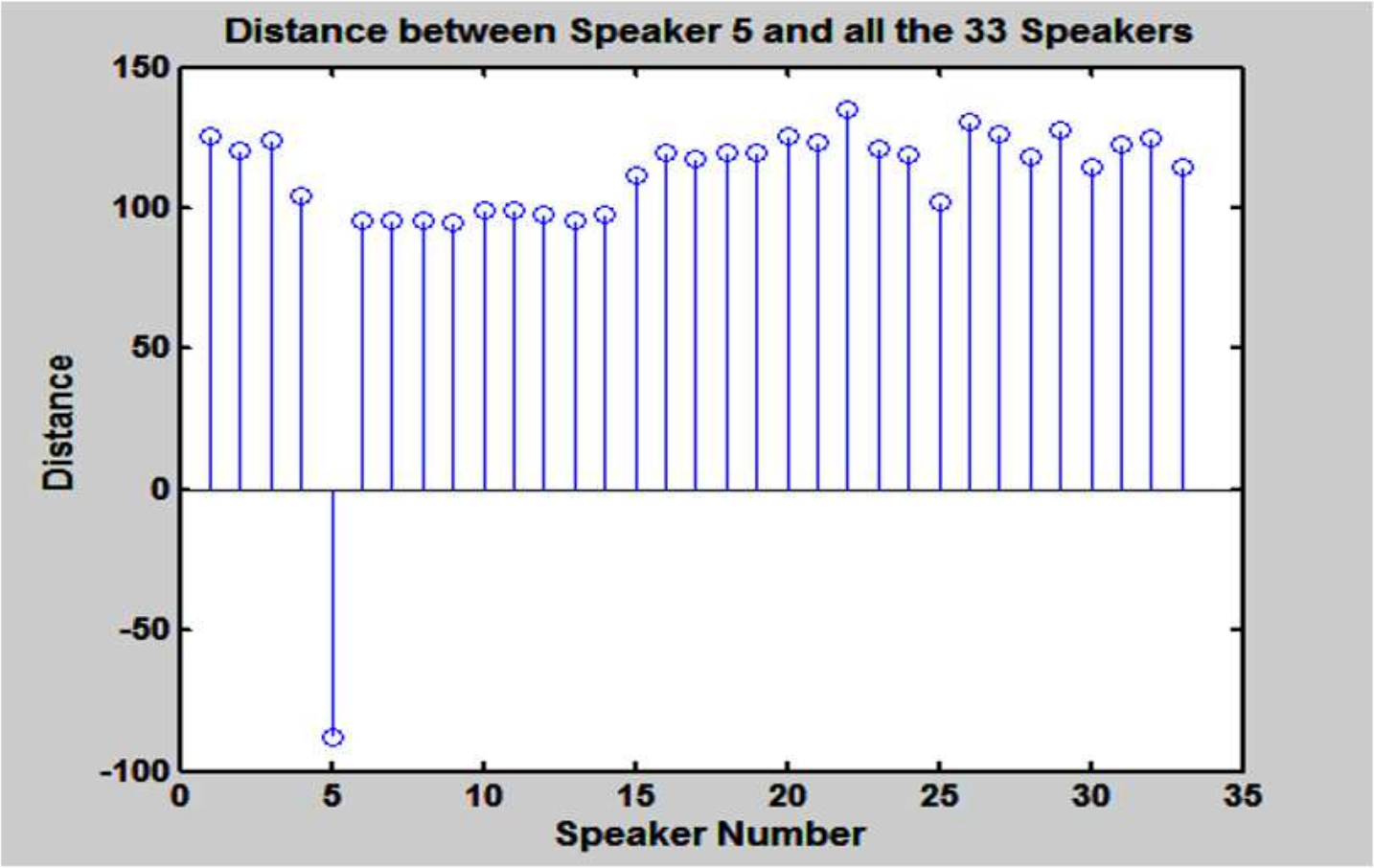

Figure 5

Output (MFCC vs Distance Metric Delta BIC)

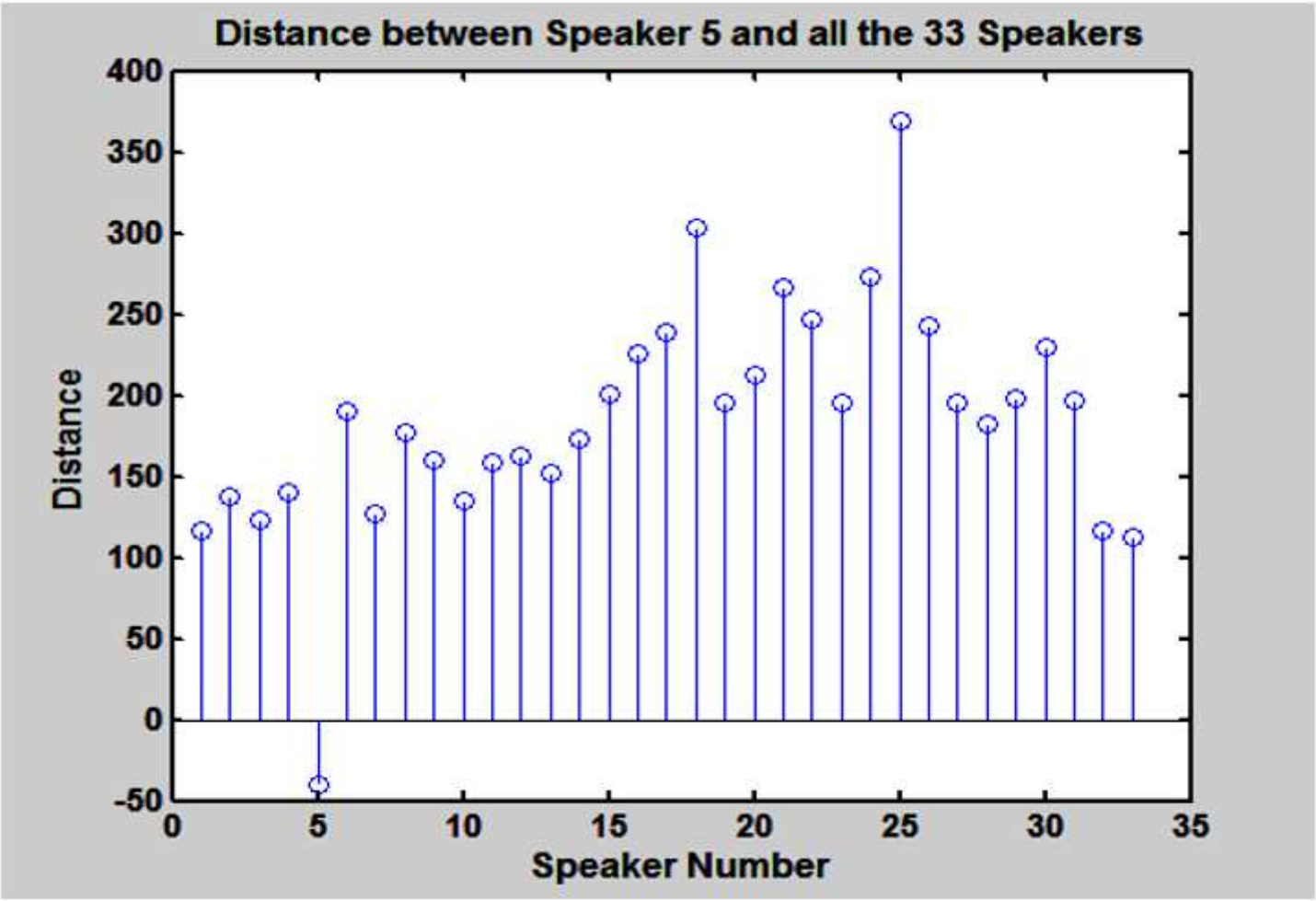

Figure 6

Output (Distance Metric Delta BIC vs proposed algorithm)

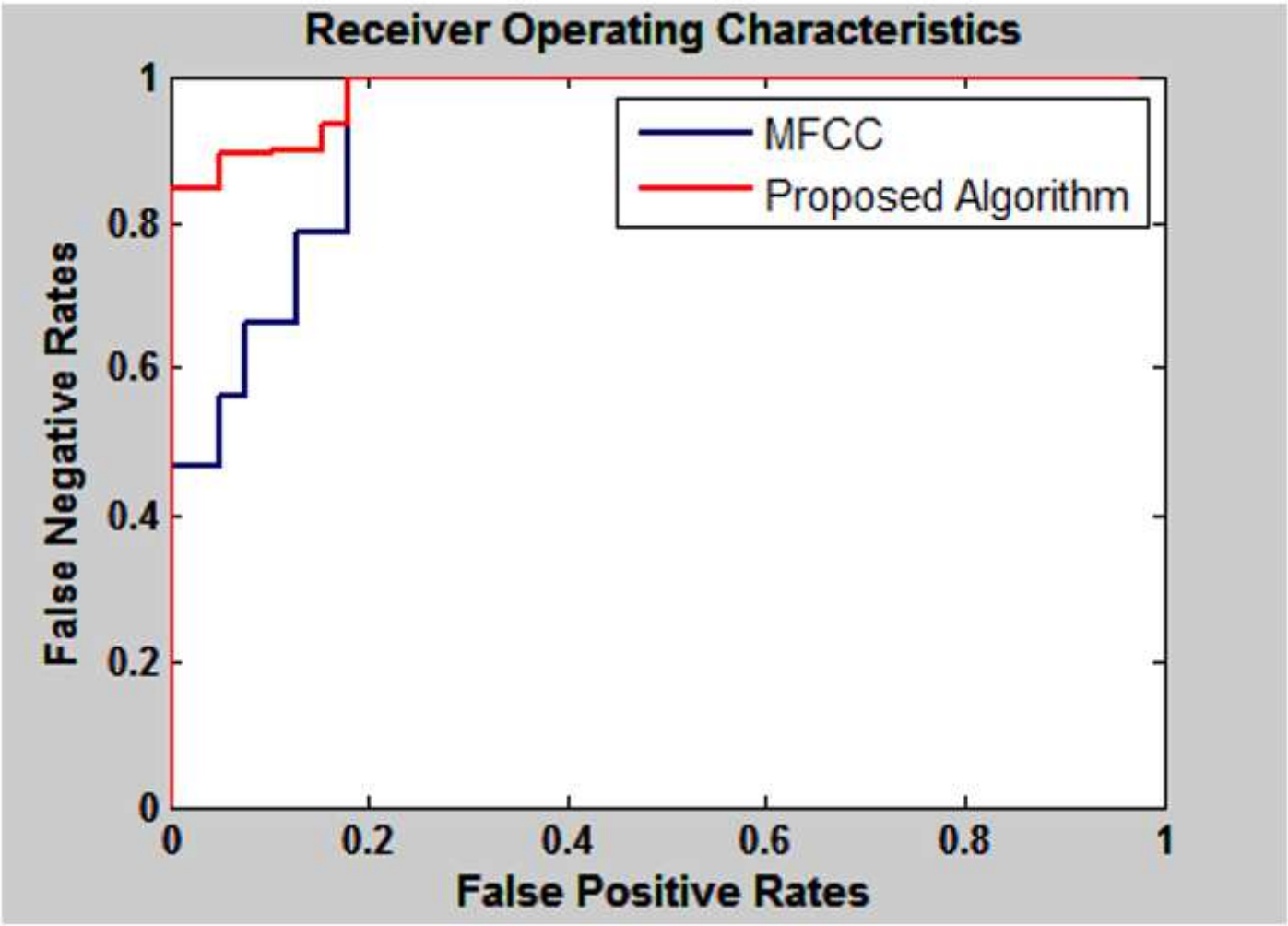

Figure 7

ROC curves for MFCC and proposed method using VSF

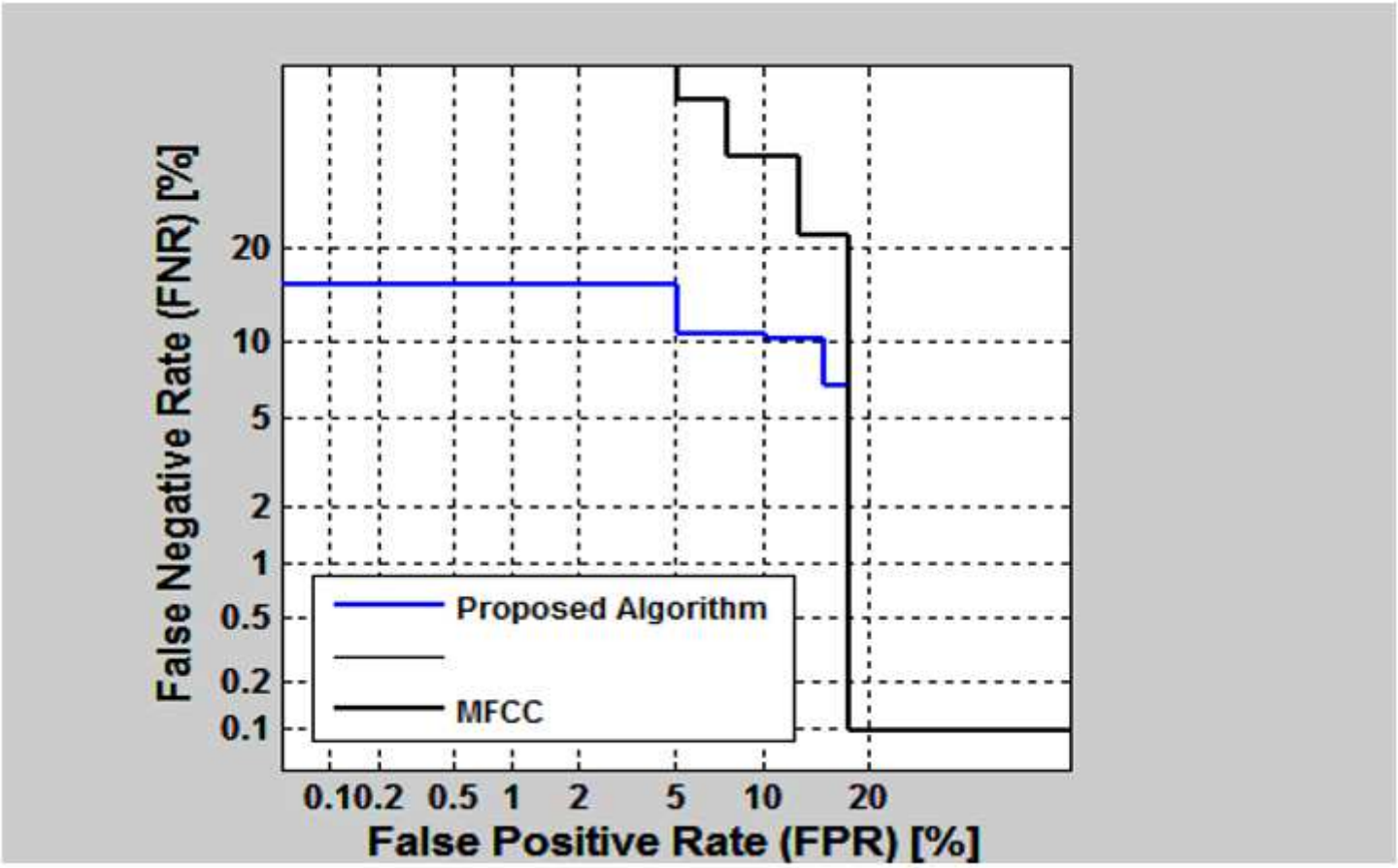

Figure 8

DET curves for BIC with MFCC and VSF algorithm.