Forensic Speaker Verification using Fractional MFCC in Noisy Environments

Mr.Mukesh D.Patil

Professor, Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra, India

Mr.Gajanan K.Birajdar

Professor, Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra, India

Abstract

Forensic speaker verification is the process of verifying the identity of a person with the available database of criminals. In forensics domain, the system may have to verify the identity of a person while talking on the telephone and in such scenarios environmental noise is inevitable leading to the false verification. Mel-frequency cepstrum coefficients (MFCC) is a widely employed feature descriptor in forensic speaker verification approaches and it achieves high performance under clean conditions. However, the performance of MFCC degrades notably in the noisy environment. To address this problem, use of fractional MFCC is proposed for forensic speaker verification in which, the Fourier transform (FT) and Discrete Cosine transform (DCT) are replaced by their fractional versions. The proposed algorithm is evaluated using the Australian Forensic Voice Comparison (AFVC) database in the presence of different types of environmental noises and reverberation condition and compared with other forensic speaker verification techniques.

Keywords Fractional Fourier Transform; MFCC · Fractional Discrete Cosine Transform · Australian Forensic Voice Comparison (AFVC) database · Forensic speaker verification.

1 Introduction

Speaker recognition is the identification task of a person and has been explored deeply over time. Speaker verification is a subset of speaker recognition and it is a process of validating the identity of a speaker by processing it's speech sample [1–3]. There are numerous applications of speaker verification in the domain of forensics and security [4]. Security forces and other government agencies which are responsible for the maintenance of the judiciary systems are prominent users of the speaker verification system [5]. In reality, the speaker verification system can be very helpful for such agencies for suspect verification with the available criminal database [6, 7].

Mel frequency cepstrum coefficients (MFCC) is most widely feature extraction approach in speaker verification in addition to the linear predictive coefficients (LPC) [3]. Traditionally, speaker verification algorithms are developed with clean speech signals without considering the noisy environment [6]. But in some scenarios like forensic speaker verification where avoiding noise is impossible because it is highly possible that a speaker is surrounded by the background noise and the system has to verify the speaker's identity. In a noisy environment, the system gets adversely affected and leads to degraded performance [6–9].

Solution to the problem of performance degradation due to the noise can be divided into three categories: (1) working on noisy signal (minimizing the noise) (2) making the feature extraction algorithm robust against the noisy environment and (3) modifying the classification strategy to achieve high accuracy (use of classifier fusion). In first case, the noise removal techniques are

implemented so that clean signal can be extracted and the rest of the process is carried on clean signal. An extra pre-processing would be needed in this case and it will make the system computationally expensive and time consuming. In the second case, sophisticated feature extraction algorithms are developed that will produce accurate and unique features capable of performing better even in the presence of noisy signals. Classifier is modified in the third case, specifically for the forensic speaker verification, i-vector is prominently used. Diken et al. [10] reviewed various strategies for feature extraction in the degraded condition for efficient recognition process.

MFCC cepstral features are developed based on the fundamentals of the human auditory system and its performance is better compared to noncepstral features [7, 10]. Although MFCC was proposed in 1980, it is still a popular choice in speaker verification and recognition processes because of its uniqueness. Recently it is successfully implemented for a wide range of applications including seizure detection [11], speech recognition [12], speaker recognition [13] and drone classification [14]. However, severe performance degradation occurs when the background noise is added in the signal as described in [15, 16]. The reason behind degradation is the masking of noise signal over the clean speech samples which leads to varying the mel-log power spectrum of clean and noisy signal [17]. As a result, the detection accuracy degrades. In the past, use of fractional Fourier transform instead of Fourier transform has enhanced the performance significantly in various applications [18–21].

Discrete wavelet transform is widely implemented in the domain of image processing but in recent years it is observed that it has promising effects is the domain of speech processing as well. Significant features can be extracted from the subbands which seems to be helpful for the verification process [7]. It solves the time frequency conflict which occurs in the Short time fourier transform [22]. As it has privilege to change the scales of time and frequency domains independently. Linear Prediction co-efficient can also be represented as Line Spectral frequencies (LSF) and has tried to gain the similar advantages of LSF which is observed in coding theory of reducing the bit error rate by approximately upto 25.

In this article, the problem of performance degradation due to the environmental noise and reverberation in forensic speaker verification process is addressed. Fractional MFCC based feature extraction approach is employed which is robust against the noise as compared to the conventional MFCC. The joint factor analysis (JFA) is used to create the total variability subspace for computation of i-vector after the MFCC feature extraction. To reduce the dimension of hyperparameters linear discriminative analysis (LDA) is used and then normalized by using probabilistic LDA (PLDA). Finally, length normalized GPLDA classifier is deployed to compute equal error rate. The proposed forensic speaker verification algorithm evaluation is carried out on Australian Forensic Voice Comparison (AFVC) database and for adding noise to clean signal QUT-Noise database is used. Fractional MFCC feature extraction technique with three different combinations are evaluated i.e. fractional Fourier transform (FrFT) with DCT, FFT with fractional DCT (FrDCT) and FrFT with FrDCT. The proposed fractional MFCC approach is compared with conventional MFCC, LSF and feature warping techniques. It is observed that robust feature extraction approach using fractional Fourier transform and fractional DCT improves the verification accuracy in the presence of noise and reverberation.

The paper is organized as follows: Section 2 describes fundamental concepts of fractional FT and DCT transforms. Additionally, MFCC and length normalized GPLDA classification is also outlined briefly. Section 3 comprises of proposed methodology of forensic speaker verification using

fractional MFCC. Simulation results and discussions are presented in Section 4. Finally, Section 5 concludes the article.

2 Preliminaries

This section describes fractional Fourier transform, Fractional DCT, MFCC and i-vector classification briefly.

2.1 Fractional Fourier Transform (FrFT)

Fractional Fourier Transform (FrFT) is the time-frequency representation of the signal and popularly explored in the domain of optics for developing filters in early 90's [23]. Fractional version of FT is obtained by shifting signal from (t, ω) to (u, v) where α denotes the degree of rotation.α can be defined as $a = \alpha\pi/2$ where α lies between 0 to 1 [19]).

The FrFT is defined as

$$f^{\alpha}(k) = \int_{-\infty}^{+\infty} k_{\emptyset}(t,u)x(t)\ dt \qquad (1)$$

where$k_{\emptyset}$ is defined as [23]:

$$k_{\emptyset}(t,u) = \begin{cases} \sqrt{\frac{1-jcot\emptyset}{2\pi}}\, e^{j\frac{t^2+u^2}{2}cot\emptyset - jutcsc\emptyset} & if\ \emptyset\ is\ not\ multiple\ of \pi \\ \delta(t-u) & if\ \emptyset\ is\ multiple\ of\ 2\pi \\ \delta(t+u) & if\ \emptyset+\pi\ is\ multiple\ of\ 2\pi \end{cases} \qquad (2)$$

2.2 Fractional discrete cosine transform (FrDCT)

Gianfranco Cariolaro et al. [24] investigated the extension of Discrete Cosine transform (DCT) with the fractional power. Let s_n be the sequence of length N. Then the DCT can be defined as [24]:

$$S_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} s_n \cos\left(2\pi \frac{(2n+1)k}{4N}\right) \qquad n = 0, \ldots . N-1$$

$$s_n = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \in_k S_K \cos\left(2\pi \frac{(2n+1)k}{4N}\right) \qquad n = 0, \ldots . N-1$$

(3)

where $\in_0$= 1 and $\in_k = \sqrt{2}$, for k ≥ 0. If we interpret the N-size sequences as column vectors s = [s, ..., s_{N-1}] ' and S = [S_0, ..., S_{N-1}] ' and denote the NxN DCT matrix by

$$C = \left\| \frac{1}{\sqrt{N}} \in_k S_K \cos\left(2\pi \frac{(2\pi-1)k}{4N}\right) \right\| \qquad (4)$$

eq. (3) gives the DCT of any signal by $S = Cs$and inverse can be obtained as $s = C^{-1}S$. Let C_α be a kernel operator for any given fraction $\alpha \in R$, that maps an N - size vector s into another N - size vector $S\alpha = C\alpha s$. C_α is a FrDCT operator and it satisfies the conditions as mentioned in [24].

2.3 Mel frequency cepstral coefficient (MFCC)

MFCC was proposed by S. Davis and P. Mermelstein [25] which is based on the perception of pitch in the human auditory system. It was observed that, for 0 to 1000 Hz the perception is linear and it rises nonlinearly with the rise in frequency. The power cepstrum is defined as the square of the modulus of the forward Fourier transform of the logarithm of the power spectrum of a signal

$$C_{pxx}(\tau) = |F\{log_{10}Gxx(\omega)\}|^2 \quad (5)$$

Mel filterbank is implemented on the cepstral coefficient of the audio signal to represent the MFCC features.

Linear scale to mel-scale frequency conversion is given as [28],

$$m = 1000 * log_2(1 + \frac{f}{1000}) \quad (6)$$

where the logarithm is being implemented above 1000 Hz frequency because below that the perception is linear.

Figure 1 depicts MFCC feature extraction steps. Speech sample is first pre-emphasised i.e high frequency components are boosted as it contains highly informative part of the speech. Then audio files are framed into 10 to 40 ms frames in order to obtain stationarity in the signal. The window is implemented on the short duration framed samples. The Fourier transform of windowed signal is computed and the Mel-filter bank is then generated. Further the triangular mel filter bank is multiplied with data which is transformed into frequency representation [26, 27]. Finally, the DCT is implemented on it so that the redundancies are being reduced and only real part of data is considered.

2.4 i-vector classifier and feature extraction

For speaker verification most commonly used classifier is i-vector classifier, proposed by Dehak et al. [29]. In this article, we have employed i-vector with length normalised Gaussian PLDA (GPLDA) classifier. Firstly, the Gaussian mixture model (GMM) creates the supervector for speech signals. Then, i-vector is used to lower the dimensionality of GMM supervector and it contains the speaker and channel variability. i-vector can be represented as [29]

$$s = m + T\omega \quad (7)$$

where m denotes mean supervector of universal background model (UBM), T is the low-rank matrix of variability in development data, and w is the i-vector. Initially heavy tailed PLDA was performing better than the GPLDA for speaker verification processes. Then length normalized GPLDA was introduced by the GarciaRomero and Espy-Wilson [30] in which the behavior of i-vectors were transformed from the heavy tailed to theGaussian as a result the performance was similar but with less complexity. Detailed description of i-vector and GPLDA classifier is explained in [7].

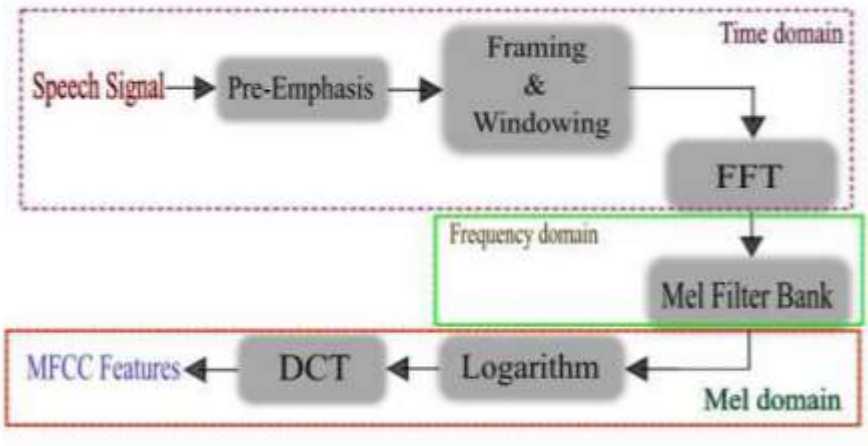

Fig. 1 MFCC feature extraction steps.

3 Proposed methodology

Forensic speaker verification using fractional FT and DCT is shown in Figure 2. The verification architecture is divided into three phases: input phase, feature extraction phase and classification phase. In the first phase, input is provided in the form of speech samples to the system. Mainly two types of speech signals are required, samples recorded in a closed room (clean environment) and samples recorded using telephone handsets (containing noise). In this work, aim is to develop robust verification algorithm in the presence of a noisy environment. The samples which are recorded using telephone handsets are added with noise and then the system is analyzed using noisy signal. The 70% of noisy samples are used for training the classifier and remaining are used for testing.

In the second phase, feature extraction process is carried out using fractional MFCC technique. Speech samples are pre-emphasized and then framed in to smaller segments. Hamming window with the length of the framed sample is implemented and then the FrFT of the windowed signal is computed as discussed above. The Mel filter bank using 32 triangular filters is implemented during the MFCC computation. Finally, in the last phase, these features are fed to the i-vector classifier for verification task. The detailed description of the classification stage is given in [7]. The performance of the proposed algorithm is evaluated using equal error rate (EER).

4 Simulation results and discussions

Australian Forensic Voice Comparison (AFVC) database provided by the Forensic Voice Comparison Laboratory, University of New South Wales, Sydney, Australia isused in this research work which is composed of 552 speakers [31, 32]. The recordings are carried out in three speaking styles: (a) pseudo-police style (b) informal telephone conversation and (c) exchange of information over telephone. In pseudo-police style, the speakers were interviewed in a studio with a clean environment depicting the interrogation in police custody. Other two styles are recorded using telephone handsets and conversation of two speakers on either side. The speech signals are sampled at a 44.1kHz frequency and 16 bit/sample resolution. The recordings are provided into small segments of 1 to 2 seconds long.

Another database employed is the QUT-Noise database to generate noisy samples. In this study, full duration pseudo-police style speech samples of 200 speakers are used to enroll the classifier and 10 seconds long samples of 200 speakers are used from informal conversation style for testing purpose.

Noise samples are down sampled to 44.1kHz to match the frequency with speech [7]. Before adding noise to speech samples Voice activity detection is implemented to remove the silent regions form the speech samples, so that the noise induced due to the non speech regions are avoided. Car type noise is added after segmenting the speech samples into 10 seconds long segment and at 5 levels of SNR from -10 to 10 with the step size of 5.

Three different feature extraction strategies are employed to study the effect of individual fractional transform on the performance of verification process. Three combinations are: (1) FrFT with DCT (2) FFT with FrDCT and (3) FrFT with FrDCT. By using such combinations, it is possible to distinguish between the effects of fractional transforms on the MFCC and eventually on the verification process. In all three combinations, the features were extracted with α starting from 0.5 to 1 with a step size of 0.05 and two intermediate values were taken i.e. 0.93 and 0.98. All other classifier parameters are set as described in [7]. For the implementation, we have used a GPU workstation with Xeon processor, 2TB HDD, 16GB DDR4 RAM, 2-K80 NVIDIA GPU processor.

4.1 Non-Reverberated

4.1.1 Analysis using FrFT and DCT

First, the verification performance is analyzed by combining FrFT and DCT. Experimental results interms of EER at different SNR values are depicted as shown in Figure 3. A common trend is observed in all the combinations where the maximum EER is observed for α equals to 0.5 and then it starts decreasing gradually and optimum results are obtained at α = 0.93. While working in fractional domain most important task is to obtain the optimum value of α suitable for operation and type of input signal available. A standard order of fractional transform which will be suitable for all application does not exist. It depends on several factors such as type of signal, type of speakers and several others. In our case the classification was done on 200 speakers data and the efficient results are obtained for α equal to 0.93.

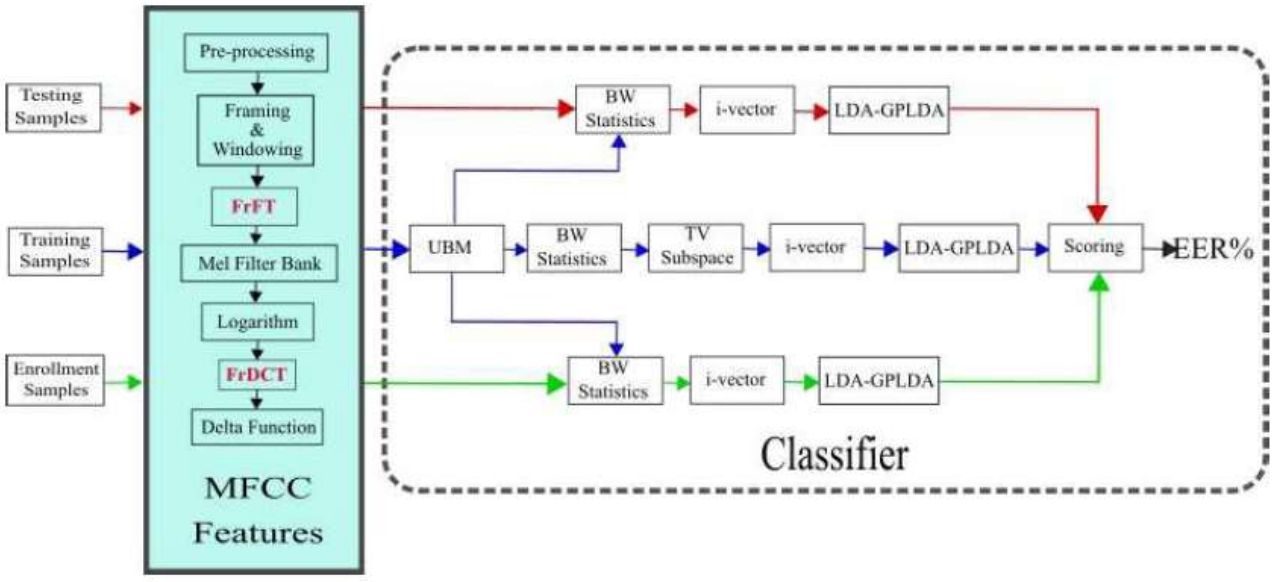

Fig. 2 Forensic speaker verification using fractional FT and DCT.

4.1.2 Analysis using FT and FrDCT

In the second experiment, combination of FT with FrDCT is evaluated. The verification rateis decreased in this case. The performance study shows that implementation of MFCC infractional FT domain gives significant enhancement in the process of speaker verification compared to

FrDCTalone. From Figure 4, it can be observed that FrFT has more significant impact on MFCC than FrDCT. In the process of speaker verification, highest accuracy is obtained when the long duration of speaker data is available as it gives maximum possibility in which a person can speak or it covers the maximum variation in a person's speaking style. For every speaker, the system forms a cluster of features which is used for matching and verifying the testing persons data. When MFCC is implemented in FrFT domain, the cluster forming is more accurate and efficient than the conventional.

4.1.3 Analysis using FrFT and FrDCT

In this experiment, results are shown for the combination of FrFT with FrDCT as shown in the Table 1. The advantages of both fractional transforms are incorporated in this type and as a result the highest enhancement is observed. The level of noise plays a vital role during the performance evaluation of MFCC, for low SNR the results are degraded throughout the fractional orders but as the signal power starts increasing the performance also enhances. The computational complexity of FrFT is M $\log_2$ M where M = 2P, P is the length of signal whereas FrDCT has P $\log_2$ P. The time required for computation of FrFT is also higher than the FrDCT but if the system has to run in real time scenario then the system would have to extract feature of a single speaker and in that case the delay will be negligible. So this makes the system suitable for the real time implementation.

4.2 Analysis using MFCC and DWT

Feature warping reduces the nonlinearity by distribution mapping to the standard deviation. Steps for the feature warping is given in the [7]. When feature warping technique is implemented on the extracted features the notable enhancement is observed, but it does not shows the uniformity for the different techniques. When the same experimental setup is used to investigate the effect of fractional transforms, the obtained result follows same pattern as the conventional transforms provide. In MFCC-FW the lowest error rate isobserved as compared to the DWT and DWT-FW shown in Figure 5 and 6. For different combinations of the fractional transform results are evaluated and the highest accuracy is achieved for the combination of FRFT-FRDCT as the benefits of both fractional transforms are incorporated in the results.

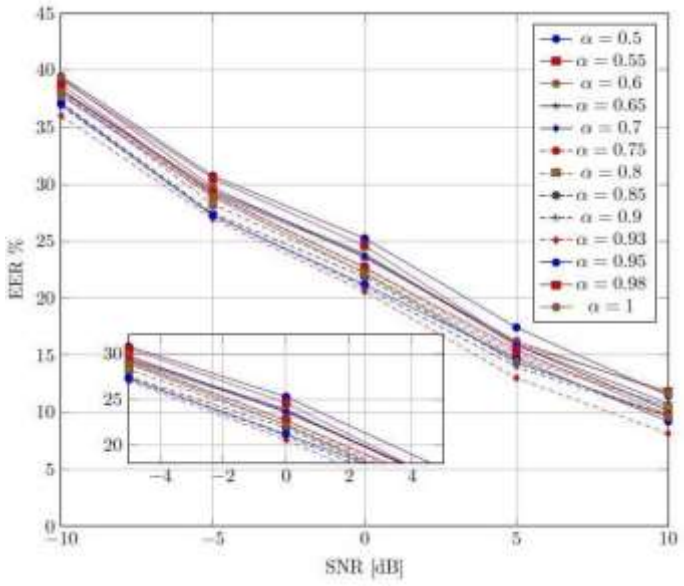

Fig. 3 Forensic speaker verification using FrFT and DCT in terms of equal error rate with varying SNR.

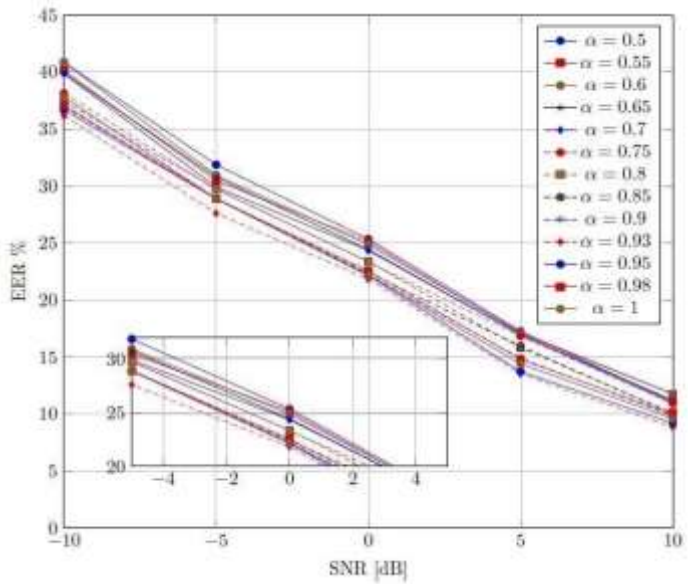

Fig. 4 Performance evaluation of system using Fourier transform and fractional DCT.

Results shown in Table 2 and Figure 7 are given for Fusion of MFCC-DWT andFusion feature warping with the conventional and fractional transforms for the order of α = 0.93 respectively. When the results of Fusion and Fusion-FW are compared we can observe that FW have major effect on the system as it reduces the EER with higher margin. But when this algorithm is implemented with the fractional version of transforms, though EER is reduced but the same trend is followed as previous results. In Fusion-FW technique we can achieve the lowest EER but this enhancement comes with the price of high computationally complex which leads to the increase in time consumption.

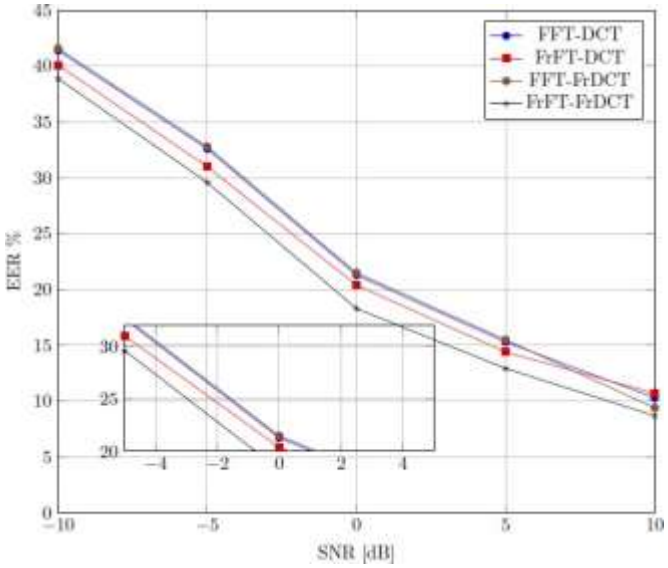

Fig. 5 Performance evaluation of system using Discrete wavelet transform.

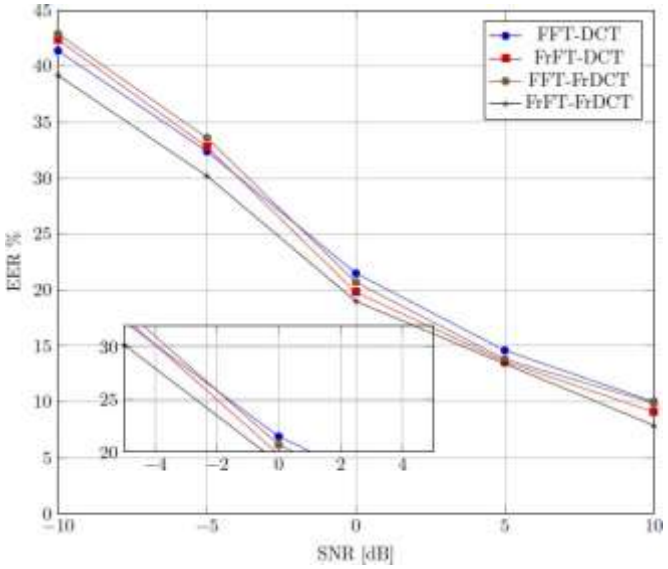

Fig. 6 Performance evaluation of system using DWT Feature Warping.

4.3 Reverberation effect on forensic speakerverification

In this section reverberation conditions are included i.e. while recording the enrollment speech it is highly probable that the recording is done in the closed room by using normal mic. Reverberation is the combination of constructive and destructive interference of the sound produced by the speaker occurred due to the closed room. In this conditions, the source of speech no longer remains the only speaker but the reflected sound also gets recorded with a specific time delay and causes thespeech hard to verify. The reverberation is added to the enrollment speech and the testing speech is kept untouched. Results are evaluated for the MFCC-FW and FUSION-FW combinations. In highly noisy environment reverberated speech is helpful for the verification algorithm but as the noise starts reducing from the speech the EER drops up to 10% in case of MFCC-FW (Table 3 and 6.4% in case of Fusion-FW.

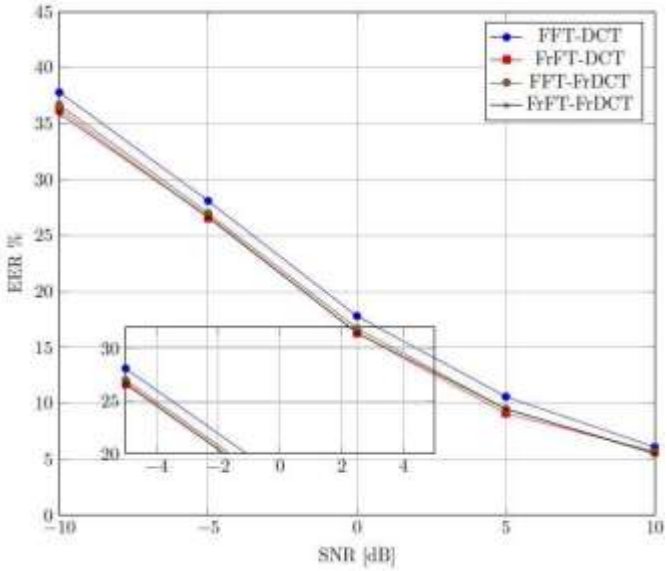

Fig. 7 Performance evaluation of system using fusion MFCC and DWT.

Table 1 Results of 200 speaker data with the combination of FrFT and FrDCT.

SNR [dB]	-10	-5	0	5	10

α=0.5	40.35	30.77	25.86	17.62	10.88
α=0.55	39.53	30.09	25.58	17.72	10.85
α=0.6	39.21	29.83	24.25	15.12	10.69
α=0.65	38.36	29.91	24.92	16.23	9.90
α=0.7	38.33	29.77	23.55	16.56	9.95
α=0.75	37.01	28.08	23.58	15.19	9.72
α=0.8	37.37	28.04	23.16	15.36	9.88
α=0.85	37.25	27.17	22.73	15.73	8.08
α=0.9	36.76	27.36	21.92	14.84	7.91
α=0.93	35.92	26.67	19.97	12.19	7.10
α=0.95	36.06	26.39	20.21	13.53	8.01
α=0.98	37.28	27.57	21.39	14.82	8.24
α=1	37.80	28.90	22.20	14.40	9.60

Table 2 Comparison of results of Fusion FW for FFT-DCT, FrFT-DCT, FFT-FrDCT and FrFT-FrDCT.

	SNR [dB]				
	-10	-5	0	5	10
FFT-DCT	33.3	25.8	17.7	6.2	3.5
FRFT-DCT	31.44	23.94	15.84	4.34	1.64
FFT-FRDCT	31.78	24.28	16.18	4.68	1.98
FRFT-FRDCT	31	23.9	16.4	4.7	2.6

4.4 Discussions

In this article, the problem of performance degradation due to the noisy environment forforensic speaker verification is addressed. The presence of noise in forensic speakerverification is unavoidable, so this study is focused on building a robust feature extractionapproach capable of performing better in a noisy environment. It is evident from aboveresults that, in MFCC andDWT feature extraction algorithm, if the transforms (FT and DCT) are replaced with its fractional versions,then the robustness of features extraction algorithmsare improved against noisy condition.

Table 3 Evaluation of reverberation e_ect with Fusion FW using FFT-DCT, FrFT-DCT, FFT-FrDCT and FrFTFrDCT.

	SNR [dB]				
	-10	-5	0	5	10
FFT-DCT	27.1	20.6	14.4	9.7	6.4
FRFT-DCT	25.11	18.61	13.01	8.51	5.41
FFT-FRDCT	26.29	19.79	13.59	8.89	5.59
FRFT-FRDCT	24.67	18.37	12.47	8.27	5.07

Robustness is dependent on the value of α and choosing the optimum α is a crucial task. While implementing the system in a real time environment, it is a difficult task to find the efficient α for a speaker, as in such cases the verification system will be verifying a single person at a time. Since the proposed algorithm is experimented with large dataset, the satisfactory value of α is obtained

resulting in better verification performance.

Table 4 shows the comparison of verification results obtained using FFT-DCT and FrFT-FrDCT with α = 0.93. During the experimentation, same set of α values are chosen for both the transforms and it is observed that the efficient results are obtained at a similar set of α value. In future, the research work can be extended to study the effect of combination of α's on MFCC. In comparison with both the transforms, FrFT has more significant effect on MFCC feature extraction as the former transforms the signal including with the noise.

Table 4 Comparison of results for FFT-DCT combination with FrFT-FrDCT.

	SNR [dB]				
	-10	-5	0	5	10
FrFT-FrDCT (_=0.93)	31	23.9	16.4	4.7	2.6
FFT-DCT	33.3	25.8	17.7	6.2	3.5

The proposed fractional MFCC based forensic speaker verification technique is compared with conventional MFCC based verification approach as described in [7]. All the experimental settings along with classifier parameters are set as defined in [7]. Table 5 shows a comparison of conventional and fractional Fusion feature warping approach. In comparison with the conventional MFCC system, enhanced performance is obtained when both the fractional transforms are implemented simultaneously with the value of α equals to 0.93. For SNR equal to 10dB the system yielded highest improvement.

Table 5 Comparison of results for conventional features with Fractional features.

	SNR [dB]				
	-10	-5	0	5	10
LSF	28.6	17.1	9.4	8.4	7.1
LSF(Reverberated)	34.2	22.7	13.1	11.5	9.0
MFCC	37.80	28.90	22.20	14.40	9.60
MFCC-FW	27.2	21.2	14.8	10.1	6.4
MFCC-FW (Reverberated)	38.9	33.2	25.6	20.2	15.4
DWT	41.4	32.4	21.5	14.6	10.01
DWT-FW	41.4	32.6	21.3	15.3	10.3
Fusion	37.8	28.1	17.8	10.6	11.1
Fusion-FW	33.3	25.8	17.7	6.2	3.5
Fusion-FW(Reverberated)	27.1	20.6	14.4	9.7	6.4
MFCC(FrFT-FrDCT)	35.92	26.67	19.97	12.19	7.10
Fusion (FrFT-FrDCT)	31.2	23.9	16.4	4.7	2.6

In Table 5, we have shown the results for features with the conventional transform and fractional trans- form. The performance of DWT features is worse as compared to all other feature extraction techniques and also feature warping couldn't improve the results. As compared to the DWT, MFCC has performed better both in a heavily degraded condition and less noisy condition of speech. Most enhanced results are obtained for the Fusion of MFCC and DWT with the feature warping as it has increased the uniqueness of speech samples sufficient larger that EER has dropped to the least. LSF seems to have constant behavior over 0dB SNR and in noisy state degradation of performance is

drastic. It is quite clear that fractional transform enhances the performance of features in a noisy environment. Fractional transforms are very helpful under noisy condition but as soon as the SNR starts improving the downfall is observed in enhancing performance. As the main function of fractional transforms is to separate out the signal from the noise the observed behavior is justified.

Forensic speaker verification is a critical task in which the accuracy is of prime importance. As Mel-frequency cepstral coefficient feature extraction technique performs poor in the presence of environmental noise, fractional Fourier transform and fractional discrete cosine trans- form is employed resulting in enhanced verification performance. FrFT and FrDCT transform in MFCC and DWT feature extraction technique with three different combinations are evaluated i.e.FrFT with DCT, FFT with FrDCT and FrFT with FrDCT. Best verification rate is reported for Fusion of MFCC and DWT with the α value of 0.93. It is also found that, best performance is obtained using the combination of FrFT with FrDCT followed by FrFT with DCT and FFT with FrDCT.

References

1. S. Furui (1997) `Recent advances in speaker recognition'Pattern Recognition Letters, Vol. 18, No. 9, pp.859-872.
2. Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre,and H. Li (2015) `Spoofing and countermeasures for speaker verification: A survey' Speech Communication, Vol.
66, pp.130-153.
3. Das, Rohan Kumar, and SR Mahadeva Prasanna (2018) 'Speaker verification from short utterance perspective: a review'IETE Technical Review, Vol.35, No.6, pp.599-617.
4. D. A. Reynolds (2002) `An overview of automaticspeaker recognition technology' Proc. IEEE InternationalConference on Acoustics, Speech, and Signal Processing(ICASSP), Vol. 4, pp.4072-4075.
5. J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf (2009) 'Forensic speaker recognition' IEEE Signal Processing Magazine, Vol. 26, No. 2, pp.95–103
6. M. I. Mandasari, M. McLaren, and D. A. van Leeuwen (2012) 'The effect of noise on modern automatic speaker recognition systems', Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.4249–4252.
7. Al-Ali, Ahmed Kamil Hasan, David Dean, BouchraSenadji, Vinod Chandran, and Ganesh Naik (2017) 'Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions' IEEE Access, Vol. 5, pp.15400–15413.
8. J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds (2007) 'Robust speaker recognition in noisy conditions' IEEE Transactions on Audio, Speech, Language Processing, Vol. 15, No. 5, pp.1711–1723.
9. S. Kim, M. Ji, and H. Kim (2010) 'Robust speaker recognition based on filtering in autocorrelation domain and sub- band feature recombination' Pattern Recognition Letters, Vol. 31, No. 7, pp.593–599.
10. Dişken, G., Tüfekçi, Z., Saribulut, L., &Çevik, U. (2017).'A review on feature extraction for speaker recognition under degraded conditions' IETE Technical Review, Vol. 34, No. 3, pp.321–332.
11. Yavuz, E., Kasapbaşı, M. C., Eyüpoğlu, C., &Yazıcı, R. (2018) 'An epileptic seizure detection system based on cepstral analysis and generalized regressionneural network' Biocybernetics and Biomedical Engineering, Vol. 38, No. 2, pp.201-216.
12. Kumar Archek Praveen, Ratnadeep Roy, Sanyog Rawat,Ashwani Kumar Yadav, Amit Chaurasia, and Raj Ku- mar Gupta (2018) 'Telugu Speech Recognition Using Combined MFCC, MODGDF Feature Extraction Techniques and MLP, TLRN Classifiers' In Soft Computing: Theories and Applications, pp. 687–696.

13. Jokinen Emma, Rahim Saeidi, Tomi Kinnunen, andPaavo Alku (2019) 'Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task' Computer Speech & Language, Vol. 53, 1–11.
14. Shi Lin, Ishtiaq Ahmad, YuJing He, and KyungHi Chang(2018) 'Hidden Markov model based drone sound recognition using MFCC technique in practical noisy environments' Journal of Communications and Networks, Vol. 20, No. 5, pp.509–518.
15. Viikki Olli and Kari Laurila (1998) 'Cepstral domain segmental feature vector normalization for noise robust speech recognition' Speech Communication, Vol.25, No. 3, pp.133– 147.
16. Chang Gwo-Ching and Yung-Fa Lai (2010) 'Performanceevaluation and enhancement of lung sound recognition sys- tem in two real noisy environments' Computer methods and programs in Biomedicine, Vol.97, No. 2, pp.141–150.
17. Yadav Ishwar Chandra, S. Shahnawazuddin and Gaydhar Pradhan (2019) 'Addressing noise and pitch sensitivity of speech recognition system through variational mode decomposition based spectral smoothing' Digital Signal Pro- cessing, Vol. 86, No.3, pp.55–64.
18. Pan Xinyu, Heming Zhao, and Yan Zhou (2015) 'Theapplication of fractional Mel cepstral coefficient in deceptive speech detection' PeerJ, Vol.3, pp.e1194–1208.
19. BhalkeDaulappaGuranna, Betsy Rajesh, and DattatrayaBormane (2017) 'Automatic Genre Classification Using Fractional Fourier Transform Based Mel Frequency Cepstral Coefficient and Timbral Features' Archives of Acoustics, Vol.42, No. 2, pp.213–222
20. Birajdar Gajanan, Vishwesh Vyawahare, and MukeshPatil (2018) 'Secure and Robust ECG Steganography Using Fractional Fourier Transform' Cryptographic and Information Security Approaches for Images and Videos, pp.541– 570.
21. Zhuo, Z. (2018) 'Novel image watermarking methodbased on FRWT and SVD', International Journal of Electronic Security and Digital Forensics, Vol. 10, No. 1, pp.97-107.
22. Tzanetakis George, Georg Essl, and Perry Cook (2001) 'Audio analysis using the discrete wavelet transform' Proc. of WSES International Conference in Acoustics and Music Theory Applications, Vol. 66, pp.1–6
23. Haldun M. Ozaktas, M. AlperKutay, and DavidMendlovic (1999) 'Introduction to the Fractional Fourier Transform and Its Applications' Advances in Imaging and Electron Physics, Vol. 106, pp.239–291
24. Cariolaro Gianfranco, Tomaso Erseghe, and PeterKraniauskas (2002) 'The fractional discrete cosine transform' IEEE Transactions on Signal Processing, Vol.50, No. 4, pp.902–911
25. Steven Davis and Paul Mermelstein (1980) 'Comparisonofparametric representations for monosyllabic word recognition in continuously spoken sentences' IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 28, No. 4, pp.357—366
26. Jeyalakshmi, C. and Revathi, A. (2018) 'Efficient speechrecognition system for hearing impaired children in classical Tamil language', Int. J. Biomedical Engineering and Technology, Vol. 26, No. 1, pp.84—100.
27. Palo, H.K., Chandra, M. and Mohanty, M.N. (2017) 'Emotion recognition using MLP and GMM for Oriya language', Int. J. Computational Vision and Robotics, Vol. 7, No. 4, pp.426—442.
28. Logan, Beth (2000) 'Mel Frequency Cepstral Coefficients for Music Modeling' International Symposium on Music In- formation Retrieval (ISMIR), Vol. 270, pp.1–11.
29. DehakNajim, Patrick J. Kenny, R´edaDehak, Pierre Dumouchel, and Pierre Ouellet (2011) 'Front-end factor analysis for speaker verification' IEEE Transactions on Audio, Speech, and Language Processing, Vol.19, No. 4, pp.788–798.
30. Garcia-Romero Daniel, and Carol Espy-Wilson (2011) 'Analysis of i-vector length normalization in speaker recognition systems' INTERSPEECH, pp.249–252
31. Morrison, Geoffrey Stewart, Philip Rose, and Cuiling Zhang (2012) 'Protocol for the collection of databases of recordings for forensic-voice-comparison research and prac- tice' Australian Journal of Forensic Sciences, Vol.44, No. 2, pp.155–167

32. Morrison G.S., Zhang C., Enzinger E., Ochoa F., Bleach D., Johnson M., Folkes B.K., De Souza S. and Cummins N., Chow D. (2015) 'Forensic database of voice recordings of 500+ Australian English speakers' http://databases.forensic-voice-comparison.net