International Journal of Engineering & Technology, 7 (2.8) (2018) 278-283

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Performance of speaker recognition system using shifted mfcc, delta spectral cepstral coefficient (DSCC) and Fuzzy techniques

Priyanka Bansal [1*], Syed Akhtar Imam [2]

[1]Research Scholar, Department of ECE,Jamia Millia Islamia, New Delhi, India
[1]Assistant Professor, Department of ECE, Manav Rachna International University, Faridabad
[2]Associate Professor, Department of ECE, Jamia Millia Islamia, New Delhi, India
**Corresponding author E-mail: priyankabansalchamp@gmail.com*

Abstract

Speech and speaker recognition systems are biometric inspired systems which are having scope in various online and offline applications. In case of biometric we ponder the variability of speech signal due to the presence of noise which greatly degrades the efficiency of Automatic Speaker Recognition (ASR) in real-world environmental circumstances. Real world speech signal is degraded by different types of noise signals like background noise, interference noise and crosstalk noise. In this paper, we have used Delta Spectrum Cepstrum Coefficient (DSCC) and Shifted MFCC with fuzzy modeling techniques to rectify the deed of ASR even in a noisy surrounding with the help of upgraded speech information which is present at high frequency in the spectral domain. The combination of fuzzy modeling and DSCC creates a firm cumulative algorithm which has reasonably high robustness to noise. Experimental results show that accuracy has enhanced by 10-20% even at 5-8dB SNR in the presence of background noise or turbulent environmental condition or in the presence of white noise. Thus proposed model has improved maturity level in comparison to obsolete methods.

Keywords: *Delta Spectrum Cepstral Coefficient (DSCC), Discrete Cosine Transform (DCT), Mel Frequency Cepstral Coefficient (MFCC), Support Vector Classifier (SVC).*

1. Introduction

Speaker acknowledgment on the basis of earthborn speech nowadays became a vast scope of interest with many explications due to the evolution of methods such as support vector classifier (SVC) with HMM , independent component analysis, GMM and much more, which conventionally uses MFCC for acoustic signal characteristic lineage. Current possession of art real-time speaker recognition (ASR) system furnish very high accuracy in the restrained surrounding when the acoustic signal is very clean, but practically the acoustical surrounding is far less amiable. Bansal P.(2017) states that the surrounding conditions include noise and repercussion within which ASR system are generally deployed, which turns out the efficiency of ASR to be poor. In our paper, we accepted multiple motivations to increase the efficiency, robustness, and accuracy of ASR by adopting shifted MFCC and delta spectral cepstral coefficient with fuzzy c-means clustering. Actually, fuzzy advances to intellectual task like speaker/speech recognition and provide many contributions to fuzzy Automatic Speaker Recognition. In general, fuzzy logic is implemented due to its high-level decision-making ability by mean of integrating acoustic data from different sources.

2. Related Work

Various preprocessing methods, feature generation methods and classifiers were suggested by the researchers to improve the rate of speaker recognition. Anzar S.M. (2016) reduced the error rate in speaker classification by adapting the inter-class variation, online and offline template updation method using vector quantization. GMM (Gaussian Mixture model) and vector quantization (VQ) were processed collectively for improving the semi-supervised learning methods. Huang Z. (2016) applied the unified learning using deep neural network to improve the performance of speaker recognition. The inter-speaker variability based knowledge transfer was defined to investigate the speaker adaptation. The unified processing model was adopted as DNN structure for effective recognition of speaker. The multi condition evaluation was provided to handle the acoustic variability for consistent and statistical evaluation.

3. Parameters and Model Used

To enhance the accuracy of the system various feature extraction techniques are used which includes MFCC, shifted MFCC and Delta Spectral Cepstral Coefficient. In the past, MFCC was the most traditional feature extraction technique possessing windowing of the signal in frequency range 130-1500HZ whereas shifted MFCC includes window ranging from 1500-2000 HZ just as to contribute more accent features, likewise delta spectral cepstral coefficient is adopted instead of delta or double delta cepstral coefficient to replace the domain with the spectral domain which provides robustness and accuracy. Aurora database proposed by Hans D.(2000) is used in our study.

MFCC

MFCC is one of the widely used feature extraction technique because it works on the concept of acoustic features. Our ear responds to sound waves in a logarithmic fashion. MFCC works on Mel unit which gives the measure of perceived throng/pitch and corresponds to nonlinear/ logarithmic scale as our auditory system perceive pitch in the same manner. This scale is linear up to 1 KHz but logarithmic above 1 KHz. MFCC adopts the functional concept of humanistic cognition sensitivity with respect to frequency/pitch. To calculate MFCC, the continuous speech signal is broken down into small frames of size 30ms with a delay of 10 ms between two consecutive frames, then edges of signal are smoothed out with the help of unequal spaced triangular filter. Mukherjee R. (2013) proposed that after smoothing of the signal FFT is performed, and then DCT is used to change the frequency domain into quefrency domain (time domain). The resultant feature is in the form of cepstrum due to which it is known as Mel frequency cepstrum coefficient.

Shifted MFCC

In MFCC technique, the windowing is done in the frequency range from 200-1500HZ using triangular filters but studies show that many accent features are covered by high-frequency range between 1500-200HZ. To gather more information, the window is shifted from range 200-1500 HZ to 1500-2000 HZ so that additional accent features can be extracted. These features are added with the feature extracted using MFCC technique.

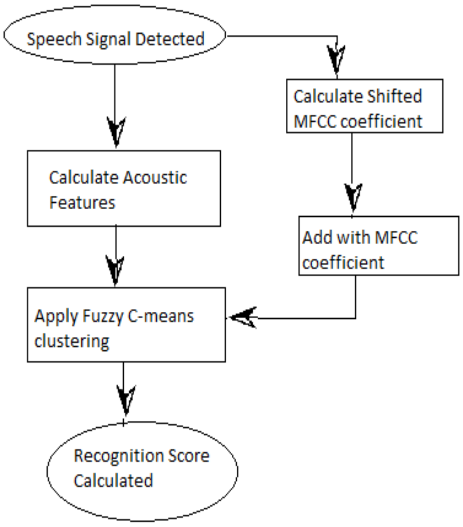

Speech Signal Detected
Calculate Shifted MFCC coefficient
Calculate Acoustic Features
Add with MFCC coefficient
Apply Fuzzy C-means clustering
Recognition Score Calculated

Fig. 1: Shifted MFCC

4. Delta Spectral Cepstrum Coefficient

Now we discuss the delta spectral cepstral coefficient approach for ASR. The motivation behind DSCC feature is non-stationary nature of the human speech signal wherein the short-period power of acoustic signal varies more rapidly than the short-period power of noise signal, noise is automatically get added due to variant environmental condition stated by Kumar Kshitiz (2011). The human ear can easily ignore the additional noise due to a huge difference between the rate of change of power of actual speech signal containing information and rate of change of power of noise signal.

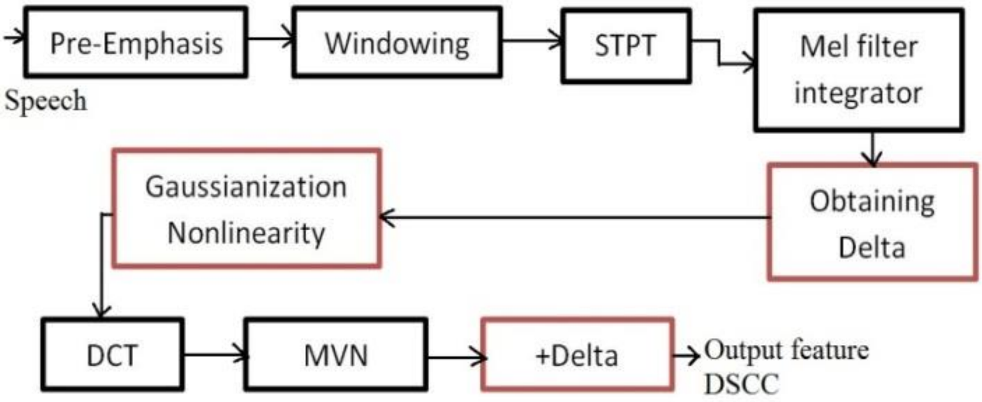
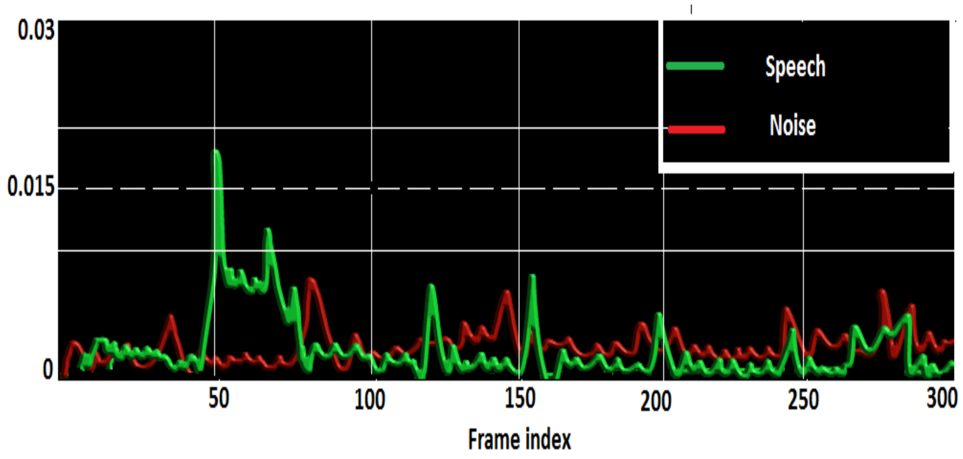


Fig. 2: Processing outline of DSCC

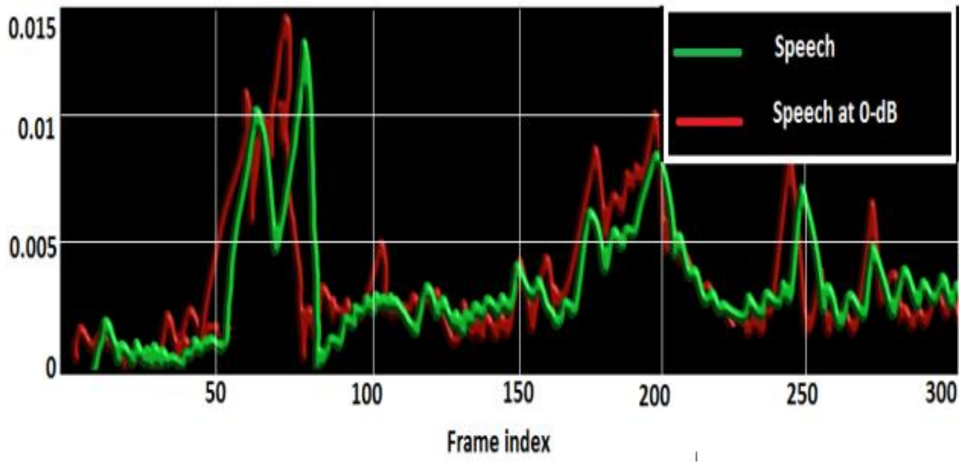
The proposed outline of delta spectral cepstral coefficient (DSCC) is explained in Fig 2. First delta operation is performed immediately after Mel filter integration after then Gaussianization operation creates a small mismatch between clean and noisy features but just after second delta operation, high mismatch is generated between clean and noisy features which makes DSCC robust towards noise signal. The delta cepstral feature for a short period cepstral sequence $C_s[n]$ is defined as

$$D_s[n] = C_s[n+m] - C_s[n-m] \quad (1)$$

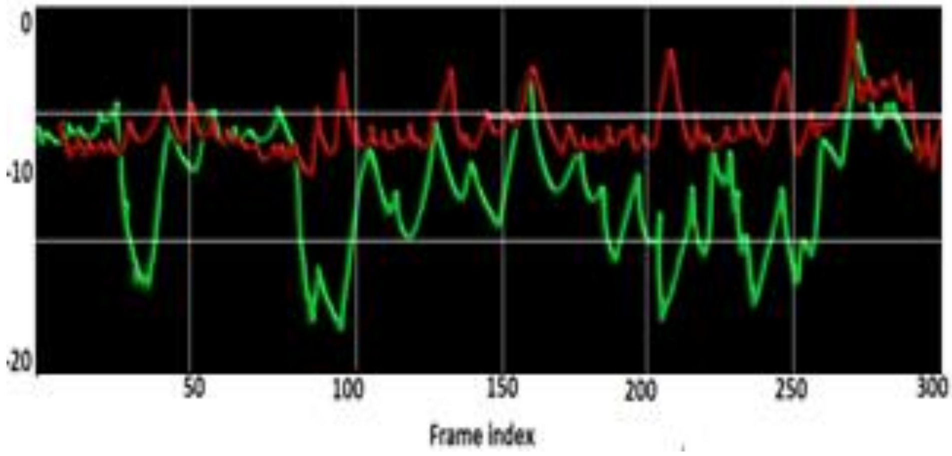
Where n represents the analysis frame index and is considered as 2 or 3 in practice. Significantly, delta operation is performed in equation (1) enhance the more variant speech signal and suppress the less variant noise in the spectral domain.


0.03
0.015
0
Speech
Noise
50
100
150
200
250
300
Frame index

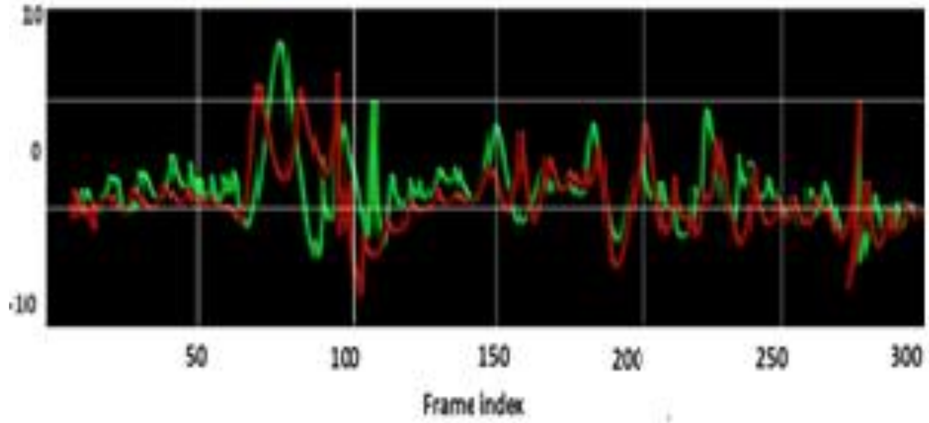
(a)


0.015
0.01
0.005
0
Speech
Speech at 0-dB
50
100
150
200
250
300
Frame index

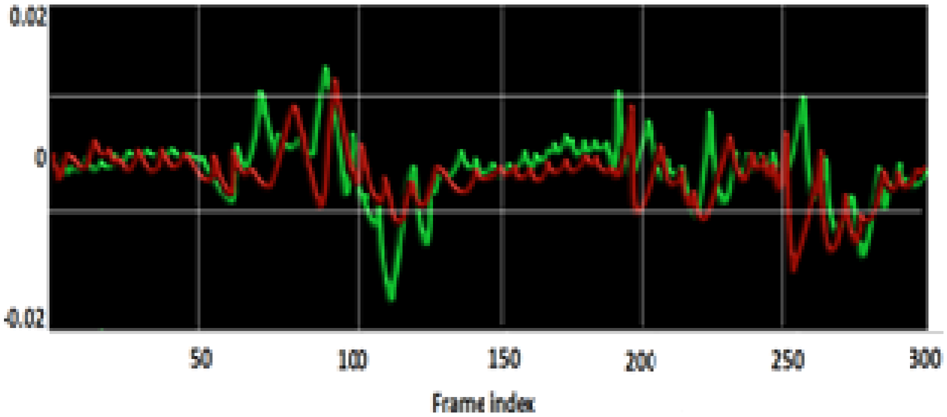
(b)


0
-10
-20
50
100
150
200
250
300
Frame index

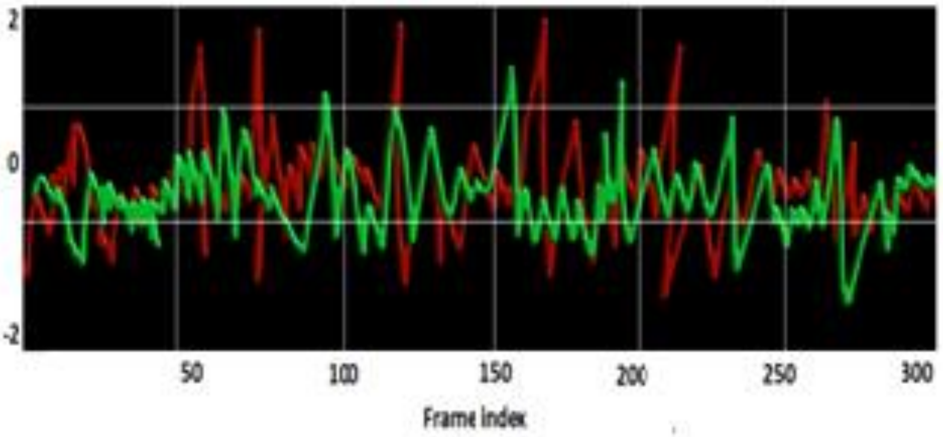
(c)


10
0
-10
50
100
150
200
250
300
Frame index

(d)



(e)



(f)

Fig. 3: (a) Mel channel short-period power(1KHz) center frequency of speech and a "real-environmental" noise signal segment(10ms frames). (b) short period power of actual speech in ideal condition and speech in "real-environmental" noise signal at 0-dB (c) logarithmic form of signals as in (b).(d) Differentiation operation applied to the speech and noise signals in (c). (e) Plots temporal operation over the speech segment in (b). (f) Plots "Gaussianization operation" to the signals as shown in (e).

Appended use of delta-cepstral coefficient (DCC) and MFCC improves ASR efficiency but results in low-grade robustness towards reverberation and noise as shown in Fig.(3). Fig 3(a) represents the short period power of a specific contextual speech segment and for the noise segment correspondingly. Fig (b) represents the short period power for clean contextual speech signal at 0-dB. Fig(c) plots the logarithmic power scale representation. The peaks remain almost the same for clean speech and noisy signal but result in more enhance mismatch between frames, which causes the noise signal to fills the curve valley. Fig(d) provides the representation of delta cepstral, which gives high-level mismatch between speech in clean and distorted condition. In order to enhance the robustness delta spectral cepstrum coefficient (DSCC) feature is used. Assuming g_i is a sequence of white Gaussian noise distributed in the form Ɲ(0,σ^2), Ƥ indicates power of an independent set of N samples, numerically equals to

$$Ƥ=1/N\sum_{i=1}^{N} g_i^{\,2}\ .$$

Where *N* represents degree of freedom of Chi-square distribution followed by Ƥ which approaches Gaussian shape with large N. Ƥ can be represented as

$$Ɛ[Ƥ] = 1/N\sum_{i=1}^{N} g_i^{\,2} = \sigma^2$$

$$\text{Var}[Ƥ] = \text{E}[\text{P}^2] - \text{E}[\text{P}]^2 = \frac{\sum_{i,j} Ɛ\left[g_i^{\,2} g_j^{\,2}\right]}{N} - \sigma^4$$

$$=1/N^2\left(\sum_i Ɛ[g_i^{\,4}]\text{-}\sum_{i,j,i\neq j} Ɛ[g_i^{\,2} g_j^{\,2}]\right) - \sigma^4 = 2\sigma^4/N$$

Means Ƥ is nearly distributed as Ɲ($\sigma^2, 2\sigma^4/N$). Where σ^2 is associated with DC power of Ƥ while variance $2\,\sigma^4/N$ is associated with AC power. Thus, total impact can be expressed as

$$Noise_{sup} = -10log_{10}\left(\frac{Power_{AC}}{power_{AC} + power_{DC}}\right)$$

$$=10\,log_{10}\left(1+\frac{N}{2}\right)\text{ dB}$$

Where $Noise_{sup}$ represents noise suppression. We consider window size of 25*ms* which creates window length N=400 for a speech sample of 16 KHz and provide $Noise_{sup}$ =23dB approximately. Hence, DSCC provides noise suppression up to a maximum limit of 23 dB as proved by Varela O. (2011).

5. Fuzzy C-Means Clustering

Fuzzy C- means clustering is an advanced version of k-means clustering. In k-means clustering a particular data set belongs to a single codebook where as in fuzzy c-means clustering a dataset can belong to multiple codebooks with some membership functions which specifies the sorting of dataset in multidimensional space. It is a major modeling technique which performs on the acoustic features such as MFCC, shifted MFCC and DSCC. FCM represents an unsupervised form of clustering in an extensive multidimensional space. Bansal P.(2015) prove that decision-making ability of fuzzy c means clustering depends upon the grade of similarity and membership function. Susan S. (2012) proves that FCM greatly provides low error valuation by means of mean square error (MSE).

6. Experimental Results

This section provides comparative analysis of automatic speaker recognition using different feature extraction techniques. Fuzzy with Delta Spectral Cepstral Coefficient technique outcomes are compared with the conventional Fuzzy, MFCC in collaboration with shifted MFCC technique. 1760 training utterances and 680 testing utterances are used as database in the form of DARPA Resource Management (DARPA-RM). The database is tested for three types of noisy signal. This includes white noise, real life environmental noise and the background noise (music noise). The performance in each case for single and multiple users (22 users) is given in fig 4.

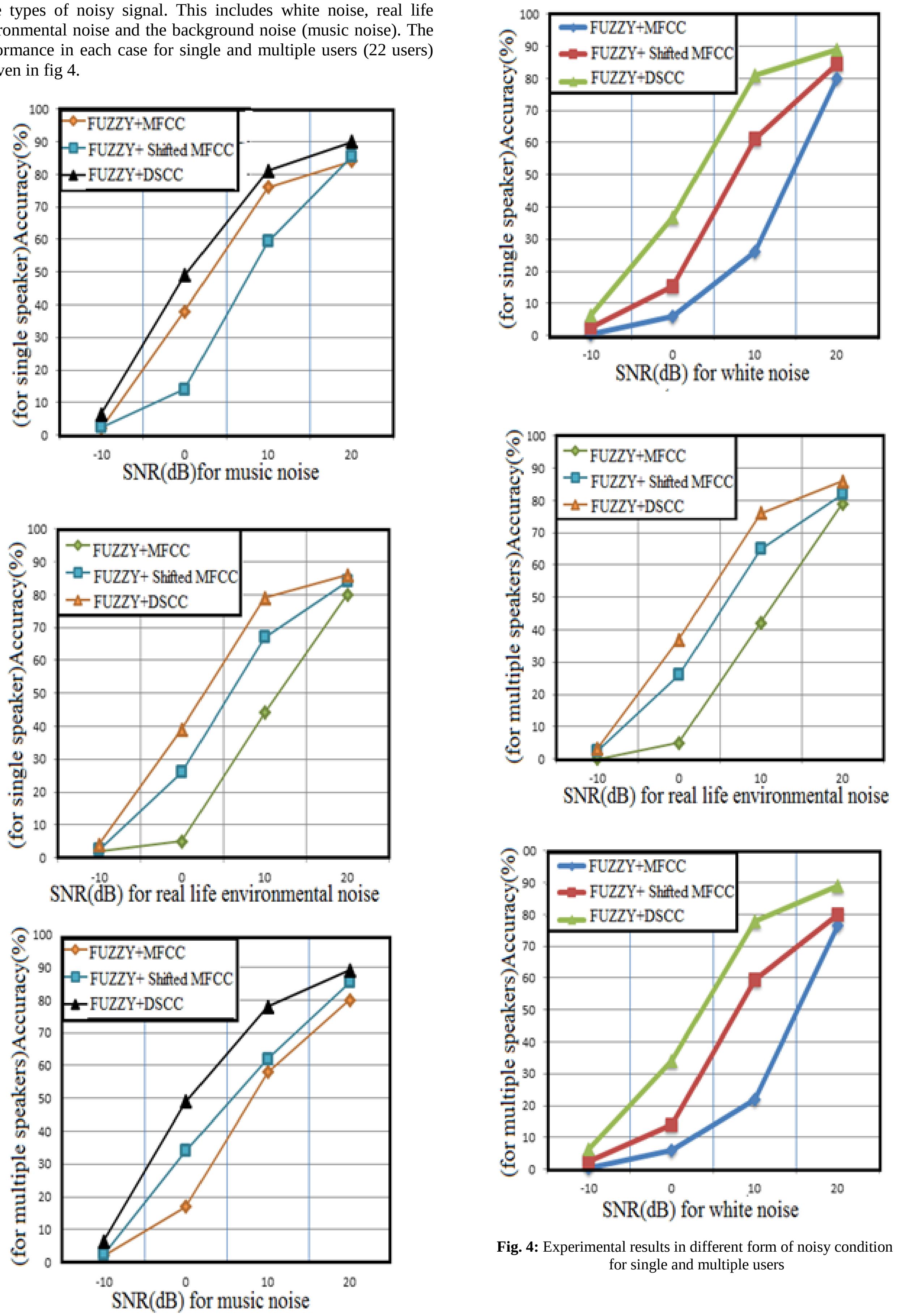
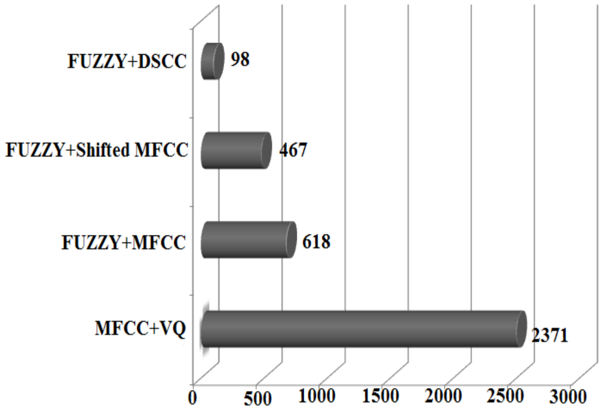


Fig. 4: Experimental results in different form of noisy condition for single and multiple users

The Experimental results show that the performance in each case is highly affected through SNR. At -10dB accuracy is almost negligible which is quite natural, increasing the SNR improves the performance as move towards 20 dB, fuzzy +DSCC gives best result in all cases because of its sustainable high robustness nature to noise including reverberation also, at 10dB DSCC provide almost 20-40% more accuracy than MFCC approach and provides 93% accuracy even in presence of noise (98% in ideal case). Fuzzy + shifted MFCC provides lower performance with 88% accuracy in the presence of noise(95%in ideal cases) but comparatively gives higher accuracy than traditional fuzzy + MFCC technique proposed by Bansal P.(2015) which provides only 87% accuracy even in an ideal case.

Fig. 5 depicts that the processing time for a segmented part of sample speech signal which specifies the speed of different algorithms based on processing time in the simulation. Fuzzy + DSCC provide 24 times faster processing time than the conventional MFCC and vector quantization technique which because VQ technique consumes a lot of time to create codebook. Hence, it is quite clear that fuzzy + DSCC has the capability of identifying speaker with very less computational complexity and can be viewed as the most promising algorithm for ASR.



Horizontal axis: processing time (ms)

Fig.5: Recognition speed of 10 min speech segment

7. Conclusion

If we talk in practical terms then cent percent accuracy cannot be achieved for automatic speaker recognition system. It results in a severe headache in the commercial and industrial area. But the results are improved because of Fuzzy + DSCC approach. These feature extraction techniques work jointly to provide enhanced degree of robustness to environmental noise. This system provides up to 40% enhanced efficiency at low SNR(7dB) with very low increase in computation/processing time. Based on our result we ensure about 7-10% increased accuracy over the work shown by Mukherjee R. (2013) and 15% increased accuracy over the work proposed by Varela O. (2011).

References

[1] Anzar S.M. (2016), Efficient online and offline template update mechanisms for speaker recognition, Computers & Electrical Engineering, (50)10-25.

[2] Bansal P.(2017), A probabilistic Feature Based SVM Model for English Speech Recognition. Journal of Engineering Technology, (6) 35 -44.

[3] Bansal P., Imam S., Bharti Roma(2015), Speaker recognition using MFCC, shifted MFCC with vector quantization and fuzzy. International Conference on Soft Computing Techniques and Implementations,

[4] Hans D.(2000), The Aurora database for speaker recognition performance. , International Conference on Spoken Language Processing.

[5] Huang Z.,Sabato Siniscalchi, Chin Hui Lee,(2016), A unified approach to transfer learning of deep neural networks with applications to speaker adaptation in automatic speech recognition, Neurocomputing, (218), 448-459.

[6] Kumar Kshitiz (2011), Delta-spectral cepstral coefficients for robust speech recognition. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 4784 – 4787.

[7] Mukherjee R.(2013), Text dependent speaker recognition using shifted MFCC . Proceedings of IEEE southeastcon, 1 – 4.

[8] A Fuzzy Nearest Neighbor Classifier for Speaker Identification. IEEE International Conference on Computational Intelligence and Communication Networks, 842 – 845.

[9] Varela O. (2011), Robust speech detection for noisy environments. IEEE transactions on Aerospace And Electronic Systems Magazine,(26) 16 – 23.

[10] S.V.Manikanthan and K.Baskaran “Low Cost VLSI Design Implementation of Sorting Network for ACSFD in Wireless Sensor Network”, CiiT International Journal of Programmable Device Circuits and Systems,Print: ISSN 0974 – 973X & Online: ISSN 0974 – 9624, Issue : November 2011, PDCS112011008.

[11] T.Padmapriya, Ms. N. Dhivya, Ms U. Udhayamathi, “Minimizing Communication Cost In Wireless Sensor Networks To Avoid Packet Retransmission”, International Innovative Research Journal of Engineering and Technology, Vol. 2, Special Issue, pp. 38-42.

[12] N.Prathima, K.Hari Kishore, “Design of a Low Power and High Performance Digital Multiplier Using a Novel 8T Adder”, International Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 3, Issue.1, Jan-Feb., 2013.