International Journal of Modern Electronics and Communication Engineering (IJMECE) ISSN: 2321-2152
Volume No.-7, Issue No.-4, July, 2019

Multitaper –MFCC features for Speech Recognition under noisy environments

P.Sunitha[1],
Research Scholar, Dept. of ECE,
JNTUK,India,
Sunitha4949@gmail.com

Dr.K.Satya Prasad[2]
Rector,VFSTR,
Guntur, India
Prasad_kodati@yahoo.co.in

Abstract:Environmental noise is a key factor which affects the performance of the speech recognition system and results in poor recognition accuracy. This paper explores a new method for speech recognition system under noisy environments. The recognition accuracy of a speech recognition system strongly depends on feature extraction and classification techniques. To overcome this problem, the proposed method uses speech enhancement module as a pre-processing operation to reduce the background noise. Later feature extraction was done on enhanced speech using low variance multi-taper spectrum estimation and further classification, training and testing was done using HTK tool kit. Performance of the proposed method was evaluated in terms of recognition accuracy which shows its superior performance when compared to other feature extraction techniques like LPCC (in time &frequency domains), MFCC under noisy and enhanced speech cases.

Keywords: Speech recognition, Feature extraction, Mel-Frequency Cepstral Coefficients (MFCC),multitaper spectrum estimation, Linear Predictive Coding (LPC),.

I. INTRODUCTION

Speech signal contains huge amount of information which gives gender classification, emotional characteristics and identity of a speech. Every speech signal and speaker has their own characteristics which were produced in their utterance. Two major issues in speech recognition are feature extraction and classification The front-end step in speech recognition is a feature extraction ,which extract significant amount of information from the speech, which plays a key role in speech recognition .This feature extraction further, affects the performance of the classifier as well as recognition system. Numerous techniques are available in the literature for feature extraction process, among them two methods are popularly used .Which are Linear Predictive Coding (LPC)[1] and Mel Frequency Cepstral Coefficients (MFCC)[2] methods. LPC is a parametric approach which is closely related to the human vocal tract produces sound. In case of LPC based feature extraction process, prediction coefficients were calculated using 13 th order LPC .Later Cepstral Coefficients and derivatives were extracted from the predicted coefficients. It works on the assumption that the signal is stationary one over a given frame and also it fails to capture the unvoiced and nasalized sounds [1].MFCC is a non-parametric approach in frequency domain based on human auditory system. MFCC feature extraction process involves, mel – filtering operation on the magnitude spectrum of windowed speech signal. Mel-filter bank consists of 20-mel-scaled log filter bank outputs. Feature vector extracted from MFCC consists of 12 MFCC,1 energy Coefficient. Compute delta and double delta coefficients of 12 MFCC and 1 energy coefficient. The feature vector extracted from MFCC consists total of 39 coefficients are extracted from each frame [2].In these two methods features are computed from the estimated spectrum and the performance depends on the accurate estimate of spectrum. In most of the speech processing applications spectrum is computed using windowed periodogram [15] through Discrete Fourier Transform (DFT).These two works well in clean conditions but poor performance in noisy conditions because of spectrum estimation by windowing. Windowing reduces the difference between the actual and estimated spectrum, but it fails to reduce the variance of the spectrum estimate. Windowing introduces large variance; hence the features computed from LPC, MFCC are also having high variance [9]. To reduce the variance replaces the spectrum which was obtained from a window by a multi-taper spectrum [3]. Varieties of techniques are available for speech/speaker recognition using multitaper spectrum analysis in literature. The gravitational and topographical representation of the terrestrial planets on the sphere can be analysed by spectral analysis. Multitaper spectral analysis allows harmonic analysis to investigate the structure of the body on the sphere [4]. Spectrum estimated from windowed period gram has large variance instead of having low-variance. One of the prominent methods to reduce the variance, windowed periodogram can be replaced by multitaper spectrum estimation method [5]. Numerous feature extraction techniques were available in the literature.Recently the multitaper MFCC features are widely used in speech/speaker applications as a front end processing i.e for feature extraction process .The accuracy of recognition system not only depends on feature extraction ,it also depends on classification and modelling which in turn depends on feature extraction. In conventional MFCC process the features are extracted from a spectrum which was obtained from a windowed periodogram [6]. This method was used to compute mel-frequency Cepstral Coefficients for speech recognition Performance of this method was tested on AURORA-2 database proves that multitaper MFCC feature extraction method performs well compared to single windowed spectrum estimation method [6]. The low-variance multitaper MFCC features can also be used in Speaker recognition and verification applications. Speaker verification done by using multi-taper MFCC

features with i-vector as a classifier [7] and experiments were conducted on NIST 2010 SRE extended list using Equal Error Rate (EER) as an evaluation metric. Improvement in results can be found rather than using a single window method. To extract useful semantics from speech, speech emotion recognition can be used and hence, improvement in the performance of speech recognition systems [11] . Multiple feature vectors and i-vector for speaker adaption process given in [12] used to improve the recognition accuracy in presence of noisy and reverberant conditions. This method was tested on a REVERB challenge 2014 corpora and the results shows that Word Error Rate has been reduced from 10% to 9.3 %[12].Multitaper spectral analysis is also used in voice disorder classification by separating the voiced and unvoiced sounds ,shows a significant improvement in results when compared to the conventional MFCC method[13].Stress recognition from speech is also an important area in research. These multi-taper MFCC features were used to recognise emotion from speech by conducting experiments on SUSAS database which gives utmost classification accuracy than conventional MFCC features[14]

The organisation of the paper is as follows, Section II provides Multitaper spectral estimation ,Section III gives proposed method for speech recognition finally section IV gives Simulation results and conclusions.

II. MULTI-TAPER SPECTRAL ESTIMATION

Due to sudden changes and sporadic behaviour, Speech signal can be modelled as a non- stationary signal. As time evolves the statistics like mean, variance, co-variance and higher order moments of a non-stationary signal changes over time. Spectral analysis plays a major rule in speech feature extraction techniques to get accurate spectrum estimation. FFT method is widely used to get power spectrum estimation in most of the speech enhancement algorithms especially in spectral subtractive type methods. The estimated power spectrum obtained by FFT is reduced by variance of the estimate and energy leakage across frequencies which create bias. To avoid leakage, multiply the signal in time domain with a suitable window which having less energy in side lobes. Type of window affects the noise estimate in speech enhancement algorithms, hence selection of desirable window which provides an accurate noise estimation plays a significant role in Speech enhancement process .Generally Hamming window is preferable with less energy in side lobes but it effects the estimate by reducing leakage but not the variance. In most of the speech processing algorithms noise estimate is obtained by using suitable windows which reduce the bias but not the variance. To alleviate the bias due to large variance one must use a Windowing function which assigns more weight to the centre of the signal rather than the end points of the signal. The variance can be reduced by taking multiple estimates from the sample which can be achieved by using tapers[15]. Tapers should be orthogonal to each other provides individual estimates with low variance. In case of multi –taper spectrum estimate selection of number of tapers and taper weights plays a significant role[16]. Various tapers are available Sleepian (Sleepian and Pollak 1960),Thomson muiti-taper(Thomson,1982),Sine tapers(Ridel and Sidernko ,1995), multi-peak multi-tapers(Hansson and Salomosson,1997) and Sinusoidal weighted Cepstrum Estimator (SWCE)(Hansson-Sandsten and Sandberg,2009).The real, unit energy sequences are Sleepian sequences having highest energy in a band width [17].Thomson multi-taper spectral estimation method uses a set of orthonormal having good leakage properties but it suffers from Eigen value solution[15].To overcome this ,Sine tapers suggested by Ridel and Sidorenko (1995),provides smaller local bias than the Sleepian sequences[16] .The use of multiple windows have several advantages over a single window i.e the variance can be reduced by using weighted number of tapers[17], Fig.3 shows the spectral variance among single window and multi-taper method using various number of tapers . This multi-taper methods has already used in Speech enhancement applications to improve the correlation between objective quality measures and subjective listening tests.Hu and Loizou [18] used these multi-tapers to get low variance spectral estimate ,further the spectrum was refined using wavelet thresholding ,Finally this was used to improve the quality of speech signal in case of highly non-stationary noise. Results shows that this method has superior performance in terms of quality measures with high correlation between subjective listening test and objective quality measures. The multi-taper method not only used in speech enhancement applications but it can be used in geographical data analysis, speaker recognition and speech recognition [18].

The multi-taper spectrum estimator is given by

$$\hat{S}^{mt}(\omega) = \frac{1}{L}\sum_{p=0}^{L-1} \hat{S}_p^{mt}(\omega) \qquad (1)$$

with

$$\hat{S}_p^{mt}(\omega) = \left|\sum_{M=0}^{N-1} \lambda(m)x(m)e^{-j\omega m}\right|^2 \qquad (2)$$

Here data length is given by N and S_p is the p [th] sine taper used for spectral estimate [12] and $\lambda(m)$ is given by

$$\lambda(m) = \sin\frac{\pi p(m+1)}{N+1}, m = 0,.N-1 \qquad (3)$$

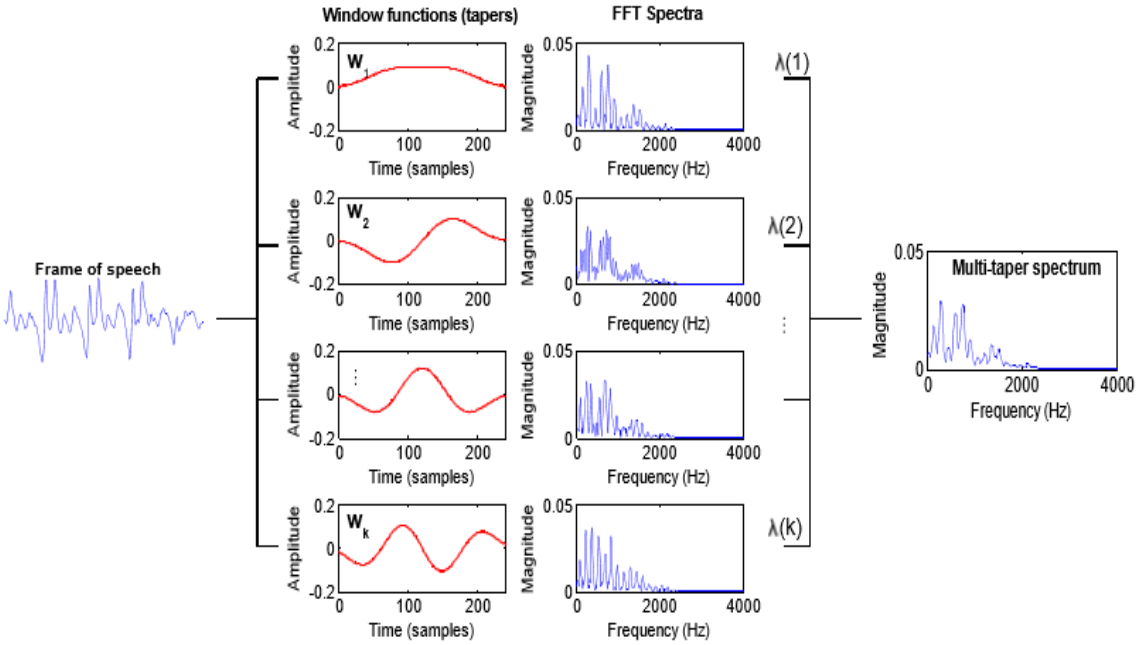


Fig.1:Multiple window method for spectrum estimation by individual windows. Multi-taper spectrum was obtained by average of all individual spectrums

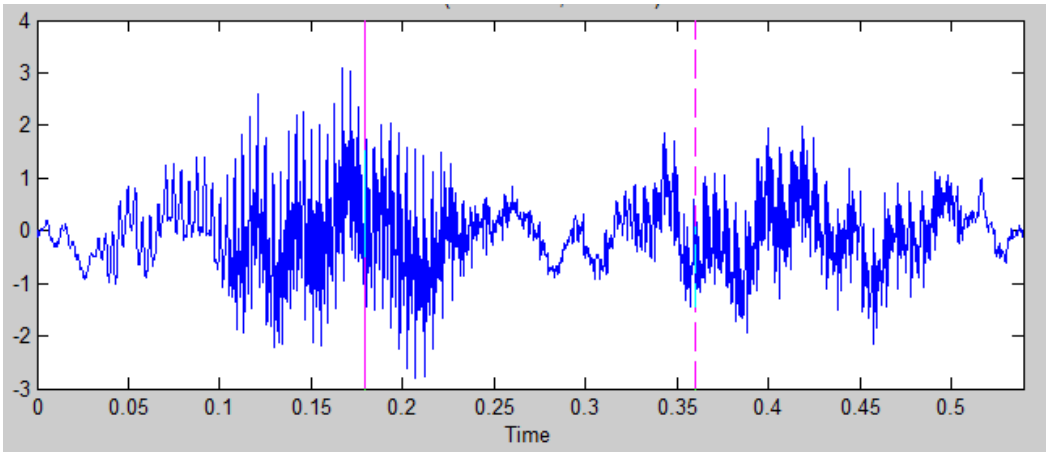
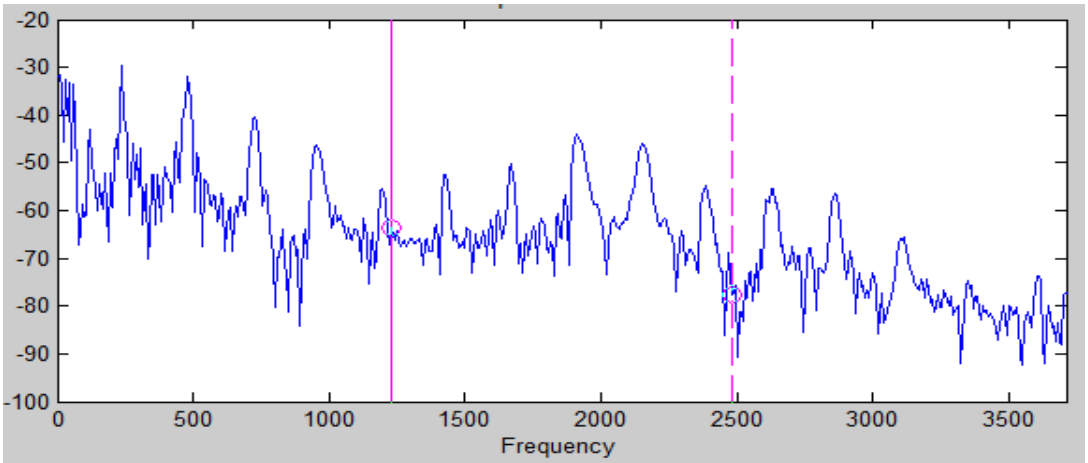
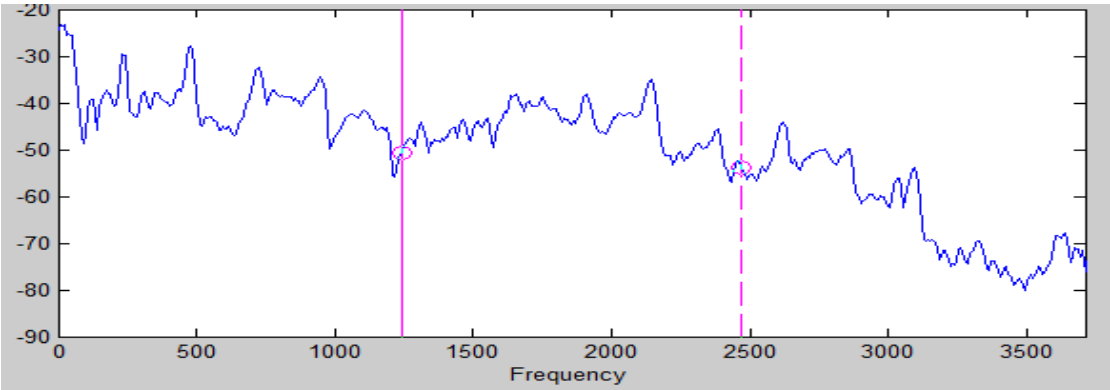


Fig. .2:Speech signal(mtlb.wav)



(a)



(b)

Fig.3 (a)Spectrum obtained by window method (b)Spectrum obtained by Multitaper method

III. PROPOSED METHOD FOR SPEECH RECOGNITION

The proposed method consists of three sections, namely the Speech enhancement section, feature extraction and recognition sections. In our proposed work, first the speech signals are enhanced from the input noisy speech using spectral subtraction process. Even though number of speech enhancement techniques are available to palliate the effect of noise, spectral subtraction method is preferable because of its relative simplicity as it involves forward and inverse transform only. Speech enhancement algorithms works on the assumption that both speech and noise are uncorrelated in additive manner. In spectral subtraction process, the noisy speech signal is divided into frames of duration 20ms to make the signal as a stationary and the spectrum is computed using STFT (Windowing +FFT). Noise spectrum is evaluated during speech absence regions. An estimate of noise spectrum is subtracted from the noisy speech spectrum to obtain a minimum mean squared error (MMSE) in estimated enhanced speech. Finally the enhanced speech signal is reconstructed using over-lap add method [19]. In Feature extraction process, features from the enhanced speech signals are computed using multitaper spectrum estimation and Mel-Frequency Cepstral Coefficients [2] .Finally the enhanced feature vectors are given as training data for recognition module. The recognition module consists of Hidden Markov Models which are trained with a training database .The mismatch between training and testing data have a great impact on recognition accuracy. To improve the recognition accuracy in presence of noise the proposed method first reduces the background noise with the use of spectral subtraction. Then speech absent segments are dropped based on voice activity detection, then feature extraction using multi-taper spectral estimation using sine tapers (no.of tapers=6) which involves spectrum estimation by multi-taper method followed by mel- filter bank, log transformation and then DCT which results in 12M_MFCC's.Compute first and second orderderivatives which results in 12 delta cepstral coefficients,12 double delta cepstralcoefficients ,3 energy coefficients(1 –energy ,1-delta energy,1-double delta),total of 39 coefficients of a feature vector in each frame for recognition.

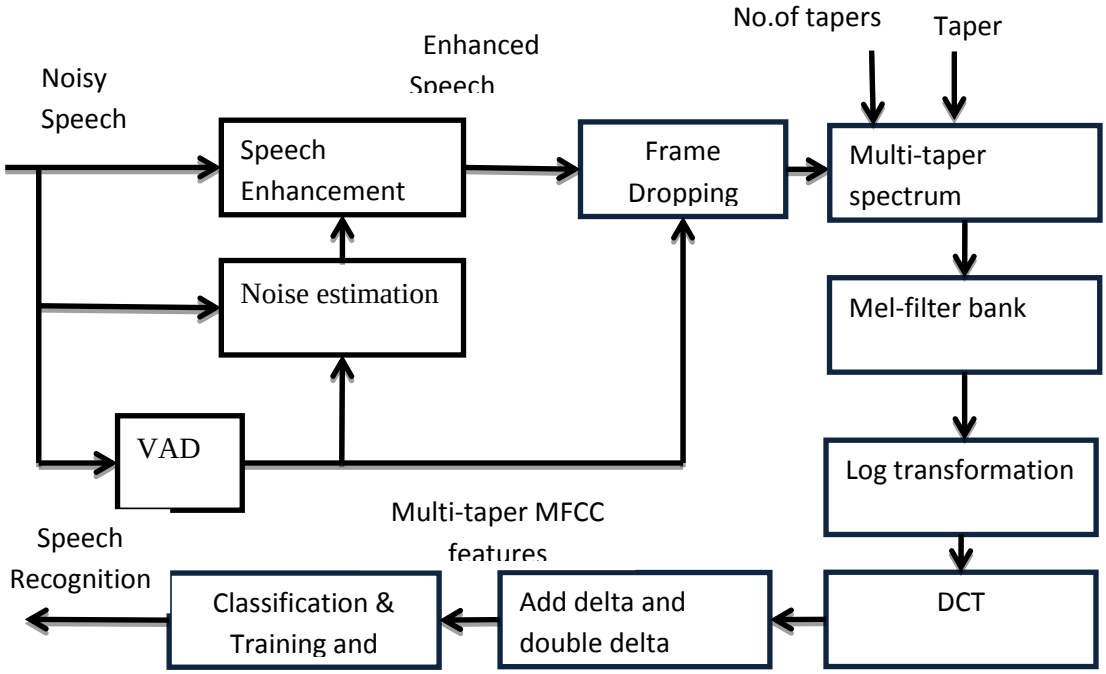


Fig. .4.:Block diagram of the proposed method

IV.Results and Conclusion: The proposed method improves the performance of speech recognition under noisy environment using Speech enhancement, feature extraction and recognition based modules. The input signals include gender of female and male speakers in presence of additive noise. In results speech recognition accuracy is evaluated for various models as TDLP, FDLP,MFCC and Multi-taper MFCC(MMFCC) with speech enhancement and without speech enhancement.The accuracy of theproposed method performs better in speech recognition in all the cases considered. The results shown in table 1.shows the recognition accuracy can be improved with the proposed method with and without speech enhancement.

Table.1:Comparison of recognition accuracy for different feature extraction methods.

Number of datasets	Without speech enhancement				With speech enhancement			
	TDLP	FDLP	MFCC	MMFCC	TDLP	FDLP	MFCC	MMFCC
1	81.09	82.99	86.99	90.99	85.59	88.99	91.99	96.19
2	83.98	85.88	89.88	93.88	88.48	91.88	94.88	99.08
3	84.73	86.63	90.63	94.63	89.23	92.63	95.63	99.83
4	82.26	84.16	88.16	92.16	86.76	90.16	93.16	97.36
5	83.83	85.73	89.73	93.73	88.33	91.73	94.73	98.93

REFERENCES

1. J.Makhoul,"Linear Predection:A tutorial Review",Proc.ofIEEE,volume 63,Issue 4,1975.
2. J.Trangol and A.Herrera,"A traditional method and multitaper to feature extraction using Mel Frequency Cepstral Coefficients",Int.Journal of Information and Electronics Engineering,Volume 5, Issue 1,2015
3. G.APreito,R.L.Parker,D.J.Thomson,F.L.Vernon and R.L.Graham,"Reducing the bias of multitaper spectrum estimates",Geophysics, Journal 0f Int, Volume 171,2007.
4. TomiKinnunen, Rahim Saeidi, Johan Sandberg and Maria Hansson-Sandsten," What Else is New Than the Hamming Window? Robust MFCCs for Speaker Recognition via Multitapering" "in Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH'10),Makuhari, Japan,September2010.
5. Mark A. Wieczorek and Frederik J. Simons,"Minimum-Variance Multitaper Spectral Estimation on the Sphere"The Journal of Fourier Analysis and Applications Volume 13, Issue 6, 2007.
6. Md. Jahangir Alam, Patrick Kenny and Douglas D. O'Shaughnessy: "A Study of Low-variance Multi-taper Features for Distributed Speech Recognition", Proc. of NOLISP, LNAI 7015, 2011.
7. Md. Jahangir Alam, TomiKinnunen, Patrick Kenny, Pierre Ouellet and Douglas D. O'Shaughnessy: "Multi-taper MFCC features for speaker verification using I-vectors", ASRU, 2011.
8. T. Kinnunen, R. Saeidi, F. Sedlak, K.A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li," Low-Variance Multi-taper MFCC Features: a Case Study in Robust Speaker Verification", IEEE Trans.on Audio, Speech and Language Processing, volume.20,Issue 7, 2012.
9. Navnath S Nehe, Raghunath and S. Holambe,"DWT and LPC based feature extraction methods for isolated word recognition", EURASIP Journal on Audio, Speech, and Music Processing 2012, 2012:7
10. Md Jahangir AlamTomiKinnunen, Patrick Kenny, Pierre Ouellet and Douglas O'Shaughnessy, ,"Multi-taper MFCC and PLP features for speaker verification using i-vectors Speech Communication "55 ,2013.
11. Y. Attabi, Md J. Alam, Pierre Dumouchel, P. Kenny, Douglas O'Shaughnessy," Multiple Windowed Spectral Features for Emotion Recognition" in Proceedings of the 38th IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP'13) , IEEE,Vancouver,Canada,May2013.
12. Md J. Alam ,Vishwa Gupta Patrick Kenny and Pierre Dumouchel "Speech recognition in reverberant and noisy environments employing multiple feature extractors and i-vector speaker adaption", EURASIP Journal on Advances in Signal Processing, 2015, 2015:50
13. Omer Eskidere and AhmetGürhanli, "Voice Disorder Classification Based on Multi-taper Mel Frequency Cepstral Coefficients Features" Computational and Mathematical Methods in Medicine, Volume 2015,
14. SalsabilBesbes and ZiedLachiri ,"Multitaper MFCC Features for Acoustic Stress Recognition from Speech", International Journal of Advanced Computer Science and Applications, Vol. 8, No. 3, 2017.
15. D. J. Thomson,"Spectrum estimation and harmonic analysis", IEEE proceeding, volume.70,Issue 9, 1982.
16. K. S. Riedel and A. Sidorenko," Minimum bias multiple taper spectral estimation".,IEEE Trans. on Signal Proc., volume 43,Issue 1, 1995.
17. Percival, D. B. and Walden, A. T.,"Spectral Analysis for Physical Applications, Multi-taper and Conventional Univariate Techniques",Cambridge UniversityPress,1993.
18. Y. Hu and P. Loizou," Speech enhancement based on wavelet thresholding the multi-taper spectrum" ,IEEE Trans. On Speech and Audio Proc., volume12, Issue 1, 2004.
19. Boll,S.F,"Suppression of acoustic noise in speech using spectral subtraction". IEEE Transactions on Acoustics Speech and Signal Processing, Volume 27,Issue 2, 1979.