Analysis of MFCC and Multitaper MFCC Feature Extraction Methods

Rupali G. Shintri
M.E.(Signal Processing)
E & TC Department ,ICOER, Pune, Maharashtra, India.

S.K. Bhatia
Professor, E & TC Department, ICOER, Pune, Maharashtra, India.

ABSTRACT
In speech & audio applications, short-term signal spectrum is often represented using mel-freuency cepstral coefficient (MFCC) computed from a windowed discrete Fourier transform (DFT). Windowing reduces spectral leakage but variance of the spectrum estimate remains high. An extension to windowed DFT is called multitaper method which uses multiple time domain windows which are called as tapers with frequency domain averaging. Then detailed statistical analysis of MFCC bias & variance is done.

For speaker verification the extracted feature is used to design a model using classifier (GMM), which implements likelihood ratio test to decide whether to accept or deny the registered speaker.

General Terms
Multitaper MFCC, EM algorithm

Keywords
Mel-frequency cepstral coefficient, multitaper, GMM, speaker verification, tapers.

1. INTRODUCTION
The speaker verification (SV) is to verify or deny a registered identity of the enrolled speaker based on a voice sample. There are two processes of speaker verification. During the training process, speaker-based feature vectors are extracted from voice signals and used to design a speaker model. During testing process, the verification system estimates a likelihood ratio [8] to differentiate between two decisions : the voice sample is of registered speaker or of false speaker . Features extracted from the speech signal are compared to a model representing the registered speaker, obtained from training process, and to some model representing potential false speakers (i.e., those are not the registered speaker). The ratio of registered speaker and false match scores is the likelihood ratio, which is then compared to matching score to decide whether to verify or deny the speaker. Fig. 1(a) & Fig.1 (b) shows the basic block diagram of speaker verification.

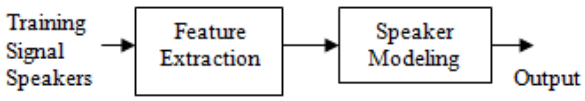

Fig .1(a) Training Stage

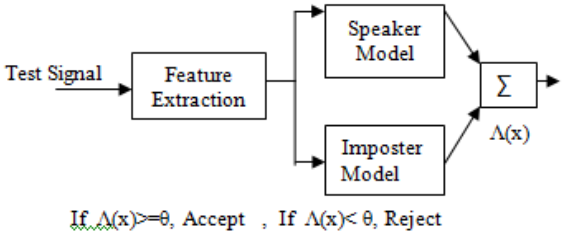

Fig. 1(b) Testing Stage

Feature Extraction consists of different process which includes speech activity detection to remove non voiced part from the signal. Then speaker specific feature information is extracted. From the source filter theory of speech production it is known that speech spectrum shape encodes information about the speaker's vocal tract shape through resonances & glottal sources via pitch harmonics. Thus some form of spectral informative features is used in most speaker verification systems. As specified in [1] Mel-frequency cepstral coefficient (MFCC), linear predictive cepstral coefficient (LPPC), perceptual linear predictive (PLP) are some spectral features. Feature extraction is the key of a speech processing. Spectral features computed from windowed DFT or Linear Predictive (LP) models are used in most of speech processing. The DFT & LP models perform wgood under clean conditions but verification accuracy degrades under different surrounding as short term spectrum tends towards the variations [2].

2. MFCC FEATURE EXTRACTION
Mel frequency can be defined as the short time power spectrum of a speech. In MFCC the are sequential frequency bands spaced on the mel scale as in [4]. The mel scale approximates the human auditory system's response. The log frequency power coefficients describe the spectral shape of the signal. Each coefficient provides a measure for the power distribution across the different sub bands in frequency domain. . This spectrum gives simple but identical representation of the spectral properties of the voice signal which is the key for identifying and verifying the voice characteristics of the speaker. A speaker voice plots may include identical sentences, spoken by the same speaker but at different times, result in a similar, but not identical sequence of MFCC matrices.

MFCC is recommended feature as it satisfies the criteria [1] of feature selection. In [4] for extracting MFCC following steps are executed: frame blocking, windowing, FFT, mel-frequency wrapping, cepstrum , mel cepstrum. Mel cepstrum is converted to time domain by, as in [4]

$$\text{Mel}(f) = 2595 * \log_{10}(1 + f/700) \quad (1)$$

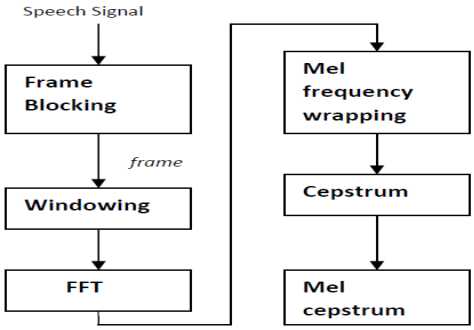

Fig. 2. Block diagram of MFCC feature extraction

From the previous studies the common MFCC implementation based on windowed DFT [1] generates spectrum with high variance, there exists an random process with each frame of the voice sample, which evaluates the frame; an autoregressive (AR) process evaluated with random inputs but estimated with fixed coefficients. For speech signals there exists a speaker dependent vocal from which the actual sounds are generated. The signal spectrum and MFCC estimation may change due to high variance of the spectrum of each frame. In speaker verification, irregularities in features are modeled & MFCC variances in the Gaussian mixture models (GMM). However, if the MFCCs extraction is done with minimum variance[2][3] the speaker and session variability of models to have minimum random variation, which will help to improve system performance.

3. MULTITAPER MFCC FEATURE EXTRACTION

Fig. 3 shows the block diagram of single & multitaper spectrum estimation MFCC feature extraction.

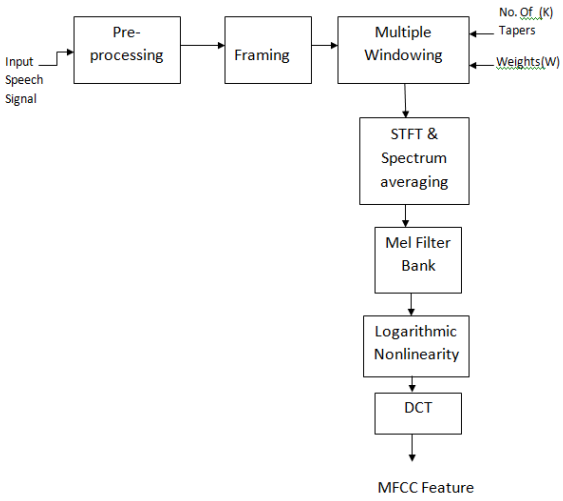

Fig.3. Block diagram of multitaper MFCC feature extraction

The pre-processing step includes pre-emphasizing, DC removal, signal normalization. Signal normalization. In framing block the speech signal is divided small frames. Frames are again divided into small duration's windows (tapers) instead of one window (Hamming). Then spectrum for each taper is estimated individually & averaged. As spectrum of each taper is uncorrelated weighted frequency domain averaging of the spectrums reduces the variance [2]. The MFCC filter bank improves Equal error rate (EER) & minimum detection cost function which indicates stable parameter setting. Then logarithmic nonlinearity is removed. then features can be normalized by any of feature normalization methods like mean & variance normalization (MVN)[7], frequency warping[6], RASTA filtering [5].

3.1 Computations of Multitaper MFCC

A hamming windowed DFT spectrum is the used for power spectrum estimation. For m-th frame & l-th frequency an MFCC estimate is given by, as in [3]

$$S(m,l) = \left|\sum_{j=0}^{N-1} w(j)s(m,j)e^{\frac{2\pi l}{N}}\right|^2 \quad (2)$$

Where l {0, 1,……l-1} denotes the frequency index, N is the frame length , s(m ,j) is the time domain speech signal & w(j) denotes the time domain window function called Taper which usually symmetric & decreases towards frame boundaries.Windowing reduces bias i.e. difference between estimated spectrum & actual spectrum but it does not reduce variance of the estimated spectrum therefore variance of MFCC. To reduce variance of estimated, replace the windowed DFT spectrum estimation by Multitaper spectrum estimate The Multi-taper spectrum estimator is given by, as in [3]

$$S(m,l) = \frac{1}{K}\sum_{p=0}^{K-1} \lambda(p) \left|\sum_{j=0}^{N-1} Wp(j)S(mj)e^{\frac{2\pi l}{N}}\right|^2 \quad (3)$$

Where N is the frame length, p w is t-th taper used the spectral estimate. K denotes the number of tapers & λ (p) is weight corresponding to the p-th taper. The tapers w p(j) are selected to be orthogonal, i.e.

$$\sum j \; W\,p\,(j) W\,q(j) = \delta pq \quad (4)$$

The multi-taper spectrum estimate is therefore obtained as weighted average of K individual spectra. The tapers in multitaper are chosen so that the estimation error in the individual sub-spectra is uncorrelated. Averaging the uncorrelated spectra gives a low variance of spectrum estimate which leads to low variance MFCC.

3.2 Choice of tapers

There are different tapers available [2][3]for spectrum estimation, like Thomson , sine and multipeak .. For cepstrum analysis, the sine tapers are applied with predefined weight. Taper are designed for specific task, Thomson tapers are used for white noise and multipeak tapers for voiced speech signal

Thomson tapers, designed for white noise, tend to perform well for any smooth spectrum. In general, the tapers are implemented to estimate faults in the subspectra which are approximately uncorrelated, which is the main factor to reduce variance. For a single voiced speech frame, like multitaper methods considered in [2], Thomson, multipeak and sine-weighted cepstrum estimator (SWCE). All the multitaper methods generates smoother spectrum compared to the Hamming method, because of variance reduction.

For a less number of tapers [2] say $K \leq 4$, all the methods maintains both the features of voice signal generated due to the vocal cord and due to the vocal tract. For a high number of tapers, say $K \geq 8$, the harmonics gets disturbed. The exact number of tapers to be used depends on the target application. In speaker recognition, both the voice source and vocal tract filter are important, thus we expect to generate the better results using small number of tapers.

4. SIGNAL MODELING

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities.GMM are commonly used as a parametric model of the probability distribution of continuous measurements or features in biometric systems, such as vocal tract related spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative Expectation- Maximization (EM) algorithm [4]

A Gaussian mixture model is weighted sum of M component Gaussian densities as given by, as in [4]

$$p(x/\lambda) = \sum_{i=0}^{M} wi\, g(x/\mu i\, , \sum i\,) \qquad (5)$$

Where x is a D-dimensional continuous valued data vector i.e. feature extracted from utterance of the speaker, wi ,i=1..,M,are the mixture weights, & g(x|μi, ∑i), i=1,……..,M, are the component Gaussian densities.

GMM are often used in biometric systems, mostly in speaker recognition system, due to their capability of representing a large class of sample distributions. As in [1] the powerful attributes of GMM is its ability to form smooth approximation to arbitrarily shaped densities.

5. RESULTS

The database consists of 4 female & 6 male Speakers. For training purpose 1 sample of each speaker is used. While testing 4 samples of each speaker are tested. Each speaker is having an input frequency as 8 KHz, then this input signal is sampled with a sampling frequency of 16 KHz. Each input signal is divided into 256 frames. The signal is first pre-processed means removing the noise, signal normalization.

The input signal of one user is tested for both the feature extraction methods that is for MFCC & Multitaper MFCC. The results are as shown below,

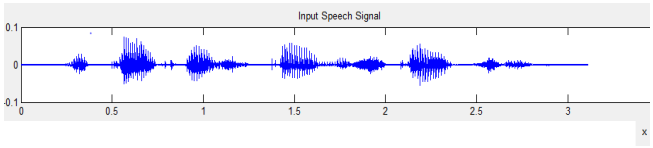

Fig.4. Input speech signal

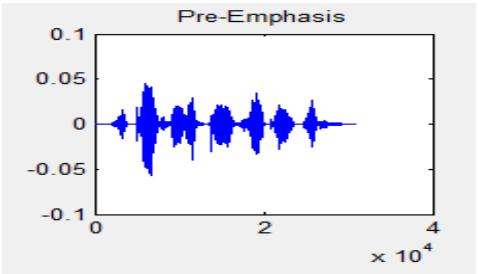

Fig.5. Pre-Emphasis of input speech signal

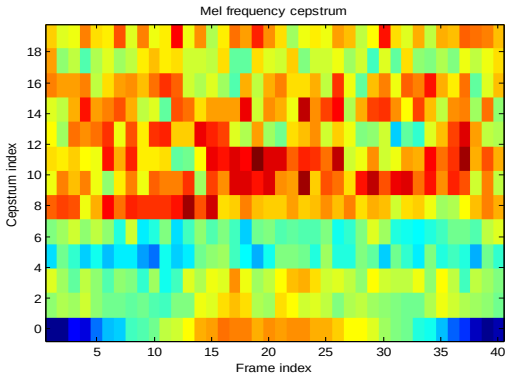

Fig.6. MFCC feature of input speech signal

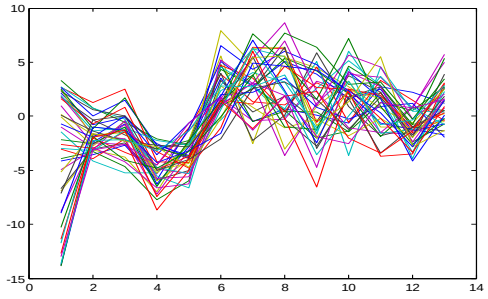

Fig .7.MFCC of input speech signal

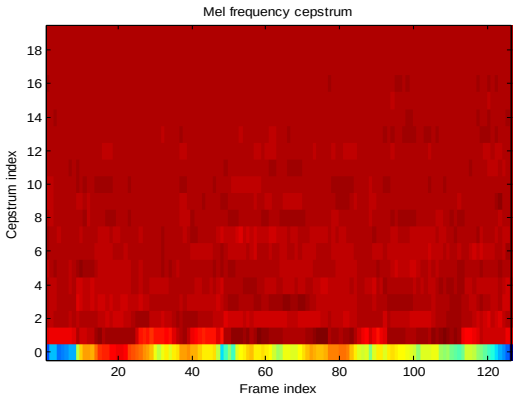

Fig.8.Multitaper MFCC feature of input speech signal

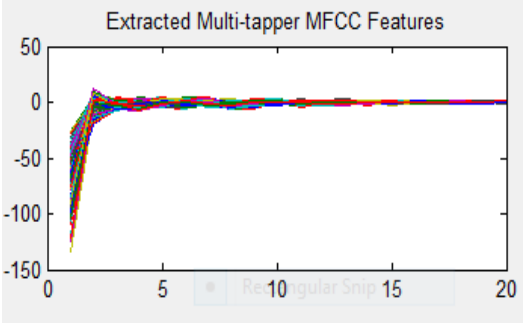

Fig.9. Multitaper MFCC of input speech signal

For variance calculation first we have estimated standard deviation & then we have to calculate variance. As per the formula,[3] variance in case of the existing method is the square of the standard deviation as only one spectrum is estimated per frame ,but in case of proposed method the

variance is calculated as square of standard deviation divided by the no. of weights of the taper i.e.,

Variance = (STD)2 / No.of taper weights (5)

In our case we have four weights of SWCE (Sine weighted Cepstral Estimate) taper. Table No.1 Shows the Variance for each speaker estimated for MFCC & Multitaper MFCC

Table 1. Variance of estimation for MFCC & Multitaper MFCC

Speaker	Variance	
Figures	MFCC	Multitaper MFCC
S1-F	102.8175	33.6556
S2-M	107.5949	47.4004
S3-F	139.8921	41.9625
S4-F	118.7642	43.7225
S5-M	101.1613	51.4503
S6-F	113.6121	58.9847
S7-M	127.2226	55.0163
S8-M	49.9537	47.6148
S9-M	112.8630	59.0576
S10-M	122.6932	54.8473

Table. No.2: Model Accuracy

Feature Extraction Method	GMM Model Accuracy (%)	Testing time For 10 speakers-four samples each (sec.)
MFCC	72.5	44.5782
Multiaper MFCC	87.5	54.1396

6. CONCLUSION

A multitaper MFCC feature extraction method extracts low variance MFCC features from the speakers voice samples. This approach exploits the better model model performance & the model of each speaker is designed with high accuracy .

7. ACKNOWLEDGMENTS

I would like to express my sincere gratitude towards my Project Guide Prof. S. K. Bhatia, for her constant support and guidance throughout the completion of this paper. I would not hesitate to thank my friends for constant help and Co-operation given to me

8. REFERENCES

[1] Kinnunen T., Li.,H. An overview of Text Independent Speaker recognition:from feature to supervectors Speechcommunication(2009),doi:10.1016/j.specom.2009 .08.009

[2] Tomi kinnuen,Rahim saeidi, Low-Variance Multitaper MFCC features: a case study in robust speakerVerification member IEEE, Manuscript IEEE ransaction in Speech & Audio processing(2012).

[3] Patrick Kenny1, Douglas O'Shaughnessy2, Study of Low-variance Multi-taper Features for Distributed Speech Recognition, INRS-EMT, University of Quebec, Montreal, Canada Speech Confrernce (2008)

[4] G.Suvarna Kumar ,K. A. Raju, Dr.MahanRao, P.Satheesh, Speaker Recognition Using GMM, et.al/International Journal Of Engineering Science &Technology Vol2 (6), 2428-2436, 2010.

[5] H. Hermansky and N. Morgan. RASTA processing of speech. IEEE Trans. on Speech and Audio Processing, 2(4):578–589, October 1994.

[6] Puming zhan,Martin westphal, Speaker Normalization Based On Frequency Warping, Article in Interactive system laboratories,Carnegie University Germany,

[7] David McCarten E6820, Comparison of Speech Normalization Techniques, Student, Columbia University March 9, 2008

[8] Douglas.A.Reynolds, Automatic Speaker Recognition :Current Approaches & Feature Trends by, MIT Lincoln Laboratories, Lexington, MA,USA.

[9] Yongxin Zhang, Adel Iskander Fahmy, Michael S. Scordilis "Speaker Verification Using Speaker-Specific Prompts" department of electrical and computer engineering, university of miami, coral gables, florida 33124

[10] Mohd Zaizu Ilyas, Member, IEEE, Salina Abdul Samad, Senior Member, IEEE, Aini Hussain, Member , IEEE and Khairul Anuar Ishak, Member, IEEE, "Speaker Verification using Vector Quantization and Hidden Markov Model", the 5th student conference on research and development –scored 2007 11-12 december 2007, malaysia

[11] Gibak Kim and Philipos C. Loizou, *Senior Member*, IEEE, "Improving Speech Intelligibility in Noise Using Environment-Optimized Algorithms", IEEE transaction on audio ,speech & language processing ,vol.18.no.8,november 2010.

[12] Alfredo Maesa1, Fabio Garzia1,2, Michele Scarpiniti1, Roberto Cusani1," Text Independent Automatic Speaker Recognition System Using Mel-Frequency Cepstrum Coefficient and Gaussian Mixture Models" journal of information security,2012,3335-340.