International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Review: MFCC & Multitaper MFCC Feature Extraction for Speaker Verification

Rupali G Shintri[1], S. K. Bhatia[2]

[1]M.E. E&TC (SP), ICOER, Pune, Maharashtra, India

[2]Assistant Professor, 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 build a model using classifier (GMM), which implements likelihood ratio test to decide whether to accept or reject the speaker.***

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

1. Introduction

Speaker verification can be divided into text dependent (Fixed words) & text independent (No fixed words) methods. In text dependent method require the speaker to provide utterances of key words or sentences ,the same text being used for both training & testing , whereas text independent method do not depend on specific text being spoken. There are several applications such as forensic & surveillance, in which predetermined key words cannot be used. Human beings can recognize speakers irrespective of the words of the utterance. Therefore, text independent methods are more attentive.

The objective of speaker verification is to accept or reject a claim identity of speaker based on voice sample. Fig. 1(a) & Fig.1(b) shows the basic block diagram of speaker verification.

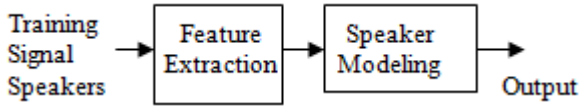
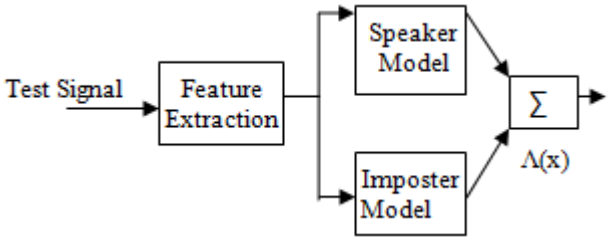


Figure 1(a): Training Stage



If Λ(x)>=θ, Accept , If Λ(x)< θ, Reject

Figure 1(b): Testing Stage

During training stage speaker dependent feature vectors are extracted from training speech signal. Different features are Frequency band analysis , Formant Frequencies, Pitch Counters, Harmonic features, cepstral coefficient , Mel-frequency cepstral coefficients, etc. This feature vectors are then modeled & compared to a model of a claimed speaker, obtained from previous enrollments & with some models representing imposter speakers(not claimed speaker) . The ratio of speaker & imposter match scores is likelihood ratio (Λ) which is then compared to a threshold (θ) to decide whether to accept or reject the speaker.

Feature Extraction consists of different process which includes speech activity detection to remove non speech portions from the signal. Then feature conveying information is extracted. From the source filter theory of speech production it is known that speech spectrum shape encodes information about the speakers vocal tract shape via resonances (formants) & glottal sources via pitch harmonics. Thus some form of spectral based features are 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 well under clean conditions but verification accuracy degrades under changes in environment & channel since short term spectrum subject to many harmful variations [2].

2. MFCC Feature Extraction

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]

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

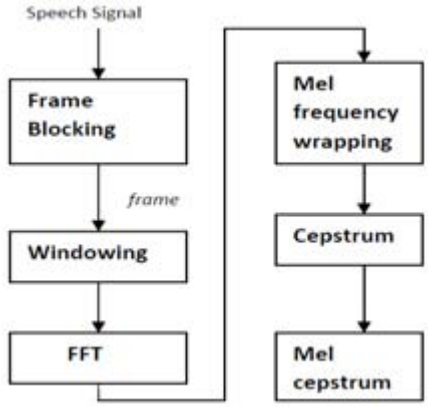


Figure 2: Block diagram of MFCC feature extraction

From statistical view, the common MFCC implementation based on windowed DFT is sometimes not suitable due to high variance of spectrum estimate. In speaker verification , uncertainty in features is modeled by the variance in the Classifiers which causes session variability in verification. However if MFCC is themselves are estimated with smaller variance[2][3], we can expect less random variations in model as well. This in turn enhances performance of verification

3. Multitaper MFCC Feature Extraction

The particular small variance method along with frequency normalization adopted is based on multitapers. Fig. 2 shows the block diagram of single & multitaper spectrum estimation MFCC feature extraction.

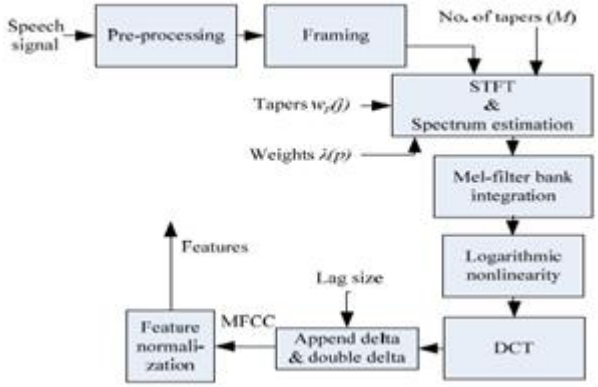


Figure 3: Block diagram of multitaper MFCC feature extraction

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 durations 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 delta & double delta coefficients are estimated , then features are normalized by any of feature normalization methods like mean & variance normalization (MVN)[7], frequency warping[6], RASTA filtering [5].

A. Compute Multitaper MFCC

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

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

Where k{0,1,…….K-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.(Hamming).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]

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

Where N is the frame length, p w is t-th taper used the spectral estimate. M denotes the number of tapers & for template is used to format your paper and style the λ (p) is weight corresponding to the p-th taper. The tapers wp(j) are selected to be orthogonal, i.e.

$$\sum j\, W\, p\,(j) W\, q(j) = \delta pq \quad (3)$$

The multi-taper spectrum estimate is therefore obtained as weighted average of M 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.

B. Choice of the Tapers

A number of different tapers have been proposed in [2][3] for spectrum estimation, such as *Thomson, sine & multipeak.* For cepstral analysis the sine tapers are applied with optimal weight. Each type of taper is designed for some type of random process; like Thomson taper is designed for flat spectra(white noise) & multipeak for peaked spectra(voiced speech)[2].

In practice the tapers are designed so that the estimation errors in the sub-spectra will be approximately uncorrelated, which is the key to reduce the variance. For a single voiced speech frame, all the three multitaper methods produce smoother spectrum compared to the Hammed method, because of variance reduction. As in [3] Thomson produces a staircase-like spectrum, multipeak with sharper peaks & sine a compromise between these two methods. For a small number of tapers all methods preserves both the harmonics & spectral envelope. For a high number of tapers, harmonics gets smeared out. The optimum number of tapers is to be dependent on the type of application [2]. In speaker verification both the voice source vocal tract filter are found to be useful, thus expecting to get best 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,

(4)

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 . Each

component density is D-variate Gaussian function of the form,

$$g\left(x/_{\mu i}, \Sigma i\right) = 1/(2\}$$

(5)

With mean vector μi & covariance matrix ∑i . The mixture weights satisfy the constraint that ,

(6)

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices & mixture weights from all component densities. These parameters are collectively represented by notation, as in (4)

$$\lambda = \{wi, \mu i, \sum i\} i = 1, \ldots, M \quad (7)$$

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. Acknowledgment

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.

References

[1] Kinnunen T., Li.,H. An overview of Text Independent Speaker recognition : from feature to supervectors Speech communication (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,CanadaSpeech 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.