International Journal of Electronics Communication and Computer Engineering
Volume 5, Issue (4) July, Technovision-2014, ISSN 2249–071X

Technovision-2014: 1st International Conference at SITS, Narhe, Pune on April 5-6, 2014

Review: GMM based Speaker Verification using MFCC Feature

Multitaper MFCC for Robust Speaker Verification

Rupali G. Shintri
M.E. E & TC (SP)
ICOER, Pune, Maharashtra, India
E-mail: rgshintri@gmail.com

Asst. Prof. S. K. Bhatia
ICOER, Pune, Maharashtra, India
Email: sukhwinderkaur1@gmail.com

Abstract **– In speech & audio applications, short-term signal spectrum is often represented using mel-frequency 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 (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.**

I. 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.

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.

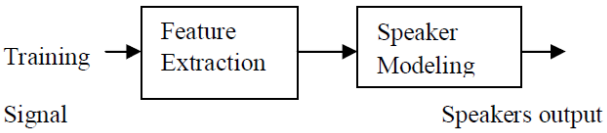
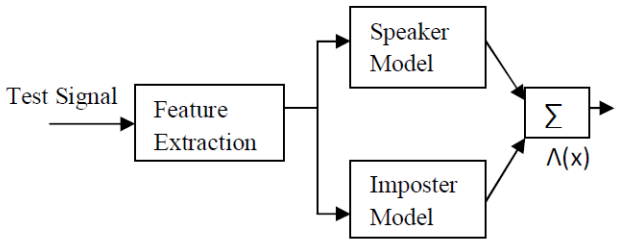


Fig.1. (a) Training Stage



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

Fig.1. (b) Testing Stage

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 speaker's vocal tract shape via resonances (formants) & glottal sources via pitch harmonics. Thus some form of spectral based 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 well under clean conditions but verification accuracy degrades under changes in environment & channel since short term spectrum subject to many harmful variations [2].

But 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, melfrequency wrapping, cepstrum, Mel cepstrum. Mel cepstrum is converted to time domain by, as in [4]

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

International Journal of Electronics Communication and Computer Engineering
Volume 5, Issue (4) July, Technovision-2014, ISSN 2249–071X

Technovision-2014: 1st International Conference at SITS, Narhe, Pune on April 5-6, 2014

From statistical view, the common MFCC implementation based on windowed DFT is suboptimal due to high variance of spectrum estimate. In speaker verification, uncertainty in features is modeled by the variance in the Classifiers (GMM) 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. Which in turn enhances performance of verification.

PROPOSED METHOD

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.

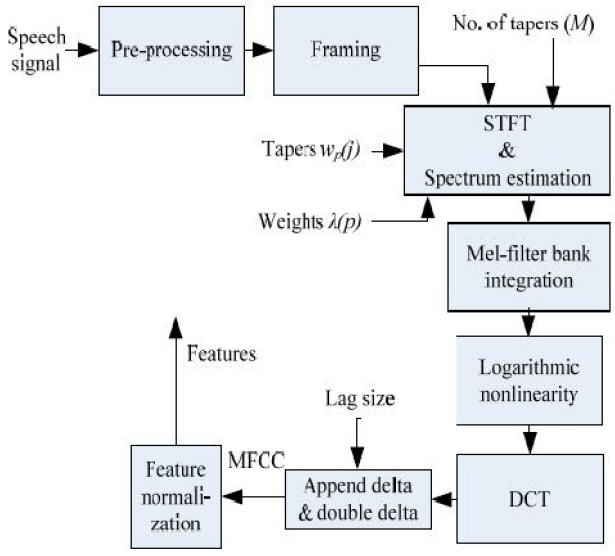


Fig.2. Block diagram of single & multitaper spectrum based MFCC feature extraction

The pre-processing step includes pre-emphasizing, DC removal, 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 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].

III. 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]

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

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.

IV. 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 staircaselike spectrum, multipeak with sharper peaks & sine a compromise between these two methods. For a small number of tapers all methods preserve 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.

V. VARIANCE ESTIMATION

To understand the bias and variance trade-off better as in [2], we consider the variance and spectral resolution of the single- and multi-taper methods. For the windowed DFT the variance is usually approximated as,

$$V[S(f)] \approx S^2(f) \quad (4)$$

The spectral resolution, that is, the frequency spacing under which two frequency components cannot be separated, is approximately Bw = 1/N for the rectangle window but Bw = 2/N for the Hamming window. Note also that it does not depend on the frame length N and thus, including more samples in a frame will *not* reduce the variance. For the multitaper spectrum estimator, the spectral resolution is approximately Bw = (K + 2)/N which is the spectral resolution parameter used in the design of the Thomson and multipeak tapers. The variance can be Approximated

$$V[S(f)] \approx \frac{1}{K} S^2(f) \quad (5)$$

This result is analogous to the well-known result that variance of the mean of sample of size K is inversely proportional to K.

Note that up to this point we have only considered variance in spectral and not MFCC domain. Intuitively it is easy to understand, that if the spectrum is estimated with low variance, the resulting MFCC vector will also have low variance. However, a general rule is that by increasing the number of tapers, we can reduce the variance of the spectrum estimate, hence making the spectrum estimate more robust across random variations.

VI. Effect of MFCC Filter Bank

In comparison of the different MFCC estimators, we evaluate speaker verification accuracy using equal error rate (EER) and minimum detection cost function (MinDCF). EER is the error rate at the threshold EER for which the miss and false alarm rates are equal: EER = Pmiss (EER) = Pfa (EER). MinDCF is used in the speaker recognition evaluations and is defined as min {CmissPmissPtar+CfaPfa (1−Ptar)}, where Cmiss = 10 is the cost of a miss (false rejection), Cfa = 1 is the cost of a false alarm (false acceptance) and Ptar = 0.01 is the prior probability of a target (true) speaker. In addition, detection error trade-off (DET) plots for the entire trade-off of false alarm and miss rates. We were also curious to see the effect of excluding the MFCC filter bank and to compute the 18 coefficients directly from the unwrapped spectrum. It is possible that the double smoothing of multitaper spectrum followed by mel-filter energy integration might be suboptimal for speaker verification where we wish to retain the spectral details in addition to the envelope.

The choice of the spectrum estimation method affects speaker verification accuracy. For each of the multitaper methods – Thomson, multipeak and SWCE – we will vary the number of tapers and contrast the result to the baseline Hamming method. EER and MinDCF,

• Multitaper methods outperform Hamming in both EER and MinDCF for a wide range of taper count (approx. $2 \leq K \leq 10$) [2]. Optimum value of K depends on the method and the objective (EER or MinDCF).

• By including the MFCC filter bank the optimum points shift to left (less tapers) in most cases. This is expected because the MFCC filter bank introduces additional averaging over multitapering. Using MFCC filter bank improves EER and MinDCF and makes the curves generally less ragged, indicating stable parameter setting.

• The performance of the three multitaper methods at their optima is close to each other [2].Thomson shows sharper local minima than multipeak and SWCE methods and gives higher error rates for large number of tapers. Both MSE, EER and MinDCF demonstrate approximately convex shapes and all the three methods give similar performance with optimized K. Secondly, for large K, MSE(Thomson) > MSE(SWCE) ' MSE(Multipeak); the same approximate ordering holds also for EER and MinDCF.

VII. 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,

$$p(x/\lambda) = \sum_{i=0}^{M} wi\, g(x/\mu i, \Sigma i) \quad (6)$$

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

$$g(x/\mu i, \Sigma i) = 1/(2\pi)^{D/2} |\Sigma i|^{1/2} \exp\{-1/2(x-\mu i)' \Sigma (x-\mu i)\} \quad (7)$$

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

$$\sum_{i=1}^{M} wi = 1 \quad (8)$$

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)

$$\lambda = \{wi, \mu i, \sum i\} \; i = 1, \ldots, M \quad (9)$$

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. EM (MAP) Algorithm for GMM: Goal of GMM is to maximize the likelihood function w.r.t. parameters.

- Initialize the means μk , Covariance Ck & mixing Coefficients Πk & evaluate initial stage of likelihood.
- E step: Evaluate responsibilities γ (Znk) using current parameter.
- M step: Re-estimate parameters μk, Ck , Πk (new) & using the current responsibilities (γ(Znk)
- Evaluate log likelihood & check for convergence of either parameters or log likelihood.
- If the convergence criterion is not satisfied go to E step.

The basic idea of the EM algorithm is, beginning with an initial model λ, to estimate a new model λ, such that $p(X|\lambda) \geq p(X|\lambda)$. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached.

VII. Conclusion

It will be possible to construct the robust speaker verification system successfully, by implementing Multitaper MFCC feature extraction method & GMM classifier, which will further help to improve the efficiency of speaker verification by reducing variance of the extracted MFCC.

References

[1] KinnunenT., 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 Lowvariance Multi-taper Features for Distributed Speech Recognition, INRS-EMT, University of Quebec, Montreal, CanadaSpeech Confrernce (2008)

[4] G.Suvarna Kumar, K.A Raju, Dr. Mahan Rao, 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 System: Current Approaches & Feature Trends by, MIT Lincoln Laboratories, Lexington, MA, USA.