Effect of Multitapered Windows on MFCC for Text Independent Speaker Recognition

Suma Paulose1st
Department of Electronics
STAS, Edapally
Kerala
sumasaji.98@gmail.com

Jinimole C G2nd
Department of Electronics
STAS, Edappally
Kerala
jinimanoj81@gmail.com

Abstract: Usually in speaker recognition systems, the short-term speech signal spectrum is often represented by the Mel-Frequency Cepstral Coefficients (MFCC).These coefficients are derived via Hamming windowed Discrete Fourier Transform(DFT).Windowing reduces the spectral leakage; the variance of the spectrum estimate will be still high. Multi tapering method is an elegant extension of the windowed DFT, which uses multiple time-domain windows (tapers) with frequency domain averaging. In this paper, we implemented a text independent speaker recognition system with multi tapered MFCC to show the effect of multi tapering. The classification is done using two different schemes namely, Gaussian Mixture Modeling (GMM) and i-vector methods. A comparison in terms of accuracy is performed with the speaker recognition system without multi taper MFCC as well.The accuracy of the recognition system is found to be improved with the multi tapered MFCC. Thus,with this improvement we can say that multi tapers are simple and robust alternative to Hamming window method and are viable method for replacing conventional MFCCs.

Keywords: MFCC, multi tapered MFCC, GMM, i-vector, Thomson multitaper

I. INTRODUCTION

Speaker recognition systems are a more reliable way of bio-metric recognition today. Speaker recognition may be of two types:speaker identification and speaker verification.In speaker verification,the test speaker's claim is verified against the set of speakers available in the data base.Hence it is a much easier task to get implemented,whereas in speaker identification,the test speaker's voice has to be identified from the set of speaker database.So,it is much more complicated to be implemented.Speaker identification may exist in two different modalities as text-dependent and text-independent.The speakers can utter certain set of words or phrases in text-dependent,but in text-independent the speakers have the freedom to speak any word or phrase of their own choice.The test speakers can be one from the database itself in closed set identification.

Two distinct operational phases are there for any of the two different modalities.They are training phase and testing phase.During training phase,models are created with speech signals of all the speakers to be identified.In testing,the speech signal of the test speaker is compared with the speaker models in the database and one with the closest match is identified as the speaker[1].In this paper,we have implemented a text-independent closed set speaker identification system of 100 speakers and accuracy rate will be compared by making use of features extracted using MFCC with hamming window and multi tapered hamming window.Two different classifiers are used for identification.

The extracted feature vectors of 100 different speakers are modelled using Gaussian Mixture Modelling(GMM)and i-vector methods and accuracy rates are compared.The Hamming time-domain window reduces the spectral leakage resulting from the convolution of the signal and window function spectra.The windowing,therefore,reduces the bias.The variance,unfortunately,remains high[2].This high variance can be lowered by making use of a Multi taper spectrum estimate instead of the Hamming window estimate in MFCC.

The rest of the paper is organized as follows:Section II describes the feature extraction using MFCC and multi taper MFCC.The speaker modelling using GMM-UBM technique and i-vector method are presented in Section III.Section IV gives the details regarding implementation of the speaker identification system and results are explained in section V.Section VI summarizes the conclusion.

II. FEATURE EXTRACTION

A.MEL FREQUENCY CEPSTRAL COEFFICIENTS(MFCC)

The Fig.1 shows the block diagram of a speaker recognition system[3].Feature extraction and speaker modeling are the main blocks.Feature extraction reduces the dimension of the data contained in the input speech signal without losing the speaker specific information.Ideal features must be robust against noise and distortion,occur frequently and naturally in speech,be easy to measure from speech signal,and be difficult to mimic etc.[4].Many different feature extraction techniques exist of which Mel Frequency Cepstral Coefficients(MFCC),Inner Hair Cell Coefficients(IHC),Linear Prediction Coefficients(PLP),and Power Normalized Cepstral Coefficients(PNCC)are the most common ones.

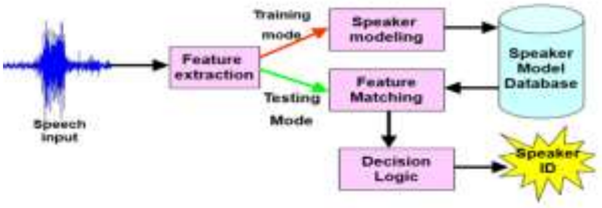

Fig.1.Block Diagram of a Speaker Recognition System

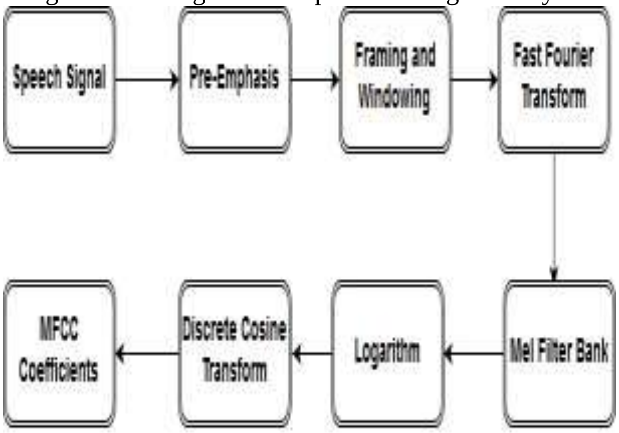

Fig.2.Block Diagram of MFCC

MFCC is one of the oldest and robust techniques used for feature extraction in speaker recognition. The block diagram of MFCC is as shown in Fig.2 [5].The MFCCs are found to be robust and reliable even if there are variations in the speakers and recording conditions. The input speech signal is framed into 25ms frames each with an overlap of 15ms.Each of these frames is multiplied by a Hamming window to get rid of the discontinuities occurring at the edges of the frames.The window function for hamming window of length n is given as in (1).

$$w(n) = [0.54 - 0.46\cos(\frac{2\pi n}{L})] \;; \quad 0 \le n \le L-1 \qquad (1)$$

where w(n)is the Hamming window and n is the total number of samples and L is the window length[3].Since acoustic perceptions does not follow the linear frequency scale,MFCC make use of a perceptual pitch scale called Mel scale for feature extraction.The equation used to convert linear scale frequency f to Mel scale frequency is given in (2)[4].

$$Mel(f) = 2595 log_{10}\left(1 + \frac{f}{700}\right) \qquad (2)$$

This single windowed Hamming window reduces the bias but these results in large variance. Hence MFCCs obtained will also have large variance. Many attempts are being done to enhance the robustness of MFCC features. One such method is to use multi tapered windows. These multiple time domain windows may be used as a solution to eliminate the problem of high variance. This is called as multi taper spectral estimation method.

B. Multitapers

Multi-taper methods reduce the variance of spectral estimates by using multiple orthogonal window functions rather than a single window. In a multi-taper spectrum estimation method the speech signal is first multiplied by not one but a family of tapers which are resistance to spectral leakage. This yields several tapered speech signals from one record. Taking the Discrete Fourier Transforms(DFTs)of each of these tapered signal several eigen spectra are produced which are combined(using a weighted averaging technique)to form the final multitaper spectral estimate. A windowed direct spectrum estimator is the most often used power spectrum estimation method for speech processing applications, such as speech and speaker recognition, and speech enhancement. The periodogram was the first(nonparametric)direct spectral estimate of the power spectral density(PSD)function. The periodogram is a biased estimate due to spectral leakage via the side lobes. It thus becomes necessary to use the method of tapering(windowing)to effectively reduce this bias.

The use of a window (also called a taper)affects the estimate by reducing leakage but it doesn't change the variance of the estimate at each frequency.A common approach to reduce the variance is taking an average across several frequencies and/or computing an average spectrum from several time epochs.Averaging across frequencies reduces spectral resolution and using multiple epochs is undesirable if the signal may be non-stationary.

The multitaper approach, first described in a seminar paper by Thomson(1982),improves the spectral estimate by addressing both leakage and variance in the estimate. In this approach, every taper k_v out of a set of K tapers is a bit different and reduces leakage of energy across frequencies. In addition, in Thomson's approach, the tapers are orthogonal and they are used to provide K orthogonal samples of the data. These samples are used to create a set of K spectral estimates that can be used to compute an average with reduced variance.

In the multi-taper method, only the first of the data tapering windows has the traditional shape. The spectra from the different tapers do not produce a common central peak for a harmonic component. Only the first taper produces a central peak at the harmonic frequency of the component. The other tapers produce spectral peaks that are shifted slightly up and down in frequency. Each of the spectra contributes to an overall spectral envelope for each component. These tapers are called S*lepian tapers* since they follow a Slepian sequence.

In the Thomson multi-taper method of spectrum estimation, a set of M orthonormal data tapers with good leakage properties is specified from the slepian sequences. Slepian sequences are defined as the real, unit-energy sequences on[0,N−1]having the greatest energy in a bandwidth W.

Slepian sequences proposed originally in,were chosen as tapers in,as these tapers are mutually orthonormal and possess desirable spectral concentration properties(i.e.,they have highest concentration of energy in the user-defined frequency interval(-W,W)).The first taper in the set of Slepian sequences is designed to produce a direct spectral estimator with minimum broadband bias(bias caused by leakage via the sidelobes).The higher order tapers ensure minimum broadband bias whilst being orthogonal to all of the lower order tapers. The first taper, resembling a conventional taper such as Hanning window, gives more weight to the center of the signal than to its ends. Tapers for larger *p(no.of tapers)*give increasingly more weight to the ends of the signal.There is no loss of information at the extremes of the signal.[16].

C.Multi tapered MFCC

The spectral leakage that occurs because of the convolution of the input speech signal and the window function spectra in MFCC can be reduced by the Hamming-type of time-domain window.The Hamming window is symmetric and the taper of such windows decreases towards the boundaries of each frame.The windowing surely reduces the bias,but the variance is still very high.Bias is actually the expected value of the difference obtained between the estimated spectrum and original spectrum.In single window estimates important parts of the signal that comes at the end portion of the spectrum may get discarded.This is the reason for the variance to get increased[6].Hence the variance of the MFCC features extracted from this spectral estimate will also be high.This high variance can be reduced by making use of a multi taper spectrum estimate instead of Hamming window DFT spectrum estimate.In multi tapering,the framed speech signal passes through different window functions and a weighted average of the individual sub-spectra is obtained as the final spectrum.The window functions or tapers are designed in such a way that the estimation errors that occurs in the individual sub-spectra are approximately uncorrelated and the variances must get reduced.A low-variance spectrum estimates and,of course a low-variance MFCC estimate will result while averaging these uncorrelated spectra [2].

In this method,only the first window will be having the original shape of the Hamming window.The spectra obtained of the remaining windows will not have a common central peak instead they produce spectral peaks that are slightly shifted in frequency[7].The basic idea behind the multi taper spectral estimation method is the analysis of the speech frames using a number of spectrum estimators(M)and each will be having a different taper.Then the final spectrum is computed as the weighted mean of each sub spectrum.By doing this,it is shown that multiple window spectral estimates are having smaller variance than single windowed spectrum estimates by a factor that approaches number of spectrum estimators(1/M).The multi taper spectrum estimator, which uses M orthogonal window functions instead of a single Hamming window, may be expressed as in (3) [15].

$$S_{MT}(m,k) = \sum_{p=1}^{M} \lambda(p) | \sum_{j=0}^{N-1} w_p(j) s(m,j) e^{\frac{2\pi jk}{N^2}} | \qquad (3).$$

where N is the frame length and $\mathbf{w_p}$**Error! No bookmark name given.**is the p^{th} data taper(p=1,2...,.M)used for estimating the spectrum.Finally,λ(p)is the weight of the p^{th} taper. The tapers $\mathbf{w_p(p)}$are typically chosen to be orthonormal. The multitaper spectrum estimate is therefore obtained as the weighted average of M individual sub-spectra.

For spectrum estimation, a number of different tapers such as Thomson, Multi-peak and Sinusoidal Weighted Cepstrum Estimator(SWCE)etc. have been proposed which are based on the Slepian tapers, peak matched multiple tapers, and sine tapers,respectively. For flat spectra, Thomson tapers and for voiced speech, Multi-peak tapers are commonly used. SWCE are used for Cepstrum analysis.

In Thomson multi taper,three different weighting schemes may be used such as uniform weights,eigenvalue weights and adaptive weights.Uniform weights are commonly used in extracting MFCC multi taper features.Higher accuracy is provided by adaptive weights when compared to the uniform and eigenvalue weighting schemes[8].

III. Speaker Recognition Methods

After extracting the features,a database is to be created containing the speaker models of all speakers that need to be identified in future comparisons.Two different modeling methods are used in this work:Gaussian Mixture Modeling(GMM)and i-vector using Probabilistic Linear Discriminatory Analysis(PLDA).

A. Gaussian Mixture Modeling(GMM)

Gaussian Mixture Modeling may be expressed as the weighted sum of multiple Gaussian distributions and is given by the (4).

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i \vec{x} \qquad (4)$$

where**Error! No bookmark name given.**is a ***D*** dimensional random vector,**Error! No bookmark name given.**)with**Error! No bookmark name given.**are component densities and $\boldsymbol{p_i}$ with i=1,2,…M**Error! No bookmark name given.**are the mixture weights [9].

The complete Gaussian mixture density is parameterized by the mean vectors,covariance matrices and the mixture weights from all component densities.These parameters are collectively represented by the notation given in (5).

$$\lambda = \{p_i, \vec{\mu}, \Sigma_i\}, \quad i = 1,2,\ldots,M \qquad (5)$$

For speaker identification,each speaker is represented by a GMM and is referred to by his/her model λ **Error! No bookmark name given.**[9].

To estimate the parameters of the GM model Expectation Maximization(EM)algorithm is used.It is an iterative algorithm consisting of two steps,Expectation(E)step and Maximization(M)step.The expectation of the likelihood function is computed in the E step and the parameter that maximizes the likelihood function is computed in the M step.For each test signal,the log likelihood score is calculated.[10].

B. i-vector and Probabilistic Linear Discriminant Analysis(PLDA)

Speaker recognition systems may be implemented using i-vectors as well and for classification PLDA may be used.The i-vector can be considered as a compact representation of a speech signal, which is obtained from a Gaussian Mixture Model(GMM)super vector.GMM is a super vector model and such a model may be decomposed into speaker independent, speaker dependent, channel dependent, and residual components.i.e.each of the speech signal may be represented by a set of low dimensional vectors[11]called as total vector or identity vector(i-vector).The basic idea of i-vector approach is that each speaker and channel-dependent GMM super vector ***M*** can be modeled as by using (6).

$$M=m+Tw+\epsilon \tag{6}$$

where ***m*** is a speaker-and channel-independent super vector,whose value is often taken from UBM super vector,***T*** is the total factor matrix with low rank,which expands a subspace containing speaker and channel-dependent information and ***w*** is a standard normal distributed vector [11].

Probabilistic linear discriminant analysis is a method of modelling the speaker and channel variability[9].For the i^{th} speaker,the i-vector $w_{i,j}$ representing the j^{th} recording can be represented as in(7).

$$w_{ij} = m + Sx_i + \varepsilon_{ij} \tag{7}$$

where $m+Sx_i$ is the speaker-dependent part,***m*** is a global mean of all i-vectors,***S*****Error! No bookmark name given.**is a set of basis vectors for the speaker subspace,representing between-speaker variability,and$\epsilon_{i,j}$ is the residual noise with covariance∑ **Error! No bookmark name given.**[9].

Given a test i-vector w_t and any target i-vector w_1 the likelihood score is given by (8).

$$S(w_1, w_t) = \frac{p(w_1, w_t|H_1)}{p(w_1|H_0)p(w_t|H_0)} \tag{8}$$

where,H_1 is the hypothesis that both i-vectors come from same speakers and H_0 is the hypothesis that they come from different speakers[9].

C. Channel Compensation Methods

The extraction of i-vectors is done in such a way that there is no distinction made between the speaker and channel variability. So the separation or removal of channel variability is taken care of before creating the classifiers for the recognition of speakers, which take i-vectors as the input features. Channel compensation methods are estimated based on the within-class and between-class variances [11].So here, a combination of two methods are used for achieving the channel compensation. These methods are the Linear Discriminant Analysis or LDA and the Within Class Covariance Normalization or WCCN.

The main purpose of using LDA is to achieve dimensionality reduction while retaining as much of speaker discriminatory information as possible and it is a supervised method [11]. WCCN is done as a pre-processing step for performing the PLDA classification [11].This helps to attenuate high variance within the same speaker class and hence helps in achieving better recognition during testing phase.

IV. Implementation

Speech signals were taken from TIMIT database,which consists of 10 speech signals each for 630 speakers.The implementation is done in MATLAB.The sampling frequency selected is 16 KHz.The duration of each of the signals in TIMIT is 2 sec to 3 sec.out of the total 630 speakers available,all the 10 speech signals from 100 speakers is taken for the implementation.Of the 10 signals of each speaker,7 were used for training and 3 for testing.Each of these signals were framed into 25ms frames with 15ms overlap.Each of these frames was weighted by a Hamming window or multitaper window method.To generate the multitapered Thomson tapers,the multitaper functions described by Kinnunen et al.were used[12].For i-vector modelling,the size of i-vector is chosen to be 100 and the number of Gaussian components selected is 32.Using the parameters zero crossing rate(ZCR)and energy,the voiced and unvoiced parts of the speech signal, were separated. To obtain the voiced speech signals, all frames with an energy greater than 0.5 or ZCR less than 100 is taken[13].

V. Results and Discussion

The accuracy rates of speaker recognition have been obtained for a text-independent speaker recognition system using GMM and i-vector methods. These systems were realized using two feature extraction methods: MFCC and multitapered MFCC. The multitapered MFCC made use of Thomson tapers with uniform window length.The performance of both is compared and the results obtained are tabulated as in Table I.

Table I. Accuracy Obtained for Voiced short utterances

Feature	GMM	i-vector
MFCC	89.33	73.66
Multitapered MFCC	93.66	75.33

From Table I it can be understood that multitapered MFCC performs better for both GMM and i-vector and also GMM outperforms i-vector.

The accuracy rates for full speech and voiced speech was tested and these accuracy rates are tabulated in Table II and from this it is clear that the performance of full speech is better for both features.

From Tables I and II it is understood that i-vector based speaker recognition systems require large amount of data to estimate its parameters.That is why the accuracy rates of i-vector modelling is less when compared to GMM.To enhance the accuracy rates,we concatenated the test speech signals and accuracy rates were found to increase [14].

By concatenating the duration of the test signals was increased to 6 sec to 9 sec.The results so obtained are tabulated in Table III.

Table II. Accuracy Obtained for full speech and voiced signals

Training	Testing	Feature	GMM	i-Vector

Training	Testing	Feature	GMM	i-Vector
Full speech	Full speech	MFCC	94.33	79.66
Voiced	Voiced	MFCC	89.33	73.66
Fullspeech	Full speech	Multitapered MFCC	97.33	81.33
Voiced	Voiced	Multitapered MFCC	93.66	75,33

Table III. Accuracy of i-vectors for long and short utterances

Feature	Shorter Utterance	Longer Utterance
MFCC	73.66	94
Multitapered MFCC	75.33	98

VI. CONCLUSION

A comparison of the text independent speaker recognition system using two different feature extraction methods were done. The accuracy obtained using multitapered MFCC was found to be more compared with the accuracy obtained using Hamming Windowed MFCC. The high variance of the Hamming windowed MFCC is reduced by making use of a multitapered Thomson window. From the analysis it is also clear that full speech performs better than voiced speech and GMM.

REFERENCES

[1] Roberto Togneri, Daniel Pullella "An Overview of Speaker Identification :Accuracy and Robustness," IEEE Circuits and Magazine, pp. 23-61, 2011

[2] Tomi Kinnunen, Rahim Saeidi, Johan Sandberg, and Maria Hansson-Sandsten, "What Else is New Than the Hamming Window? Robust MFCCs for Speaker Recognition via Multitapering".

[3] Siddhant C. Joshi, A.N.Cheeran. "MATLAB Based Feature Extraction Using Mel Frequency Cepstrum Coefficients for Automatic speech Recognition", International Journal of Science, Engineering and Technology Research, 3 2014; 1820-1823

[4] Tomi Kinnunen, Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors", Speech Communication 2010; 52, 12-40.

[5] Namratha Dave, "Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition", International Journal for Advanced Research in Engineering and Technology, 2013.

[6] Santosh v. Chapaneri, Deepak D. J ayaswal, "Multi-Taper Spectral Features for Emotion Recognition from Speech", International Conference on Industrial Instrumentation and Control, 2015, 1044-1049.

[7] Md Jahangir Alam, Tomi Kinnunen, Patrick J. Kenny,Pierre Ouellet, Douglas O'Shanaughnessy, "Multitaper MFCC and PLP Featuresfor speaker verification using i-vectors", Speech Communication ,2013,55,237-255.

[8] Omer Eskidere and Ahmet Gurhanli, Voice Disorder classification based on Mel Frequency Cepstral Coefficients Features", Computational and Mathematical Methods in Medicine, Hindawi Publishing Corporation, 2015.

[9] Padmanabhan Rajan, Anton Afanasyev, Ville Hautamaki,Tomi Kinnunan, "From Single to Multiple Enrollment i-vectors : Practical PLDA Scoring Variants for Speaker Verification", Digital Signal Processing, Vol. 31, pp. 93-101, 2014

[10] Tomi Kinnunen, Haizhou Lib, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors," Elsevier Speech Communication, vol.52, no. 1, pp. 12-40, 2009.

[11] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet "Front-End Factor Analysis for Speaker Verification," Transactions on Audio, Speech and language Processing, vol. 19, no.4, pp. 788-798, May 2011

[12] T. Kinnunen, R. Saeidi, F. Sedl´ak et al., "Low-variance multitaper MFCC features : a case study in robust speaker verification," IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 7, pp. 1990–2001, 2012.

[13] Dominic Mathew, V.D. Devassia and Tessamma Thomas,"A K-means Clustering Algorithm for Frequency Estimation and Classification of Speech Signals", IEEE Conference/ICSIP 2006

[14] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason. "I-vector based speaker recognition on short utterances," Proceedings of the 12th Annual Conference of the International Speech communication Association, pp. 2341-2344. International Speech Communication, 2014

[15] Yazid Attabi1, Md Jahangir Alam, Pierre Dumouchel , Patrick Kenny, Douglas O'Shaughnessy. "Multiple Windowed Spectral Features For Emotion Recognition" .

[16] Md Jahangir Alam, Tomi kinnunen, Patrick Kenny, Pierre Ouellet, Douglas O'Shaughnessy. "Multitaper MFCC and PLP Features for Speaker Verification Using i-Vectors", Article in Speech Communication , February 2013.