A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition

Md Sahidullah, *Student Member, IEEE*, and Goutam Saha, *Member, IEEE*

***Abstract*—In this letter, we propose a novel family of windowing technique to compute mel frequency cepstral coefficient (MFCC) for automatic speaker recognition from speech. The proposed method is based on fundamental property of discrete time Fourier transform (DTFT) related to differentiation in frequency domain. Classical windowing scheme such as Hamming window is modified to obtain derivatives of discrete time Fourier transform coefficients. It is mathematically shown that this technique takes into account slope of power spectrum and phase information. Speaker recognition systems based on our proposed family of window functions are shown to attain substantial and consistent performance improvement over baseline single tapered Hamming window as well as recently proposed multitaper windowing technique.**

***Index Terms*—Differentiation in frequency, power spectrum estimation, speaker recognition, tapered window, mel-frequency cepstral coefficients (MFCC).**

I. Introduction

MEL FREQUENCY cepstrum coefficient (MFCC) extraction schemes use discrete Fourier transform (DFT) for calculating short-term power spectrum of speech signal. During this process, Hamming or Hanning window is applied to raw speech frames in order to reduce spectral leakage effect. These windows have reasonable sidelobe and mainlobe characteristics which are required for DFT computation. However, there exist various other window functions which also have good behavior in terms of certain parameters of their frequency responses [1]. In practice, selecting the optimal window function for speech processing application is still an open challenge [2]. Recently, alternatives of Hamming window have drawn attention of the researchers [3], [4]. For example, performance of speaker recognition systems based on MFCC, extracted using multitaper window function, are shown comparatively robust than existing single tapered Hamming window based approach [5].

In this work, we propose a simple time domain processing of speech after it is multiplied with a standard window. The processing is based on well-known *difference in frequency* property of discrete time Fourier transform [6], and it can be easily integrated with standard window during DFT computation. Due to the proposed modification, we inherently compute derivative of Fourier transform. Power spectrum is computed from those differentiated Fourier coefficients. There are evidences that speaker discriminating attribute is present in slope of power spectrum [7] as well as in phase information [8], [9]. In this letter, we have mathematically shown that our proposed technique integrates both slope and phase information with magnitude spectrum. Therefore, it can be hypothesized that the speech feature extraction from these modified Fourier coefficients will give better recognition performance. We have evaluated the performance in multiple databases for speaker verification (SV) task, and consistent performance improvement is achieved over Hamming window based baseline system.

The rest of the letter is organized as follows. In Section II, we describe the proposed windowing scheme and its features. In addition to that, the effect of newly introduced window in power spectrum computation is mathematically analyzed. Experimental results are shown in Section III. Finally, the letter is concluded in Section IV.

II. Proposed Windowing Method

A. Design of Proposed Window Function

Let $x(n)$ be a windowed speech frame of length N and its DTFT is given by, $X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n}$. Now, differentiating $X(e^{j\omega})$ w.r.t. ω, a relationship can be obtained between DTFT of $nx(n)$ and $X(e^{j\omega})$. This is known as *differentiation in frequency* property [6], [10]. We can express DTFT of $nx(n)$ as $\hat{X}(e^{j\omega}) = j(dX(e^{j\omega}))/(d\omega)$. As DFT coefficients $X(k)$ are samples of DTFT at $\omega = (2\pi k)/(N)$, DFT of $nx(n)$ are discrete samples of $\hat{X}(e^{j\omega})$ at $\omega = (2\pi k)/(N)$. Therefore, $\hat{X}(k) = \hat{X}(e^{j\omega})|_{\omega=(2\pi k)/(N)}$ are the DFT coefficients of $nx(n)$.

Since $x(n)$ is a windowed speech frame, it can be represented as $w(n)s(n)$, where $s(n)$ is raw speech frame and $w(n)$ is window function. We propose new window function as $\hat{w}(n) = nw(n)$. The windowed speech frame is then represented as $\hat{x}(n) = \hat{w}(n)s(n)$.

From generalization of differentiation in frequency property, we can write that, for an integer τ, DTFT of $n^{\tau}x(n)$ is $j^{\tau}(d^{\tau}X(e^{j\omega}))/(d^{\tau}\omega)$. Therefore, the proposed window function of τ-th order window can be written as $n^{\tau}w(n)$. Standard Hamming window can be viewed as *zero order window* of proposed family. The window functions are shown in Fig. 1 for first and second order along with Hamming window. Note that in contrast to frequently used window functions, the newly introduced family of window functions is asymmetric and non-tapered.

Manuscript received June 11, 2012; revised August 27, 2012; accepted December 06, 2012. Date of publication December 20, 2012; date of current version January 04, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Anna Scaglione.

The authors are with the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India (e-mail: sahidullahmd@gmail.com; gsaha@ece.iitkgp.ernet.in).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2012.2235067

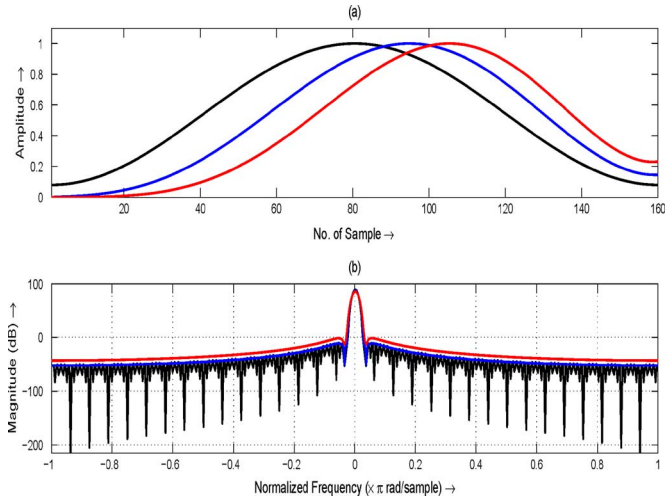

Fig. 1. Comparison of Hamming window (black) with first (blue) and second (red) order differentiation based window in (a) time domain and (b) frequency domain for a window of size 160 samples. Amplitude of all the window functions are normalized to one for visual clarity.

TABLE I
PERFORMANCE METRICS OF VARIOUS WINDOW FUNCTIONS. SEQUENCE LENGTH IS 160 SAMPLES I.E., 20 MS FOR SAMPLE RATE 8 KHZ

Window	Leakage Factor	Relative Sidelobe Attenuation	Mainlobe Width
Hamming	0.04%	-42.6 dB	0.015625
$\tau = 1$	0.06%	-42.6 dB	0.017578
$\tau = 2$	0.17%	-37.9 dB	0.018555

B. Characteristics of the Proposed Window Function

Commonly, the effectiveness of a window function is judged by different performance metrics [1]. In order to evaluate the performance of the window in DFT computation, various performance metrics are computed prior to the application of this window function in speech feature extraction. We have calculated three widely used performance evaluation metrics: spectral leakage factor, relative sidelobe attenuation, and mainlobe width (-3 dB) of the Hamming and proposed windows of different orders. The results are shown in Table I for window size of 160 samples. It can be observed that with the increase of order, the spectral leakage increases and sidelobe attenuation decreases to some extent which have minor effect in recognition performance. However, considerable increase in mainlobe width will help to estimate smooth power spectrum, and that is expected to improve recognition performance [10].

C. Effect of the Proposed Window in Power Spectrum Computation

In this subsection, we find out a mathematical connection between power spectrum of proposed windowed speech frame and power spectrum of original Hamming windowed speech frame.

Let us assume that power spectrum of Hamming windowed signal is given by $P(\omega)$, and power spectrum of the proposed window is $\hat{P}(\omega)$. Therefore, $P(\omega) = H^2(\omega) = |X(e^{j\omega})|^2$ and $\hat{P}(\omega) = \hat{H}^2(\omega) = |(dX(e^{j\omega}))/(d\omega)|^2$, where $H(\omega)$ and $\hat{H}(\omega)$ are magnitude spectrum of two signals respectively. Now, since $X(e^{j\omega})$ can be decomposed into a real, $X_R(\omega)$ and imaginary, $X_I(\omega)$ part, the slope of magnitude spectrum of Hamming windowed speech signal can be written as,

$$\begin{aligned}\frac{dH(\omega)}{d\omega} &= \frac{d\left|X(e^{j\omega})\right|}{d\omega} \\ &= \frac{X_R(\omega)}{\sqrt{X_R^2(\omega)+X_I^2(\omega)}}\frac{dX_R(\omega)}{d\omega} \\ &\quad + \frac{X_I(\omega)}{\sqrt{X_R^2(\omega)+X_I^2(\omega)}}\frac{dX_I(\omega)}{d\omega} \end{aligned} \tag{1}$$

On the other hand, magnitude spectrum of the modified signal can be written as,

$$\begin{aligned}\hat{H}(\omega) &= \left|\frac{dX(e^{j\omega})}{d\omega}\right| \\ &= \sqrt{\left(\frac{dX_R(\omega)}{d\omega}\right)^2 + \left(\frac{dX_I(\omega)}{d\omega}\right)^2}\end{aligned} \tag{2}$$

Now, if we consider that $(dX_R(\omega))/(d\omega) = a(\omega)\cos\varphi(\omega)$ and $(dX_I(\omega)/d\omega) = a(\omega)\sin\varphi(\omega)$, then $a(\omega) = \sqrt{((dX_R(\omega))/(d\omega))^2 + ((dX_I(\omega))/(d\omega))^2}$ and $\varphi(\omega) = \tan^{-1}((dX_I(\omega))/(d\omega)/(dX_R(\omega))/(d\omega)) = \tan^{-1}(dX_I(\omega)/dX_R(\omega))$. Therefore, from (2),

$$a(\omega) = \hat{H}(\omega) \tag{3}$$

On the other hand, if we put $(X_R(\omega)/\sqrt{X_R^2(\omega)+X_I^2(\omega)}) = \cos\phi(\omega)$ and $(X_I(\omega)/\sqrt{X_R^2(\omega)+X_I^2(\omega)}) = \sin\phi(\omega)$ in (1) we get,

$$\begin{aligned}\frac{dH(\omega)}{d\omega} &= \frac{X_R(\omega)}{\sqrt{X_R^2(\omega)+X_I^2(\omega)}} \times a(\omega)\cos\varphi(\omega) \\ &\quad + \frac{X_I(\omega)}{\sqrt{X_R^2(\omega)+X_I^2(\omega)}} \times a(\omega)\sin\varphi(\omega) \\ &= a(\omega)[\cos\phi(\omega)\cos\varphi(\omega) + \sin\phi(\omega)\sin\varphi(\omega)] \\ \therefore \frac{dH(\omega)}{d\omega} &= a(\omega)\cos[\phi(\omega)-\varphi(\omega)]\end{aligned} \tag{4}$$

where $\phi(\omega) = \tan^{-1}(X_I(\omega))/(X_R(\omega))$. Therefore, from (3) and (4), we get,

$$\hat{H}(w) = \frac{dH(\omega)}{d\omega} \times \sec[\phi(\omega)-\varphi(\omega)] \tag{5}$$

Finally, we can write the final expression of the output power spectrum $\hat{P}(\omega)$ as,

$$\hat{H}^2(\omega) = \frac{1}{4P(\omega)}\left[\frac{dP(\omega)}{d\omega}\right]^2 \times \sec^2\left[\phi(\omega)-\varphi(\omega)\right]. \tag{6}$$

The term $(dP(\omega))/(d\omega)$ in (6) corresponds to the slope of the power spectrum of the Hamming windowed speech at frequency ω. Hence, as a consequence of power spectrum computation from derivative of Fourier transform, we obtain a modified power spectrum which is related to the slope of original power spectrum. Apart from it, the newly formulated power spectrum is also related to phase spectrum of the signal $\phi(\omega)$. Using a more complicated computation, it can also be shown that the higher order version of proposed differentiation window (e.g., for $\tau > 1$) will compute power spectrum with higher order derivative of $P(\omega)$.

TABLE II
DATABASE DESCRIPTION (CORETEST SECTION) FOR THE PERFORMANCE EVALUATION OF VARIOUS WINDOW FUNCTIONS

	SRE 2001	SRE 2004	SRE 2006
Target Models	74♂, 100♀	246♂, 370♀	354♂, 462♀
Test Segments	2038	1174	3735
Total Trial	22418	26224	51068
True Trial	2038	2386	3616
Impostor Trial	20380	23838	47452

The modified DFT magnitude coefficients are nothing but the samples of $\hat{H}(\omega)$ at $\omega = (2\pi k)/(N)$. Therefore, mel cepstrum computation using proposed window integrates the slope of power spectrum, phase, and of course, power spectrum of the signal. It is expected that the speech feature will be more efficient compared to the standard cepstrum which is solely based on power spectrum.

III. EXPERIMENTAL SETUP AND RESULTS

A. Speaker Recognition Setup

1) Database: SV experiments are conducted on multiple large population NIST corpora for obtaining statistically significant results. We have chosen SRE 2001, SRE 2004, and SRE 2006. The database descriptions for current experiments are briefly shown in Table II.

2) Feature Extraction: 38 dimensional MFCC feature vectors has been extracted for different types of window functions using 20 filters linearly spaced in Mel scale from speech frames of size 20 ms (with 50% overlap). Detailed explanation of used MFCC computation technique is available in [7].

3) Classifier Description: The performances of speaker recognition systems have been evaluated using Gaussian mixture model-universal background model (GMM-UBM) based classifier [11]. The speech data for UBM training are taken from development data of SRE 2001 and training section of SRE 2003 for the evaluation of SRE 2001 and SRE 2004 respectively. The number of mixtures is set at 256 for these experiments. Here, gender dependent GMM clusters are initialized using binary split based vector quantization. The final UBM parameters are estimated using EM algorithm. Target models are created by adapting only the means of the UBM with relevance factor 14. During the score computation, top-5 Gaussians of corresponding background model per each frame are considered.

For the evaluation of SRE 2006, the GMM-UBM system is trained with 512 mixtures of gender dependent UBM with complete one side training data of SRE 2004 (i.e., 246 male and 370 female utterances). zt-score normalization is performed on raw score of GMM-UBM system. Normalization data is obtained from one side section of SRE 2004. Experiments are also conducted using classifiers based on GMM supervector and support vector machine (GSV-SVM) [12]. This is based on the same UBM of GMM-UBM system. The negative examples of SVM are obtained from the same data used for UBM preparation. Experiments are also carried out with nuisance attribute projection (NAP) based channel compensation technique [13]. Channel factors are obtained using the speech signals of SRE 2004. All together, 699 utterances of 101 male and 905 utterances of 142 female are utilized to train the NAP projection matrix of co-rank 64.

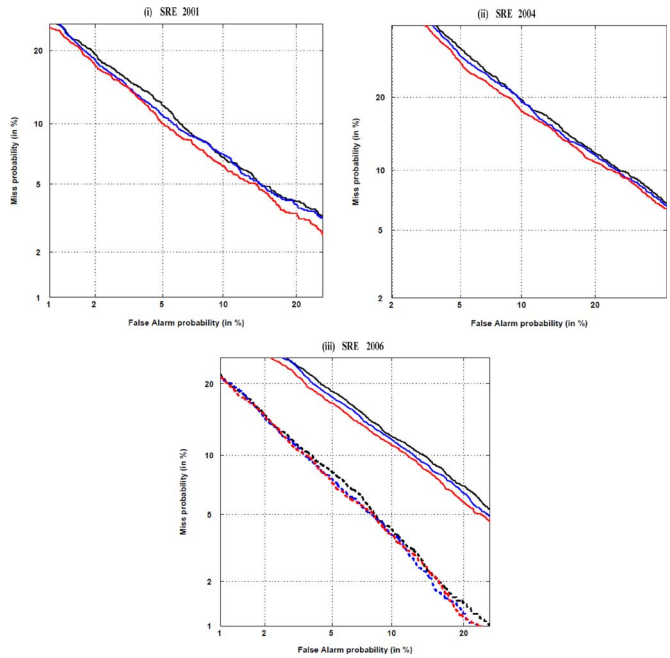

Fig. 2. DET plots of different window based systems (Black: Hamming, Blue: first order, Red: second order) are shown for (i) SRE 2001, (ii) SRE 2004, (iii) SRE 2006. In subfigure (iii), the dotted lines show results for GSV-SVM system with NAP.

B. Results

Speaker recognition experiments are carried out with different window function keeping other blocks identical i.e., pre-processing, feature extraction and classification are precisely same for all various window based systems. We first evaluate the performance on SRE 2001 and SRE 2004 with classical GMM-UBM system. The performance of proposed windows (first and second order) are compared with single tapered Hamming window as well as recently proposed multitaper window. The performance has been evaluated with multipeak taper of size (denoted by k in Table III) 6 and 12 as mentioned in [14], [5]. The results are shown in Table III and corresponding detection error trade-off (DET) plots are shown in Fig. 2(i) and (ii). Equal error rate (EER) and minimum detection cost function (minDCF) of SV systems based on newly proposed window functions are consistently better for both the databases. In comparison with baseline Hamming window based system, we have obtained 0.6% and 7.74% relative improvement in EER, and 0.26% and 5.59% relative improvement in minDCF for SRE 2001. In contrast, for SRE 2004, the relative improvements in EER are 1.96% and 4.26%, and for minDCF these are 1.15% and 3.45%. Interestingly, we have observed that multitaper windowing techniques do not give better performance as compared to the proposed method. SV experiments are also conducted with an existing phase based feature: modified group delay function (MODGDF) [9] and the score level classifier fusion is performed with the MFCC based systems. It has been observed that the performance obtained

TABLE III
SV PERFORMANCE ON NIST SRE 2001 AND NIST SRE 2004 USING VARIOUS WINDOW FUNCTIONS FOR GMM-UBM BASED SYSTEM

Window Type	NIST SRE 2001		NIST SRE 2004	
	EER (in %)	minDCF $\times$ 100	EER (in %)	minDCF $\times$ 100
Hamming	8.2434	3.5763	14.9629	6.3231
Multitaper ($k = 6$)	8.0471	3.5778	15.2501	6.4363
Multitaper ($k = 12$)	10.9372	4.6606	18.0196	7.2271
First Order	8.1943	3.5672	14.6694	6.2501
Second Order	7.6055	3.3763	14.3255	6.1050

TABLE IV
SV PERFORMANCE ON NIST SRE 2006 USING VARIOUS SYSTEMS FOR DIFFERENT WINDOW FUNCTION

Window Type	GMM-UBM		GSV-SVM		GSV-SVM (with NAP)	
	EER (in %)	minDCF $\times$ 100	EER (in %)	minDCF $\times$ 100	EER (in %)	minDCF $\times$ 100
Hamming	11.4493	4.4702	8.8471	4.0330	6.6419	3.1161
Multitaper ($k = 6$)	11.6981	4.5493	9.0705	4.2211	6.8886	3.2725
Multitaper ($k = 12$)	14.2971	5.2299	11.430	5.0286	8.2416	3.9699
Proposed First Order	10.9856	4.3521	8.3792	4.0233	6.2503	3.0961
Proposed Second Order	10.7559	4.2627	8.3242	3.9555	6.1359	3.0646

with the fusion of proposed MFCCs and MODGDF based systems are better than the combination of baseline MFCC and MODGDF based systems.

In Table IV, the performance is shown for different classifiers on SRE 2006. Also, in this case, we have achieved consistent and reasonable performance improvement for proposed window based SV systems. The DET plots are shown in Fig. 2(iii) for both GMM-UBM and GSV-SVM (with NAP) system. It can easily be interpreted that SV systems based on the proposed window functions are consistently better than Hamming window based baseline system. It is also observed that performances of second order window based systems are better than first order window based systems.

IV. CONCLUSION

In this letter, we have focused on the usage of a class of window functions by which more effective speech feature can be computed. The newly formulated feature represents the power spectrum of the original spectrum as well as its derivative. In addition to that, it also integrates phase information which is also relevant for speaker recognition. Speaker recognition system based on proposed windowing schemes are evaluated on different NIST databases. We have achieved consistent performance improvement over baseline Hamming window based technique on various combinations of classifiers and databases.

REFERENCES

[1] F. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[2] J. O. Smith, *Spectral Audio Signal Processing*. Burlingame, CA: W3K, 2011.

[3] M. Mottaghi-Kashtiban and M. Shayesteh, "New efficient window function, replacement for the hamming window," *Signal Process.*, vol. 5, no. 5, pp. 499–505, Aug. 2011.

[4] Y. Wang, "An effective approach to finding differentiator window functions based on sinc sum function," *Circuits, Syst. Signal Process.*, vol. 31, no. 5, pp. 1809–1828, Oct. 2012.

[5] J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P. Flandrin, and P. Borgnat, "Multitaper estimation of frequency-warped cepstra with application to speaker verification," *IEEE Signal Process. Lett.*, vol. 17, no. 4, pp. 343–346, Apr. 2010.

[6] A. Oppenheim, S. Willsky, and S. Nawab, *Signals and Systems*, 2nd ed. New Delhi, India: PHI Learning, 2009.

[7] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition," *Speech Commun.*, vol. 54, no. 4, pp. 543–565, May 2012.

[8] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining mfcc and phase information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1085–1095, May 2012.

[9] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 190–202, Jan. 2007.

[10] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, And Applications*, 4th ed. Upper Saddle River, NJ: Pearson, 2007.

[11] D. Reynolds, T. Quatieri, and D. R. B. , "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[12] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[13] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2006)*, May 2006, vol. 1, p. I.

[14] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-variance multitaper mfcc features: A case study in robust speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 1990–2001, Sept. 2012.