Efficient Window for Monolingual and Crosslingual Speaker Identification using MFCC

B. G. Nagaraja
Department of Electronics and Communication Engineering
GM Institute of Technology
Davangere, India
nagarajbg@gmail.com

H. S. Jayanna
Department of Information Science and Engineering
Siddaganga Institute of Technology
Tumkur, India
jayannahs@gmail.com

***Abstract*— In this paper an experimental evaluation of the various windowing techniques using mel-frequency cepstral coefficient (MFCC) for monolingual and crosslingual speaker identification is demonstrated. The set of windows presented here allows a tradeoff between main lobe bandwidth and side lobe ripple decay. The speaker identification study is conducted using randomly selected 50 speakers from IITG Multi-variability speaker recognition (IITG-MV) database, MFCC feature and Gaussian mixture model (GMM)-universal background model (UBM) classifier. Speaker identification system based on various windowing techniques shown to have considerably improved performance over baseline Hamming window technique.**

Keywords*— Speaker identification; window; monolingual; crosslingual; MFCC

I. INTRODUCTION

Speaker identification aims at recognizing the speakers from their voice [1]. Speaker identification is a one-to-many comparison, ie., system identifies a speaker from a database of *N* known speakers. Depending on the mode of operation, speaker identification can be either text-dependent or text-independent [2]. In the former case, the speaker must speak a given phrase known to the system, which can be fixed or prompted. In the latter case, the system does not know the phrase spoken by the speaker. Speaker identification can be performed in monolingual, crosslingual and multilingual mode [3]. In monolingual speaker identification, training and testing languages for a speaker are the same whereas in crosslingual speaker identification, training is done in one language (say *A*) and testing is done in different language (say *B*). In multilingual speaker identification, speaker specific models are trained in one language and tested with multiple languages.

Most of the state-of-the-art speaker identification systems work within the single language environment (monolingual). People have an ability to learn more than one language [4]. For instance, in India more than 50 languages are officially recognized. A person in a multilingual country usually speaks more than one language. Therefore, identifying a speaker in this context is an issue. This paper addresses monolingual and crosslingual speaker recognition using 15 seconds of train and test speech data [5].

[978-1-4799-3506-2/13/$31.00 ©2013 IEEE]

The MFCC is the most widely used features for speaker recognition [6]. MFCC is a real cepstrum derived from the windowed discrete Fourier transform (DFT). Windowing is applied to raw speech frames in order to reduce the spectral leakage effect [7]. Recently, many researchers have focused on the design of new windows for various applications in signal processing area [7] [8] [9] [10]. Tomi kinnunen et al. demonstrated the use of multi-taper MFCC features for speaker verification task in [8]. The basic idea in multi-tapering is to pass the analysis frame through the multiple window functions and then estimate the weighted mean of individual sub-spectra to obtain the final spectrum. The experimental results on NIST-2002 and NIST-2008 databases shows that multi-tapers outperform conventional single-window hamming technique.

A novel family of windowing method to obtain MFCC features for speaker recognition was proposed in [7]. The proposed window technique was based on basic property of discrete time Fourier transform (DTFT) related to differentiation in the frequency domain. Speaker recognition experiments on different NIST databases (SRE-2001, SRE-2004 and SRE-2006) showed that the proposed window based MFCC technique achieved consistently improved performance over baseline Hamming window based MFCC method. In [11], the extended Kalman filter (EKF) algorithm was used for obtaining the new window having equal main lobe width and smaller side lobe peak compared to the Hamming window. It was observed that the new window function has at least 5.6 to 12 dB less side lobe peak compared to the Hamming window, while offering smaller or the same main lobe width.

In this paper, we study the effects of different windowing methods on monolingual and crosslingual speaker identification performance using MFCC features. The different types of windows presented in this work allow a trade-off between main lobe bandwidth and side lobe ripple decay in comparison with the baseline Hamming windowing method. The remainder of the paper is organized as follows: Section II describes the windowing methods used for the study. Database for the study, Feature extraction using MFCC and speaker modeling using GMM-UBM technique are presented in Section III. Section IV gives experimental results. Finally, Summary and conclusions of this study and scope for the future work are mentioned in Section V.

II. WINDOWING METHODS

Let $F = (f[0],\ f[1], \dots ,\ f[N-1])^T$ denote one frame of speech of N samples. Windowed DFT spectrum estimate can be expressed as:

$$\hat{S}(f) = \left|\sum_{n=0}^{N-1} w[n]\, f[n]\, e^{-i2\pi fn/N}\right|^2 \quad (1)$$

where *w[n]* is the time-domain window function. In this work four different types of windows are used.

- Window 1: The Hamming window is widely used in speech applications and it has the shape of

$$w_1[n] = 0.54 - 0.46\, cos\left(\frac{2\pi n}{M}\right) \quad (2)$$

where M is the window order.

- Window 2: A new computationally efficient window for signal spectrum analysis was proposed in [9]. The new window was obtained by adding the third harmonic of the cosine function to the Hamming window, and finding the suitable amplitudes of DC term, the cosine function, and its third harmonic to reduce the peak side-lobe amplitude.

$$w_2[n] = a_0 - a_1 \cos\left(\frac{2\pi n}{M}\right) - a_3 \cos\left(\frac{6\pi n}{M}\right) \quad (3)$$

where a_0 = 0.5363 - 0.14/M, a_1 = 0.996 - a_0 and a_3 = 0.04. The time domain of window 1 and window 2 functions are shown in Fig. 1.

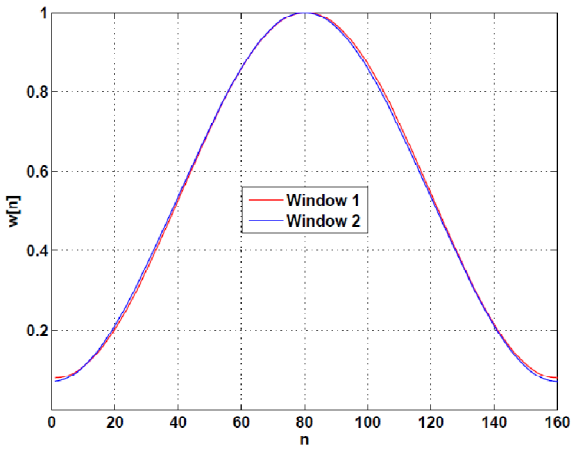

Fig. 1. window 1 and window 2 in the time domain for M = 160.

- Window 3: The conventional MFCC realization based on windowed (hamming) DFT may not yield good performance due to the high variance of the spectrum estimation [8]. To reduce the variance, multi-taper spectrum estimator can be used is as follows [8]:

$$\hat{S}(f) = \sum_{j=1}^{K} \lambda(j) \left|\sum_{n=0}^{N-1} w_j[n]\, f[n]\, e^{-i2\pi fn/N}\right|^2 \quad (4)$$

Here K represents the number of multi-tapers used. $W_j = \left[w_j(0), w_j(1), \dots, w_j(N-1)\right]^T$ is the multi-taper weights and j = *1, 2, ..., K*, are used with corresponding weights $\lambda(j)$. In this work sine-weighted cepstrum estimator (SWCE) is used with *K=6*. The sine tapers are defined as [12]

$$w_3[n] = \sqrt{\frac{2}{N+1}}\, sin\left(\frac{\pi j(n+1)}{N+1}\right);\ n = 0,1,..N-1 \quad (5)$$

Fig. 2 shows the window 1 and window 3 (sine tapers with K=6) representation in frequency domain.

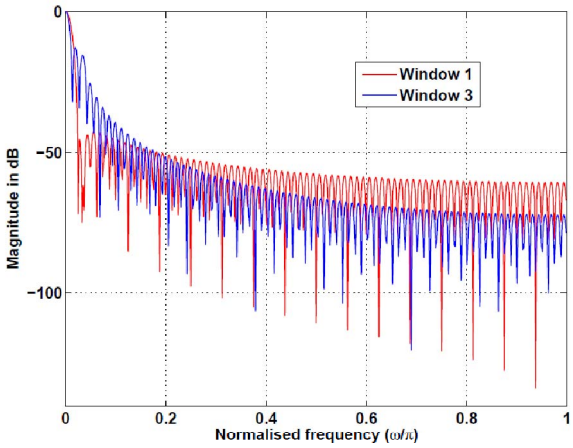

Fig. 2. Frequency domain for a window of size 160 samples.

- Window 4: Based on the DTFT differentiation in frequency, a new window function was proposed in [7]. The proposed window function of the τ^{th} order window can be described as

$$w_4[n] = n^{\tau}\, w_1[n] \quad (6)$$

This work concentrates for τ = 2. Fig. 3 shows the window 1 and window 4 (τ = 1 and τ = 2) representation in frequency domain. It is observed that window 4 exhibits a considerable increase in mainlobe width.

III. EXPERIMENTAL SETUP

A. Database for the Study

Speaker identification experiments are carried out on the subset of the IITG-MV database which is collected in a set up having five different sensors, two different environments, two different languages and two different styles [13]. The recording was done in the office (controlled environment) and hostel rooms, laboratory and corridors etc. (uncontrolled environments). The speech signal was sampled at 16 kHz and stored with 16 bits resolution. The recording was done in Indian English and favorite language of the speaker which may be one of the Indian languages like Hindi, Kannada, Tamil, Oriya, Assami, Malayalam and so on [14]. For the present work, we consider randomly selected 50 speakers set of IITG-MV database (headphone speech data), which include 30-male and 20-female speakers.

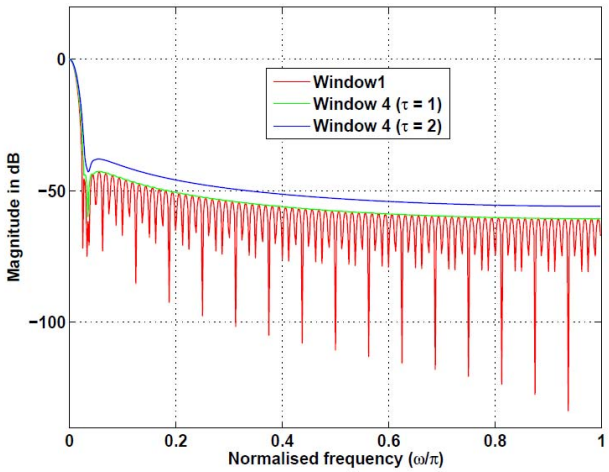

Fig. 3. Frequency domain for a window of size 160 samples.

B. Features

Speech recordings were re-sampled to 8 kHz with 16 bits resolution and pre-emphasized (0.97). Frame duration of 20 msec and a 10 msec for overlapping are considered. A mel-warping is performed using 22 triangular band pass filters followed by discrete cosine transform (DCT). A 13-dimensional MFCC feature vectors are finally obtained (excluding 0^{th} coefficient).

C. Classifier

In this work GMM-UBM technique is used as a classifier. For building the UBM, we have used 1 hour of speech data from 138 speakers of YOHO database. The speaker specific models were created by adapting only the mean vectors of the UBM using maximum a posteriori (MAP) adaptation algorithm. The parameters of the GMM models (mean vector covariance matrix, mixture weights) were estimated using expectation maximization (EM) algorithm. We have modeled speakers by using GMMs with 8, 16, 32, 64 and 128 mixtures.

IV. EXPERIMENTAL RESULTS

In this section monolingual and crosslingual speaker identification results are presented. In all our experiments, the speaker set (50 speakers) and amount of speech data (15 seconds) are kept constant to make a relative comparison of the performance of speaker identification using different windowing methods. Note: (1) *A/B* indicates training with language *A* and testing with language *B*.

A. Monolingual Speaker Identification

The performance of different windows for monolingual speaker identification are given in Table I and II. The speaker identification system trained and tested with the English language (E/E) gives the highest performance of 68% for window 4. The performance of the speaker identification trained and tested with Hindi language (H/H) is 72% for window 4.

TABLE I. PERFORMANCE OF MONOLINGUAL SPEAKER IDENTIFICATION SYSTEM (E/E).

Method	Gaussian Mixtures				
	8	**16**	**32**	**64**	**128**
Window 1	32	42	52	50	64
Window 2	38	38	44	48	64
Window 3	36	42	46	50	58
Window 4	36	48	46	56	68

TABLE II. PERFORMANCE OF MONOLINGUAL SPEAKER IDENTIFICATION SYSTEM (H/H).

Method	Gaussian Mixtures				
	8	**16**	**32**	**64**	**128**
Window 1	46	54	62	68	68
Window 2	46	56	56	62	60
Window 3	38	52	62	64	70
Window 4	44	52	54	68	72

B. Crosslingual Speaker Identification

The performance of different windows for crosslingual speaker identification are given in Table III and IV. The window 3 system gives the highest recognition performance of 52% for training in English and testing in Hindi language (E/H). Similarly, the window 3 system gives the highest recognition performance of 44% for a system trained in Hindi and tested in English language (H/E).

TABLE III. PERFORMANCE OF CROSSLINGUAL SPEAKER IDENTIFICATION SYSTEM (E/H).

Method	Gaussian Mixtures				
	8	**16**	**32**	**64**	**128**
Window 1	30	34	40	46	44
Window 2	30	34	36	42	46
Window 3	30	30	42	44	52
Window 4	26	30	34	44	46

TABLE IV. PERFORMANCE OF CROSSLINGUAL SPEAKER IDENTIFICATION SYSTEM (H/E).

Method	Gaussian Mixtures				
	8	**16**	**32**	**64**	**128**
Window 1	24	24	30	34	38
Window 2	22	24	26	30	36
Window 3	20	24	42	44	44
Window 4	26	30	32	34	42

Some of the observations we made from the monolingual and crosslingual results are as follows:

1. The window 4 yields good recognition in the monolingual speaker identification experiments. The improvement in performance may be due to the considerable increase in main lobe width that in turn helps in smooth power spectrum estimation [7].
2. The window 3 yields better identification in the crosslingual speaker identification experiments. This may be due to the use of multiple windows (multi-tapers) that reduce the variance of the MFCC features and thus making the spectrum less sensitive to the noise [8].

3. It was observed that the results are better for monolingual experiments than the crosslingual. This may be due to the variation in fluency and word stress when the same speaker speaks different languages and also due to different phonetic and prosodic patterns of the languages [15].

V. CONCLUSION

In this paper we have compared the performance of four different windowing methods using MFCC for monolingual and crosslingual speaker identification. The results indicate that window 3 and window 4 based system can be used for improving the speaker identification performance. In order to study the robustness of these windows, needs to be verified with different languages, different data sizes and large amount of speaker set.

REFERENCES

[1] B.S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, Vol. 64(4), pp. 460–475, Apr. 1976.

[2] D.A Reynolds, and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Trans. Speech and Audio Processing*, 3,pp. 72-83, 1995.

[3] P.H. Arjun, "Speaker Recognition in Indian Languages: A Feature Based Approach", *Ph.D.dissertation*, Indian Institute of Technology Kharagpur, India, 2005.

[4] U. Halsband, "Bilingual and multilingual language processing," *J. Physiology-Paris 99*, pp. 355-369, 2006.

[5] H.S. Jayanna, "Limited data Speaker Recognition", *Ph.D.dissertation*, Indian Institute of Technology, Guwahati, India, 2009.

[6] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Trans. Audio, Speech and Language Process.*, vol. 20, no. 4, pp. 1085-1095, 2012.

[7] Md Sahidullah, and Goutam Saha, "A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition," *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 149-152, Feb. 2013.

[8] T. Kinnunen, R. Saeidi, F. Sedlak, K.A. Lee, J. Sandberg, M.H. Sandsten, and H. Li, "Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification*," IEEE Trans. Audio, Speech and Language Process.* 20, pp. 1990-2001, 2012.

[9] M. Mottaghi-Kashtiban, and M. Shayesteh, "New efficient window function, replacement for the hamming window," *IET Signal Processing*, vol. 5, no. 5, pp. 499505, Aug. 2011.

[10] Y. Wang, "An effective approach to finding differentiator window functions based on sinc sum function," *Circuits, Syst. Signal Process*., vol. 31, no. 5, pp. 18091828, Oct. 2012.

[11] M.G. Shayesteh, M. Mottaghi-Kashtiban, "FIR filter design using a new window function," *Proc. IEEE*, 16th International Conference on Digital Signal Processing, pp. 1-6, Jul. 2009.

[12] Md. Jahangir Alam, T. Kinnunen, P. Kenny, P. Ouellet, D.D. O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors," *Speech Communication*, vol. 55, no. 2, pp. 237-251, 2013.

[13] B.C. Haris, G. Pradhan, A. Misra, S. Shukla, R. Sinha, S.R.M. Prasanna, "Multivariability speech database for robust speaker recognition," *Proc. IEEE, Communications (NCC),* 2011 National Conference, pp. 1-5, Jan. 2011.

[14] G. Pradhan, S.R.M. Prasanna, "Significance of vowel onset point information for speaker verification," *Int. Jr. of computer & comm. tech*, vol. 2, no. 6, pp. 56-61, Feb. 2011.

[15] G. Durou, "Multilingual text-independent speaker identification," *Proc. Multi-lingual Interoperability in Speech Technology (MIST)*, Leusden, Netherlands, pp. 115-118, 1999.