Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 237–251

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Multitaper MFCC and PLP features for speaker verification using i-vectors

Md Jahangir Alam [a,b,*], Tomi Kinnunen [c], Patrick Kenny [b], Pierre Ouellet [b], Douglas O'Shaughnessy [a]

[a] *INRS-EMT, Montreal, Canada*
[b] *CRIM, Montreal, Canada*
[c] *School of Computing, University of Eastern Finland (UEF), Joensuu, Finland*

Received 29 March 2012; received in revised form 23 July 2012; accepted 10 August 2012
Available online 31 August 2012

Abstract

In this paper we study the performance of the low-variance multi-taper Mel-frequency cepstral coefficient (MFCC) and perceptual linear prediction (PLP) features in a state-of-the-art i-vector speaker verification system. The MFCC and PLP features are usually computed from a Hamming-windowed periodogram spectrum estimate. Such a *single-tapered* spectrum estimate has large variance, which can be reduced by averaging spectral estimates obtained using a set of different tapers, leading to a so-called *multi-taper* spectral estimate. The multi-taper spectrum estimation method has proven to be powerful especially when the spectrum of interest has a large dynamic range or varies rapidly. Multi-taper MFCC features were also recently studied in speaker verification with promising preliminary results. In this study our primary goal is to validate those findings using an up-to-date i-vector classifier on the latest NIST 2010 SRE data. In addition, we also propose to compute robust perceptual linear prediction (PLP) features using multitapers. Furthermore, we provide a detailed comparison between different taper weight selections in the Thomson multi-taper method in the context of speaker verification. Speaker verification results on the telephone (det5) and microphone speech (det1, det2, det3 and det4) of the latest NIST 2010 SRE corpus indicate that the multi-taper methods outperform the conventional periodogram technique. Instead of simply averaging (using uniform weights) the individual spectral estimates in forming the multi-taper estimate, weighted averaging (using non-uniform weights) improves performance. Compared to the MFCC and PLP baseline systems, the *sine-weighted cepstrum estimator* (SWCE) based multi-taper method provides average relative reductions of 12.3% and 7.5% in equal error rate, respectively. For the *multi-peak* multi-taper method, the corresponding reductions are 12.6% and 11.6%, respectively. Finally, the *Thomson* multi-taper method provides error reductions of 9.5% and 5.0% in EER for MFCC and PLP features, respectively. We conclude that both the MFCC and PLP features computed via multitapers provide systematic improvements in recognition accuracy.

Keywords: Speaker verification; Multi-taper spectrum; Feature extraction; i-Vectors; MFCC; PLP

1. Introduction

Useful information extraction from speech has been a subject of active research for many decades. Feature extraction (or *front-end*) is the first step in an automatic speaker or speech recognition system. It transforms the raw acoustic signal into a compact representation. Since feature extraction is the first step in the chain, the quality of the subsequent steps (modeling and classification) strongly depends on it. The *mel-frequency cepstral coefficient* (MFCC) (Davis and Mermelstein, 1980) and *perceptual linear prediction* (PLP) (Hermansky, 1990) front-ends have been dominantly used in speech and speaker recognition systems and they demonstrate good performance in

* Corresponding author. Address: INRS-EMT, University of Quebec, 800, de La Gauchetière West, Suite 6900, Montréal (Québec), Canada H5A 1K6. Tel.: +1 514 228 7012x3030; fax: +1 514 875 0344.

E-mail addresses: alam@emt.inrs.ca (M.J. Alam), tkinnu@cs.joensuu.fi (T. Kinnunen), Patrick.Kenny@crim.ca (P. Kenny), Pierre.Ouellet@crim.ca (P. Ouellet), dougo@emt.inrs.ca (D. O'Shaughnessy).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.08.007

both applications. The MFCC and PLP parameterization techniques aim at computing the speech parameters similar to the way how a human hears and perceives sounds (Davis and Mermelstein, 1980). Since these features are computed from an estimated spectrum, it is crucial that this estimate is accurate. Usually, the spectrum is estimated using a *windowed periodogram* (Harris, 1978) via the discrete Fourier transformation (DFT) algorithm. Despite having low bias, a consequence of the data tapering (windowing) is increased estimator variance. Therefore, MFCC or PLP features computed from this estimated spectrum have also high variance. One elegant technique for reducing the spectral variance is to replace a windowed periodogram estimate with a *multi-taper* spectrum estimate (Sandberg et al., 2010; Thomson, 1982; Riedel and Sidorenko, 1995).

In the multi-taper spectral estimation method, a set of orthogonal tapers is applied to the short-time speech signal and the resulting spectral estimates are averaged (possible with nonuniform weights), which reduces the spectral variance. As each taper in a multi-taper technique is pairwise orthogonal to all the other tapers, the windowed signals provide statistically independent estimates of the underlying spectrum. The multi-taper method has been widely used in geophysical applications and, in multiple cases, it has been shown to outperform the windowed periodogram. It has also been used in speech enhancement applications (Hu and Loizou, 2004) and, recently, in speaker recognition (Kinnunen et al., 2010, in press; Sandberg et al., 2010; Alam et al., 2011) with promising preliminary results. The preliminary experiments of Kinnunen et al. (2010) and Sandberg et al. (2010) were reported on the NIST 2002 and 2006 SRE corpora using a lightweight Gaussian mixture model–universal background model (GMM–UBM) system (Reynolds et al., 2000) and generalized linear discriminant sequence support vector machine (GLDS-SVM) without any session variability compensation techniques. The recent results of Kinnunen et al. (in press), using multi-taper MFCC features only, were reported on NIST 2002 and 2008 SRE corpora using GMM–UBM, GMM-SVM and *joint factor analysis* (JFA) (Kenny et al., 2007a, 2007b) classifiers.

In this paper, our aims are, firstly, to study whether the improvements obtained using multi-taper MFCC features in (Kinnunen et al., 2010, in press; Sandberg et al., 2010) translate to a state-of–the-art speaker verification task. Secondly, we propose to use multi-taper PLP features in an *i-vector* speaker verification system as we have found that the performance of PLP features (HTK version of PLP, also denoted as revised PLP (RPLP) in (Honig, 2005)) can outperform MFCC accuracy in speaker verification, and thirdly, we provide a comparison of the performance of using uniform average versus weighted average to get the final multi-taper spectral estimate in a Thomson multi-taper method, in the context of speaker verification. Proper selection of weights is an important design issue in multi-taper spectrum estimation. Even though (Kinnunen et al., 2010, in press; Sandberg et al., 2010; Alam et al., 2011) extensively compare different types of taper windows, their weight selection was *not* addressed. Therefore, in this work, we provide detailed comparison between different taper weight selections in the popular Thomson multi-taper method. The recent i-vector model (Dehak et al., 2011; Kenny, 2010; Senoussaoui et al., 2011) includes elegant inter-session variability compensation, with demonstrated significant improvements on the recent NIST speaker recognition evaluation corpora. Since i-vectors already do a good job in compensating for variabilities in the speaker model space, one may argue that improvements in the front-end may not translate to the full recognition system. This is the question which we address in this paper. In the experiments, we use the latest NIST 2010 SRE benchmark data with the state-of-the-art i-vector configuration. To this end, we utilize a completely gender independent i-vector system based on mixture *probabilistic linear discriminant analysis* (PLDA) model of Senoussaoui et al. (2011). In this paper, similar to Senoussaoui et al. (2011)), we also use a gender independent i-vector extractor and then form a mixture PLDA model by training and combining two gender dependent models, where the gender label is treated as a latent (or hidden) variable.

2. Multi-taper spectrum estimation

A windowed direct spectrum estimator is the most often used power spectrum estimation method in speech processing applications. For the *m*th frame and *k*th frequency bin an estimate of the windowed periodogram can be expressed as:

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

where $k \in \{0, 1, \ldots, K-1\}$ denotes the frequency bin index, N is the frame length, $s(m,j)$ is the time domain speech signal and $w(j)$ denotes the time domain window function, also known as *taper*. The taper, such as the Hamming window, is usually symmetric and decreases towards the frame boundaries. Eq. (1) is sometimes called *single-taper*, *modified* or *windowed periodogram*. If $w(j)$ is a rectangular or uniform taper, Eq. (1) is called a *periodogram*. Fig. 1 presents time- and frequency-domain plot of the Hamming window.

Windowing reduces the bias, i.e., expected value of the difference between the estimated spectrum and the actual spectrum, but it does not reduce the variance of the spectral estimate (Kay, 1988) and therefore, the variance of the MFCC features computed from this estimated spectrum remains large. One way to reduce the variance of the MFCC or PLP estimator is to replace the windowed periodogram estimate by a so-called *multi-taper* spectrum estimate (Sandberg et al., 2010; Thomson, 1982; Riedel and Sidorenko, 1995). It is given by

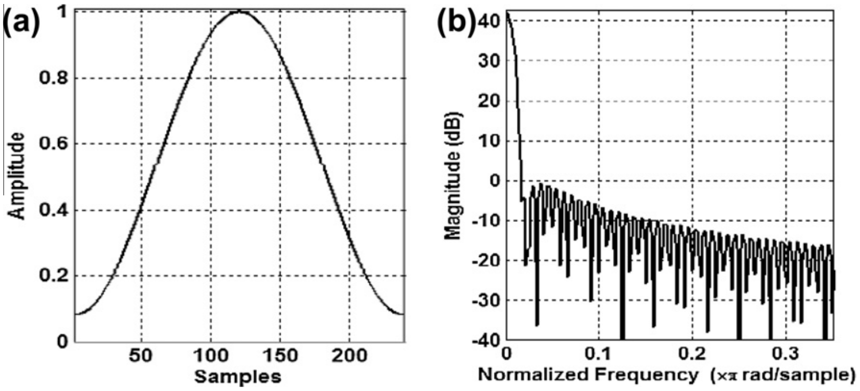

Fig. 1. Hamming window for $N = 256$, in (a) time domain and (b) frequency domain.

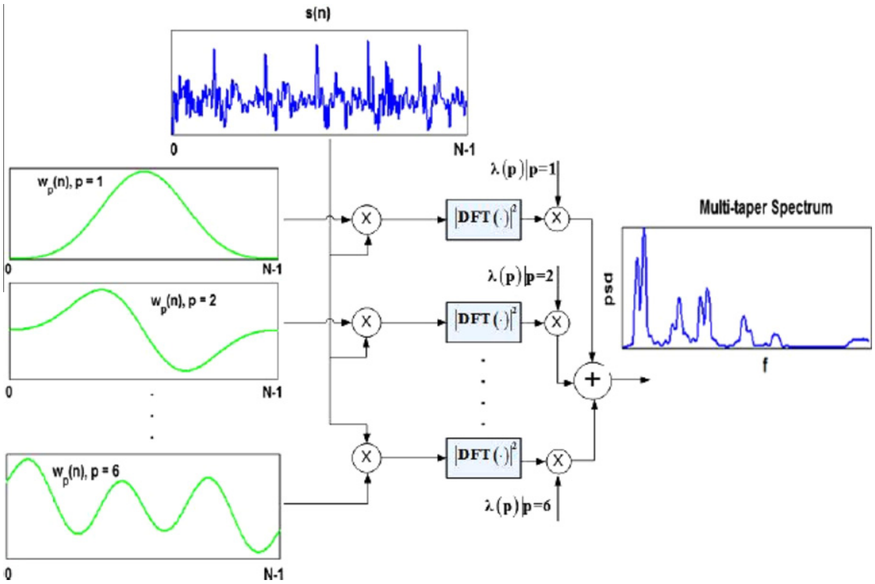

Fig. 2. Block diagram of multi-taper spectrum estimation method.

$$\hat{S}_{MT}(m,k) = \sum_{p=1}^{M} \lambda(p) \left| \sum_{j=0}^{N-1} w_p(j) s(m,j) e^{-\frac{2\pi i k}{N}} \right|^2, \tag{2}$$

where N is the frame length and w_p is the pth data taper ($p = 1, 2, \ldots, M$) used for the spectral estimate $\hat{S}_{MT}(\cdot)$, also known as the pth *eigenspectrum*. Finally, M denotes the number of tapers and $\lambda(p)$ is the weight of the pth taper. The tapers $w_p(j)$ are typically chosen to be orthonormal so that, for all p and q,

$$\sum_j w_p(j) w_q(j) = \delta_{pq} = \begin{cases} 1, & p = q \\ 0, & \text{otherwise.} \end{cases}$$

The multi-taper spectrum estimate is therefore obtained as the weighted average of M individual sub-spectra. Eq. (1) can be obtained as a special case of Eq. (2) when $p = M = 1$ and $\lambda(p) = 1$. Fig. 2 illustrates the multi-taper spectrum estimation process using $M = 6$ tapers.

The idea behind multi-tapering is to reduce the variance of the spectral estimates by averaging M direct spectral estimates, each with a different data taper. If all M tapers are pairwise orthogonal and properly designed to prevent leakage, the resulting multi-taper estimates outperform the windowed periodogram in terms of reduced variance, specifically, when the spectrum of interest has high dynamic range or rapid variations (McCoy et al., 1998). Therefore, the variance of the MFCC and PLP features computed via this multi-taper spectral estimate will be low as well. The underlying detail of the multi-taper method is similar to Welch's modified periodogram (Kay, 1988), it, however, focuses only on one frame rather than forming a time-averaged spectrum estimate over multiple frames. 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. The so-called *Slepian tapers* that underlie the Thomson multi-taper method (Thomson, 1982) are

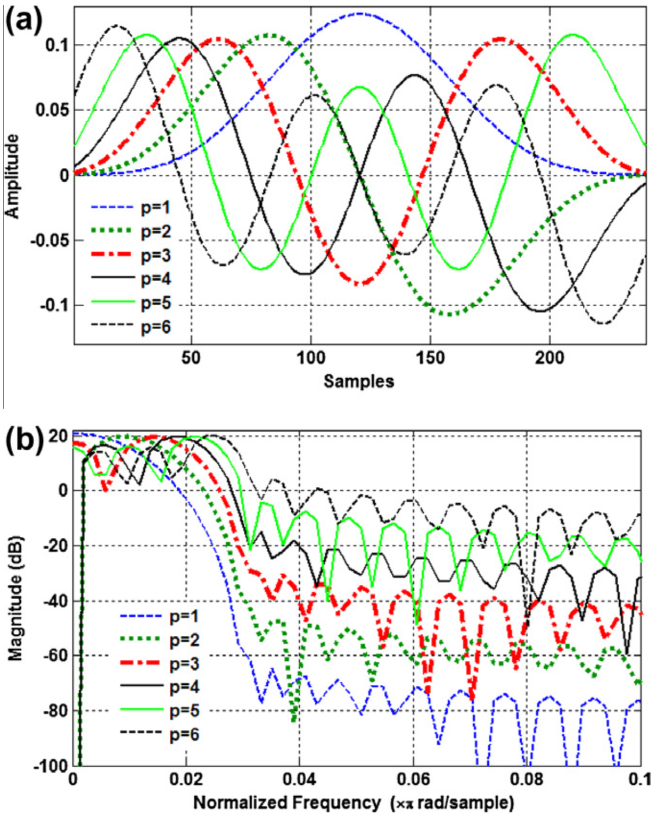

Fig. 3. Thomson multi-tapers for $N = 256$, $M = 6$ in (a) time and (b) frequency domains.

illustrated in Fig. 3 for $M = 6$ both in time and frequency domains.

2.1. Choice of the tapers and the taper weights

The choice of taper has a significant effect on the resultant spectrum estimate. The objective of the taper is to prevent energy at distant frequencies from biasing the estimate at the frequency of interest. Based on the Slepian tapers (also called *discrete prolate spheroidal sequence*, DPSS) (Slepian and Pollak, 1960) and the *sine* tapers (Riedel and Sidorenko, 1995), various multi-taper methods have been proposed in the literature for spectrum estimation, such as Thomson multi-taper (Thomson, 1982), SWCE (sinusoidal weighted cepstrum estimator) multi-taper (Hansson-Sandsten and Sandberg, 2009) and Multi-peak multi-taper (Hansson and Salomonsson, 1997). For completeness, we briefly review each method in the following.

2.1.1. Thomson multi-taper method

In the Thomson multi-taper method of spectrum estimation (Thomson, 1982), a set of M orthonormal data tapers with good leakage properties is specified from the *Slepian sequences* (Slepian and Pollak, 1960). Slepian sequences are defined as the real, unit-energy sequences on $[0, N-1]$ having the greatest energy in a bandwidth W. Slepian tapers can be shown to be the solutions to the following eigenvalue problem,

$$\boldsymbol{A} w_j^p = v^p w_n^p, \tag{3}$$

where $0 \leqslant n \leqslant N-1$, $0 \leqslant j \leqslant N-1$, $\boldsymbol{A}$ is a real symmetric matrix, $0 < v^p \leqslant 1$ is the pth eigenvalue corresponding to the pth eigenvector w_n^p known as the Slepian taper. The elements of the matrix $\boldsymbol{A}$ are given by $a_{nj} = \frac{\sin 2\pi W(n-j)}{\pi(n-j)}$, where W is the half-frequency bandwidth (or one sided bandwidth).

Slepian sequences (or DPSS), proposed originally in (Slepian and Pollak, 1960), were chosen as tapers in (Thomson, 1982) 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 give increasingly more weight to the ends of the signal. There is no loss of information at the extremes of the signal. In the experiments of Kinnunen et al. (2010, in press) and Sandberg et al. (2010), uniform weights were applied to obtain the final Thomson multi-taper estimate. That is, $\lambda(p) = 1/M$. Even though (Kinnunen et al., 2010, in press; Sandberg et al., 2010) reported increased speaker verification accuracy when the standard windowed periodogram was replaced by the Thomson multi-taper, the question of weight selection in the Thomson method was not addressed. We hypothesize that recognition accuracy might be further increased by allowing non-uniform weighting in the Thomson method. In order to compensate for the increased energy loss at higher order tapers the uniform weights can be replaced with the weights corresponding to either the eigenvalues of the Slepian tapers, i.e., $\lambda(p) = v^p$ or, alternatively, adaptive weights obtained as $\lambda(p) = 1/\sum_{q=1}^{p} v^q$ (Thomson, 1982, 1990). The different weighting schemes used in the Thomson multi-taper method are illustrated in Fig. 5 for $M = 6$ tapers including the weights used in the multi-peak (Hansson and Salomonsson, 1997) and the SWCE (Hansson-Sandsten and Sandberg, 2009) methods.

2.1.2. SWCE multi-taper

The Thomson multi-taper method requires solving an eigenvalue problem of Eq. (3) and does not have a closed-form expression for the tapers. A simpler set of orthonormal tapers that has such a closed-form expression is the set of the *sine* tapers (see Fig. 4(c)) given by Riedel and Sidorenko (1995):

$$w_p(j) = \sqrt{\frac{2}{N+1}} \sin\left(\frac{\pi p(j+1)}{N+1}\right), \quad j = 0, 1, \ldots, N-1. \tag{4}$$

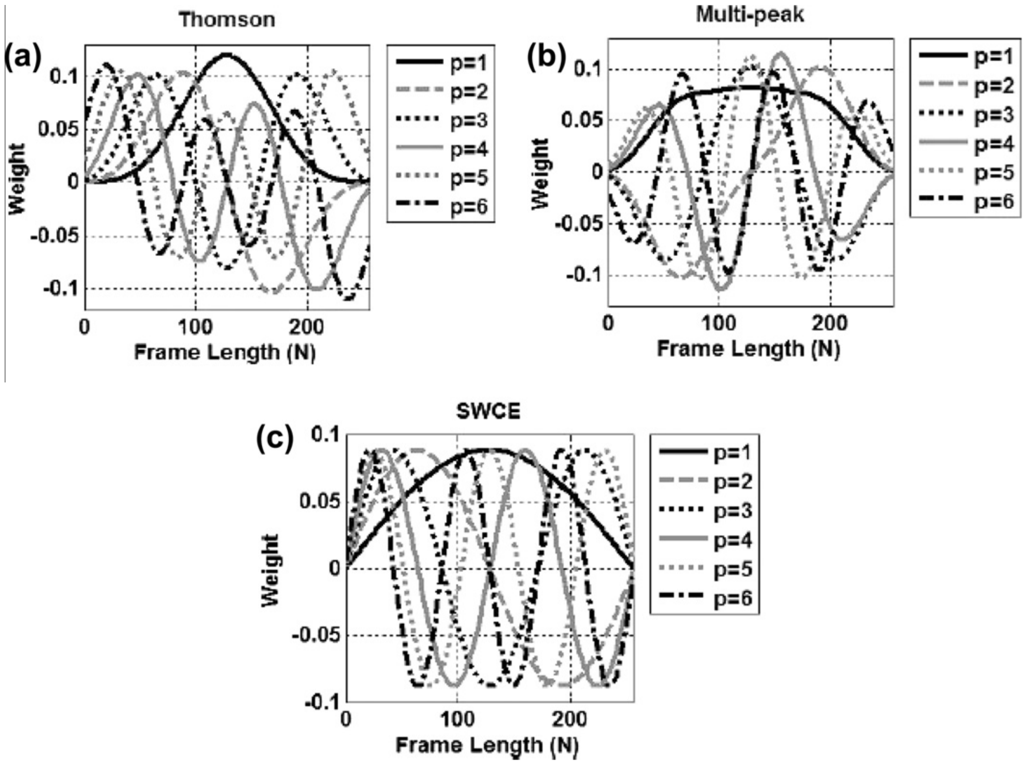

Fig. 4. (a) Six Slepian tapers in the Thomson method, (b) multi-peak tapers in the multi-peak method, and (c) *sine* tapers for SWCE method, for $N = 256$.

The sine tapers achieve a smaller local bias (the bias due to the smoothing by the mainlobe) than the Slepian tapers at the expense of sidelobe suppression (Riedel and Sidorenko, 1995; McCoy et al., 1998). The first taper in the set of *sine* tapers produces a direct spectral estimator with minimum local bias and the higher order tapers ensure minimum local bias whilst being orthogonal to all of the lower order tapers.

In the SWCE method (Hansson-Sandsten and Sandberg, 2009), the *sine* tapers are applied with optimal weighting for cepstrum analysis. The weights used in the SWCE method (see Fig. 5) have the following closed-form expression (Hansson-Sandsten and Sandberg, 2009):

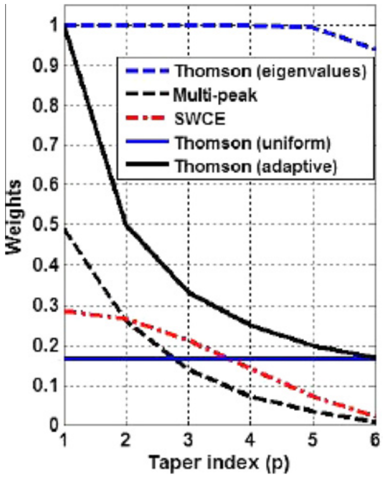

Fig. 5. Weights used in multi-taper spectrum estimation methods for six tapers.

$$\lambda(p) = \frac{\cos\left(\frac{2\pi(p-1)}{M/2}\right) + 1}{\sum_{p=1}^{M}\left(\cos\left(\frac{2\pi(p-1)}{M/2}\right) + 1\right)}, \quad p = 1, 2, \ldots, M. \tag{5}$$

2.1.3. Multi-peak multi-taper

In (Hansson and Salomonsson, 1997), a multi-taper method, dubbed as *peak matched multiple windows* (PMMW), was proposed for peaked spectra to obtain low bias at the frequency peak as well as low variance of the spectral estimate. Here, similar to Kinnunen et al. (2010)), we denote this method as the *multi-peak* method and the tapers (or windows) as the multi-peak tapers. The multi-peak tapers are obtained as the solution of the following generalized eigenvalue problem:

$$\boldsymbol{R}_{B'} w_j = v_j \boldsymbol{R}_Z w_j, \quad j = 1, 2, \ldots, N, \tag{6}$$

where $\boldsymbol{R}_{B'}$ is the $(N \times N)$ Toeplitz covariance matrix of the assumed spectrum model defined by Hansson and Salomonsson (1997):

$$S_s(f) = \begin{cases} e^{-\frac{2C|f|}{10\log_{10}(e)}} & |f| \leqslant B'/2 \\ 0 & |f| > B'/2, \end{cases}$$

with $C = 20$ dB and a predetermined interval of width B' outside of which spectral leakage is to be prevented, $\boldsymbol{R}_Z$ is the Toeplitz covariance matrix, chosen for decreasing the leakage from the sidelobes of the tapers, of the following frequency penalty function:

$$S_Z(f) = \begin{cases} G & |f| > B'/2 \\ 1 & |f| \leqslant B'/2 \end{cases},$$

where $G = 30$ dB (Hansson and Salomonsson, 1997). The eigenvectors corresponding to the M largest eigenvalues of (6) are used as multi-peak tapers for the multi-peak method and the weights for the tapers can be found from the M largest eigenvalues of (6) as:

$$\lambda_p = \frac{v_p}{\sum_{p=1}^{M} v_p}, \quad p = 1, 2, \ldots, M.$$

Six multi-peak tapers and the weights corresponding to these tapers are shown in Figs. 4(b) and 5, respectively.

2.2. Variance reduction by multitapering

The use of multiple orthogonal windows can have several advantages over the use of any single window (Percival and Walden, 1993; Walden et al., 1994; Wieczorek and Simons, 2005, 2007; McCoy et al., 1998). In particular, the energy of a single band-limited window always nonuniformly covers the desired concentration region, which results in some data being statistically over- or underrepresented when forming the spectral estimate (Wieczorek and Simons, 2005, 2007). In contrast, the cumulative energy of the multiple orthogonal windows more uniformly covers the concentration region. Since the spectral estimates that result from using orthogonal tapers are uncorrelated, a multi-taper average (or weighted average) of these possesses a smaller estimation variance than the single-tapered spectrum estimates.

The variance of an estimator $\hat{\theta}$ measures how much variability an estimator has around its mean (i.e., expected) value and is defined as (Kay, 1988; Djuric and Kay, 1999):

$$\mathrm{var}(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2],$$

where $E[\cdot]$ is the expectation operator. A 'good' estimator is one that makes some suitable trade-off between low bias and low variance.

A multi-taper spectrum estimator is somewhat similar to averaging the spectra from a variety of conventional tapers such as Hamming and Hann tapers. But in this case, there will be strong redundancy as the different tapers are highly correlated (all the tapers have a common time-domain shape). Unlike conventional tapers, the M orthonormal tapers used in a multi-taper spectrum estimator provide M statistically independent (hence uncorrelated) estimates of the underlying spectrum. The weighted average of the M individual spectral estimates $\hat{S}_{MT}(m,k)$ then has smaller variance than the single-tapered spectrum estimates $\hat{S}_d(m,k)$ by a factor that approaches $1/M$, i.e., $\mathrm{var}(\hat{S}_{MT}(m,k)) \approx \frac{1}{M}\mathrm{var}(\hat{S}_d(m,k))$ (McCoy et al., 1998).

The reduction in the variance of the spectrum ordinates between using single taper (e.g., Hamming window) and multi-taper methods is illustrated in Fig. 6. Spectral variance reduction using multi-taper methods has been addressed by many researchers, including in Kay (Kay, 1988; Sandberg et al., 2010; Thomson, 1982; Riedel and Sidorenko, 1995; Hansson-Sandsten and Sandberg, 2009; Hansson and Salomonsson, 1997; Thomson, 1990; Percival and Walden, 1993; Walden et al., 1994; Wieczorek and Simons, 2005, 2007; McCoy et al., 1998). The objective of our paper is to apply multi-taper methods to compute MFCC and PLP features for speaker verification using i-vectors and compare their performance with the Hamming window-based baseline MFCC and PLP systems.

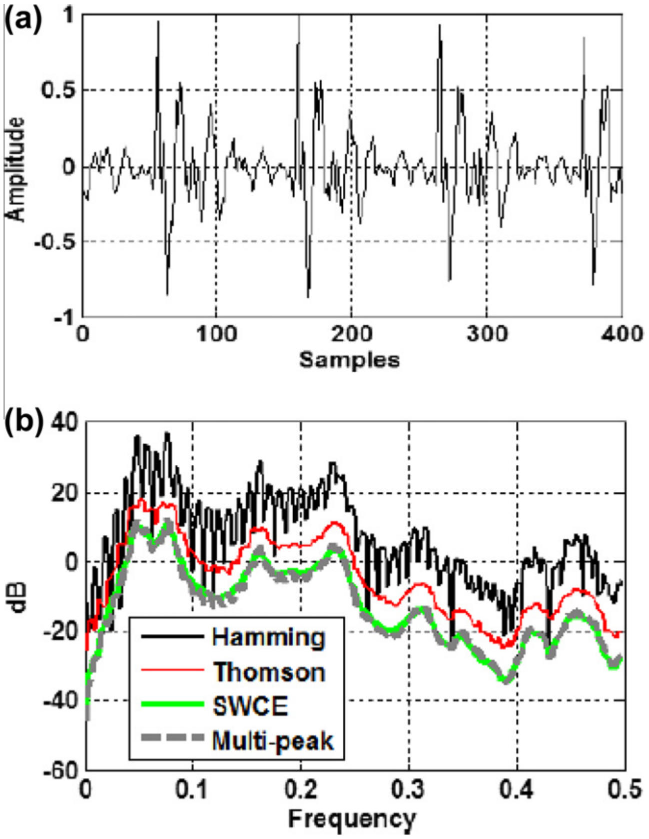

Fig. 6. (a) Speech signal, (b) estimated spectrum by the single taper (Hamming) and the multi-taper methods. Sampling frequency is 16 kHz, frame length 25 ms and number of tapers used for the multi-taper methods is 6.

3. Multi-taper MFCC and PLP feature extraction

The two most widely used forms of speech parameterizations are the mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980) and the perceptual linear prediction (PLP) coefficients (Hermansky, 1990). Figs. 7 and 8 present the generalized block diagrams of MFCC and PLP feature extraction processes, respectively. MFCC extraction begins with pre-processing (DC removal and pre-emphasis using a first-order high-pass filter with transfer function $H(z) = 1 - 0.97 ^{*} z^{-1}$). Short-time Fourier transform (STFT) analysis is then carried out using a single taper (e.g., Hamming) or multi-taper technique, and triangular Mel-frequency integration is performed for auditory spectral analysis. The logarithmic nonlinearity stage follows, and the final static features are obtained through the use of discrete cosine transform (DCT).

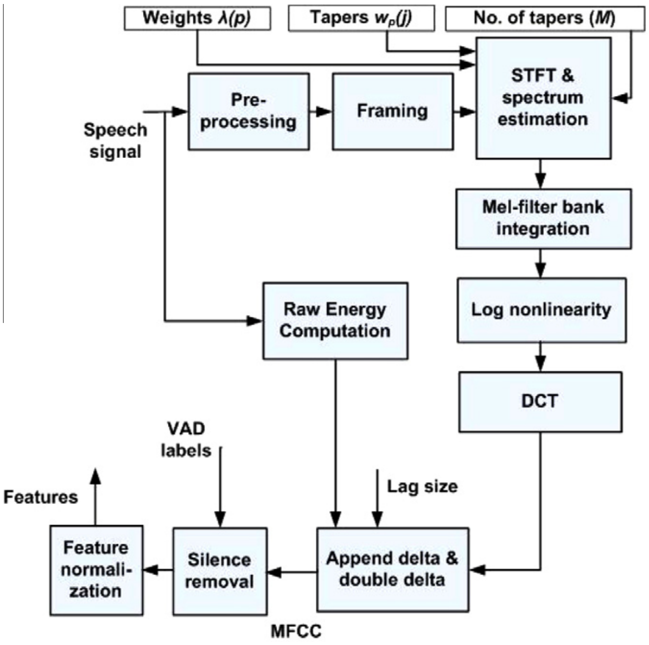

Fig. 7. Generalized block diagram for the single taper and multi-taper spectrum estimation-based MFCC feature extraction.

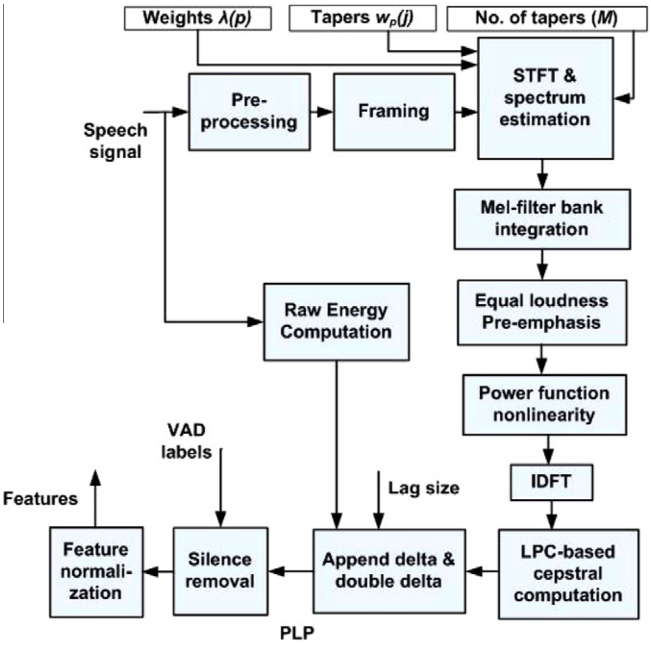

Fig. 8. Generalized block diagram for the single taper and multi-taper spectrum estimation-based PLP feature extraction.

PLP processing, which is similar to MFCC processing in some ways, begins with STFT analysis followed by critical-band integration using trapezoidal frequency-weighting functions. In contrast to MFCC, pre-emphasis is performed based on an equal-loudness curve after frequency integration. The nonlinearity in PLP is based on the power-law nonlinearity proposed by Hermansky (1990). After this stage, inverse discrete Fourier transform (IDFT) is used for obtaining a perceptual autocorrelation sequence following the linear prediction (LP) analysis. Cepstral recursion is also usually performed to obtain the final features from the LP coefficients (Gold and Morgan, 2000). Here, for PLP feature extraction, we follow HTK-based processing (Young et al., 2006), in which, for auditory frequency analysis, a Mel filterbank is used instead of a trapezoidal-shaped bark filterbank.

After extracting the static MFCC or PLP features, augmented with the log energy of the frame, the delta and double delta features are computed using the following regression formula:

$$\Delta c(m,t) = \frac{\sum_{q=1}^{L_{lag}} q(c(m+q,t) - c(m-q,t))}{2\sum_{q=1}^{L_{lag}} q^2}, \tag{7}$$

where m is the frame index, t is the cepstral index, L_{lag} represents the window lag size, and $c(m, t)$ is the tth cepstral coefficient of the mth frame. Nonspeech frames are removed using our voice activity detector (VAD) labels. For telephone speech, the VAD labels are produced by a Hungarian phoneme recognizer (Matejka, 2006; ABC System) and for microphone speech, VAD labels are generated using a GMM-based VAD by training one GMM for nonspeech and another one for speech (CRIM System). Final features are obtained after appending the delta and double delta features and normalizing the features using a short-time Gaussianization (STG) method (Xiang et al., 2002; Pelecanos and Sridharan, 2001).

There is a limit to the number of tapers that can be used in multi-taper spectrum analysis for the computation of the MFCC or PLP features. Specifically, spectral leakage increases with each taper in the sequence. For a time-bandwidth product $tbp = 2NW$ from 3 to 5, a usual range for the number of tapers $M = 2^{tbp\text{-}1}$ is from 4 to 16, where N is the taper length and W is the design interval expressed as $W = (M+1)/2(N+1)$. The optimal number of tapers for our recognition task is found to be $M_{opt} = 6$. Since speech recognition and speaker recognition systems share similar front-ends, we first determined the optimum number of tapers for speech recognition by doing a series of recognition experiments by ranging M from 4 to 10 (Alam et al., 2011) and applying the optimum value ($M_{opt} = 6$) to the speaker verification task. Interestingly, in the recent extensive speaker verification experiments on NIST 2002 and NIST 2008 corpora using three independently constructed speaker verification systems (Kinnunen et al., in press), the optimum range for M was found to be $3 \leqslant M \leqslant 8$ with a recommended value of $M = 6$. Therefore, in this study we fix $M = 6$ and focus on studying the i-vector recognizer accuracy across the multiple conditions available in the NIST 2010 SRE data.

4. Speaker verification using i-vector framework

Given two recordings of speech in a speaker detection trial, each assumed to have been uttered by a single speaker, are both speech utterances produced by the same speaker or by two different speakers? Speaker verification is the implementation of this detection task. Speaker detection provides a scalar valued match score for each trial, where a large score favors the target hypothesis (i.e., same speaker hypothesis) and a small score favors the non-target hypothesis (i.e., different speaker hypothesis). In the NIST speaker recognition evaluations (SREs), non-target trials may be male, female, or mixed but target trials, by definition, cannot have mixed gender. Real world deployment of a gender dependent speaker recognition system is not straightforward and typically involves making a premature hard-decision based on a gender detector output. Recently, in (Senoussaoui et al., 2011), an i-vector system based on probabilistic linear discriminant analysis (PLDA) is introduced, where a mixture of gender-dependent models (i.e., a male PLDA model and a female PLDA model) is used to compute the likelihood ratio scores for speaker verification. This system avoids the need for explicit gender detection. Here, we adopt this gender-independent speaker recognition system for the speaker verification experiments. An i-vector speaker verification system consists of three steps, extraction of i-vectors, generative PLDA modeling of the i-vectors and, finally, likelihood ratio computation (or scoring). We review these shortly in the following.

4.1. Extraction of i-vectors

i-Vector extractors have become the state-of-the-art technique in the speaker verification field. An i-vector extractor represents entire speech segments as low-dimensional feature vectors called i-vectors (Dehak et al., 2011; Kenny, 2010; Brümmer and de Villiers, 2010). The i-vector extractors studied in (Dehak et al., 2011; Kenny, 2010; Brümmer and de Villiers, 2010) are – according to long traditions in speaker verification research following NIST SRE evaluation protocol – gender-dependent and they are followed by gender-dependent generative modeling stages. In this paper, however, we use a gender-*independent* i-vector extractor, as shown in Fig. 9, trained on both microphone and telephone speech. The universal background model (UBM) used in this i-vector extractor is also gender-independent. The advantage of a gender-independent system is simplified system design as separate female and male detectors do not need to be constructed. In order to handle telephone as well as microphone speech, the dimension of the i-vectors is reduced from 800 to 200 using ordinary linear discriminant analysis (LDA). The purpose of applying length normalization is to Gaussianize the distribution of the i-vectors so that a simple Gaussian PLDA model can be used instead of the heavy-tailed PLDA model (Garcia-Romero and Espy-Wilson, 2011), i.e., PLDA models with heavy-tailed prior distributions (Kenny, 2010). A heavy-tailed PLDA is 2–3 times slower than the Gaussian PLDA.

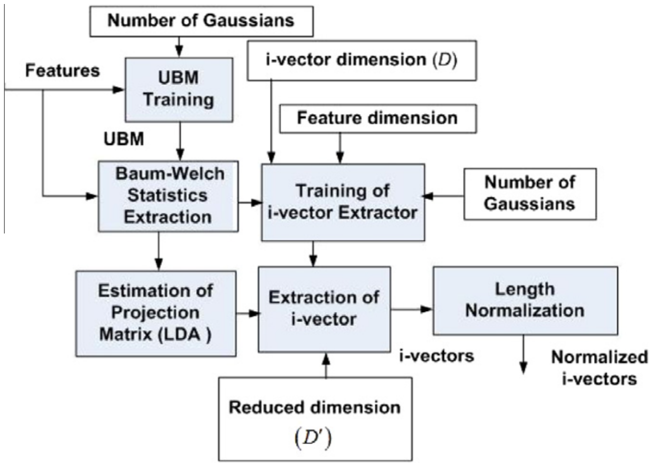

Fig. 9. Gender-independent i-vector extractor.

4.2. Generative PLDA model for i-vectors

In a generative PLDA model, the i-vectors, denoted by $\boldsymbol{i}$, are assumed to be distributed according to (Kenny, 2010):

$$\boldsymbol{i} = \boldsymbol{V}\boldsymbol{y} + \boldsymbol{m} + \varepsilon, \tag{8}$$

where the *speaker variable*, $\boldsymbol{y}$ is Gaussian distributed and its value is common to all segments of a given speaker, $\boldsymbol{m}$ is the mean vector, $\boldsymbol{V}$ is a fixed hyper-parameter matrix and ε is the residual assumed to be Gaussian. Usually $\boldsymbol{m}$, $\boldsymbol{V}$ and the residual covariance matrix are taken to be gender-dependent, which is optimal for NIST conditions. Probability calculations with this model involve a Gaussian integral that can be evaluated in closed form (Kenny, 2010).

4.3. Likelihood ratio computation

In a speaker verification task, given a pair of i-vectors $z = (i_1, i_2)$, the likelihood ratio is computed as:

$$\frac{P(z|H_1)}{P(z|H_0)} = \frac{P(z|H_1)}{P(i_1)P(i_2)}, \tag{9}$$

where the target hypothesis H_1 indicates that both i_1 and i_2 share the same speaker variable y (i.e., $y_1 = y_2$) and the non-target hypothesis indicates that the i-vectors were generated from different speaker variables y_1 and y_2. Because i_1 and i_2 can be considered independent under the non-target hypothesis H_0, $P(z|H_0)$ factorizes as $P(i_1)P(i_2)$. In this work, we use a gender-independent likelihood ratio computation framework as described in (Senoussaoui et al., 2011).

5. Experiments

5.1. Experimental Setup

We conducted experiments on the trial lists from the extended *core–core* condition of the NIST 2010 speaker recognition evaluation (SRE) corpus. To evaluate the performance of our speaker recognition systems we used the following evaluation metrics: equal error rate (EER), and the new normalized minimum detection cost function (minDCF$_{\text{new}}$). EER corresponds to the operating point with equal miss and false alarm rates whereas minDCF$_{\text{new}}$ correspond to the evaluation metrics for the NIST SRE 2010 protocols. The normalized detection cost function DCF_n, used to measure the performance of a speaker recognition system for application specific costs and priors, is defined as:

$$DCF_n = \frac{C_{Miss}P(Miss|Target)P_{Target} + C_{FA}P(FA|Non\text{-}target)(1 - P_{Target})}{\min\{C_{Miss}P_{Target}, C_{FA}P_{Non\text{-}target}\}}, \tag{10}$$

where C_{Miss} and C_{FA} represent the costs of miss and false alarm, respectively. Further, P_{Target} and $P_{Non\text{-}target} = 1 - P_{Target}$ are the prior probabilities of the target and non-target trial, respectively. For NIST 2010 SRE, cost values $C_{Miss} = C_{FA} = 1$ and $P_{Target} = 0.001$ are used. The normalized minimum detection cost function (minDCF$_{\text{new}}$) is the minimum of DCF_n over the threshold that determines $P(FA)$ and $P(Miss)$.

The *relative improvement* (RI) in performance (either EER or minDCF$_{\text{new}}$) of the multi-taper systems over the corresponding baseline system is calculated as:

$$RI = \frac{R_{baseline} - R_{mt}}{R_{baseline}} * 100\%, \tag{11}$$

where $R_{baseline}$ and R_{mt} represent, respectively, the results of the baseline and the multi-taper systems.

Based on the single taper (e.g., Hamming window) and multi-taper MFCC and PLP features, we developed four speaker verification systems as shown in Table 2. Our baseline systems are based on the Hamming windowed MFCC and PLP features. For the Thomson (Thomson, 1982), Multi-peak (Hansson and Salomonsson, 1997) and SWCE (Hansson-Sandsten and Sandberg, 2009) methods, as mentioned in Table 2, MFCC features are computed from the multi-taper spectrum estimates described in Section 2. We report results on all of the principal subconditions (telephone speech and microphone speech) of the NIST 2010 SRE for the baseline and multi-taper systems.

Table 1
Evaluation conditions (*extended core–core*) for the NIST 2010 SRE task.

Condition	Task
det1	Interview in training and test, same mic.
det2	Interview in training and test, different mic.
det3	Interview in training and normal vocal effort phone call over tel. channel in test.
det4	Interview in training and normal vocal effort phone call over mic channel in test.
det5	Normal vocal effort phone call in training and test, different tel.

Table 2
Single-taper and multi-taper MFCC and PLP feature-based speaker verification systems.

System	Description
Hamming (baseline)	MFCC and PLP features are computed from the Hamming windowed spectrum estimate.
SWCE	MFCC and PLP features are computed from the sinusoidal weighted (i.e., *sine* tapered) spectrum estimate (Hansson-Sandsten and Sandberg, 2009).
Multi-peak	MFCC and PLP features are computed from the multi-taper spectrum estimate using multi-peak tapering (Hansson and Salomonsson, 1997).
Thomson	MFCC and PLP features are calculated from the multi-taper spectrum estimates with dpss tapering (Thomson, 1982) and adaptive weights.

5.1.1. Feature Extraction

For our experiments, we use 20 static MFCC or PLP features (including the log energy) augmented with their delta and double delta coefficients, making 60-dimensional MFCC (PLP) feature vectors. MFCC and PLP features are extracted following the procedures shown in Figs. 6 and 7, respectively, with a frame shift of 10 ms. Delta and double features are calculated using a 5-frame window (i.e., ±2 frame lag) for the baseline and the multi-taper systems. Nonspeech frames are then removed using pre-computed VAD labels using algorithms mentioned in Section 3. For feature normalization, we apply the short-time Gaussianization (STG) technique (Xiang et al., 2002; Pelecanos and Sridharan, 2001) over a 300-frame window.

5.1.2. Training the universal background model (UBM)

We train a gender-independent, full covariance universal background model (UBM) with 2048-component Gaussian mixture models (GMMs) by pooling all training features together. NIST SRE 2004 and 2005 telephone data (420 female speakers and 307 male speakers in 305 hours of speech) are used for training the UBM. Normally, to train

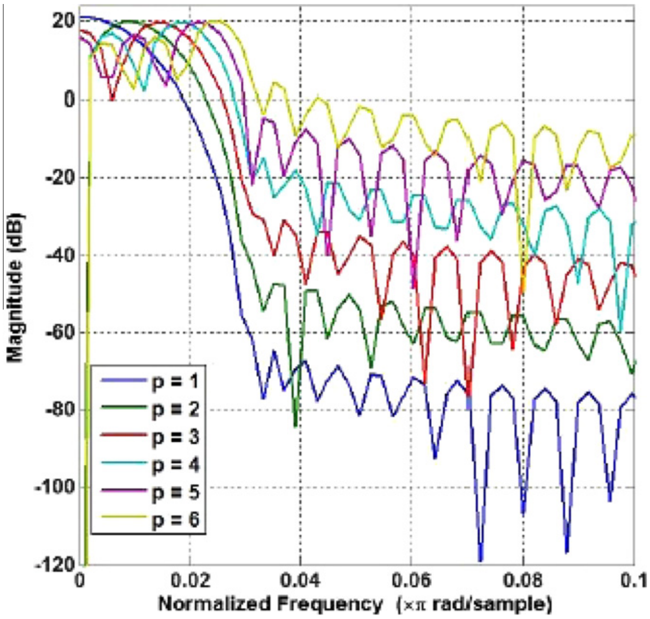

Fig. 10. Frequency domain plot of six ($M = 6$) Slepian tapers, p is the taper index. Attenuation in the side-lobes decreases for higher order tapers.

a gender-independent UBM by pooling all the training data, the pooled data should be balanced over the subpopulations, i.e., male and female, telephone and microphone. If the pooled data are not balanced then the final model may be biased towards the dominant subpopulations (Reynolds).

In this work, our gender-independent UBM is trained from NIST SRE 2004 and 2005 telephone data that include more female trials than male. Therefore, the verification results for female trials should be better than that of the male trials. But our obtained results (for the baseline Hamming and multi-taper systems) depict that the verification results (in terms of EER, $\text{minDCF}_{\text{old}}$, and $\text{minDCF}_{\text{new}}$) for male trials are consistently better than that for female trials, so the trained UBM is not biased towards the female trials. It should be mentioned here that, in this work, the data used for training a gender-independent i-vector extractor includes female trials 1.3 times of the male trials.

Training an UBM from a balanced set of female-male trials or inclusion of microphone data (NIST SRE 2005 microphone and/or NIST SRE 2006 microphone data) with the telephone data for training UBM did not help our system to improve recognition performance but increased the UBM training time considerably. The possible reasons why including microphone data to UBM or training an UBM from a balanced set of female-male trials did not help our systems could be: Firstly, we have more telephone data (approximately 10 times of microphone data) than the microphone data for training the i-vector extractor and consequently more i-vectors from telephone

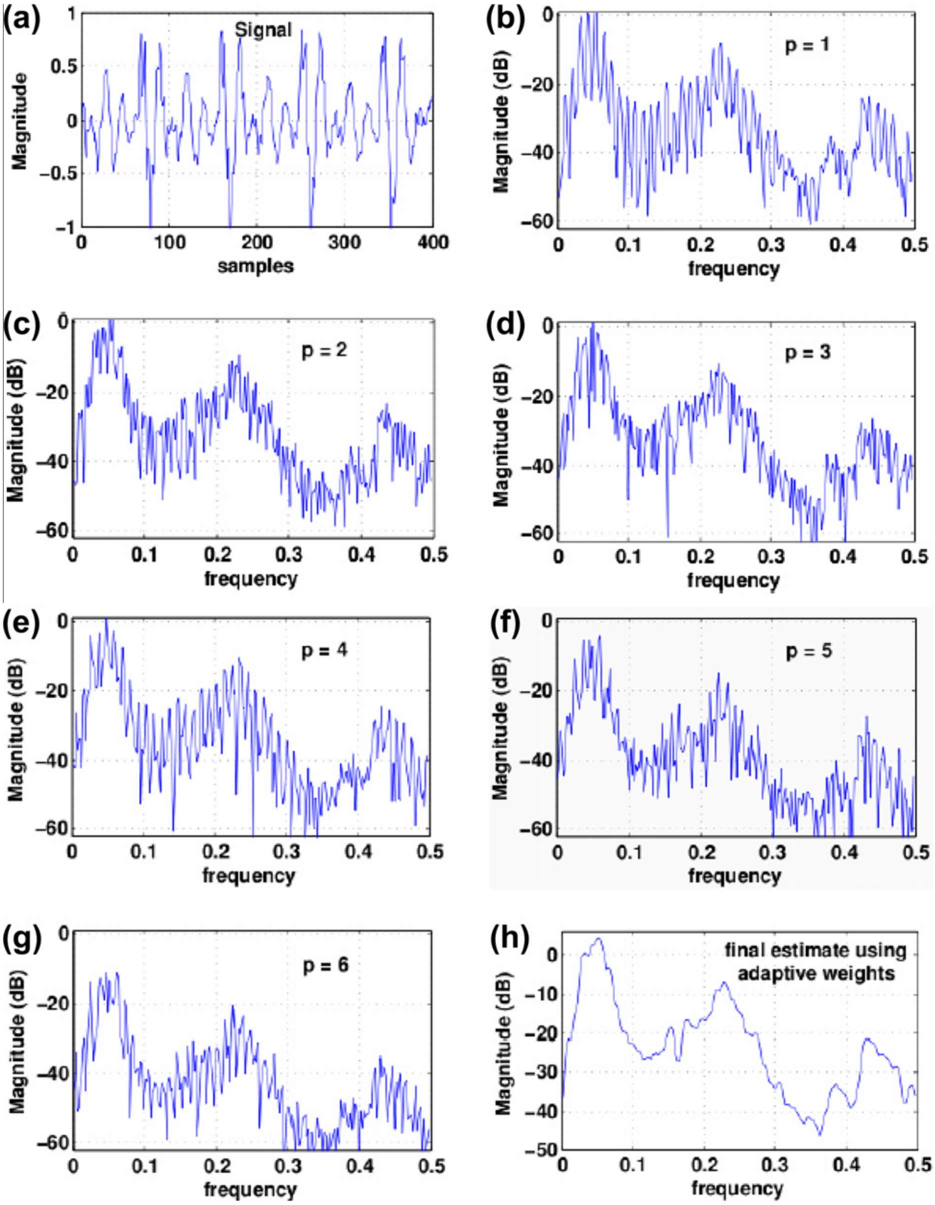

Fig. 11. Multi-taper spectral estimates when adaptive weights are applied to the individual estimates (b–g) to get the final estimate (h) of a 25 ms duration speech signal (a).

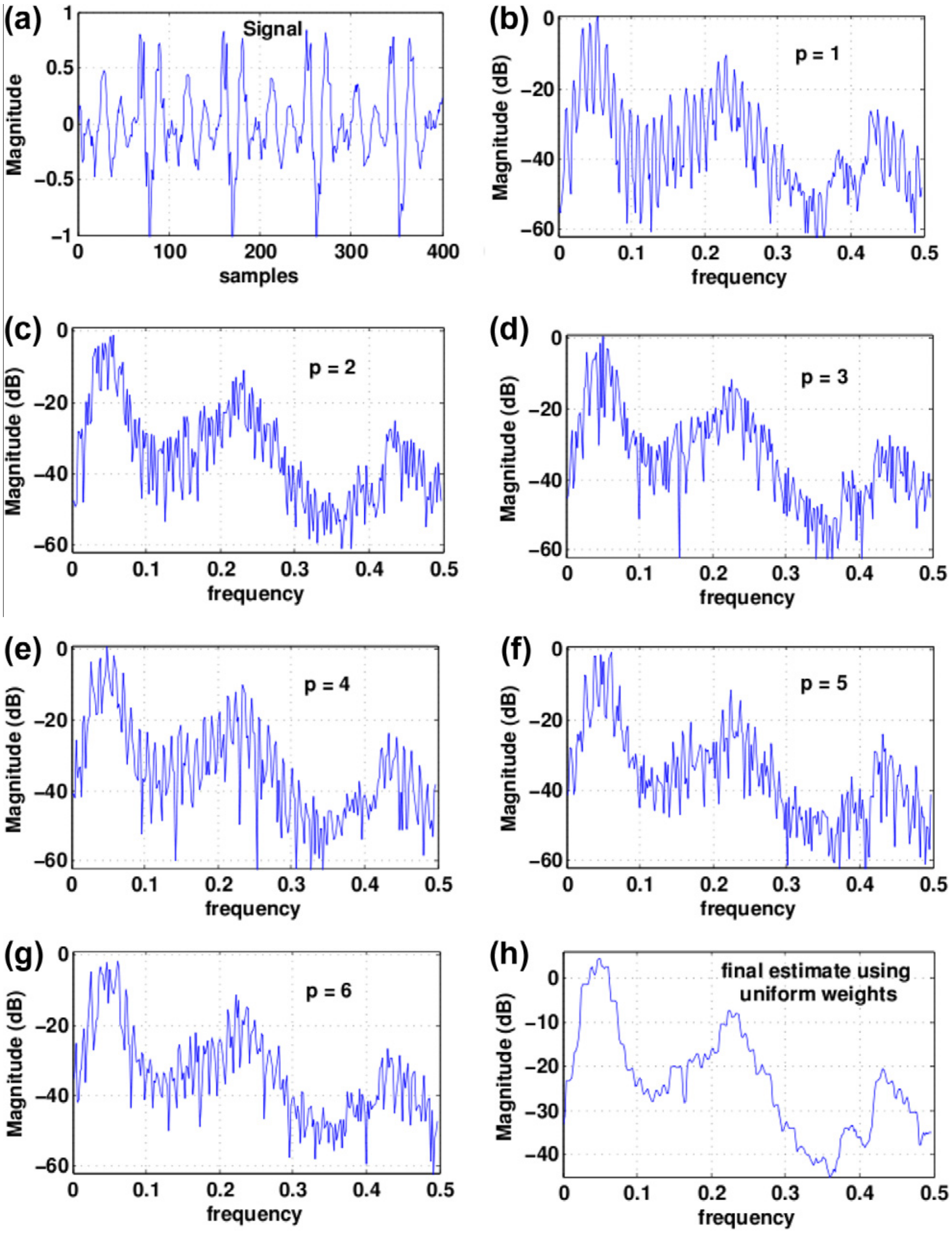

Fig. 12. Multi-taper spectral estimates when uniform weights ($1/M$) are applied to the individual estimates (b–g) to get the final estimate (h) of a 25 ms duration speech signal (a).

data than that from microphone data for training the PLDA models. Moreover, to handle both the microphone and telephone speech, we use ordinary linear discriminant analysis where the between-class scatter matrix is estimated from all telephone training data and the within-class scatter matrix is estimated using all telephone and microphone training, as described in Section 5.1.3, to reduce the dimensionality of the i-vectors from 800 to 200 (Senoussaoui et al., 2010). Secondly, the ratio of female to male utterances in the database is approximately 1.3:1 and therefore, we have more i-vectors from female utterances from training the PLDA models.

Note also that, for the baseline Hamming and the multi-taper systems, we use same data sets for training the UBM and other components of the system.. The only difference between the baseline and multi-taper systems is in the spectrum estimation method.

5.1.3. Training and extraction of i-vectors

A block diagram of the i-vector extractor used in this paper is shown in Fig. 9. Our gender-independent i-vector extractor is of dimension 800. After training the gender-independent UBM, we train the i-vector extractor using the Baum-Welch (BW) statistics extracted from the following data: LDC release of Switchboard II – phase 2 and phase 3, Switchboard Cellular – part 1 and part 2, Fisher data, NIST SRE 2004 and 2005 telephone data, NIST SRE 2005 and 2006 microphone data and NIST SRE

2008 interview development microphone data. Fisher data used in this work are Fisher English. In order to reduce the i-vector dimensionality, a linear discriminant analysis (LDA) projection matrix is estimated from the BW statistics by maximizing the following objective function:

$$B_{LDA} = \arg\max_{B} \frac{|B^T \Sigma_b B|}{|B^T \Sigma_w B|}, \qquad (12)$$

where B is the LDA transformation matrix, Σ_b and Σ_w represent the between- and within-class scatter matrices, respectively. The optimization problem in (8) is equivalent to finding the eigenvectors φ corresponding to the largest eigenvalues η of the following generalized eigenvalue problem:

$$\Sigma_b \varphi = \eta \Sigma_w \varphi, \qquad (13)$$

For the estimation of Σ_b we use all telephone training data excluding the Fisher data and Σ_w is estimated using all telephone and microphone training data excluding the Fisher data. We choose only speakers with more than four utterances for the estimation of LDA transformation matrix. Dimensionality reduction via LDA helps to handle microphone speech as well as telephone speech (Senoussaoui et al., 2010). An optimal reduced dimension of 200 is determined empirically.

We then extract 200-dimensional i-vectors for all training data excluding Fisher data by applying this transformation matrix on the 800-dimensional i-vectors. For the test data, first BW statistics and then 200-dimensional i-vectors are extracted following a similar procedure using the same projection matrix. We also normalize the length (using 2-norm) of the i-vectors to gaussianize the i-vectors distribution (Garcia-Romero and Espy-Wilson, 2011).

5.1.4. Training the PLDA model

We train two PLDA models, one for the males and another for females. These models were trained using all the telephone and microphone training i-vectors; then we combine these PLDA models to form a mixture of PLDA models in i-vector space as described in (Senoussaoui et al., 2011). For both of the models, the fixed hyper-parameter $\boldsymbol{V}$ is a full rank matrix of dimension 200. For training the PLDA models we choose only speakers with more than four utterances.

5.2. Results and discussion

5.2.1. Use of uniform versus non-uniform weights in multi-tapering

Usually, in a multi-taper spectrum estimation method, the final spectrum is obtained by averaging (using uniform weights, $1/M$) over the M tapered subspectra. In (Kinnunen et al., 2010; Sandberg et al., 2010), for the Thomson multi-taper method, the individual spectra were averaged to obtain the final estimate. Only the first taper ($p = 1$) in the multi-taper method produces a central peak at the harmonic frequency of the component while the other tapers ($p > 1$) produce spectral peaks that are shifted slightly up or down in frequency. The information lost at the extremes of the first taper is included and indeed emphasized in the subsequent tapers. As can be seen from Fig. 10, attenuation in the side-lobes decreases with each taper in the sequence, i.e., spectral leakage increases for the higher-order tapers. If uniform weights are applied to get the final spectrum estimate, the energy loss at higher-order tapers will be high. In order to compensate for this increased energy loss, a weighted average (using non-uniform weights) is used instead of simply averaging the individual estimates. In (Thomson, 1982), the weights are changed adaptively to optimize the bias/variance tradeoff of the estimator. Figs. 11 and 12 provide a comparison of the multi-taper spectral estimates when uniform & non-uniform weights are applied, respectively. Table 3 presents a comparison of the use of uniform and non-uniform weights (eigenvalue as the weight, EVW) and adaptive weight (AW) computed from the eigenvalues) in the Thomson multi-taper method, in the context of speaker verification. The speaker verification results suggest that non-uniform weights, specifically, the adaptive weights, should be preferred.

Table 3
Comparison of Speaker verification results (EER %) using a mixture PLDA model for the Thomson multi-taper method when uniform weights (UW), Eigenvalues as the weights (EVW) and adaptive weights (AW) are used to obtain the final spectrum estimate. The results of the baseline Hamming system are also included for comparison purposes. . For each condition, the minimum value is highlighted with boldface. We have 60-dimensional MFCC features, a 256-component UBM and 800-dimensional i-vector extractor with dimension reduced to 150.

EER (%)					
Gender	Condition	Thomson			Baseline Hamming
		UW	EVW	AW	
Female	det1	2.4	**2.1**	**2.1**	2.4
	det2	4.5	4.4	**4.2**	4.6
	det4	3.9	3.7	**3.4**	3.9
	det3	3.1	**2.9**	**2.9**	3.6
	det5	**3.2**	3.4	**3.2**	4.0
Male	det1	1.6	1.6	**1.0**	1.5
	det2	3.0	2.7	**2.5**	3.1
	det4	2.4	2.2	**1.9**	2.6
	det3	3.5	3.3	**2.8**	4.1
	det5	2.7	2.5	**2.4**	3.2

5.2.2. Performance evaluation of multi-taper MFCC and PLP features

To evaluate and compare the performance of the systems in Table 2, we conducted experiments using both telephone and microphone speech on the extended core–core condition of the NIST SRE 2010 task. The results are reported for five evaluation conditions corresponding to detection (det) conditions 1–5, as shown in Table 1, as specified in the evaluation plan (National Institute of Standards and Technology).

Fig. 13 presents EERs for the Hamming (baseline) and multi-taper MFCC systems both for the female and male trials. For all the MFCC-based systems, minDCF$_{\text{new}}$ is

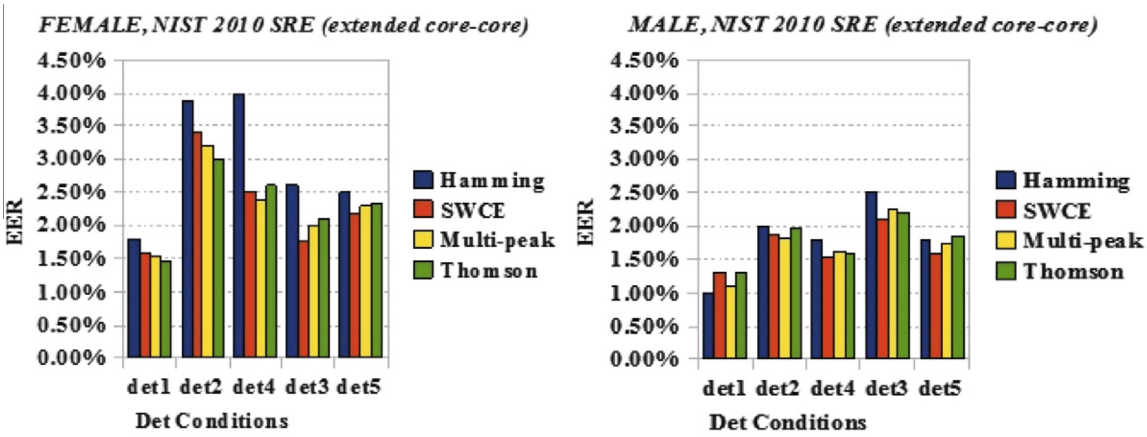

Fig. 13. Male and female det1 to det5 speaker verification results for the baseline Hamming window system and multi-taper systems, measured by EER: 60-dimensional MFCCs with log-energy, deltas and double deltas, UBM with 2048 Gaussians, 800-dimensional i-vectors with reduced dimension of 200.

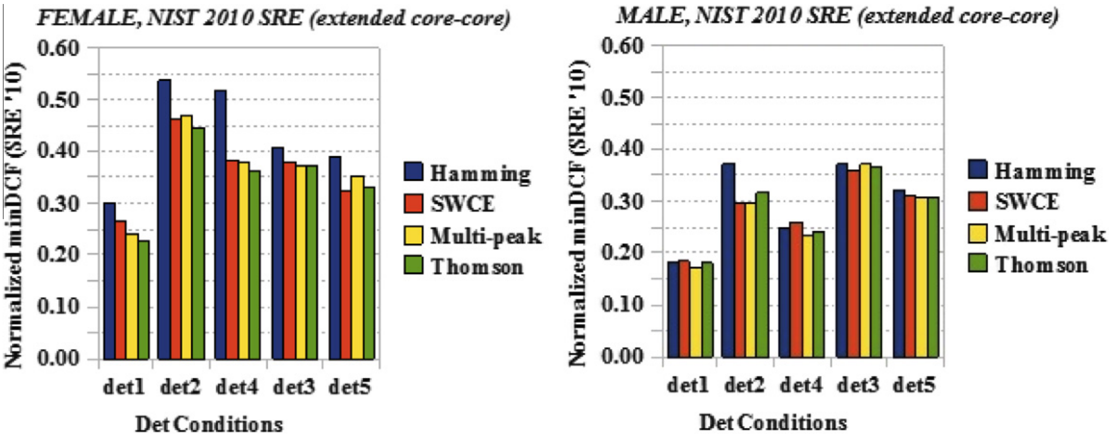

Fig. 14. Same as Fig. 13 but for minDCF$_{\text{new}}$.

shown in Fig. 14, for the male and female trials. In terms of both metrics, EER, and minDCF$_{\text{new}}$, multi-taper MFCC systems outperform the baseline MFCC system. Compared to the baseline (Hamming) MFCC system, average relative improvements (female–male, det1–det5), as shown in Table 3, obtained by the multi-taper systems are as follows:

Relative improvements of the SWCE MFCC system are 12.2%, and 9.7% in EER, and minDCF$_{\text{new}}$, respectively. The multi-peak system provides relative improvements of 12.6%, and 15.4% in EER, and minDCF$_{\text{new}}$, respectively. The corresponding improvements for the Thomson method are 17.1%, and 11.9%.

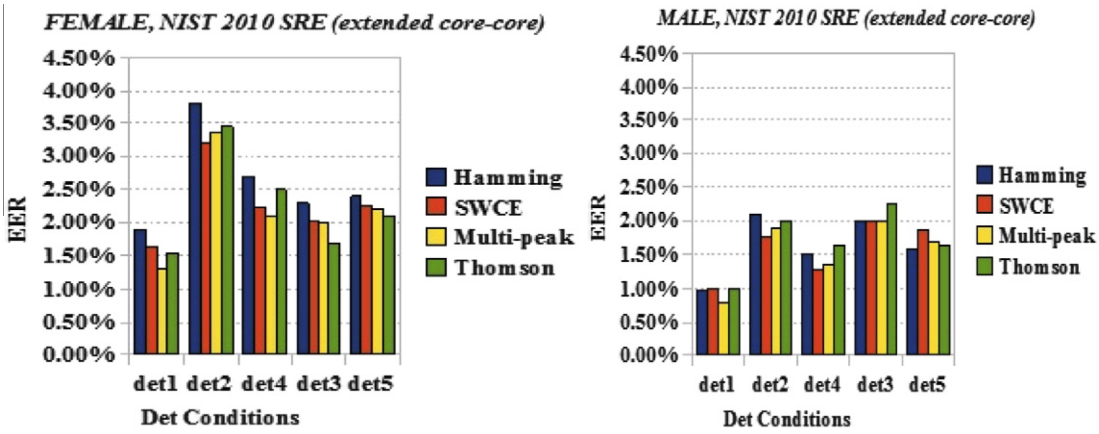

Fig. 15. Male and female det1 to det5 speaker verification results for the baseline Hamming window system and multi-taper systems, measured by EER: 60-dimensional PLP with log-energy, deltas and double deltas, UBM with 2048 Gaussians, 800-dimensional i-vectors with dimension reduced to 200.

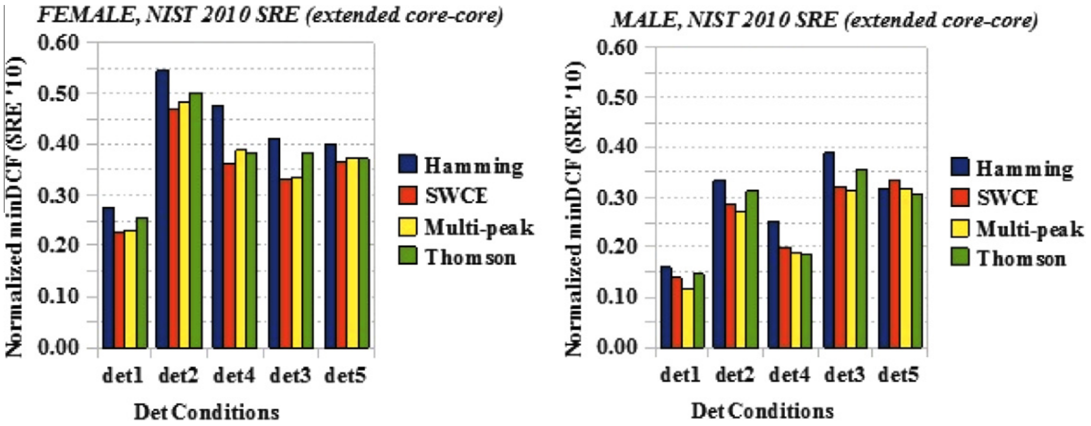

Fig. 16. Same as Fig. 16 but for minDCF$_{new.}$

Figs. 15 and 16 present EER and minDCF$_{new}$ values, respectively, for the Hamming (baseline) and multi-taper PLP systems both for the male and female trials. In the case of female trials, all the multi-taper PLP systems yield systematically less errors in comparison to the baseline PLP in terms of all the evaluation measures. For male trials, the multi-peak and SWCE PLP systems provide higher accuracy in the first four det conditions (1, 2, 3, 4). The results for the det5 condition for both systems are close to the baseline. Compared to the baseline PLP, the Thomson PLP system also performs better except in the det3 and det4 conditions in EER for the male trials.

Compared to the Hamming PLP system, average relative improvements (female–male, det1–det5), as shown in Table 3, obtained by the multi-taper PLP systems are as follows. Relative improvements of SWCE, Multi-peak and Thomson PLP systems are 7.5%, 11.6% and 5.0% in terms of EER, and 14.4%, 16.2% and 10.1% in terms of minDCF$_{new.}$

Although all three multi-taper variants outperformed the baseline Hamming method, considering the performances of both of the front-ends (i.e., MFCC and PLP), the SWCE and multi-peak systems are preferred.

In the multi-taper spectrum estimators, data are more evenly weighted and they have a reduced variance compared to single-tapered direct spectrum estimates. It is straightforward to choose the weights used in constructing the multi-taper estimate in order to minimize the estimation variance.

Table 4
Average relative improvement in both female and male trials in det1 to det5 conditions obtained by the multi-taper systems over the baseline system. The larger the relative improvement, the more effective the improvement due to multi-tapering. For each evaluation metric (EER or minDCF$_{new}$) and for each front-end (MFCC or PLP) the maximum value is highlighted with boldface.

Average relative improvement (male–female, det1–det5)						
	SWCE		Multi-peak		Thomson	
	MFCC	PLP	MFCC	PLP	MFCC	PLP
EER	12.3	7.5	**12.6**	**11.6**	9.5	5.0
minDCF$_{new}$	9.7	14.4	11.5	**16.2**	**11.9**	10.1

6. Conclusion

In this paper we used multi-taper spectrum estimation approaches for low-variance MFCC and PLP feature computation and compared their performances, in the context of i-vector speaker verification, against the conventional single-taper (Hamming window) technique. In a Thomson multi-taper method, instead of uniform weights, use of non-uniform weights, specifically adaptive weights, can bring improvement in speaker recognition. Experimental results on the telephone and microphone portion of the NIST 2010 SRE task indicate that multi-tapering using sine or multi-peak or Slepian tapers outperforms the baseline single-taper method in most cases. Among the three multi-taper methods, the multi-peak and the SWCE MFCC systems outperformed the Thomson method (if uniform weights are chosen), which agrees well with the results of Kinnunen et al. (2010, in press). However, if non-uniform weights (e.g., eigenvalues) are used in the Thomson method, from Table 4 it is observed that the Thomson MFCC system can outperform the other two multi-taper MFCC systems. The number of tapers was set to 6 according to (Kinnunen et al., 2010, in press; Alam et al., 2011) without additional optimizations on the i-vector speaker verification system. The largest relative improvements over the baseline were observed for conditions involving microphone speech. Overall, the multi-taper method of MFCC and PLP feature extraction is a viable candidate for replacing the baseline MFCC and PLP features.

Acknowledgements

The work of T. Kinnunen was supported by the Academy of Finland (Project No. 132129). We would like to thank the anonymous reviewers for their comments that helped to improve the content of this paper.

References

ABC System description for NIST SRE 2010.

Alam, J., Kenny, P., O'Shaughnessy, D., 2011. A study of low-variance multi-taper features for distributed speech recognition. In: Proc. NOLISP. LNAI, vol. 7015, pp. 239–245.

Alam, M.J., Kinnunen, T., Kenny, P., Ouellet, P., O'Shaughnessy, D., 2011. Multitaper MFCC features for speaker verification using i-vectors. In: Proc. ASRU, pp. 547–552.

Brümmer, N., de Villiers, E., 2010. The speaker partitioning problem. In: Proc. Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic.

The CRIM System for the 2010 NIST Speaker Recognition Evaluation.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28 (2), 357–366.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. 19 (4), 788–798.

Djuric, P.M., Kay, S.M., 1999. Spectrum Estimation and Modeling. Digital Signal Processing Handbook. CRC Press LLC.

Garcia-Romero, D., Espy-Wilson, Carol Y., 2011. Analysis of i-vector length normalization in speaker recognition systems. In: Proc. Interspeech, Florence, Italy, pp. 249–252.

Gold, B.G., Morgan, N., 2000. Speech and Audio Signal Processing: Processing and Perception of Speech and Music. John Wiley & Sons, Inc., New York.

Hansson, M., Salomonsson, G., 1997. A multiple window method for estimation of peaked spectra. IEEE Trans. Signal Process. 45 (3), 778–781.

Hansson-Sandsten, M., Sandberg, J., 2009. Optimal cepstrum estimation using multiple windows. In: Proc. ICASSP, pp. 3077–3080.

Harris, F., 1978. On the use of windows for harmonic analysis with the discrete Fourier transform. Proc. IEEE 66 (1), 51–84.

Hermansky, H., 1990. Perceptual linear prediction (PLP) analysis of speech. J. Acoust. Soc. Amer. 87 (4), 1738–1752.

Honig, Florian, Stemmer, George, Hacker, Christian, Brugnara, Fabio, 2005. Revising perceptual linear prediction (PLP). In: Proc. Interspeech, pp. 2997–3000.

Hu, Y., Loizou, P., 2004. Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans. Speech Audio Process. 12 (1), 59–67.

Kay, S.M., 1988. Modern Spectral Estimation. Prentice-Hall, Englewood Cliffs, NJ.

Kenny, P., 2010. Bayesian speaker verification with heavy tailed priors. In: Proc. Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007a. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans. Audio Speech Lang. Process. 15 (4), 1435–1447.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007b. Speaker and session variability in GMM-based speaker verification. IEEE Trans. Audio Speech Lang. Process. 15 (4), 1448–1460.

Kinnunen, T., Saeidi, R., Sandberg, J., Hansson-Sandsten, M., 2010. What else is new than the Hamming window? Robust MFCCs for speaker recognition via multitapering. In: Proc. Interspeech, pp. 2734–2737.

Kinnunen, T., Saeidi, R., Sedlak, F., Lee, K.A., Sandberg, J., Hansson-Sandsten, M., Li, H., 2012. Low-variance multitaper MFCC features: A case study in robust speaker verification. IEEE Trans. Audio Speech Lang. Process. 20 (7), 1990–2001.

Matejka, Pavel, Burget, Lukáš, Schwarz, Petr, Cernocký, Jan "Honza", 2006. Brno University of technology system for NIST 2005 language recognition evaluation. In: Proc. IEEE Odyssey 2006 Speaker and Language Recognition Workshop, pp. 57–64.

McCoy, E.J., Walden, A.T., Percival, D.B., 1998. Multi-taper spectral estimation of power law processes. IEEE Trans. Signal Process. 46 (3), 655–668.

National Institute of Standards and Technology, NIST Speaker Recognition Evaluation. Available from: <http://www.itl.nist.gov/iad/mig/tests/sre/>.

Pelecanos, J., Sridharan, S., 2001. Feature warping for robust speaker verification. In: Proc. Speaker Odyssey: The Speaker Recognition Workshop, Crete, Greece, pp. 213–218.

Percival, D.B., Walden, A.T., 1993. Spectral Analysis for Physical Applications, Multitaper and Conventional Univariate Techniques. Cambridge University Press.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Process. 10 (1), 19–41.

Reynolds, D.A., 2008 Universal Background Models. Encyclopedia of Biometric Recognition. Springer, Available from: <http://www.ll.mit.edu/mission/communications/ist/publications/0802_Reynolds_Biometrics_ UBM.pdf>.

Riedel, K.S., Sidorenko, A., 1995. Minimum bias multiple taper spectral estimation. IEEE Trans. Signal Process. 43 (1), 188–195.

Sandberg, J., Hansson-Sandsten, M., Kinnunen, T., Saeidi, R., Flandrin, P., Borgnat, P., 2010. Multitaper estimation of frequency-warped cepstra with application to speaker verification. IEEE Signal Process. Lett. 17 (4), 343–346.

Senoussaoui, M., Kenny, P., Dehak, N., Dumouchel, P., 2010. An i-vector extractor suitable for speaker recognition with both microphone and telephone speech. In: Proc Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic.

Senoussaoui, M., Kenny, P., Brummer, N., de Villiers, E., Dumouchel, P., 2011. Mixture of PLDA models in I-vector space for gender independent speaker recognition. In: Proc. Interspeech, Florence, Italy, pp. 25–28.

Slepian, D., Pollak, H.O., 1960. Prolate spheroidal wave functions, Fourier analysis and uncertainty – I. Bell System Tech. J. 40, 43–63.

Thomson, D.J., 1982. Spectrum estimation and harmonic analysis. Proc. IEEE 70 (9), 1055–1096.

Thomson, D.J., 1990. Quadratic-inverse spectrum estimates: Applications to paleoclimatology. Phys. Trans. Roy. Soc. Lond. A 332, 539–597.

Walden, A.T., McCoy, E.J., Percival, D.B., 1994. The variance of multitaper spectrum estimates for real Gaussian processes. IEEE Trans. Signal Process. 2, 479–482.

Wieczorek, M.A., Simons, F.J., 2005. Localized spectral analysis on the sphere. Geophys. J. Internat. 162, 655–675.

Wieczorek, M.A., Simons, F.J., 2007. Minimum-variance multi-taper spectral estimation on the sphere. J. Fourier Anal. Appl. 13 (6), 665–692.

Xiang, B., Chaudhari, U., Navratil, J., Ramaswamy, G., Gopinath, R., 2002. Short-time Gaussianization for robust speaker verification. In: IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 681–684.

Young, S.J. et al., 2006. HTK Book, 3.4 ed. Entropic Cambridge Research Laboratory Ltd.