Multi-taper MFCC Features for Speaker Verification using I-vectors

Md Jahangir Alam [#1], Tomi kinnunen [*2], Patrick Kenny [#3], Pierre Ouellet[#4], Douglas O'Shaughnessy [#5]

[#1,3,4] *CRIM, Montreal, Canada*
[#1,5] *INRS-EMT, Montreal, Canada*
[1] Jahangir.Alam@crim.ca, [3] Patrick.Kenny@crim.ca, [4] Pierre.Ouellet@crim.ca
[5] dougo@emt.inrs.ca

[*] *School of Computing, University of Eastern Finland*
Joensuu, Finland
[2] tkinnu@cs.joensuu.fi

***Abstract*—This paper studies the low-variance multi-taper mel-frequency cepstral coefficient (MFCC) features in the state-of-the-art speaker verification. The MFCC features are usually computed using a Hamming-windowed DFT spectrum. Windowing reduces the bias of the spectrum but variance remains high. Recently, low-variance multi-taper MFCC features were studied in speaker verification with promising preliminary results on the NIST 2002 SRE data using a simple GMM-UBM recognizer. In this study our goal is to validate those findings using a up-to-date i-vector classifier on the latest NIST 2010 SRE data. Our experiment on the telephone (det5) and microphone speech (det1, det2, det3 and det4) indicate that the multi-taper approaches perform better than the conventional Hamming window technique.**

I. Introduction

Useful information extraction from speech has been a subject of active research for many decades. Feature extractor (or front-end) is the first step in an automatic speaker or speech recognition system which transforms a raw signal into a compact representation. Since feature extraction is the first step in the chain, the quality of later steps (modelling and classification) strongly depends on it. The MFCC features are the most popular in speech and speaker recognition systems and they demonstrate good performance in speech and speaker recognition. The MFCC representation is an approximation of the structure of the human auditory system [1]. Since MFCC features are computed from an estimated spectrum, it is crucial that this estimate is accurate. Usually, the spectrum is estimated using a windowed periodogram [16]. Despite having low bias, a windowed periodogram has large variance and therefore, MFCC features computed from this estimated spectrum have also high variance. One elegant technique for reducing the variance is to replace a windowed periodogram estimate with a multi-taper spectrum estimate [8, 9, 10].

The multi-taper method reduces the variance of the spectral estimates by using multiple time-domain window functions or tapers rather than a single taper. The multi-taper method has been widely used in geophysical applications and has been shown in multiple cases to outperform the windowed periodogram. It has also been used in speech enhancement application [2] and, recently, in speaker recognition [3] with promising preliminary results. The preliminary experiments of [3, 8] were reported on the NIST 2002 and 2006 SRE corpora using lightweight Gaussian mixture model–universal background model (GMM-UBM) system [17] and generalized linear discriminant sequence without any session variability compensation techniques.

In this paper, our aim is to study whether the improvements in [3, 8] translate to state-of–the-art speaker verification. The recent i-vector model [4, 5, 6] includes elegant inter-session variability compensation, with demonstrated significant improvements on the recent NIST speaker recognition corpora. Since i-vector does already 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 state-of-the-art i-vector configuration. To make the system gender independent, recently, a mixture Probabilistic Linear Discriminant Analysis (PLDA)-based speaker verification system has been proposed in [6] to deal with gender dependent problem. In this paper, 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.

This paper is organized as follows: Background information on low-variance spectrum estimators is given in section II. Section III provides a description of the fundamental components of our speaker recognition system. Section IV provides a detailed description about the experimental setup used throughout all the experiments and presents results on the NIST 2010 SRE task. Conclusion is drawn in section V.

II. Low-variance Spectrum Estimation

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

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

where $k \in \{0,1,.....,K-1\}$ 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 called a taper, which usually is symmetric and decreases towards the frame boundaries (e.g., Hamming). Eq. (1) is sometimes called a single taper, modified or windowed periodogram. If $w(j)$ is a boxcar function, eq. (1) is called the periodogram. Windowing reduces the bias, i.e., difference between the estimated spectrum and the actual spectrum, but it does not reduce the variance of the spectral estimate [7] and therefore, the variance of the MFCC features computed from this estimated spectrum is also large. One way to reduce the variance of the MFCC estimator is to replace a windowed periodogram estimate by a so-called multi-taper spectrum estimate [8, 9, 10]. The multi-taper spectrum estimator is given by:

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

where N is the frame length, w_p is the pth data taper used for the spectral estimate $\hat{S}_{MT}(\cdot)$, which is also called the pth eigenspectrum, M denotes the number of tapers and $\lambda(p)$ is the weight corresponding to the pth taper. The tapers $w_p(j)$ are chosen to be orthonormal, i.e.,

$$\sum_j w_p(j)\,w_q(j) = \delta_{pq}$$

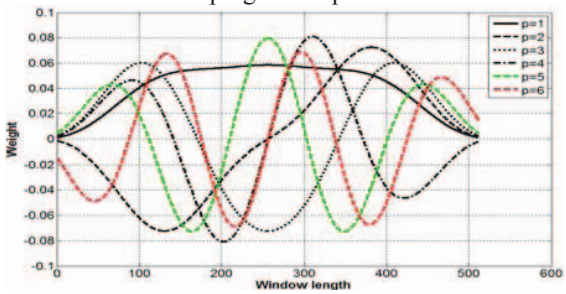
The multi-taper spectrum estimate is therefore obtained as the weighted average of M individual sub-spectra. Eq. (1) can be obtained as a special case of eq. (2) when $M=1$ and $\lambda(p)=1$. Tapers in the multi-taper method are chosen so that the estimation errors in the individual sub-spectra are uncorrelated. Averaging these uncorrelated spectra gives a low variance spectrum estimate and, consequently low variance MFCC estimate as well. The underlying philosophy of multi-taper method is similar to Welch's modified periodogram [7], it, however, focuses only one frame rather taking a time-averaged spectrum over multiple frames.

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. Various tapers have been proposed in the literature for spectrum estimation. A good set of M orthonormal data tapers with good leakage properties are specified from Slepian sequences (also called discrete prolate spheroidal sequences (dpss)) [9]. Another orthogonal family of tapers are the *sine* tapers given by [10]:

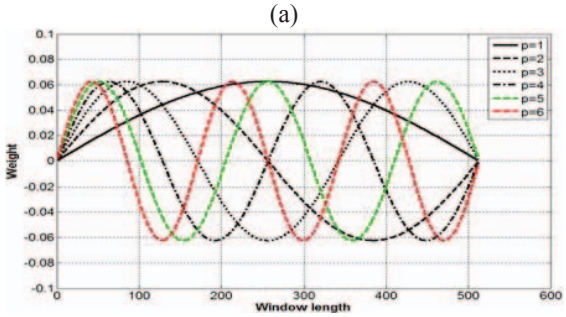
$$w_p(j) = \sqrt{\frac{2}{N+1}}\sin\left(\frac{\pi p(j+1)}{N+1}\right),\ j = 0,1,....,N-1.$$

In [11] the sine tapers are applied with optimal weighting for cepstrum analysis and multi-peak tapers are designed for peaked spectra in [12]. Fig. 1 presents the multi-peak, the sine and the Thomson tapers used for multi0taper spectrum estimation. As there exists a number of different multi-tapers to choose from, it may not be clear which multi-taper suits well for modelling speech signal. In this paper, our goal is to do a comparative evaluation of various multi-taper methods for MFCC estimation and compare their performance with the conventional Hamming window-based MFCC estimation, in the context of speaker verification [6].

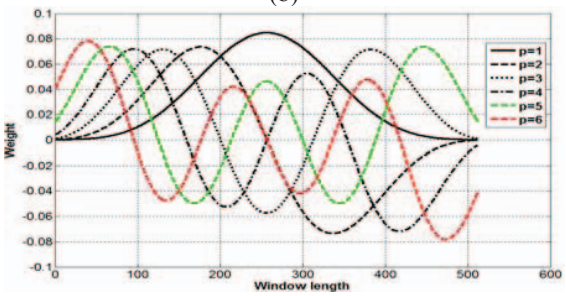
Fig. 2 shows the generalized block diagram of MFCC-DCC (MFCC- Delta Cepstral Coefficients (DCC)) computation from the multi-taper spectrum estimates. The pre-processing step may include pre-emphasizing, DC removal and signal normalization. As we mentioned above, the Hamming-windowed spectrum estimates can be obtained as a special case of the multi-taper spectrum estimation method. To compute MFCC features from single taper (or window) spectrum estimates we use $M=1$, $\lambda(1)=1$, and $w_1(j)$ is the Hamming window. After extracting MFCC-DCC features we then remove the silence frames using our VAD (Voice Activity Detector) label files. Finally, MFCC-DCC features are normalized using feature warping technique (e.g., STG (Short-time Gaussianization), STMVN (Short-time Mean and Variance Normalization)) on speech only frames. In this paper we use STG feature warping technique.



(a)



(b)



(c)

Figure 1. Plot of three types of widely used tapers for multi-taper spectrum estimation. (a) The multi-peak tapers, (b) the *sine* tapers, and (c) the Thomson tapers or Slepian sequences. Window length is 512, p is the taper number.

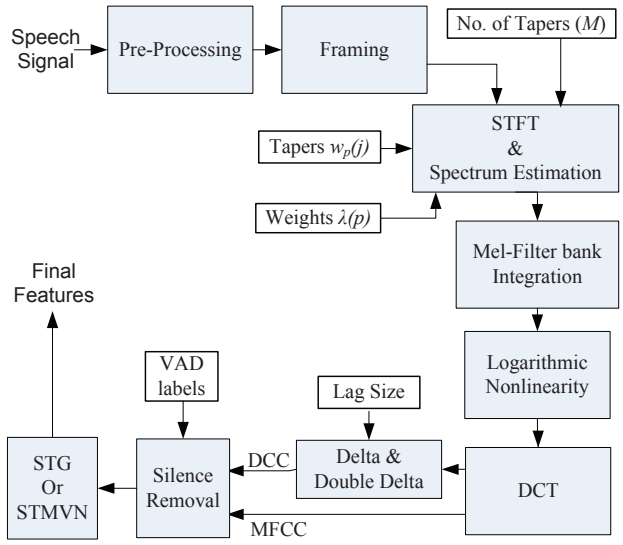


Figure 2. Generalized block diagram for the single taper and multi-taper spectrum estimation-based MFCC–DCC feature extraction.

III. Speaker Recognition System

Given two recordings of speech in a speaker detection trial, each assumed to have been uttered by a single speaker, are both speech utterances uttered by the same speaker or by two different speakers? Speaker verification is the direct implementation of this detection task. Speaker detection provides a scalar score by processing given speech recordings. A more positive score favors the target hypothesis (i.e., same speaker hypothesis) and a more negative score favors non-target hypothesis (i.e., different speaker hypothesis). Non-target trials may be male, female, or mixed but target trials, by definition, cannot have mixed gender. Similar to [6], in this paper, we are interested in the case where no gender information are provided and where there may be mixed non-target trials [6].

In the following sub-section we will provide a brief description of the fundamental components of our speaker verification system.

A. I-vector extraction

I-vector extractors have become the state-of-the-art technique in the speaker verification field. They convert an entire speech recording into a low dimensional feature vectors called i-vectors [4, 5, 14]. The i-vector extractors explained in [4, 5, 14] are gender dependent and are followed by gender dependent generative modeling stages. In this paper, we use a gender-independent i-vector extractor and a mixture of male and female Probabilistic Linear Discriminant Analysis (PLDA) models, where the gender label is treated as a latent variable [6].

B. Generative models for i-vectors

We construct a speaker detector using a generative PLDA model for a pair of i-vectors in a trial [6]. The model assumes that the i-vectors were produced by simple random processes. In the model, the pair of i-vectors z_1, z_2 is produced as follows:

$$\left.\begin{aligned} z_1 &= \mathbf{V}y_1 + x_1 \\ z_2 &= \mathbf{V}y_2 + x_2 \end{aligned}\right\}, \tag{3}$$

where the hidden speaker variables, y_1, y_2, are d-dimensional vectors sampled from a continuous multivariate between-speaker distribution. $y_1 = y_2$, for target trials whereas for non-target trials, y_1 and y_2 are sampled independently. The hidden channels, x_1, x_2 are D-dimensional and sampled from a continuous multivariate within-speaker distribution. Normally $d \le D$, but in our experiments $d = D$. The between- and within-speaker distributions are either normal or heavy-tailed [5], the $d \times D$ matrix $\mathbf{V}$ is a fixed hyper-parameter and z_1, z_2 are observed variables. There are two types of hidden variables: (i) the continuous nuisance variables: x_1, x_2, y_1, y_2 and (ii) the variable of interest to be inferred, i.e., the trial type, which can have the discrete values target (T) or non-target (N) [6].

1) Gender modeling: Let, g_1, g_2 represent the genders of the speakers that produced z_1, z_2 which take the values male (M) or female (F). For a target trial, $g_1 = g_2$, while for a non-target trial they may be different [6].

The generative model needs priors for all hidden variables. The priors for the continuous hidden variables are the within and between speaker distributions mentioned above. In this paper the prior for the trial type is not needed. The priors for the gender labels are trial-type dependent and are defined as [6]:

$$P_M = P(MM|T) \quad P_F = P(FF|T)$$
$$Q_{MM} = P(MM|N) \quad Q_{FF} = P(FF|N)$$
$$Q_{MF} = P(MF|N) \quad Q_{FM} = P(FM|N),$$

where the event $g_1 = M$ and $g_2 = F$ is denoted by MF, $P_M + P_F = 1$, and $Q_{MM} + Q_{MF} + Q_{FF} + Q_{FM} = 1$. Equi-probable priors are used for our case, as in [6]. These priors take values of 0 or 1 in the limiting case of given gender labels.

C. Scoring

For gender-independent scoring of the model we assume to have the following available gender-dependent likelihoods [6]: For targets:

$$P(z_1, z_2|MM, T)$$
$$P(z_1, z_2|FF, T),$$

and for non-targets:

$$P(z_1, z_2|MM, N) = P(z_1|M)P(z_2|M) \tag{4}$$
$$P(z_1, z_2|FF, N) = P(z_1|F)P(z_2|F). \tag{5}$$

The independence assumption in (4) and (5) holds when the model parameters are assumed known at the scoring time [5, 14, 6]. In a more fully Bayesian treatment, the uncertainty in the estimates of the model parameters is taken into account during scoring [15].

If the continuous hidden variables have normal distributions, all these likelihoods can be computed in closed form [14], and, if heavy-tailed distributions are considered then they can be approximated [5]. The above mentioned likelihoods can be expressed by the following likelihood ratios [6]:

$$R_M = \frac{P(z_1, z_2 | MM, T)}{P(z_1 | M) P(z_2 | M)} \tag{6}$$

$$R_F = \frac{P(z_1, z_2 | FF, T)}{P(z_1 | F) P(z_2 | F)} \tag{7}$$

$$G_i = \frac{P(z_i | M)}{P(z_i | F)}, \tag{8}$$

where $\log R_M$ and $\log R_F$ are gender dependent speaker verification scores for the male and female, respectively. $\log G_i$ can be used as a gender discrimination score [6].

The gender-independent likelihood ratio $\bar{R}$ can be obtained by marginalizing over the gender variables as [6]:

$$\begin{aligned} \bar{R} &= \frac{P(z_1, z_2 | T)}{P(z_1, z_2 | N)} \\ &= \frac{P_M P(z_1, z_2 | MM, T) + P_F P(z_1, z_2 | FF, T)}{\sum_{g_1, g_2} Q_{g_1 g_2} P(z_1 | g_1) P(z_2 | g_2)}. \end{aligned} \tag{9}$$

Applying eq. (4) to eq. (7) in eq. (9), $\bar{R}$ can be expressed in terms of the likelihood ratios and priors as [6]:

$$\bar{R} = \frac{P_M}{Q_{MM}} S_M R_M + \frac{P_F}{Q_{FF}} S_F R_F, \tag{10}$$

where

$$S_M = \frac{Q_{MM} G_1 G_2}{Q_{MM} G_1 G_2 + Q_{MF} G_1 + Q_{FM} G_2 + Q_{FF}} \tag{11}$$

$$S_F = \frac{Q_{FF}}{Q_{MM} G_1 G_2 + Q_{MF} G_1 + Q_{FM} G_2 + Q_{FF}}. \tag{12}$$

IV. Experiments

A. *Experimental setup*

We conducted experiments on the *coreext-coreext* condition of the NIST 2010 SRE extended list. To evaluate the performance of our speaker recognition systems we used following evaluation metrics: the Equal Error Rate (EER), the old normalized minimum detection cost function (DCF_{Old}) and the new normalized minimum detection cost function (DCF_{New}). DCF_{Old} and DCF_{New} correspond to the evaluation metric for the NIST SRE in 2008 and 2010, respectively. For our baseline Hamming method, MFCC features are computed from Hamming-windowed spectrum estimates using HCopy of the HTK toolkit. For the Thomson [9], multi-peak [12] and sine weighted cepstrum estimator (SWCE) [11] methods, MFCC features are computed from the multi-taper spectrum estimates as described in section II.

1) Feature Extraction: For our experiments, we use 20 MFCC features (including 0^{th} cepstral coefficient (c0)) augmented with their delta and double delta coefficients (DCC), making 60 dimensional MFCC-DCC feature vectors. For the baseline Hamming method we use log-energy instead of c0. The analysis frame length is 30 ms (for baseline, it is 25 ms) with a frame shift of 10 ms. Baseline system uses pre-emphasis whereas multi-taper systems do not. Delta and double coefficients were calculated using a 2-frame window for baseline and a 1-frame window for the multi-taper systems. Then Silence frames are removed according to the VAD labels. After that we apply Short-time Gaussianization (STG) which uses a 300-frame window. For the baseline system we use HTK-based front-end and for the multi-taper systems MFCC features are extracted using the MATLAB. We chose HTK-based baseline because using the same configuration as the multi-taper systems we also developed a baseline speaker verification system where we calculated MFCC-DCC features from the Hamming windowed spectrum estimates, but the performance of that system was not as good as the HTK-based baseline system.

2) GMM-UBM training: We train a gender-independent, full covariance Universal Background Model (UBM) with 256 component Gaussian Mixture Models (GMMs). NIST SRE 2004 and 2005 telephone data were used for training the UBM for our system.

3) Training and extraction of i-vectors: Our gender-independent i-vector extractor is of dimension 800. After training gender-independent GMM-UBM, we train the i-vector extractor using the Whitened Baum-Welch (WBW) 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 2005 microphone data and NIST SRE 2008 interview development microphone data. In order to reduce the i-vectors dimension, a Linear Discriminant Analysis (LDA) projection matrix is estimated from the WBW statistics by maximizing the following objective function:

$$P_{LDA} = \arg\max_P \frac{\left| P^T \Sigma_b P \right|}{\left| P^T \Sigma_w P \right|},$$

where Σ_b and Σ_w represent the between- and within-class scatter matrices, respectively. For the estimation of Σ_b we use all telephone training data excluding Fisher data and Σ_w is estimated using all telephone and microphone training data excluding Fisher data. An optimal reduced dimension of 150 is determined empirically.

Then we extract 150 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 WBW statistics and then 150 dimensional i-vectors are extracted following the similar procedure using the same projection

matrix. We also normalize the length of the i-vectors, as it has been found that normalizing the length of the i-vectors after mapping by the estimated LDA projection matrix helps Gaussian PLDA model to give the same results as the heavy-tailed PLDA model [13] i.e., PLDA model with heavy-tailed prior distributions [5]. Heavy-tailed PLDA is 2 or 3 times slower than the Gaussian PLDA.

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. For both of the models, the fixed hyper-parameter **V** is a full rank matrix of dimension $d = 150$.

B. *Results*

The effectiveness of the mixture PLDA model has been shown in [6]. In this paper we use mixture PLDA model-based speaker verification system to compare the performance of the following 4 systems:

i) Baseline: For this system, MFCC-DCC features were computed from the Hamming windowed spectrum estimates by using HCopy of the HTK toolkit.

ii) SWCE: Here, MFCC-DCC features are calculated from the multi-taper spectrum estimates with sine tapering [11] and number of tapers used is 6.

iii) Multi-peak: We computed MFCC-DCC features for this system from the multi-taper spectrum estimates with multi-peak tapering [12] and number of tapers used is 6.

iv) Thomson: Here, MFCC-DCC features are calculated from the multi-taper spectrum estimates with dpss tapering [9] and number of tapers used is 6.

The number of tapers was set to $M = 6$ in all verification experiments. This selection is based on the preliminary GMM-UBM results on the NIST 2002 SRE corpus [3].

To evaluate and compare the performance of the above mentioned systems we conducted experiments using telephone speech and microphone speech on the extended core-core condition of NIST SRE 2010 task. Results are reported for five evaluation conditions correspond to det conditions 1-5 (as shown in table I) in the evaluation plan [18].

TABLE I: Evaluation conditions (*coreext-coreext*) 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 II: Male and female det1 to det5 speaker verification results using a mixture PLDA model for the baseline Hamming window system and multi-tapers systems, measured by EER. For each row the best EER is in boldface. Positive relative improvement (RI) indicates reduction in EER whereas negative RI indicates an increase in EER.

		EER (%) / RI (%)			
		Base-line	**SWCE**	**Multi-peak**	**Thomson**
Female	det1	2.5	2.2 / 12.0	**1.8 / 28**	2.3/8.0
	det2	5.1	4.0 / 21.6	**3.8 / 25.5**	4.3/ 15.7
	det4	3.9	**3.4 / 12.8**	3.5 / 10.3	3.8 / 2.6
	det3	3.3	**2.8 / 15.2**	2.9 / 12.1	3.0 / 9.1
	det5	3.4	**2.9 / 14.7**	3.0 / 11.8	3.2 / 5.9
Male	det1	1.6	1.5 / 6.3	**1.2 / 25.0**	1.7 / -6.3
	det2	2.7	**2.4 / 11.1**	2.6 / 3.7	2.9 / -7.4
	det4	2.4	2.3 / 4.2	**2.0 / 16.7**	2.4 / 0.0
	det3	3.2	3.5 / -9.4	**3.1 / 3.1**	3.5 / -9.4
	det5	2.6	**2.4 / 7.7**	2.5 / 3.8	2.7 / -3.8

TABLE III: speaker verification results using a mixture PLDA model for the baseline Hamming window system and multi-tapers systems, measured by normalized minimum DCF (DCF_{Old}). For each row the best DCF_{Old} is in boldface. Positive relative improvement (RI) indicates reduction in DCF_{Old} whereas negative RI indicates an increase in DCF_{Old}.

		DCF_{Old} / RI (%)			
		Base-line	**SWCE**	**Multi-peak**	**Thomson**
Female	det1	0.12	0.12/0.0	**0.09/25.0**	0.12 / 0.0
	det2	0.24	0.20 / 16.7	**0.19/20.8**	0.22/8.3
	det4	0.20	**0.16 / 20.0**	**0.16/20.0**	0.18/10.0
	det3	0.18	0.17 / 5.6	**0.15/16.7**	0.18/0.0
	det5	**0.16**	**0.16 / 0.0**	**0.16 / 0.0**	0.17 / 0.0
Male	det1	0.07	**0.06 / 14.3**	0.07 / 0.0	0.07 / 0.0
	det2	0.14	**0.11 / 21.4**	0.12/14.3	0.14 / 0.0
	det4	0.11	**0.09 / 18.2**	**0.09/18.2**	0.11 / 0.0
	det3	**0.14**	0.15 / -7.1	0.15/-7.1	0.16/-14.3
	det5	**0.13**	**0.13 / 0.0**	0.14/-7.7	0.15/-15.4

Table II presents EERs for all the systems and relative improvements (RI) by the multi-taper systems compared to the baseline system. From Table II it has been observed that the Multi-peak multi-taper system was consistently better than the baseline system. The SWCE system does not perform well only in det3 (male) case. The Thomson system performs well on female data but its performance degrades on male data. Tables III and IV depict the normalized minimum DCF (DCF_{Old}) and normalized minimum DCF (DCF_{New}), respectively, for all the systems considered in this paper. In terms of DCF_{Old} Multi-peak multi-taper method outperform all other methods in female, det1 to det5 case but in the case of male, det1 to det5 (except det3, when baseline give better DCF_{Old}) the performance of the SWCE system is better. Compared to baseline system the Thomson system performs well only in the female, det2 and det4 cases.

In terms of DCF_{New} (as shown in Table IV), SWCE and Multi-peak multi-taper systems outperform the baseline and the Thomson system in most of the cases. The Thomson

system, compared to baseline system, perform well only in det2 and det4 cases (both male and female). So after observing all the results we can come to a conclusion that the multi-taper-based spectrum estimation technique (SWCE and Multi-peak) can be an alternative to the Hamming windowed spectrum estimation technique to compute MFCC–DCC features for speaker verification system.

TABLE IV: speaker verification results using a mixture PLDA model for the baseline Hamming window system and multi-tapers systems, measured by normalized minimum DCF (DCF_{New}). For each row the best DCF_{New} is in boldface. Positive relative improvement (RI) indicates reduction in DCF_{New} whereas negative RI indicates an increase in DCF_{New}.

		DCF_{New} / RI (%)			
		Base-line	**SWCE**	**Multi-peak**	**Thomson**
Female	det1	0.40	0.38/5.0	**0.34/15.0**	0.41/-2.5
	det2	0.67	**0.56/16.4**	**0.56/16.4**	0.61/9.0
	det4	0.57	**0.50/12.3**	**0.50/12.3**	0.56/1.8
	det3	0.56	**0.52/7.1**	0.56/0.0	0.59/-5.3
	det5	**0.47**	0.48/-2.1	0.52/-10.6	0.53/-12.8
Male	det1	**0.25**	0.26/-4.0	0.26/-4.0	**0.25 / 0.0**
	det2	0.49	0.41/16.3	**0.40/18.4**	0.45/8.2
	det4	0.37	0.36/2.7	**0.32/13.5**	0.36/2.7
	det3	0.53	0.51/3.8	**0.49/7.5**	0.55/-3.8
	det5	**0.46**	**0.46 / 0.0**	0.47/-2.2	0.50/-8.7

V. Conclusion

In this paper we used three multi-taper spectrum estimation approaches for low variance MFCC computation and compared their performances, in the context of mixture PLDA models in gender-independent i-vector space for speaker verification, to the conventional single taper technique. Experimental results on the telephone and microphone portion of the NIST 2010 SRE task indicate that multi-taper methods outperform the baseline single taper method. Among the three tapers, the multi-peak and the SWCE outperformed the Thomson which agrees well with the preliminary GMM-UBM results of [3]. The number of tapers was set to 6 according to [3] without doing 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 feature extraction is a viable candidate for replacing the baseline MFCC feature.

VI. Acknowledgements

The work of T. Kinnunen was supported by the Academy of Finland (project no. 132129).

References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, August 1980.

[2] Y. Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Trans. On Speech and Audio Proc.,* vol. 12, No. 1, pp. 59-67, January 2004.

[3] T. Kinnunen, R. Saeidi, J. Sandberg, M. Hansson-Sandsten, "What Else is New Than the Hamming Window? Robust MFCCs for Speaker Recognition via Multitapering", *Proc. Interspeech 2010*, pp. 2734--2737, Makuhari, Japan, Sept. 2010.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *in IEEE Trans. on Audio, Speech and Language Processing,* vol. 19, No. 4, pp. 788-798, May, 2011.

[5] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in Proceedings of the Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic, Jun. 2010.

[6] M. Senoussaoui, P. Kenny, N. Brummer, E. de Villiers, and P. Dumouchel, ''Mixture of PLDA models in I-vector space for gender independent speaker recognition,'' *to appear in the Proceed. of INTERSPEECH 2011,* Florence, Italy, August 2011.

[7] S. M. Kay, *Modern Spectral Estimation*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[8] 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 Processing Letters*, vol. 17, no. 4, pp. 343–346, April 2010.

[9] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proc. of the IEEE*, vol. 70, no. 9, pp. 1055–1096, Sept 1982.

[10] K. S. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation," *IEEE Trans. on Signal Proc.*, vol. 43, no. 1, pp. 188–195, Jan 1995.

[11] M. Hansson-Sandsten and J. Sandberg, "Optimal cepstrum estimation using multiple windows," in *Proc.ICASSP 2009*, pp. 3077–3080, Taipei, Taiwan, April 2009.

[12] M. Hansson and G. Salomonsson, "A multiple window method for estimation of peaked spectra," *IEEE T. on Sign. Proc.*, vol. 45, no. 3, pp. 778–781, Mar. 1997.

[13] D. Garcia-Romero, and Carol Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proceedings of Interspeech*, Florence, Italy, Aug. 2011.

[14] N. Brümmer, and E. de Villiers, "The speaker partitioning problem," in *Proceedings of the Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, June, 2010.

[15] J. Villalba, and N. Brümmer, "Towards fully Bayesian speaker recognition: Integrating out the between speaker covariance," in *Proceedings of Interspeech,* Florence, Italy, Aug, 2011.

[16] F. Harris,"On the use of windows for harmonic analysis with the discrete Fourier transform," Proceedings of the IEEE, vol. 66, no. 1, pp. 51-84, January 1978.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing,* vol. 10, no. 1, pp. 19-41, Jan, 2000.

[18] National Institute of Standards and Technology, NIST Speaker Recognition Evaluation, http://www.itl.nist.gov/iad/mig/tests/sre/.