Low-variance Multitaper Mel-Frequency Cepstral Coefficient Features for Speech and Speaker Recognition Systems

Md Jahangir Alam[1,2], Patrick Kenny[2], Douglas O’Shaughnessy[1],

[1]INRS-EMT, University of Quebec, Montreal, Canada
[2]CRIM, Montreal, Canada
{Janagir.Alam, Patrick.Kenny}@crim.ca, dougo@emt.inrs.ca

Abstract. In this paper we investigate low-variance multi-taper spectrum estimation methods to compute the mel-frequency cepstral coefficient (MFCC) features for robust speech and speaker recognition systems. In speech and speaker recognition, MFCC features are usually computed from a single tapered (e.g., Hamming-window) direct spectrum estimate, i.e., the squared magnitude of the Fourier transform of the observed signal. Compared with the periodogram, a power spectrum estimate that uses a smooth window function, such as Hamming window, can reduce spectral leakage. Windowing may help to reduce spectral bias, but variance often remains high. A multi-taper spectrum estimation method that uses well-selected tapers can gain from the bias-variance trade-off, giving an estimate that has small bias compared with a single taper spectrum estimate but substantially lower variance. Speech recognition and speaker verification experimental results on the AURORA-2 and AURORA-4 corpora and the NIST 2010 speaker recognition evaluation (SRE) corpus (telephone as well as microphone speech), respectively, show that the multi-taper methods perform better compared to the Hamming-windowed spectrum estimation method. In a speaker verification task, compared to the Hamming window technique, the sinusoidal weighted cepstrum estimator (SWCE), Multi-peak, and Thomson multi-taper techniques provide a relative improvement of 20.25%, 18.73% and 12.83%, respectively, in Equal Error Rate (EER).

Keywords: Speech recognition, speaker verification, multitaper spectrum, AURORA-2. NIST 2010 SRE

1 Introduction

Extraction of useful information from speech signal has been a subject of active research for many decades [24-26]. A feature extractor (or front-end), first step in an automatic speech or speaker recognition system, transforms a raw speech signal into a compact representation. Since feature extraction is the first step in the chain, the quality of later steps (modelling and pattern matching/classification) strongly depends on it [30, 33, 48]. The mel-frequency cepstral coefficient (MFCC) features are the most popular in speech and speaker recognition systems, where they demonstrate good performance [1, 24, 38]. The MFCC representation is an approximation of how the human auditory system works [1]. Since MFCC features are computed from an estimated spectrum, the variance of these features is greatly influenced by the variance the spectral estimate of the observed speech signal. Hence, an accurate estimation of the speech power spectrum is crucial [48, 51]. Direct spectrum estimation (also known as nonparametric spectrum estimation) based on a window function, such as Hamming window, is the most often used spectrum estimation method for speech processing applications [2]. Windowing that gives more weight to the center of the signal than to its ends helps to alleviate bias due to leakage in the periodogram but variance remains large [11, 56], and therefore, MFCC features computed from this estimated spectrum have also high variance [3-4]. Hence, the focus of this work is to find spectrum estimators with low variance. One elegant technique for reducing the variance is the multi-taper spectrum estimate [4-6], that uses a variety of orthogonal tapers, some of which give more weight to the ends of the signal, trading off bias and variance.

Multi-taper methods reduce the variance of spectral estimates by using multiple orthogonal window functions rather than a single window. In a multitaper spectrum estimation method the speech signal is, first, multiplied by not one but a family of tapers which are resistance to spectral leakage. This yields several tapered speech signals from one record. Taking the Discrete Fourier Transforms (DFTs) of each of these tapered signal several *eigenspectra* are produced which are combined (using a weighted averaging technique) to form the final multitaper spectral estimate. The multi-taper method has already proved useful in applications, such as geophysical data analysis [5-6, 12-13], and noise reduction [2]. Recently, multi-tapering has been applied in [3, 20] for speaker recognition using a simple Gaussian mixture model-universal background model (GMM-UBM) [21], generalized linear discriminant sequence-support vector machine (GLDS-SVM), joint factor analysis (JFA) [22-23] and i-vectors-Probabilistic Linear Discriminant (PLDA) [30, 51] classifiers with promising preliminary results, but not yet in speech recognition. In this paper, our aim is to compute MFCC features from a multi-taper spectral estimate for robust speech recognition under additive and convolutive noise conditions on the AURORA-2 small vocabulary connected digits task [7] and the AURORA-4 large vocabulary continuous speech recognition (LVCSR) task [60]. Similar to [30, 51], we also wish to validate the findings of [3, 20] using a state-of-the-art i-vector classifier on the telephone and microphone speech of the latest NIST 2010 Speaker Recognition Evaluation (SRE) data. Robustness of the multitaper MFCC features under additive noise conditions has been analysed in [3, 20].

Even though speaker and speech recognition systems use often the same acoustic features (e.g., MFCCs and the perceptual linear prediction coefficients (PLPs) computed from a Hamming windowed direct spectrum estimates), the goals in these two tasks are nearly the opposites of each other. In speaker verification, one is interested in the speaker's voice timbre independent of words being spoken, whereas in speech recognition one is interested in words being spoken independent of the speaker. The modeling techniques in these problems are quite different and consequently, spectrum estimator parameters optimized using a full recognition system in one of these tasks does not automatically guarantee increased recognition accuracy in the other task. Indeed, we have found several front-ends that outperform conventional MFCCs in speech recognition **[**40, 41**]** but not in speaker recognition based on an i-vectors-PLDA classifier. It has also been found in [41] that feature extractors that are robust to additive noise environments perform much worse in microphone speech or under reverberation. The motivation for presenting extensive experimental results (on speaker and speech recognition tasks) for the low-variance multitaper MFCCs is to show how it performs in different applications as well as under various environmental mismatch conditions.

2 Low Variance Multi-taper Spectrum Estimation

A windowed direct spectrum estimator is the most often used power spectrum estimation method for speech processing applications, such as speech and speaker recognition, and speech enhancement. The periodogram [49] was the first (non-parametric) direct spectral estimate of the power spectral density (PSD) function. The periodogram is a biased estimate due to spectral leakage via the sidelobes. It thus becomes necessary to use the method of tapering (windowing) to effectively reduce this bias [11, 50]. For the m-th frame and k-th frequency bin an estimate of the windowed periodogram can be expressed as:

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

where $k\in\{0,1,...,K-1\}$ denotes the frequency bin index, N is the frame length, $s\left(m,j\right)$ is the time domain speech signal and $w\left(j\right)$ 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 $w\left(j\right)$ is a rectangular window, i.e., $w\left(j\right)=1$, then eq. (1) is called the periodogram.

Although windowing reduces the bias (the difference between the estimated spectrum and the actual spectrum) due to leakage, it does not reduce the variance of the spectral estimate [8] and therefore, the variance of the mel frequency cepstral coefficient (MFCC) features computed from this estimated spectrum is also large [27-28]. A multitaper spectrum estimate, as a replacement of the windowed periodogram estimate, can be used to reduce the variance of the MFCC features [4-6]. The multi-taper spectrum estimator, which uses M orthogonal window function rather than a single window, can be expressed as

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

where N is the frame length, w_{p} is the p-th data taper used for the spectral estimate $\hat{S}_{MT}(\cdot)$, which is also called the p-th eigenspectrum, M denotes the number of tapers and $\ddot{e}\left(p\right)$ is the weight corresponding to the p-th taper. The tapers $w_{p}\left(j\right)$ are chosen to be concentrated in the frequency domain so that their broadband bias (bias caused by spectral leakage due to sidelobes) is as small as possible [11, 28, 52]. If the tapers are pairwise orthonormal, i.e.,

$$\sum_j w_p\left(j\right) w_q\left(j\right)=\delta_{pq}=\begin{cases}1, & \text{for p = q}\\ 0, & \text{for p} \neq \text{q}\end{cases},$$

then their corresponding spectrum estimates are approximately uncorrelated [52]. The multi-taper spectrum estimate is therefore obtained as the weighted average of M individual sub-spectra. The windowed periodogram can be obtained as a special case of eq. (2) when $M=1$ and ë$\left(p\right)=1$. Averaging the M uncorrelated spectra gives a low variance spectrum estimate and, consequently, low variance MFCC estimate as well.

Fig. 1 (a & b) and 2 (a & b) present a comparison of single window (e.g., Hamming) and multi-taper (e.g., Thomson [5]) with four tapers in the time and frequency domains, respectively. Unlike conventional data tapers, the orthogonal tapers used in producing multitaper spectrum estimates are uncorrelated. It is observed from fig. 2 (a) and fig. 3 that only the first taper from the DPSS tapers has the traditional window shape such as Hanning window (see fig. 3), and gives more weight to the center of the signal than to its end. Tapers for $p>1$ give increasingly more weight to the ends of the signal. Fig. 2 (c) presents the taper energy (sum of the squared tapers normalized by the number of tapers M) as a function of j (sample index) for different number of tapers (M = 1-4).

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 and the other tapers produce spectral peaks that are shifted slightly up and down in frequency (see fig. 2 (b)). Each of the spectra contributes to an overall spectral envelope for each component.

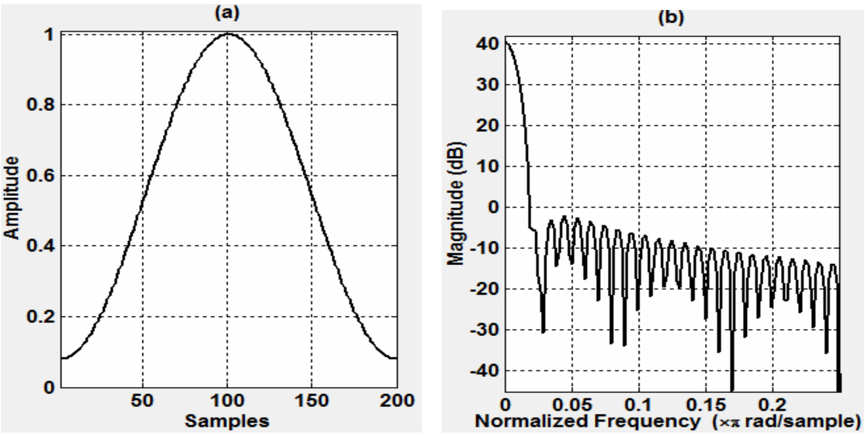
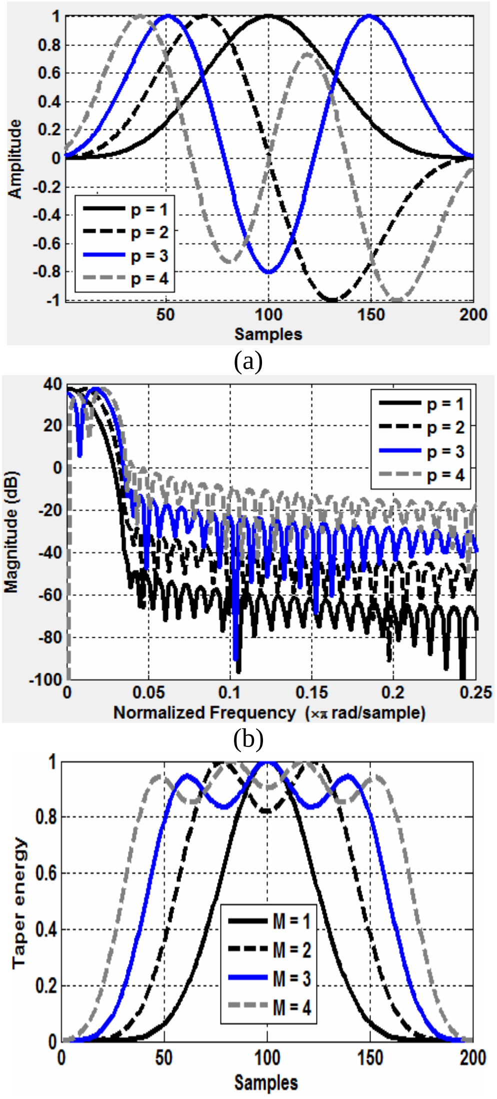


Fig. 1. Time and frequency domain plots of a Hamming window (a) Time domain, (b) frequency domain. Hamming window gives more weight to the center of the signal than to its ends and provides a relative side-lobe attenuation of 42.7 dB.



(a)

(b)

(c)

Fig. 2. Time and frequency domain plots of the first four Slepian tapers used in the Thomson multitaper method, (a) time domain, (b) frequency domain. (c) The taper energy as a function of j (sample index) for different numbers of tapers (M = 1-4).

The use of multiple orthogonal windows can have significant advantages over the use of any single window [11, 12, 13, 28]. In particular, the energy of a single band-limited window always non-uniformly covers the desired concentration region, which results in some data being statistically over- or under-represented when forming the spectral estimate. 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 somewhat uncorrelated, a multi-taper average of these possesses a smaller estimation variance [53]. The windowed periodogram is an inconsistent estimator because its variance does not decrease with increasing N (window size in samples). The multitaper estimator, in turn, has smaller variance than the windowed periodogram estimator and is a consistent estimator provided the number of tapers M increases with N [29].

Various tapers have been proposed in the literature for spectrum estimation. A good set of M orthonormal data tapers with good leakage properties is given by the Slepian sequences (also called discrete prolate spheroidal sequences (dpss)), which are a function of a prescribed mainlobe width [2, 5]. Another family of tapers is the *sine* tapers, which are very easy to compute and are pairwise orthogonal, and can be given by [6]:

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

The multiplicative constant makes the tapers orthonormal (i.e., have unit norm and are orthogonal). The *sine* tapers are applied with optimal weighting for cepstrum analysis (called Sinusoidal Weighted Cepstrum Estimator (SWCE)) in [9] and in [10] the multi-peak tapers are designed for peaked spectra. A brief description of all the three tapers can also be found in [51]. In this paper, we use the Thomson multitaper [5], the SWCE [9], and the Multi-peak multitaper spectrum estimator to compute the low variance MFCC features for speaker and speech recognition systems.

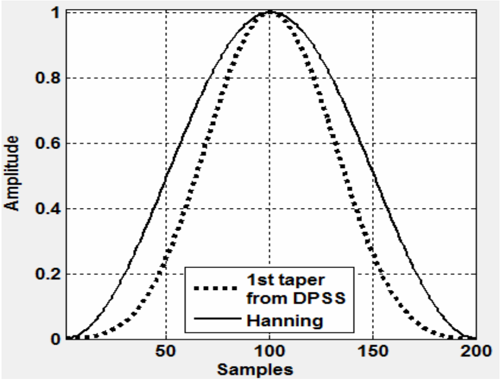


Fig. 3. A comparison of the shape of the conventional Hanning window and the first taper from the Slepian (or DPSS) tapers used in Thomson multitaper method. The 1st DPSS taper has been normalized to have a maximum height of unity for comparison purposes**.**

3 Feature Extraction

For a speaker/speech recognition system, as for any classification task, feature extraction is necessary to extract relevant information that is both good at discriminating different speaker/speech classes and insensitive to factors irrelevant to speaker/speech recognition. Currently, the most widely used speech features both in speaker and speech recognition systems are the Mel-frequency cepstral coefficient (MFCC) [1] and perceptual linear predictive (PLP) features [39]. In this paper, we describe only the MFCC feature extraction process.

Fig. 4 shows the generalized block diagram of the mel frequency cepstral coefficients (MFCCs) feature extraction from the single and multi-taper spectrum estimates. As we mentioned above, the single tapered (e.g., Hamming-windowed) direct 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 $M=1$, ë$(1)=1$, and $w_1(j)$is the Hamming window.

A detailed description of the various steps of the MFCC feature computation framework can be found in [25, 31-32, 33-34, 38, 59]. In this paper, for completeness, we briefly describe the various steps of the MFCC feature extraction process.

The MFCC feature extraction process begins with the pre-processing of the speech signal. In this step, the DC offset of the speech signal is removed and the signal spectrum is pre-emphasized by approximately 20 dB per decade to flatten the spectrum of the speech signal. The pre-emphasis filter, a first order finite impulse response (FIR) filter with a transfer function of the form $H(z) = 1\text{-}0.97z^{-1}$, is used to offset the negative spectral slope of the voiced speech signal to improve the efficiency of the spectral analysis [24-25, 34]. The speech signal is then decomposed into a series of overlapping (to ensure better temporal continuity in the transform domain) small duration segments called frames. Commonly used frame length and frame shift are 20-30 ms and 10 ms respectively for a speech/speaker recognition task because the positions of the articulators do not change much in the period of a frame length.

After being partitioned into frames, each frame is multiplied by a single window (such as Hamming window) or multiple window (such as Thomson multitaper) function prior to the spectral analysis to reduce the effect of discontinuity introduced by the framing process. The power spectrum of the speech signal is estimated using the single taper (e.g., Hamming window) or multitaper direct spectral estimator, i.e., by computing the squared magnitude of the discrete Fourier transform (DFT) of a segment of the speech signal.

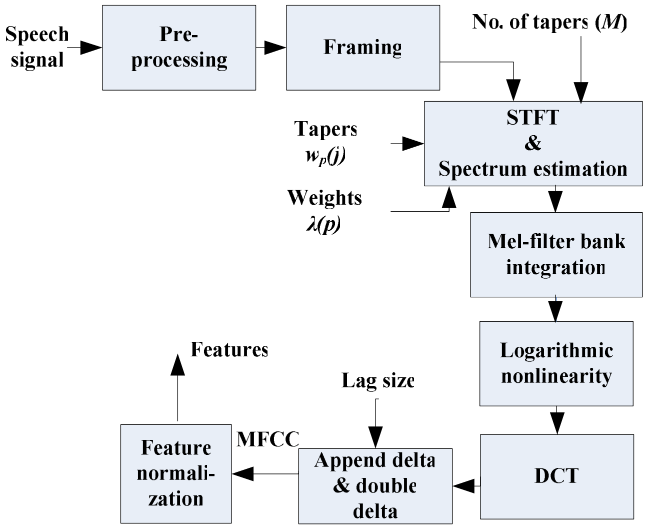


Fig. 4. Generalized block diagram for the single and multi-taper spectrum estimation-based Mel-frequency cepstral coefficient (MFCC) features extraction. For the computation of MFCC features from single-taper (or window) spectrum estimates, we use $M=1$, ë$(1)=1$, and $w_1(j)$is the Hamming window.

The spectrum of the speech signal is then filtered by a group of triangle bandpass filters that simulate the characteristics of a human's ear. These windows are called the Mel windows and the filtering process is called Mel filtering. The Mel filtering is to model the human auditory system that perceives sound in a nonlinear frequency binning [35]. While the Mel filtering approximates the nonlinear characteristics of the human auditory system in frequency, the natural logarithmic nonlinearity (or power function nonlinearity [39-41]) deals with the loudness nonlinearity. It approximates the relationship between a human's perception of loudness and the sound intensity [36]. The DCT is applied on the log Mel filterbank coefficients to generate the cepstral coefficients [37]. Only the lower order coefficients (usually the first 12 or 13 coefficients for speech recognition and the first 19 or 20 coefficients for speaker recognition) are used for speech/speaker recognition, hence a dimension reduction is achieved. Another benefit of the DCT is that the generated cepstral coefficients are less correlated, which is beneficial for diagonal covariance matrix modeling. In addition to static cepstral coefficients, the delta and double delta coefficients are used both in speaker and speech recognition systems to model the temporal dynamics in the speech signal. The 1st order dynamic coefficients, i.e., delta coefficients, are calculated from the static MFCCs $c(m,\tau)$ as

$$\Delta c\left(m,\tau\right)=\frac{\sum_{q=1}^{Q} q\left(c(m+q,\tau)-c(m-q,\tau)\right)}{2\sum_{q=1}^{Q} q^{2}},$$

where m is the frame index, τ is the cepstral index, Q represents the window lag size, and $c\left(m,\tau\right)$ is the τ-th cepstral coefficient of the m-th frame. The 2nd order dynamic coefficients, i.e., double delta coefficients, $\Delta\Delta c\left(m,\tau\right)$ can be calculated from $\Delta c\left(m,\tau\right)$ following the same procedure.

After appending delta and double delta features with the static cepstral coefficients, the final features (39-dimensional for speech recognition [1, 24-26] and 60-dimensional speaker verification [42-44]) are then normalized using a feature normalization method. Unlike speech enhancement and feature compensation methods that aim to recover the clean speech coefficients, the feature normalization method usually modifies certain statistics of features, e.g., means and variances, scale, statistical distribution to some reference values or distributions [24-25, 45, 46-47, 54-55]. A reason for doing so is that the statistics of speech features are changed when the speech signal is distorted by noise. Features in a speech recognition task are usually normalized over the whole utterance using the cepstral mean subtraction (CMS) or using the mean and variance normalization method (MVN) [54-55]. In speaker recognition, the 60-dimensional feature vector is usually subject to a short-time feature normalization technique using a sliding window of 3 s [42-45].

4 Speech & Speaker Recognition Frameworks

4.1 Speech Recognition Framework

For the recognition task we use the Hidden Markov Model Toolkit (HTK) speech recognizer [31]. In the experiments for the connected digits task we use a simple Hidden Markov Model (HMM)-based system with 16 states per word model, 3 Gaussian components per state. Most speech recognition systems use statistical models, such as hidden Markov models (HMMs), to represent the basic speech patterns (generally referred to as acoustic units) required by the recognizer [57, 58]. For the LVCSR task on the AURORA-4 corpus, all experiments employed state-tied crossword speaker-independent triphone acoustic models with 4 Gaussian mixtures per state. A single-pass Viterbi beam search-based decoder was used along with a standard 5K lexicon and bigram language model with a prune width of 250 [60, 61].

For our experiments, we use 13 Mel-frequency cepstral coefficient (MFCC) features (including log energy) augmented with their delta and double delta coefficients, making 39-dimensional MFCC feature vectors. The analysis frame length is 25 ms with a frame shift of 10 ms. The delta and double features were calculated using a 3-frame and 2-frame lag window, respectively. For the baseline (i.e., Hamming) and multi-taper methods (Thomson, Multi-peak and SWCE), the MFCC features are normalized using the conventional mean and variance (MVN) [54-55] normalization technique over the whole utterance.

4.1.1 Speech Recognition Corpus

The AURORA-2 small vocabulary connected digits (English) [7] and the AURORA-4 large vocabulary continuous speech recognition (LVCSR) [60] corpora are used for the speech recognition task. In the AURORA-2 corpus, there are two training sets (clean training set and multi-condition training set) and three test sets (test sets A, B and C). The clean training set consists of 8440 clean speech recordings only from 55 male and 55 female adult speakers [7]. The multi-condition training consists of 8440 utterances of both clean and noisy speech split into 20 subsets. The 20 subsets represent 4 different noise scenarios (subway, babble, car and exhibition hall) at 5 different signal-to-noise ratios (SNRs) (20, 15, 10, 5, and 0 dB). Test set A is composed of 28028 speech recordings with conditions matched to the multi-condition training set, test set B is composed of 28028 speech recordings with non-matched background noise (restaurant, street, airport and train-station) and test set C is composed of 14014 speech utterances with partly matched background noise and non-matched convolutional noise (MIRS (modified intermediate reference system) filtered subway and street noise). The speech recordings in the three test sets are uttered by 52 female and 52 male speakers. The clean training set constitutes mismatched training/testing conditions whereas the multi-condition training set constitutes much more matched training/testing conditions [7].

The AURORA-4 LVCSR corpus is divided into 3 sets, namely, training, development (dev test) and evaluation (eval or test) sets. This task is often referred to as the 5k closed vocabulary task, i.e., there are no out of vocabulary words (OOVs) in the evaluation set. The training set contains 7138 utterances from 83 speakers, totaling 14 hours of speech data. 14 evaluation sets were defined in order to study the degradations in speech recognition performance due to microphone conditions, filtering and noisy environments. Each of the filtered versions of the evaluation set recorded with Sennheiser microphone and secondary microphone was selected to form the two eval sets. The remaining 12 subsets were defined by randomly adding each of the 6 noise types (car, babble, restaurant, street traffic, airport, and train-station noises) at randomly chosen SNR between 5 and 15 dB for each of the microphone types as mentioned above. The goal was to have an equal distribution of each of the 6 noise types and the SNR with an average SNR of 10 dB [60]. Each of the test sets contains 166 utterances from 8 speakers, totaling 20.69 minutes of speech data. The 14 test sets are grouped into the following 4 families [60, 61]:

1. Test set A: clean speech in training and test, same channel (set 1)
2. Test set B: clean speech in training and noisy speech in test, same channel (sets 2-7)
3. Test set C: clean speech in training and test, different channel (set 8)
4. Test set D: clean speech in training and noisy speech in test, different channel (sets 9-14).

The number inside the brackets represents the test set number defined in the AURORA-4 corpus.

4.2 Speaker Verification Framework

The i-vector framework for speaker verification has set a new performance standard in the research field. The i-vector extractor converts an entire speech recording into low-dimensional feature vectors called i-vectors [15-17]. The i-vector speaker verification framework used in this paper consists of the following stages: i-vector extraction, generative modeling of i-vectors, and scoring or likelihood ratio computation as described in [14]. A detailed description of the speaker verification framework used in this paper can be found in [14, 30].

4.2.1 Speaker Verification Corpus

We performed experiments on the NIST 2010 speaker recognition evaluation (SRE) corpus [19]. Our focus is on the five evaluation conditions (microphone and telephone speech, det1 to det5) [19] of both female and male (of different ages) parts of the *extended core-core* task. All of the speech recordings in the NIST 2010 SRE task are in English, though English was not the first language of some of the speakers [19]. Both the training and test utterances used English conversational speech recorded over multiple recording sessions: speech recorded, over ordinary (wired or wireless) telephone channels, over a room microphone channel, over a room microphone channel from a interview scenario. The training data consist of the LDC release of Switchboard II - phase 2 and phase 3, Switchboard Cellular - part 1 and part 2, Fisher English data, NIST SRE 2004 and 2005 telephone data, NIST SRE 2005 and 2006 microphone data and NIST SRE 2008 interview development microphone data giving 57620 recordings (31874 from female and 25746 from male speakers). The test data for the NIST 2010 task comprises of 21586 recordings (female utterances = 11845, male utterances = 9741). All the training data are used for training the i-vector extractor. LDA transformation matrix and PLDA models are trained on the same data as the i-vector extractor, except for the Fisher English data that were excluded.

4.2.2 Feature Extraction and UBM training

For our experiments, we use 20 MFCC features (including the 0th cepstral coefficient) augmented with their delta and double delta coefficients, making 60-dimensional MFCC feature vectors. The analysis frame length is 30 ms with a frame shift of 10 ms. Delta and double coefficients are calculated using a 2- and 1-frame lag window, respectively. Then silence frames are removed according to the VAD labels extracted using the algorithms described in [42-43]. We apply a short-time Gaussianization (STG) technique [46-47] to normalize the feature vectors using a 300-frame window. 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.

4.2.3 Training and extraction of i-vectors

Our gender-independent i-vector extractor is of dimension 800. After training the gender-independent GMM-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 English data, NIST SRE 2004 and 2005 telephone data, NIST SRE 2005 and 2006 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 BW statistics by maximizing the following objective function:

$$\mathrm{P}_{\mathrm{LDA}} = \underset{\mathrm{P}}{\operatorname{argmax}} \frac{\left|\mathrm{P}^{\mathrm{T}}\Sigma_{\mathrm{b}}\mathrm{P}\right|}{\left|\mathrm{P}^{\mathrm{T}}\Sigma_{\mathrm{w}}\mathrm{P}\right|},$$

where P represents the LDA transformation matrix, 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 BW 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 to gaussianize the i-vectors distribution so that we can use a Gaussian PLDA model instead of a heavy-tailed PLDA model [18], i.e., PLDA model with heavy-tailed prior distributions [16].

4.2.4 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 [14]. The PLDA models are trained on all the training data except the Fisher data, resulting in 1686 female speakers in 720 hours of speech and 1294 male speakers 540 hours of speech.

5 Results & Discussion

5.1 Speech Recognition Results

We use word accuracy (%) as a performance evaluation measure for comparing the recognition performances of the multi-taper spectrum estimation methods to that of the single taper technique. The baseline (i.e., single taper) and multi-taper systems considered here for performance evaluation are shown in table 5.

The number of tapers for the multi-taper method is chosen according to $\mathrm{M}{=}2^{\mathrm{tbp}\text{-}1}$ where tbp represents time-bandwidth product. A usual range for tbp is 3 to 5 [51]. In order find an optimal number of tapers for the multi-taper methods, we extracted multi-taper MFCC features for the AURORA-2 connected digits task with tapers starting from 4 to 10, and we have found experimentally that MFCC features extracted from multi-taper spectrum estimators with M = 6 give better speech recognition performance in terms of word accuracy (%).

Table 1. Average (0-20 dB) word accuracy as percentage for test sets (a) A, (b) B, and (c) C in clean training condition on the AURORA-2 corpus. The higher the word accuracy means the performance of the system is better. For each column the best result is in boldface. Experimental setup: 39-dimensional MFCC features (MFCC_E_D_A (13-dimensional static MFCCs (including the log energy)+ delta coefficients + double delta coefficients)), 16 states HMM per word model, 3 Gaussian components per state.

(a)

	Word accuracy (%)				
	Subway	Babble	Car	Exhibition	Average
Hamming	63.77	66.85	63.23	**63.95**	64.45
SWCE	64.70	68.87	64.77	63.63	65.49
Multi-peak	**65.32**	69.26	**65.08**	63.94	**65.90**
Thomson	64.16	**69.34**	64.99	63.10	65.40

(b)

	Word accuracy (%)				
	Restaurant	Street	Airport	Train-station	Average
Hamming	68.88	65.64	69.78	65.13	67.36
SWCE	70.23	66.70	70.89	66.56	68.59
Multi-peak	**70.78**	**67.17**	**71.35**	66.79	**69.02**
Thomson	69.83	66.72	70.67	**67.08**	68.58

(c)

	Word accuracy (%)		
	Subway (MIRS)	Street(MIRS)	Average
Hamming	58.11	60.97	59.54
SWCE	58.57	**62.45**	60.51
Multi-peak	**59.66**	62.33	**60.99**
Thomson	58.35	62.22	60.28

Tables 1 (a-c) present the average word accuracy (averaged over 0-20 dB SNRs) for test sets A, B and C on the AURORA-2 connected digits task, respectively, in clean training conditions. Tables 2 (a-c) present the average word accuracy (averaged over 0-20 dB SNRs) for test sets A, B and C on the AURORA-2 connected digits task, respectively, in multi-condition training. Multi-taper methods perform better than the baseline single taper technique in all the cases except one. In multi-condition training and for test set B, a Hamming windowed spectrum estimation method provides better word accuracy than the multi-taper methods.

Table 3 presents the word accuracy (in %) for test sets A, B, C, and D on the AURORA-4 large vocabulary continuous speech recognition (LVCSR) task. It is observed from table 3 that the multitaper methods performed better than the baseline Hamming system in the case of test sets A and C, i.e., under clean and channel mismatch conditions. Under additive noise distortions, i.e., in the case of test sets B and D, the Thomson multitaper method provides better word accuracy than all other methods considered in this work. In the LVCSR task the Thomson method performed the best. For this LVCSR task, the performances of the SWCE and Multipeak multitaper methods were slightly worse than that of the baseline system under additive noise conditions. The possible reason could be that we did not optimize the parameters of the multitaper methods, e.g., the number of tapers, separately for LVCSR task on a development test set. The optimal number of tapers that has been obtained using a development test set for AURORA-2 small vocabulary connected digits task was also applied for the LVCSR task.

Table 2. Average (0-20 dB) word accuracy as percentage for test sets (a) A, (b) B, and (c) C, respectively, in multi-condition training on the AURORA-2 corpus. The higher the word accuracy indices better performance. For each column the best result is in boldface. Experimental setup: 39-dimensional MFCC features (MFCC_E_D_A (13-dimensional static MFCCs (including the log energy) + delta coefficients + double delta coefficients)), 16 states HMM per word model, 3 Gaussian components per state.

(a)

Word accuracy (%)					
	Subway	Babble	Car	Exhibition	Average
Hamming	85.69	88.58	90.69	**88.90**	88.46
SWCE	86.23	89.20	90.77	88.67	**88.72**
Multi-peak	85.85	**89.31**	**90.81**	88.73	88.68
Thomson	**87.21**	88.83	90.58	87.62	88.56

(b)

Word accuracy (%)					
	Restaurant	Street	Airport	Train-station	Average
Hamming	**88.37**	**88.49**	**90.86**	**89.00**	**89.18**
SWCE	87.74	87.98	90.22	89.18	88.78
Multi-peak	88.14	88.05	90.53	89.27	89.00
Thomson	87.13	87.66	89.98	88.75	88.38

(c)

Word accuracy (%)			
	Subway (MIRS)	Street(MIRS)	Average
Hamming	84.60	86.96	85.78
SWCE	85.53	86.98	86.26
Multi-peak	85.16	**87.13**	86.14
Thomson	**86.25**	86.72	**86.49**

Table 3. Average word accuracy as a percentage for test sets A, B, C, and D on the AURORA-4 LVCSR corpus. The higher the word accuracy indices better performance. For each column the best result is in boldface. Experimental setup: 39-dimensional MFCC features (MFCC_E_D_A (13-dimensional static MFCCs (including the log energy) + delta coefficients + double delta coefficients)); all experiments employed state-tied cross-word speaker independent triphone HMM models with 4 Gaussian mixtures per state.

	Word accuracy (%)			
	A	B	C	D
Hamming	88.80	52.2	75.87	39.76
SWCE	88.91	51.07	75.91	39.09
Multi-peak	88.79	51.67	**79.71**	39.66
Thomson	**89.21**	**53.49**	75.86	**43.27**

5.2 Speaker Verification Results

We conducted speaker verification experiments on the *extended core-core* condition of the NIST 2010 SRE extended list. For the performance evaluation of the single-taper (e.g., Hamming window) and multi-taper spectrum estimation-based speaker verification systems (see table 5) we used three evaluation metrics: the traditional Equal Error Rate (EER), which constrains false alarm (when a non-target is accepted as a target) and miss (when a target is rejected) error to be the same, the old normalized minimum detection cost function (minDCF_old), which weighs false alarm errors as ten times as costly as miss errors, and the new normalized minimum detection cost function (minDCF_new), which weighs false alarm errors as 1000 times more costly than miss errors. minDCF_old and minDCF_new correspond to the evaluation metric for the NIST SRE in 2008 and 2010, respectively [19].

Results are reported for five evaluation conditions corresponding to det conditions 1-5 (as shown in table 4) in the evaluation plan [19]. The target (the specified speaker is speaking in the test segment) and non-target (or Impostor) trials (the specified speaker is not speaking in the test segment) of each det condition are presented in table 4. Table 6 presents (a) EERs, (b) minDCF_old, and (c) minDCF_new, respectively, for the baseline and multi-taper systems both for the female and male trials. In terms of the EER, and minDCF_old, minDCF_new multi-taper systems perform better than the baseline system. Average relative improvements (female-male, det1-det5) obtained by the sinusoidal weighted cepstrum estimator (SWCE) system are 20.25%, 17.87%, and 10.85% in EER, minDCf_old and minDCF_new, respectively. The Multi-peak and Thomson multitaper systems provide relative improvements over the baseline of EER = 18.73%, minDCF_old = 15.26%, minDCF_new = 9.63% and EER = 12.83%, minDCF_old = 11.05%, minDCF_new = 5.01%, respectively. The relative improvements obtained by the multitaper systems over the baseline are presented in fig. 5.

Table 4: Evaluation conditions (*extended core-core*) for the NIST 2010 SRE task and the target and non-target (impostor) trials to the corresponding det conditions.

Condition	**Task**	**Target trials (Female/male)**	**Non-target (Impostor) trials (Female/male)**
det1	Interview in training and test, same Mic.	2326 / 1978	449138 / 346857
det2	Interview in training and test, different Mic.	8152 / 6932	1573948 / 1215586
det3	Interview in training and normal vocal effort phone call over Tel channel in test.	2031 / 1958	334438 / 303412
det4	Interview in training and normal vocal effort phone call over Mic channel in test	1886 / 1751	392467 / 364308
det5	Normal vocal effort phone call in training and test, different Tel	3704 / 3465	233077 / 175873

Table 5: Single-taper and multi-taper MFCC feature-based speaker & speech recognition systems.

System	**Description**
Baseline	MFCC features are computed from the Hamming windowed direct spectrum estimate.
SWCE	MFCC features are computed from the sinusoidal weighted (i.e., *sine* tapered) spectrum estimate [9].
Multi-peak	MFCC features are computed from the multi-taper spectrum estimate using multi-peak tapering [10].
Thomson	MFCC features are calculated from the multi-taper spectrum estimates with dpss tapering [5].

The underlying details of a multi-taper spectrum estimator are somewhat similar to averaging the spectra from a variety of conventional tapers, e.g., Hamming, Hann, and Blackman windows. In this case, there will be strong redundancy as the different tapers are highly correlated as they have a similar time-domain shape. Unlike these conventional tapers, the M orthogonal tapers used in a multitaper spectrum estimator provide M statistically independent (hence uncorrelated) estimates of the underlying spectrum.

Table 6. Speaker verification results (female and male, det1 to det5) for the baseline and multitaper systems in terms of the evaluation metrics: (a) EER, (b) minDCF$_{old}$, and (c) minDCF$_{new}$, respectively. The lower the EER, minDCF$_{old}$, or minDCF$_{new}$ the better is the performance of the system. For each column the best result is in boldface. Experimental parameters: 60-dimensional MFCC features (MFCC_0_D_A (20-dimensional static MFCCs (including the 0th cepstral coefficient) + delta coefficients + double delta coefficients)), 256-mixture component UBM, 800-dimensional i-vector extractor, LDA reduced dimension is 150.

(a)

		EER (%)			
		Baseline	**SWCE**	**Multi-peak**	**Thomson**
Female	det1	2.40	**1.80**	2.10	2.06
	det2	4.60	**3.80**	3.90	4.36
	det4	3.90	3.50	**3.40**	3.66
	det3	3.60	2.90	3.00	**2.86**
	det5	4.00	**3.00**	3.30	3.43
Male	det1	1.50	**1.20**	1.30	1.57
	det2	3.10	2.60	**2.30**	2.73
	det4	2.60	**2.00**	2.20	2.17
	det3	4.10	3.10	**3.00**	3.25
	det5	3.20	2.50	**2.20**	2.45

(b)

		minDCF_old			
		Baseline	**SWCE**	**Multi-peak**	**Thomson**
Female	det1	0.11	**0.09**	0.10	0.10
	det2	0.22	**0.19**	**0.19**	0.21
	det4	0.19	**0.16**	**0.16**	0.17
	det3	0.19	**0.15**	0.16	0.17
	det5	0.20	**0.16**	**0.16**	0.17
Male	det1	0.08	**0.07**	**0.07**	**0.07**
	det2	0.14	**0.12**	**0.12**	0.13
	det4	0.11	**0.09**	0.10	0.10
	det3	0.17	**0.15**	**0.15**	0.16
	det5	0.18	0.14	**0.13**	0.14

(c)

		minDCF_new			
		Baseline	**SWCE**	**Multi-peak**	**Thomson**
Female	det1	0.38	**0.34**	0.35	0.37
	det2	0.60	**0.56**	**0.56**	0.59
	det4	0.54	0.50	**0.49**	0.51
	det3	**0.55**	0.56	**0.55**	0.61
	det5	0.57	0.52	**0.49**	0.50
Male	det1	0.30	**0.26**	0.28	0.27
	det2	0.47	**0.40**	0.42	0.46
	det4	0.38	**0.32**	0.35	0.38
	det3	0.59	**0.49**	0.52	0.56
	det5	0.56	0.47	**0.44**	0.46

The average of the M individual spectral estimates $\hat{\mathrm{S}}_{\mathrm{MT}}(\mathrm{m,k})$ then has smaller variance than the single tapered direct spectrum estimates $\hat{\mathrm{S}}_{\mathrm{d}}(\mathrm{m,k})$ by a factor that approaches $1/\mathrm{M}$ [27-28, 56], i.e.,

$$\mathrm{var}\left(\hat{\mathrm{S}}_{\mathrm{MT}}(\mathrm{m,k})\right) \approx \frac{1}{\mathrm{M}} \mathrm{var}\left(\hat{\mathrm{S}}_{\mathrm{d}}(\mathrm{m,k})\right).$$

Fig. 6 demonstrates a reduction in the variance by the multitaper method compared to the Hamming windowed periodogram estimate.

The advantages in multitaper spectrum estimators are that no Fourier resolution is sacrificed and there is no loss of information at the extremes of the data. While information near the bounds is indeed lost with the first taper, it is included and indeed emphasized in the subsequent tapers (see fig. 2). Also multi-taper methods result in a reduction of the variance of the spectral estimate (hence MFCC feature) by weighted averaging the individual estimate of each taper.

Determination of the suitable tapers $\mathrm{w}_{\mathrm{p}}(\mathrm{j})$ (p is the taper index and j is the time or sample index) used in a multitaper method (Thomson, Multi-peak or SWCE multitaper method) and the weights $\lambda(\mathrm{p})$ corresponding to these tapers does not increase the system complexity because they can be pre-computed. Since in a multitaper method the final spectrum estimate is obtained by taking weighted average of M individual spectral estimates, compared to the classical single tapered direct spectral estimate, multi-tapering results in a small increase in processing time as shown in table 7. The execution time is calculated using MATLAB on a Intel (R) core (TM) i7 CPU having clock speed 2.93 GHz. The benefit of using the SWCE multitaper method over the other two multitaper methods is that it has a closed form mathematical expression for the computation of the tapers and the weights corresponding to these tapers [9, 51].

Table 7. Execution time of the classical Hamming windowed direct spectrum estimator and multi-tapered direct spectrum estimator for a speech signal having an average duration of approximately 300 seconds. Execution time reported in this table is obtained by averaging the execution times obtained for 100 speech signals of different durations taken from the NIST SRE 2006 microphone speech data.

Average duration of the speech signal	**Windowed periodogram estimator**	**Multitaper estimator**
300 seconds (5 minutes)	0.85 seconds	1.89 seconds

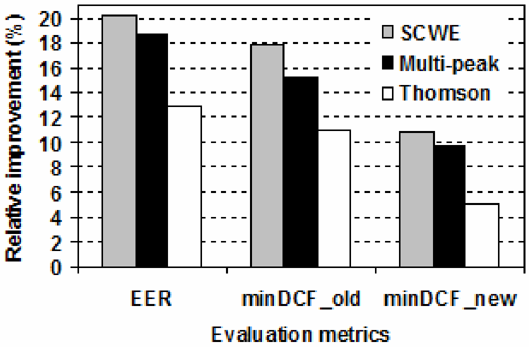


Fig. 5. Relative improvements (in %, averaged over the all det condition and over female-male trails) obtained by the multitaper methods over the baseline Hamming windowed direct spectrum estimation method in terms of the EER, $\mathrm{DCF}_{\mathrm{old}}$, and $\mathrm{DCF}_{\mathrm{new}}$. The higher the relative improvement the better the system is. The relative improvement is the ratio of the difference in performances between the baseline and multitaper system to the performance of the baseline system.

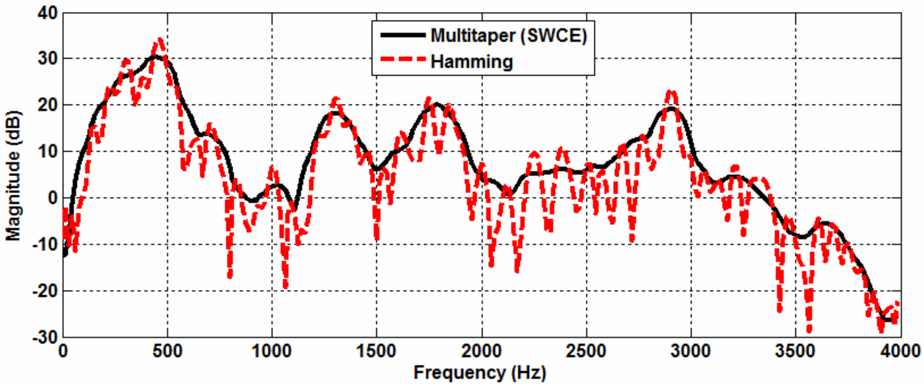


Fig. 6. Single tapered (e.g., Hamming) spectrum estimate and multi-tapered (with six tapers) spectrum estimate demonstrating large variance reduction. Sampling frequency of the speech signal is 8 kHz, analysis frame length = 25 msec with a frame shift of 10 msec. The variance of an estimator measures how much variability an estimator has around its mean (i.e., expected) value.

6 Conclusion

In multi-taper spectrum estimation, data are more evenly weighted and it has a reduced variance when compared to single-taper (e.g., Hamming window) estimates. In this paper we have used three multi-taper spectrum estimation approaches for low-variance Mel-frequency cepstral coefficient (MFCC) features computation and compared their performances with the single window technique, in the context of speech recognition and speaker verification. Experimental speech recognition and speaker verification results on the AURORA-2 small vocabulary task, AURORA-4 large vocabulary task, and NIST 2010 speaker recognition evaluation (SRE) corpora, respectively, depict that an improvement in recognition performance can be obtained by applying multi-tapering, with only a small computational load and processing time increase, compared to the overall recognition process. The relative improvements obtained in speech recognition and speaker verification using multi-taper MFCC features are very encouraging. Therefore, multi-taper methods can be an alternative to the conventional single window technique for the estimation of low variance MFCC features for speech and speaker recognition.

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(4), pp. 357–366 (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(1), pp. 59-67 (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. Interspeech, Makuhari, Japan, pp. 2734-2737 (2010).
4. J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P. Flandrin, P. Borgnat: Multitaper estimation of frequency-warped cepstra with application to speaker verification. IEEE Signal Processing Letters, vol. 17(4), pp. 343–346 (2010).
5. D. J. Thomson: Spectrum estimation and harmonic analysis. IEEE proceeding, vol. 70(9), pp. 1055–1096 (1982).
6. K. S. Riedel and A. Sidorenko: Minimum bias multiple taper spectral estimation. IEEE Trans. on Signal Proc., vol. 43(1), pp. 188–195 (1995).
7. H. G. Hirsch and D. Pearce: The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condition. ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium, France (2000). online: http://aurora.hsnr.de/aurora-2/publications.html
8. S. M. Kay: Modern Spectral Estimation. Englewood Cliffs, NJ: Prentice-Hall (1988).
9. M. Hansson-Sandsten and J. Sandberg: Optimal cepstrum estimation using multiple windows. IEEE ICASSP, Taipei, Taiwan, pp. 3077–3080 (2009).
10. M. Hansson and G. Salomonsson: A multiple window method for estimation of peaked spectra. IEEE Trans. on Sign. Proc., vol. 45(3), pp. 778–781 (1997).
11. Percival, D. B. and Walden, A. T.: Spectral Analysis for Physical Applications, Multitaper and Conventional Univariate Techniques: Cambridge University Press (1993).
12. Prieto, G. A., R. L. Parker, D. J. Thomson, F. L. Vernon, and R. L. Graham: Reducing the bias of multitaper spectrum estimates. Geophys. J. Int., vol. 171, pp. 1269-1281 (2007).
13. Wieczorek, M. A. and Simons, F. J.: Localized spectral analysis on the sphere. Geophys. J. Int. vol. 162, pp. 655–675 (2005).
14. M. Senoussaoui, P. Kenny, N. Brummer, E. de Villiers, P. Dumouchel: Mixture of PLDA models in I-vector space for gender independent speaker recognition. Interspeech, Florence, Italy, August (2011).
15. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet: Front-end factor analysis for speaker verification. IEEE Trans. on Audio, Speech and Language Processing, vol. 19(4), pp. 788-798 (2011).

16. P. Kenny: Bayesian speaker verification with heavy tailed priors. The Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic, June (2010).
17. N. Brümmer, E. de Villiers: The speaker partitioning problem. The Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic, June (2010).
18. D. Garcia-Romero, and Carol Y. Espy-Wilson: Analysis of i-vector length normalization in speaker recognition systems. Interspeech 2011, Florence, Italy, Aug.(2011).
19. National Institute of Standards and Technology, NIST 2010 Speaker Recognition Evaluation Plan, http://www.itl.nist.gov/iad/mig/tests/spk/2010/index.html.
20. T. Kinnunen, R. Saeidi, F. Sedlak, K.A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li: Low-Variance Multitaper MFCC Features: a Case Study in Robust Speaker Verification. IEEE Transactions on Audio, Speech and Language Processing, vol. 20(7), pp. 1990-2001 (2012).
21. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, vol. 10(1), pp. 19-41 (2000).
22. Kenny, P., Boulianne, G., Ouellet, P. and P. Dumouchel: Joint factor analysis versus eigenchannels in speaker recognition. IEEE Transactions on Audio, Speech and Language Processing, vol. 15 (4), pp. 1435-1447 (2007).
23. Kenny, P., Boulianne, G., Ouellet, P. and P. Dumouchel: Speaker and session variability in GMM-based speaker verification. IEEE Transactions on Audio, Speech and Language Processing, vol. 15 (4), pp. 1448-1460 (2007).
24. Douglas D. O'Shaughnessy: Invited paper: Automatic speech recognition: History, methods and challenges. Pattern Recognition, vol. 41(10), pp. 2965-2979 (2008).
25. Douglas O'Shaughnessy: Speech communications - human and machine (2. ed.), IEEE Press, I-XXV, 1-547 (2000).
26. B. Kotnik, D. Vlaj, Z. Kacic, B. Horvat: Robust MFCC feature extraction algorithm using efficient addictive and convolutional noise reduction procedures. Proc. ICSLP, pp. 445-448 (2002).
27. Djuric, P. M., Kay, S. M.: Spectrum Estimation and Modeling. Digital Signal Processing Handbook, CRC Press LLC (1999).
28. Walden, A. T., McCoy, E. J., and Percival, D. B.: The variance of multitaper spectrum estimates for real Gaussian processes. IEEE Trans. Signal Process., vol. 2, pp. 479–482 (1994).
29. Md. Jahangir Alam, Patrick Kenny, Douglas D. O'Shaughnessy: A Study of Low-variance Multi-taper Features for Distributed Speech Recognition. Proc. of NOLISP, LNAI 7015, pp. 239-245 (2011).
30. Md. Jahangir Alam, Tomi Kinnunen, Patrick Kenny, Pierre Ouellet, Douglas D. O'Shaughnessy: Multi-taper MFCC features for speaker verification using I-vectors. ASRU, pp. 547-552 (2011).
31. S. J. Young et al., HTK Book, Entropic Cambridge Research Laboratory Ltd., 3.4 edition (2006). online: http://htk.eng.cam.ac.uk/.
32. Shing-Tai Pan, Chih-Chin Lai and Bo-Yu Tsai, The Implementation of Speech Recognition Systems on FPGA-Based Embedded Systems with SOC Architecture, International Journal of Innovative Computing, Information and Control, vol.7(11), pp.6161-6176 (2011).
33. T. Kinnunen, and H. Li: An overview of text-independent speaker recognition-from features to supervectors. Speech Comm., vol. 52(1), pp. 12-40 (2010).
34. J. W. Picone: Signal modeling techniques in speech recognition. Proc. IEEE, vol. 81, pp. 1215-1247 (1993).
35. X. D. Huang, A. Acero, and H. W. Hon: Spoken language processing: A guide to theory, algorithm, and system development. Prentice-Hall, 2001.
36. G. von Bekesy: Experiments in Hearing. McGraw-Hill, 1960.
37. A. V. Oppenheim and R. W. Schafer: Digital signal processing. Prentice-Hall, 1975.
38. F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier,

T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, D. A. Reynolds: A Tutorial on Text-Independent Speaker Verification. EURASIP Journal on Applied Signal Processing, vol. 4, pp. 430–451 (2004).

39. H. Hermansky: Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America, vol. 87 (4), pp. 1738-1752 (1990).
40. C. Kim and R. M. Stern: Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring. Proc. IEEE ICASSP, pp. 4574-4577 (2010).
41. M. J. Alam, P. Kenny, D. O'Shaughnessy: Robust Feature Extraction for Speech Recognition by Enhancing Auditory Spectrum. Proc. INTERSPEECH, Portland, Oregon, September (2012).
42. P. Kenny, P. Ouellet, and M. Senoussaoui: The CRIM system for the 2010 NIST Speaker Recognition Evaluation, April (2010).
43. ABC (Agnitio BUT and CRIM) system description for NIST Speaker Recognition Evaluation, June (2010).
44. N. Dehak et at.: MIT-CSAIL Spoken Language Systems and Lincoln Labs NIST SRE systems, (2010).
45. Alam, J., Ouellet, P., Kenny, P., O Shaughnessy, D.: Comparative Evaluation of Feature Normalization Techniques for Speaker Verification. Proc NOLISP, LNAI 7015, pp. 246-253 (2011).
46. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. In: Proc. Speaker Odyssey: the Speaker Recognition Workshop, Crete, Greece, pp. 213–218 (2001)
47. Xiang, B., Chaudhari, U., Navratil, J., Ramaswamy, G., Gopinath, R.: Short-time Gaussianization for robust speaker verification. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, Orlando, Florida, USA, pp. 681–684 (2002).
48. T. Kinnunen: Spectral Features for Automatic Text-independent Speaker Recognition. Licentiate’s thesis, University of Joensuu, Finland, December (2003).
49. Schuster, A.: On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena, Terr. Magn, vol. 3, pp. 13-41 (1898).
50. Priestley, M.B.: Spectral Analysis and Time Series, I. & II, Academic Press, London, UK (1981).
51. M. J. Alam, T. Kinnunen, P. Ouellet, P. Kenny, D. O'Shaughnessy: Multitaper MFCC and PLP features for Speaker Verification using I-vectors. accepted for publication in Speech Comm. (2012). DOI :10.1016/j.specom.2012.08.007
52. R. W. Komm, Y. Gu, F. Hill: Multitaper Spectral Analysis and Wavelet Denoising Applied to Helioseismic Data. Astrophysical Journal, vol. 519, pp. 407-421 (1999).
53. Wieczorek, M. A. and Simons, F. J.: Minimum Variance Multitaper Spectrum Estimation on the sphere. Journal of Fourier Analysis and Applications. vol. 13 (6), pp. 665–692 (2007).
54. Furui, S.: Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoustics, Speech Signal Process., vol. 29 (2), pp. 254–272 (1981).
55. Chen, C.-P. and Bilmes, J. : MVA processing of speech features. Technical Report UWEETR-2003-0024, EE Department, University of Washington, USA (2003).
56. E. J. Coy, A. T. Walden, D. B. Percival: Multitaper Spectral Estimation of Power Law Process. IEEE Trans. on Signal Processing, vol. 46 (3), pp. 655-668 (1998).
57. L. Rabiner and B. H. Juang: Fundamentals of Speech Recognition. Prentice-Hall (1993).
58. J. Makhoul and J. Schwartz: State of the art in continuous speech recognition. in Voice Communication between Humans and Machines., D. Roe and J. Wilpon, Eds. Washington, DC: National Academy Press, pp. 165–188 (1994).
59. A. Ezeiza, K. Lopez de Ipina, C. Hernandez, N. Barosso: Enhancing the Feature Extraction Process for Automatic Speech Recognition with Fractal dimensions. Cognitive Computation Journal, July (2012). DOI: 10.1007/s12559-012-9165-0.

60. N. Parihar, J. Picone, D. Pearce, H.G. Hirsch: Performance analysis of the Aurora large vocabulary baseline system, Proceedings of the European Signal Processing Conference, Vienna, Austria, (2004)

61. S.-K. Au Yeung, M.-H. Siu: Improved performance of Aurora-4 using HTK and unsupervised MLLR adaptation, Proceedings of the Int. Conference on Spoken Language Processing, Jeju, Korea, (2004)