# Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification

Tomi Kinnunen, *Member, IEEE*, Rahim Saeidi, *Member, IEEE*, Filip Sedlák, Kong Aik Lee, Johan Sandberg, Maria Hansson-Sandsten, *Member, IEEE*, and Haizhou Li, *Senior Member, IEEE*

***Abstract*—In speech and audio applications, short-term signal spectrum is often represented using mel-frequency cepstral coefficients (MFCCs) computed from a windowed discrete Fourier transform (DFT). Windowing reduces spectral leakage but variance of the spectrum estimate remains high. An elegant extension to windowed DFT is the so-called multitaper method which uses multiple time-domain windows (tapers) with frequency-domain averaging. Multitapers have received little attention in speech processing even though they produce low-variance features. In this paper, we propose the multitaper method for MFCC extraction with a practical focus. We provide, first, detailed statistical analysis of MFCC bias and variance using autoregressive process simulations on the TIMIT corpus. For speaker verification experiments on the NIST 2002 and 2008 SRE corpora, we consider three Gaussian mixture model based classifiers with universal background model (GMM-UBM), support vector machine (GMM-SVM) and joint factor analysis (GMM-JFA). Multitapers improve MinDCF over the baseline windowed DFT by relative 20.4% (GMM-SVM) and 13.7% (GMM-JFA) on the interview-interview condition in NIST 2008. The GMM-JFA system further reduces MinDCF by 18.7% on the telephone data. With these improvements and generally noncritical parameter selection, multitaper MFCCs are a viable candidate for replacing the conventional MFCCs.**

***Index Terms*—Mel-frequency cepstral coefficient (MFCC), multitaper, small-variance estimation, speaker verification.**

## I. Introduction

FEATURE extraction is the key function of a speech processing front-end. Spectral features computed from the windowed discrete Fourier transform (DFT) [1] or linear prediction (LP) models [2] are used in most of the front-ends. The DFT and LP models perform reasonably well under clean conditions but recognition accuracy degrades severely under changes in environment and channel since the short-term spectrum is subjected to many harmful variations. In this paper, we focus on one of the most successful techniques, the *mel-frequency cepstral coefficients* (MFCCs), that were introduced three decades ago [3] and are extensively used in speaker and language recognition, automatic speech recognition, emotion classification, audio indexing and, with certain modifications, even in speech synthesis and conversion applications. There is no doubt that the way we derive MFCC features has great impact on the performance of many speech processing applications.

There have been many attempts to enhance the robustness of MFCC features. Several techniques have demonstrated effective ways to normalize the MFCC features by using the statistics of the MFCC temporal trajectory. For example, cepstral mean and variance normalization (CMVN) [4], RASTA filtering [5], temporal structure normalization [6], feature warping [7], and MVA processing [8] are commonly used for enhancing MFCC robustness against additive noises and channel distortions. The specific configuration and order of chaining them depends, however, on the target application. Such techniques obtain the statistics either from the run-time signals themselves or from some training data. Therefore, they require either delayed processing or offline modeling. In this paper, we would like to study a new way to derive MFCC features, with which we reduce the MFCC estimation variance without relying on any statistics beyond a speech frame.

From a statistical point of view, the common MFCC implementation based on windowed DFT is suboptimal due to high *variance* of the spectrum estimate [10]. To elaborate on this, imagine that, for every short-term speech frame there exists an underlying *random process* which generates that particular frame; an example would be an autoregressive (AR) process driven with random inputs but with fixed coefficients. For speech signals, we imagine that there exists a speaker- and phoneme-dependent vocal tract configuration from which the actual speech sounds are generated from. A spectrum estimator with high variance then implies that, for the *same* underlying random process (e.g., two non-overlapping parts of the very same vowel sound), the estimated spectra and MFCCs may vary considerably.

In speaker verification [11], uncertainty in features is modeled by the variances in the Gaussian mixture models (GMMs) [12] and, recently, by subspace models of speaker and session variabilities in a supervector space [13]–[19]. However, if the

Manuscript received January 27, 2012; revised March 11, 2012; accepted March 15, 2012. Date of publication April 03, 2012; date of current version May 07, 2012. The work of T. Kinnunen was supported by the Academy of Finland (project no. 132129) and the works of R. Saeidi and H. Li were supported by the Nokia Foundation. Computing services from *CSC—IT Center for Science* were used for the speaker verification experiments (project no uef4836). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nestor Becerra Yoma.

T. Kinnunen and F. Sedlák are with the School of Computing, University of Eastern Finland, FI-80101 Joensuu, Finland (e-mail: tkinnu@cs.joensuu.fi; fsedlak@cs.joensuu.fi).

R. Saeidi is with the School of Computing, University of Eastern Finland, FI-80101 Joensuu, Finland, and also with the Radboud University Nijmegen, 6500 HC Nijmegen, The Netherlands (e-mail: rahim@cs.joensuu.fi; rahim.saeidi@let.ru.nl).

K. A. Lee and H. Li are with the Human Language Technology, Institute for Infocomm Research (I2R), Singapore 138632 (e-mail: kalee@i2r.a-star.edu.sg; hli @i2r.a-star.edu.sg).

J. Sandberg is with Nordea Bank, DK-0900 Copenhagen, Denmark, and also with the Centre for Mathematical Sciences, Lund University, SE-221 00 Lund, Sweden (e-mail: sandberg@maths.lth.se).

M. Hansson-Sandsten is with Mathematical Statistics, Centre for Mathematical Sciences, Lund University, SE-221 00 Lund, Sweden (e-mail: sandberg@maths.lth.se; sandsten@maths.lth.se).

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

Digital Object Identifier 10.1109/TASL.2012.2191960

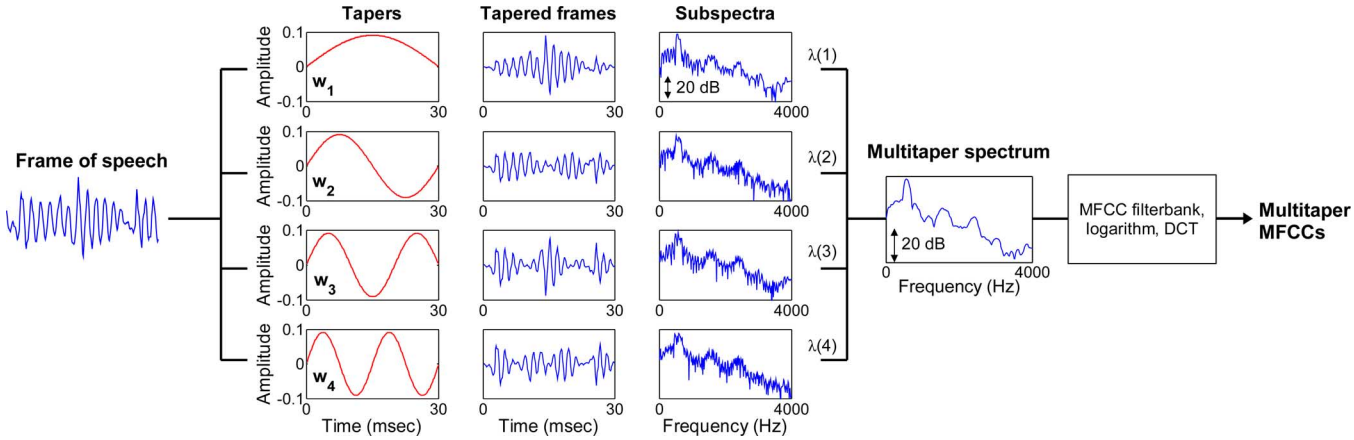


Fig. 1. Multitaper method uses multiple windows (tapers) leading to different subspectra, whose weighted average forms the spectrum estimate and leads to robust MFCCs. For visualization, spectra are shown in dB scale but computations are based on linear values. The tapers are from the SWCE method [9].

MFCCs themselves are estimated with smaller variance, one should expect the subsequent speaker and session variability models to exhibit less random variation as well. Using low-variance spectrum estimators has already been demonstrated to enhance performances of voice activity detection (VAD) [20], [21], speech enhancement [22], and speech recognition [23], to give a few examples.

The particular small-variance method adopted in this paper is based on *multitapers*, as illustrated in Fig. 1. The multitaper method [24]–[27], as a simple and elegant extension of the conventional windowed DFT, uses multiple window functions (a.k.a. *tapers*) with weighted frequency-domain averaging to form the spectrum estimate [24], [25], [27]. The tapers are designed to give approximately uncorrelated spectrum estimates so that averaging them reduces the variance. More specifically, the side-lobe leakage effect in the conventional windowed DFT is partially suppressed by multitapering [28], [29]. The well-known *Welch's method* [30] is a special case of the multitaper technique with identically shaped but time-shifted tapers. Thus, in Welch's method, the subspectra are uncorrelated because they are computed from different segments. The multitapers applied in this paper, in contrast, are fully overlapping in time but their shapes are designed so that they have only small overlap in frequency domain [10], [24]. Conceptually, multitapering also shares some similarity with smoothing the DFT estimate using frequency-domain convolution (e.g., [10]), but generally these are not mathematically equivalent.

The multitaper method of spectrum estimation was introduced around the same time as the MFCCs [24] but has found little use in speech processing so far [22], [31], [32]. This might be due to previously unstudied statistical properties of multitaper MFCCs and availability of different multitaper variants to choose from [9], [24]–[26]. Additionally, due to mostly theoretically focused treatments of the topic [10], [24], [25], practitioners may have had difficulties in implementing and choosing the control parameters in a typical recognition application.

Since the statistical properties of the multitaper MFCCs—briefly summarized in Section III—are recently analyzed [27] and further, we got encouraging preliminary speaker verification results in [33], we were curious to explore the technique further. In Section IV, we carry out detailed evaluation of multitaper bias and variance using simulated random processes on the TIMIT corpus. Importantly, in Sections V and VI we extend and complement the preliminary GMM-UBM results of [33] using two high-performance classifiers, GMM supervector with support vector machine (GMM-SVM) [13], [34] and GMM with joint factor analysis technique (GMM-JFA) including integrated speaker and intersession variability modeling [15], [35], [36]. To sum up, the main purpose of this paper is to review, collect and extend our recent work on the use of multitapers in speech processing with application to speaker verification. We provide sample implementation and recommendations for setting the control parameters.

## II. Computing the Multitaper MFCCs

Let $\mathbf{x} = [x(0) \dots x(N-1)]^{\mathrm{T}}$ denote one frame of speech of $N$ samples. The most popular spectrum estimate in speech processing, windowed discrete Fourier transform (DFT), is given by

$$\hat{S}(f) = \left| \sum_{t=0}^{N-1} w(t)x(t)\mathrm{e}^{-i2\pi tf/N} \right|^2 \qquad (1)$$

where $i = \sqrt{-1}$ is the imaginary unit and $f = 0, 1, \dots, N-1$ denotes the discrete frequency index. Here $\mathbf{w} = [w(0) \dots w(N-1)]^{\mathrm{T}}$ is a time-domain window function which usually is symmetric and decreases towards the frame boundaries. In this study, we choose the most popular window in speech processing, the *Hamming* window, with $w(t) = 0.54 - 0.46\cos(2\pi t/N)$.

From a statistical perspective, the use of a Hamming-type of window reduces the *bias* of the spectrum estimate, i.e., how much the estimated value $\hat{S}(f)$ differs from the true value $S(f)$, on average. But the estimated spectrum still has high variance. To reduce the variance, *multitaper* spectrum estimator [10], [24], [26] can be used:

$$\hat{S}(f) = \sum_{j=1}^{K} \lambda(j) \left| \sum_{t=0}^{N-1} w_j(t)x(t)\mathrm{e}^{-i2\pi tf/N} \right|^2 . \qquad (2)$$

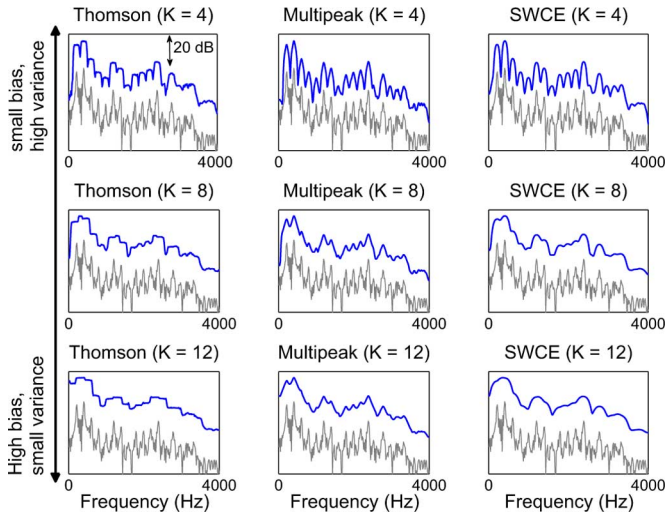


Fig. 2. Typical multitaper spectra for the methods used in this paper. The lower thin lines (gray) show the Hamming-windowed DFT spectrum as a reference. The spectra have been shifted by 20 dB for visualization.

Here, $K$ multitapers $\mathbf{w}_j = [w_j(0)\dots w_j(N-1)]^{\mathrm{T}}$, where $j = 1,\dots,K$, are used with corresponding weights $\lambda(j)$. The multitaper estimate is therefore obtained as a weighted average of $K$ subspectra (Fig. 1). The windowed DFT (1) is obtained as a special case when $K = 1$ and $\lambda = 1$.

### A. Choice of the Tapers

A number of different tapers have been proposed for spectrum estimation, such as *Thomson* [24], *sine* [25], and *multipeak* tapers [26]. For cepstrum analysis, the sine tapers are applied with optimal weighting in [9]. Each type of taper is designed for some given type of (assumed) random process; as an example, Thomson tapers are designed for flat spectra (white noise) and multipeak tapers for peaked spectra (such as voiced speech). In practice, many multitapers work well even though designed for another process. For instance, the Thomson tapers [24], designed for white noise, tend to perform well for any smooth spectrum.

In general, the tapers are designed so that the estimation errors in the subspectra will be approximately uncorrelated, which is the key to variance reduction. It is out of the scope of this paper to describe the details of finding optimal tapers. For theoretical treatment, we point the reader to [10], [24] while [9], [25], [26], [37] provide more concise discussions. In the Appendix of this paper, we point to practical MATLAB implementations. In short, the solution is obtained from an eigenvalue problem where the eigenvectors and -values correspond to the tapers and their weights, respectively. The tapers considered in this paper are all computed offline without any data-adaptive training process and applied to all speech utterances.

Fig. 2 shows, for a single voiced speech frame, examples of the three multitaper methods considered in this study, Thomson [24], multipeak [26] and sine-weighted cepstrum estimator (SWCE) [9]. Each panel shows the multitaper spectrum (upper thick line) along with Hamming-windowed DFT estimate (lower thin line). All the three multitaper methods produce smoother spectrum compared to the Hamming method, because of variance reduction. Thomson produces a staircase-like spectrum, multipeak a spectrum with sharper peaks and SWCE a compromise between these two methods. In this example, for a small number of tapers, say $K \leq 4$, all the three methods preserve both the harmonics (due to the voice source) and the spectral envelope (due to the vocal tract). For a high number of tapers, say $K \geq 8$, the harmonics gets smeared out. The optimum number of tapers is expected to depend on the target application. In speaker recognition, both the voice source and vocal tract filter are found to be useful; thus, we expect to get the best results using a relatively small number of tapers.

### B. Computational Complexity and Periodogram Smoothing

The windowed periodogram in (1) can be computed using fast Fourier transform (FFT) of complexity $\mathcal{O}(N \log N)$. Since the multitaper estimator (2) requires $K$ FFTs, the complexity of the direct implementation is $\mathcal{O}(KN \log N)$, which might become a critical design consideration under low-resource platforms. Luckily, when the tapers are sinusoids as in [25] and the SWCE method [9], complexity can be reduced. Indeed, the $j$th sine taper can be written using Euler's formula as

$$w_j(t) = \sin(2\pi f_j t) = \frac{1}{2i}\{e^{i2\pi f_j t} - e^{-i2\pi f_j t}\}. \tag{3}$$

Thus, DFT of the windowed data segment $x(t)w_j(t)$ is

$$\mathcal{F}\{x(t)w_j(t)\} = \frac{1}{2i}\{X(f-f_j) - X(f+f_j)\} \tag{4}$$

where $\mathcal{F}\{\cdot\}$ denotes the DFT operator and $X(f) = \mathcal{F}\{x(n)\}$. Substituting this to the multitaper spectrum estimator (2) and simplifying leads to

$$\hat{S}(f) = \frac{1}{4}\sum_{j=1}^{K}\lambda(j)\Big\{|X(f-f_j)|^2 + |X(f+f_j)|^2 - 2X(f-f_j)X^*(f+f_j)\Big\}.$$

This consists of computing $X(f)$ by one FFT, followed by the three frequency-domain smoothing terms of complexity $\mathcal{O}(KN)$, thus totaling $\mathcal{O}(N \log N + KN)$ steps. Since typically $K \ll N$, this is usually faster than the direct implementation (2).

A popular method for producing smooth spectrum estimates, *periodogram smoothing*, is to convolve the unwindowed (raw) periodogram with a suitable frequency-domain smoothing kernel, which has also complexity $\mathcal{O}(N \log N + KN)$. Note that the first sum term in (5) is, in fact, the convolution of $|X(f)|^2$ with kernel $\{\lambda_1, \lambda_2, \dots, \lambda_K\}$. However, because of the two additional terms, the methods are not equivalent. In our speaker verification experiments, we will also provide experiments with periodogram smoothing.

## III. Bias and Variance of Multitaper Estimators

To understand the bias and variance tradeoff better, we consider the variance and spectral resolution of the single- and

multi-taper methods. For the windowed DFT (1), the variance is usually approximated as [10]

$$V\left[\hat{S}(f)\right] \approx S^2(f). \tag{5}$$

The spectral resolution, that is, the frequency spacing under which two frequency components cannot be separated, is approximately $B_w = 1/N$ for the rectangle window but $B_w = 2/N$ for the Hamming window. Note also that (5) does not depend on the frame length $N$ and thus, including more samples in a frame will *not* reduce the variance.

For the multitaper spectrum estimator (2), the spectral resolution is approximately $B_w = (K+2)/N$ which is the spectral resolution parameter used in the design of the Thomson [24] and multipeak [26] tapers. The variance can be approximated as

$$V\left[\hat{S}(f)\right] \approx \frac{1}{K} S^2(f). \tag{6}$$

This result is analogous to the well-known result that variance of the mean of sample of size $K$ is inversely proportional to $K$ [4, p. 82]. The formula (6) is approximately valid also for the Welch's method [30] with 50% overlap between the windows [10].

Note that up to this point we have only considered variance and bias in spectral and not MFCC domain. Intuitively, it is easy to understand, that if the spectrum is estimated with low bias and low variance, the resulting MFCC vector will also have low bias and variance. Using vector notation, the MFCC vector $\mathbf{c}$ is related to the (true) spectrum vector $\mathbf{s} = [S(0) \dots S(N-1)]^{\mathrm{T}}$ by $\mathbf{c} = (1/M)\mathbf{\Phi}^{\mathrm{H}} \log(\mathbf{M}\mathbf{s})$, where $M$ is the number of filters in the filter-bank $\mathbf{M} \in \mathbb{R}^{M \times N}$, the logarithm operates element-wise and $\mathbf{\Phi}$ is the $M$-by-$M$ Fourier matrix with the $(a,b)^{\mathrm{th}}$ element: $\mathbf{\Phi} \triangleq \{\mathrm{e}^{-i2\pi(a-1)(b-1)/M}\}_{ab}$. Bias of cepstral coefficients has been studied in [38], whereas approximate bias in MFCCs $\mathsf{B}[\hat{\mathbf{c}}]$ can be written as [27]:

$$\mathsf{B}[\hat{\mathbf{c}}] \approx \frac{1}{M}\mathbf{\Phi}^{\mathrm{H}} \left( \log\left(\frac{\mathbf{M}\mathsf{E}[\hat{\mathbf{s}}]}{\mathbf{M}\mathbf{s}}\right) - \frac{\mathsf{diag}(\mathbf{M}\mathsf{V}\,[\hat{\mathbf{s}}]\mathbf{M}^{\mathrm{T}})}{2\,(\mathbf{M}\mathsf{E}[\hat{\mathbf{s}}])^2} \right). \tag{7}$$

Here, the division operates element-wise, $\hat{\mathbf{s}} = [\hat{S}(0) \dots \hat{S}(N-1)]^{\mathrm{T}}$ denotes the estimated spectrum using multitapers (2), $\mathsf{E}[\hat{\mathbf{s}}]$ denotes the expected value of $\hat{\mathbf{s}}$, $\mathsf{V}[\hat{\mathbf{s}}]$ denotes the covariance matrix of the spectrum estimate and $(\cdot)^{\mathrm{H}}$ stands for conjugate transpose. Both the expected value $\mathsf{E}[\hat{\mathbf{s}}]$ and the covariance matrix $\mathsf{V}[\hat{\mathbf{s}}]$ (see [10] and [27] for details) depend on the covariance matrix $\mathbf{R}$ of the random process and hence, on the true spectrum $\mathbf{s}$.

The covariance matrix of the estimated MFCC vector using multitapers can be approximated as [27]

$$\mathsf{V}[\hat{\mathbf{c}}] \approx \frac{1}{M^2}\mathbf{\Phi}^{\mathrm{H}} \frac{\mathbf{M}\mathsf{V}[\hat{\mathbf{s}}]\mathbf{M}^{\mathrm{T}}}{\mathbf{M}\mathsf{E}[\hat{\mathbf{s}}]\mathsf{E}[\hat{\mathbf{s}}]^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}}\mathbf{\Phi}. \tag{8}$$

The bias and variance of the MFCC estimator depend on the true spectrum of the process. As this is usually unknown, it is impossible to use these formulas directly. However, a general rule is that by increasing the number of tapers, we can reduce the variance of the spectrum estimate, hence making the spectrum estimate more robust across random variations. As an example, Fig. 3 shows the sample mean and standard deviations of conventional and multitaper MFCCs for one speech utterance in the NIST 2002 corpus. The mean vectors in this example differ mostly by an additive constant, whereas the variances of the higher order MFCCs (beyond the 5th MFCC) are significantly reduced due to multitapering.

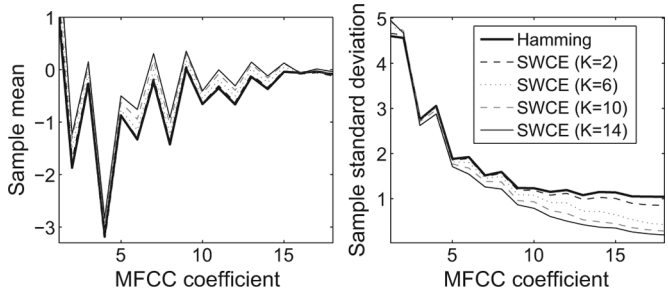


Fig. 3. Multitapers help in reducing variance of the higher order MFCCs, but without modifying the mean value much.

## IV. Numerical Evaluation of Bias and Variance of Multitaper MFCC Estimators

### A. Monte Carlo Computation of Bias and Variance Using Known Autoregressive Models

We would like to know how much the estimated MFCCs differ from the true MFCCs. As is common in the evaluation of nonparametric spectrum estimators, we consider a parametric model with known parameters as a ground truth. Due to their success in spectral modeling of speech signals [2], we consider autoregressive $\mathrm{AR}(p)$ random process

$$x(t) = -\sum_{m=1}^{p} a_m x(t-m) + \varepsilon(t) \tag{9}$$

where $\{a_m\}_{m=1}^{p}$ are the known AR coefficients and $\varepsilon(t) \sim \mathcal{N}(0,1)$ are independent and identically distributed (i.i.d.) samples of the driving white noise sequence. The corresponding $\mathrm{AR}(p)$ spectrum (sampled at discrete data points $f = 0, 1, \dots, N-1$) is given by

$$S_{\mathrm{AR}(p)}(f) = \frac{1}{\left|1 + \sum_{m=1}^{p} a_m \exp(-i2\pi f m/N)\right|^2}. \tag{10}$$

Thus, given the known parameters $\{a_m\}_{m=1}^{p}$, we can simulate a specific realization $\mathbf{x}$ of the random process using (9). Applying windowed DFT, multitaper or any other spectrum estimator on $\mathbf{x}$ produces an estimate $\hat{S}(f)$ of the spectrum (10). Depending on the random input $\varepsilon(t)$ in (9), the estimate will be different each time. We are concerned in how the estimated MFCC vector $\hat{\mathbf{c}}$ (computed from $\hat{S}(f)$) differs from the ground-truth MFCC vector $\mathbf{c}^{\mathrm{AR}(p)}$ [computed from (10)] on average. To this end, we consider the three well-known descriptors of any estimator—bias, variance, and mean square error (MSE):

$$\mathsf{B}[\hat{\mathbf{c}}] = \mathsf{E}[\hat{\mathbf{c}}] - \mathbf{c}^{\mathrm{AR(p)}} \tag{11}$$

$$\mathsf{V}[\hat{\mathbf{c}}] = \mathsf{E}[\hat{\mathbf{c}}^2] - (\mathsf{E}[\hat{\mathbf{c}}])^2 \tag{12}$$

$$\mathsf{MSE}[\hat{\mathbf{c}}] = \mathsf{E}\left[\left(\hat{\mathbf{c}} - \mathbf{c}^{\mathrm{AR(p)}}\right)^2\right] \tag{13}$$

where we introduced shorthand notation $\mathbf{z}^2 = \mathsf{diag}(\mathbf{z}\mathbf{z}^{\mathrm{T}})$ for vector $\mathbf{z}$. MSE further links the bias and variance as $\mathsf{MSE}[\hat{\mathbf{c}}] = \mathsf{B}[\hat{\mathbf{c}}]^2 + \mathsf{V}[\hat{\mathbf{c}}]$. To compute the bias and variance for a single random process (one set of $a_k$s), we approximate the expectations of random vector $\mathbf{z}$ in (11)–(13) using sample mean as $\mathsf{E}[\mathbf{z}] \approx (1/N_{\mathrm{MC}}) \sum_{r=1}^{N_{\mathrm{MC}}} \mathbf{z}_r$. Here, $N_{\mathrm{MC}}$ is the number of random Monte Carlo draws and $\mathbf{z}_r$ corresponds to the vector of the $r$th random draw. We fix $N_{\mathrm{MC}} = 30000$ for which we found the values of (11)–(13) converged so that the Monte Carlo error can be considered negligible.

### B. Summarizing Bias, Variance and MSE

Note that above bias, variance and MSE are defined for a *single* random process (one set of $a_k$s). Depending on the choice of the coefficients or the order of the AR model $(p)$, one gets different conclusions. As an overall measure, therefore, we are interested on the average bias, variance and MSE over a large number of different random processes (different set of $a_k$s and different AR model order $p$). This resembles a typical speaker recognition setting where inferences about speaker identity are drawn over a large number of speech frames.

The average MSE vector is given by

$$\boldsymbol{\mu}_{\mathsf{MSE}} = \frac{1}{N_{\mathrm{P}}} \sum_{n=1}^{N_{\mathrm{P}}} \mathsf{MSE}[\hat{\mathbf{c}}_n] \tag{14}$$

where $\mathsf{MSE}[\hat{\mathbf{c}}_n]$ indicates MSE (13) of the $n$th random process out from a collection of $N_{\mathrm{P}}$ random processes. We are also interested in whether the difference in the means are statistically significant. To this end, we also compute the confidence interval of the mean for each of the individual coefficients. By denoting the individual dimensions of $\mathsf{MSE}[\hat{\mathbf{c}}_n]$ and $\boldsymbol{\mu}_{\mathsf{MSE}}$ by $\mathsf{MSE}_n(q)$ and $\mu_{\mathsf{MSE}}(q)$, respectively, we compute the confidence intervals as $\mu_{\mathsf{MSE}}(q) \pm 1.96\sqrt{\sigma^2_{\mathsf{MSE}}(q)/N_{\mathrm{P}}}$ where the MSE variance is given by

$$\sigma^2_{\mathsf{MSE}}(q) = \frac{1}{N_{\mathrm{P}} - 1} \sum_{n=1}^{N_{\mathrm{P}}} \left(\mathsf{MSE}_n(q) - \mu_{\mathsf{MSE}}(q)\right)^2 \tag{15}$$

for each MFCC feature indexed by $q = 1, 2, \ldots, 18$. The confidence interval signifies that, with 95% certainty, the true mean value falls within the confidence bounds. Regarding bias and variance, their means with associated confidence intervals can be similarly computed.

### C. Obtaining the Reference AR Models and MFCCs

To simulate speech-like AR random processes, we obtain the AR coefficients $a_k$ from real speech utterances rather than handcrafting them. To this end, we pick the common SA1 utterance ("*She had your dark suit in greasy wash water all year*") from a total number of 77 speakers (59 male, 18 female) from the training section of the Western dialect (DR7) on the TIMIT corpus. We use the corpus annotations to locate phone boundaries (excluding short phonemes less than 11.25 ms in duration), and compute the AR coefficients of each phone. This set, consisting of $N_{\mathrm{P}} = 2849$ phones, is representative of all American English phonemes and phoneme groups. For consistency with the following speaker recognition experiments, we resample the utterances down to 8 kHz. To avoid favoring AR models of a particular fixed order, we adapt the AR model order $(p)$ differently for each phoneme. To this end, we use the well-known Schwarz's Bayesian criterion (SBC) [39] implemented in the toolbox of [40]. We set the search limits for the optimum model order as $[p_{\min} = 1, p_{\max} = 40]$.

The MFCCs are extracted using procedure similar to the speaker verification experiments (see Section V-C). To exclude the effects of application-dependent feature normalizations, we measure the distortions in the lowest 18 base MFCCs, excluding the DC coefficient $c_0$. In the recognition experiments (Section VI), however, we utilize a complete front-end with additional RASTA filter, delta features and cepstral mean/variance normalization (CMVN) as a normal practice in speaker verification.

### D. Results

We first compare the average biases, variances and MSEs of the windowed DFT (Hamming) and SWCE multitaper $(K = 4)$ MFCC estimators in Fig. 4; the results for Thomson and multipeak tapers were similar to SWCE and were excluded for visual clarity. Regarding bias of the first cepstral coefficient $c_1$, both methods have a large negative bias which is, interestingly, larger in magnitude for Hamming. Regarding the other coefficients, both methods yield positive bias. Hamming introduces generally less bias but SWCE clearly reduces variance of all MFCCs by a wide margin. Regarding MSE of the lowest MFCCs ($c_1$ through $c_3$), SWCE yields smaller MSEs but the differences are not significant due to overlapping confidence intervals. However, the intermediate and higher order MFCCs produce significantly smaller MSEs.

Next we compare bias, variance and MSE integrated over all the 18 MFCC coefficients in Fig. 5 as a function of taper count for all the four estimators. Computations are similar as in Fig. 4 but we replace the vector quantities in (11)–(13) by their corresponding $L_1$-norms $\|\cdot\|_1$, i.e., the sum of the (absolute value of) individual elements. In the case of bias, we display $L_1$-norm of the *squared* bias. This is natural because $\mathsf{MSE}[\hat{\mathbf{c}}] = \mathsf{B}[\hat{\mathbf{c}}]^2 + \mathsf{V}[\hat{\mathbf{c}}]$, which helps in better interpreting the relative contributions of bias and variance terms to MSE.

According to Fig. 5, there is a large positive bias for all the four spectrum estimators. This bias is generally larger for the multitaper estimators in comparison to windowed DFT, as expected. But the variance of all three multitaper estimators is significantly smaller than that of Hamming-windowed DFT. For the biases, Hamming $<$ multipeak $<$ SWCE $<$ Thomson, but the order is reversed for the variances. The compromise measure, MSE, shows nicely convex behavior; for small number of tapers $K$, the large variance dominates over squared bias, leading to high MSE. For large $K$, similarly, the squared bias dominates and increases MSE. The smallest MSE values are obtained at $K = 4$ for Thomson and SWCE and at $K = 6$ for multipeak. Behavior of the MSE values suggests that a suitable number of tapers for Thomson might be smaller compared to multipeak and SWCE.

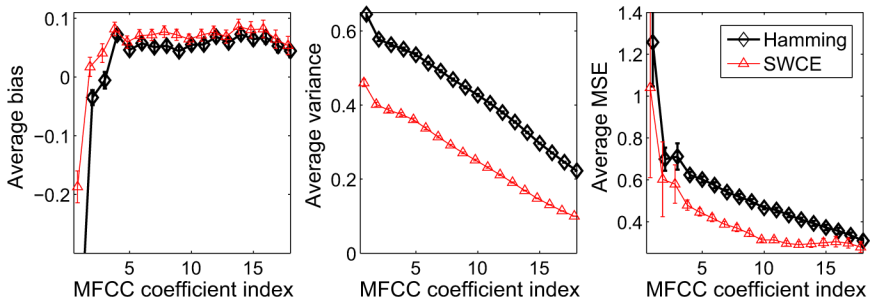


Fig. 4. Average MFCC bias, variance and MSE ($\mathrm{MSE} = \mathrm{bias}^2 + \mathrm{variance}$) of Hamming and SWCE ($K = 4$) estimators over 2849 different AR random processes of varying order. Bias, variance and MSE of each random process are computed using 30 000 Monte Carlo draws. The error bars indicate 95% confidence interval of the mean. For visual clarity, the results for Thomson and multipeak are excluded. While multitapers slightly increase bias for most coefficients, the variance of each coefficient is significantly reduced. The MSE improvements are most prevalent for coefficients $c_3$ through $c_{16}$.

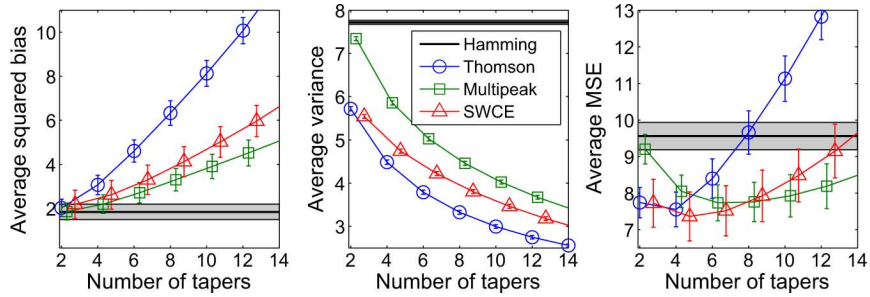


Fig. 5. Similar to Fig. 4 but for squared bias, variance and MSE integrated over all the 18 MFCCs as a function of taper count. The results are shown for $K = 2, 4, \ldots, 14$ tapers; for visual clarity, multitapers are slightly offset in horizontal direction.

To sum up, the results in Figs. 4 and 5 clearly indicate that multitapers reduce the variance of the MFCC estimates which is useful from generalization point of view. From these application-independent statistical MFCC estimator analyses, suitable $K$ might be on the range $2 \leq K \leq 8$ for typical speech applications utilizing MFCCs, although it will certainly depend on the task at hand; we will now proceed to our target application, speaker verification.

## V. Speaker Verification Setup

### A. Corpora and Classifiers

For the speaker verification experiments, we utilize two different corpora and three classifiers (Table I). We use the NIST 2002 speaker recognition evaluation (SRE) corpus for extensive exploration of control parameters and effect of additive noise. For this, we employ a lightweight *Gaussian mixture model with universal background model* (**GMM-UBM**) method [12] with *test normalization* (T-norm) [41]. The same system was used in our recent studies [33], [43]–[45]. Here we use it for choosing the type of multitaper variant for the more expensive NIST 2008 experiments. All the data in NIST 2002 contains telephone conversations collected over the cellular network.

We then verify our findings on an independent, more recent and much larger NIST 2008 SRE corpus which includes telephone, interview and auxiliary microphone data. For the experiments on NIST 2008 data, we employ two classifiers which were developed in participation of the past two NIST SRE campaigns [46]. The first system (**GMM-SVM**) uses Gaussian mean *supervectors* with support vector machine (SVM) [13] and *nuisance attribute projection* (NAP) technique [14], [34] for channel compensation. Zero normalization followed by T-norm (ZT-norm) [42] is used for score normalization. The

TABLE I
Details of the Evaluation Corpora and the Three Classifiers (UBM=Univ. Background Model; JFA=Joint Factor Analysis; NAP=Nuisance Attribute Projection; SWB=Switchboard)

| | **NIST 2002** | **NIST 2008** | |
|---|---|---|---|
| Speakers | 139 ♂, 191 ♀ | 1092 ♂, 1649 ♀ | |
| Gen. trials | 2982 | 15,345[a] | |
| Imp. trials | 39,259 | 56,792[b] | |
| Type of data | telephone | telephone, interview, mic. | |
| Training dur. | 2 min | 3–5 min | |
| Test dur. | 15–45 sec | 3–5 min | |
| | **GMM-UBM** [12] | **GMM-SVM** [13] | **GMM-JFA** [15], [35], [36] |
| Spec. subtraction | Yes | No | No |
| Gaussians per gender | 1024 | 512 | 512 |
| Intersession compens. | – | NAP [34] | JFA [15], [35], [36] |
| Background data | SRE 01 | SRE 04, 05, 06, MIXER5 | SRE 04 |
| Eigenchannel data | – | SRE 04, 06, MIXER5 | SRE 04, 05, 06, MIXER5 |
| Eigenvoice data | – | – | SRE 05, 06, SWB |
| Diag. model | – | – | SRE 04 |
| Score normalization | T-norm [41] SRE 01 | ZT-norm [42] SRE 05, 06 | TZ-norm [42] SRE 05, 06 |

[a] 11540 det1, 1105 det4, 1472 det5, 1228 det7.

[b] 22641 det1, 10636 det4, 6982 det5, 16533 det7.

second system (**GMM-JFA**) is a widely recognized high performance system, which uses *joint factor analysis* (JFA) technique [15], [16], [35] for integrated intersession and speaker variability modeling in the GMM supervector space. For score normalization, we use T-norm followed by Z-norm (TZ-norm).

For the recognition experiments under additive noise degradation, we contaminate the test utterances with factory noise while the background, cohort and target models are kept untouched. In [47] we found *spectral subtraction* [48] to be useful under additive noise degradation and it is thus included in the NIST 2002 experiments. We also did preliminary evaluation on the NIST 2008 data but the improvement was not systematic, and, given the added computational overhead, we decided not to include it to the NIST 2008 experiments.

### B. Performance Evaluation

In comparison of the different MFCC estimators, we evaluate speaker verification accuracy using equal error rate (EER) and minimum detection cost function (MinDCF). EER is the error rate at the threshold $\theta_{\mathrm{EER}}$ for which the miss and false alarm rates are equal: $\mathrm{EER} = P_{\mathrm{miss}}(\theta_{\mathrm{EER}}) = P_{\mathrm{fa}}(\theta_{\mathrm{EER}})$. MinDCF is used in the NIST speaker recognition evaluations and is defined as $\min_\theta\{C_{\mathrm{miss}}P_{\mathrm{miss}}(\theta)P_{\mathrm{tar}} + C_{\mathrm{fa}}P_{\mathrm{fa}}(\theta)(1 - P_{\mathrm{tar}})\}$, where $C_{\mathrm{miss}} = 10$ is the cost of a miss (false rejection), $C_{\mathrm{fa}} = 1$ is the cost of a false alarm (false acceptance) and $P_{\mathrm{tar}} = 0.01$ is the prior probability of a target (true) speaker. In addition, we show selected detection error tradeoff (DET) plots [49] for the entire tradeoff of false alarm and miss rates.

### C. Feature Extraction

For the baseline *Hamming* method, we compute the MFCCs using the typical procedure [4]: Hamming window (frame duration 30 ms and hop 15 ms), DFT spectrum estimate using windowed periodogram (1), 27-channel mel-frequency filterbank, logarithmic compression and discrete cosine transform (DCT). We retain the lowest 18 MFCCs, excluding the energy coefficient $c_0$. For *Thomson* [24], *multipeak* [26], and *sine-weighted cepstrum estimator* (SWCE) [9] methods, the steps are the same, except that the spectrum is estimated using (2). In preliminary experiments, we found the frequently used pre-emphasis filter $H(z) = 1 - 0.97z^{-1}$ to degrade accuracy and it is therefore turned off in all the experiments.

After the 18 base MFCCs are extracted, we apply RASTA filter [5] and append the $\Delta$ and $\Delta^2$ coefficients, implying 54-dimensional features. We then discard the nonspeech frames using an energy-based voice activity detector (VAD) and carry out utterance-level cepstral mean and variance normalization (CMVN). RASTA and CMVN are used for mitigating linear channel distortions.

We were also curious to see the effect of excluding the MFCC filterbank and to compute the 18 coefficients directly from the unwarped spectrum. We hypothesized that the double smoothing of multitaper spectrum followed by mel-filter energy integration might be suboptimal for speaker verification where we wish to retain the spectral details in addition to the envelope. We address this hypothesis on the NIST 2002 corpus in Section VI-A.

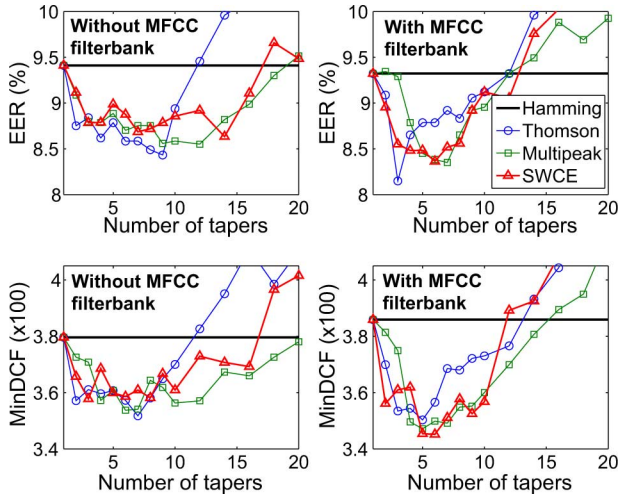


Fig. 6. Effects of the number of tapers and MFCC filterbank to different methods.

## VI. Speaker Verification Results

### A. GMM-UBM System on the NIST 2002 SRE Corpus

We first study how the choice of the spectrum estimation method affects speaker verification accuracy. For each of the multitaper methods—Thomson, multipeak and SWCE—we vary the number of tapers and contrast the result to the baseline Hamming method. EER and MinDCF, for both with and without MFCC filterbank, are shown in Fig. 6 where the horizontal (black) line represents the baseline. We observe the following:

- Multitaper methods outperform Hamming in both EER and MinDCF for a wide range of taper count (approx. $2 \leq K \leq 10$). Optimum value of $K$ depends on the method and the objective (EER or MinDCF).
- By including the MFCC filterbank the optimum points shift to left (less tapers) in most cases. This is expected because the MFCC filterbank introduces additional averaging over multitapering. Using MFCC filterbank improves EER and MinDCF and makes the curves generally less ragged, indicating stable parameter setting.
- The performance of the three multitaper methods at their optima are close to each other. Thomson shows sharper local mimima than multipeak and SWCE methods and gives higher error rates for large number of tapers.

The trends in Fig. 6 are, interestingly, in a reasonable agreement with Fig. 5. Both MSE, EER, and MinDCF demonstrate approximately convex shapes and all the three methods give similar performance with optimized $K$. Second, for large $K$, $\mathrm{MSE(Thomson)} > \mathrm{MSE(SWCE)} \gtrapprox \mathrm{MSE(Multipeak)}$; the same approximate ordering holds also for EER and MinDCF.

We next study the accuracy under additive factory noise corruption. Based on Fig. 6, for each method, we set the number of tapers to give both small EER and MinDCF. For the nonwarped case (no MFCC filterbank) we set the values to $K = 8$ (Thomson), $K = 10$ (multipeak), and $K = 7$ (SWCE). For the warped frequency case (MFCC filterbank included), in turn, we set the values to $K = 3$ (Thomson), $K = 5$ (multipeak), and

TABLE II
RESULTS UNDER FACTORY NOISE CORRUPTION ON THE NIST 2002 CORPUS CORRESPONDING TO THE RIGHT-HAND SIDE PLOTS (MFCC FILTERBANK INCLUDED) OF FIG. 7. IN EACH ROW, THE ERROR COUNTS SIGNIFICANTLY DIFFERING FROM THE BASELINE HAMMING, USING MCNEMAR'S TEST AT 95% CONFIDENCE LEVEL, ARE INDICATED FOR BOTH GENUINE (•) AND IMPOSTOR (†) TRIALS

| SNR | Equal error rate (EER, %) | | | | MinDCF (×100) | | | |
|---|---|---|---|---|---|---|---|---|
| (dB) | Hamming | Thomson | Multip. | SWCE | Hamming | Thomson | Multip. | SWCE |
| Orig. | 9.32 | 8.15 • † | 8.45 • † | 8.36 • † | 3.86 | 3.53 • † | 3.47 • † | 3.45 • |
| 20 | 9.73 | 8.79 • † | 8.62 • † | 8.69 • † | 3.91 | 3.73 • † | 3.62 • | 3.56 • † |
| 10 | 10.41 | 9.85 † | 9.66 † | 9.62 † | 4.30 | 4.20 • † | 4.11 † | 4.03 • |
| 0 | 11.53 | 11.50 | 11.44 | 11.32 | 5.04 | 5.02 • † | 4.93 • † | 4.76 † |
| -10 | 17.17 | 16.52 † | 15.86 • † | 15.96 • † | 7.38 | 7.04 • † | 6.72 • † | 6.49 • |

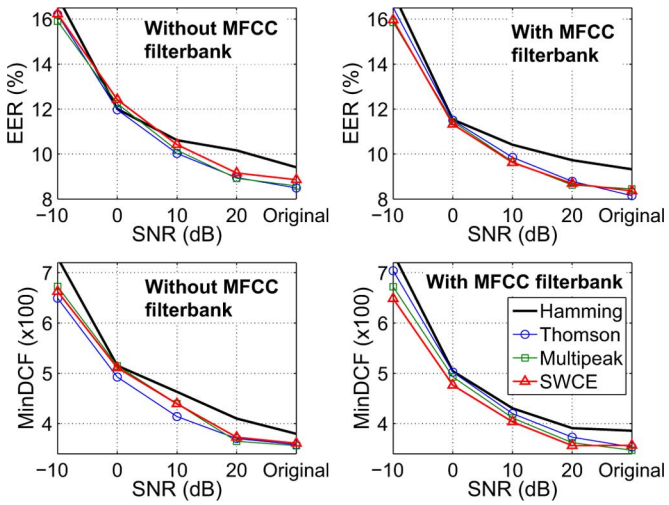


Fig. 7. Effect of factory noise under different signal-to-noise ratios.

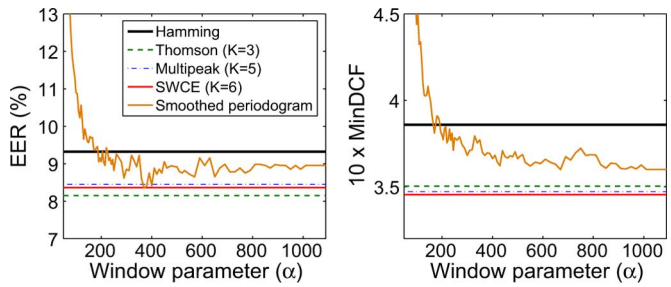


Fig. 8. Periodogram smoothing on the NIST 2002 corpus. As a reference, results for Hamming, Thomson, multipeak, and SWCE method are also given. Smoothing is performed by convolving the unwindowed periodogram with a Gaussian kernel with parameter $\alpha$ controlling the kernel width.

$K = 6$ (SWCE). The results, as a function of SNR, are given in Fig. 7. The following can be observed:

- Accuracy of all methods drops as SNR decreases, as expected. Multitapers outperform Hamming in nearly all cases (the exception occurs at 0 dB but the EER difference is not statistically significant, see Table II).
- In the noisy cases (SNR $\leq$ 20 dB), Thomson performs best on average when mel-warping is not applied; for the mel-warped case, SWCE performs the best.
- MFCC filterbank improves both EER and MinDCF.

Table II further displays the exact error values for the mel-warped case. We also carry out McNemar's significance testing with 95% confidence level at both operating points [4], [50]. In 27 out of 30 cases, the difference between the multitaper and the baseline is significant in at least one of the error types.

### B. Comparison with Periodogram Smoothing

Due to its popularity in other application domains, we are interested in the performance of periodogram smoothing, i.e., convolution of $|X(f)|^2$ with a frequency-domain smoothing kernel. As discussed in [10], choice of the kernel (in particular, its bandwidth) is not easy but typically requires trial-and-error for a given application. To this end, we convolve the unwindowed periodogram estimate with a Gaussian window[1] $w(n) = \exp\{-(1/2)(\alpha(n/(N/2)))^2\}$, where $N$ and $\alpha$ are the size and width of the window, respectively. The width of the window is inversely related to the value of $\alpha$; a larger value of $\alpha$ produces a narrower window. The result is displayed in Fig. 8 for the same configuration as Fig. 6 for the mel-warped case. As a reference, we show the optimized results for Hamming, Thomson, multipeak, and SWCE methods from Fig. 6.

By optimizing $\alpha$, periodogram smoothing outperforms the baseline Hamming method, but it does not outperform any of the multitaper methods. For $\alpha \approx 400$ (for which the effective number of nonzero samples in the kernel is about 4), EER is close to those of the SWCE and multipeak methods, but for the primary metric of speaker recognition evaluations, MinDCF, multitapers perform better.

### C. Experiments on the NIST 2008 SRE Corpus

Due to expensive nature of NIST 2008 experiments, we fix as many parameters as we think reasonable. We choose to use SWCE method with MFCC filterbank based on observations from Fig. 7. We first verify our observations regarding suitable number of tapers, using the GMM-SVM system. The EER and MinDCF values in Figs. 9 and 10 indicate that, even though setting depends on the data condition, the optima are always achieved with $3 \leq K \leq 8$. This range agrees well with the NIST 2002 GMM-UBM result in Fig. 6, *which has a completely different classifier, implementation details and choice of data sets*. SWCE outperforms Hamming for a wide range of $K$ and therefore, the exact setting does not appear very critical.

In one of the sub-conditions (det4), the baseline Hamming outperforms multitaper. One reason might be non-optimal selection of datasets for channel compensation in this subcondition; the error rates for both Hamming and SWCE are higher than those in the other three conditions.

[1]Choice of window is less important than its bandwidth [10]. We use Matlab's gausswin$(N, \alpha)$ command, with $N = 512$.

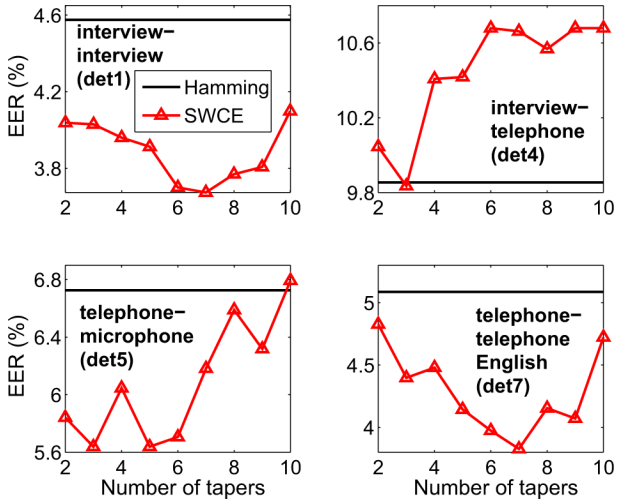


Fig. 9. Equal error rates (EER) on the NIST 2008 core task.

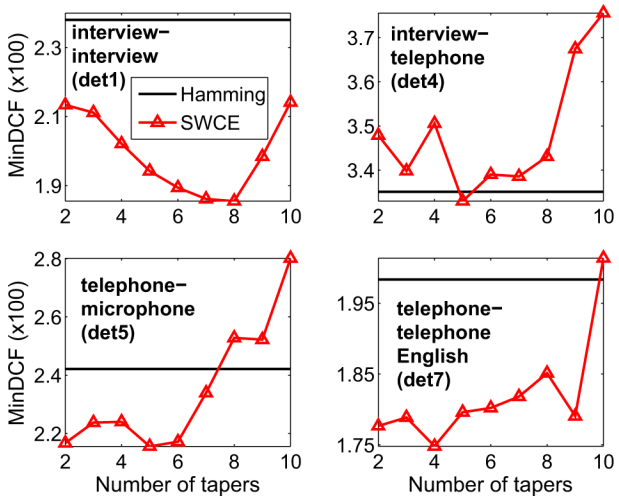


Fig. 10. MinDCF values on the NIST 2008 core task.

In the following, we fix $K = 6$ for all the four subconditions and for both genders. Gender-pooled results (without any score calibration) for each subcondition are given in Table III for both recognizers. Additionally, Fig. 11 displays the DET plot for the interview-interview condition (det1). From Table III, we observe the following;

- Except for det4 in GMM-SVM system, SWCE systematically outperforms Hamming in both EER and MinDCF.
- For GMM-SVM, det7 task was observed the highest EER improvement of 21.9%, while det1 task the highest MinDCF improvement of 20.4%. The largest overall improvements are in det1 where both metrics decrease by about 20%.
- For GMM-JFA, det4 task was observed the highest EER improvement of 16.9%, while det7 task the highest MinDCF improvement of 18.7%.

The DET plots in Fig. 11 further confirm that both recognizers benefit from multitapering over a wide range of operating points. Our JFA result is roughly on the same range as other similar systems, such as the full JFA system in [19, Table IX]. The *i-vector* system in [19] outperforms our JFA result on det7

TABLE III
RESULTS ON THE DIFFERENT SUB-CONDITIONS OF THE NIST SRE 2008 CORE TASK (SHORT2–SHORT3) USING $K = 6$ TAPERS. det1=INTERVIEW TRAINING AND TESTING; det4= INTERVIEW TRAINING, TELEPHONE TESTING; det5= TELEPHONE TRAINING, NON-INTERVIEW MICROPHONE TESTING; det7= TELEPHONE TRAINING AND TEST INVOLVING ENGLISH LANGUAGE ONLY

| | EER (%) | | | 100 ×MinDCF | | |
|---|---|---|---|---|---|---|
| | Hamm. | SWCE | Impr. (%) | Hamm. | SWCE | Impr. (%) |
| **GMM-SVM** recognizer | | | | | | |
| det1 | 4.58 | 3.70 | 19.1 | 2.38 | 1.89 | 20.4 |
| det4 | 9.85 | 10.68 | −8.4 | 3.35 | 3.39 | −1.2 |
| det5 | 6.73 | 5.71 | 15.2 | 2.42 | 2.17 | 10.3 |
| det7 | 5.09 | 3.97 | 21.9 | 1.98 | 1.80 | 9.1 |
| **GMM-JFA** recognizer | | | | | | |
| det1 | 5.36 | 4.73 | 11.8 | 2.95 | 2.55 | 13.7 |
| det4 | 7.51 | 6.24 | 16.9 | 3.14 | 2.82 | 10.2 |
| det5 | 6.79 | 5.91 | 13.0 | 2.45 | 2.23 | 9.1 |
| det7 | 3.58 | 3.48 | 2.9 | 1.58 | 1.28 | 18.7 |

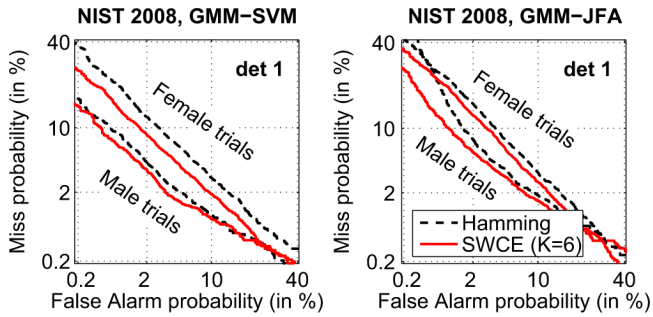


Fig. 11. DET plots for the interview-interview data (det1) on NIST 2008. The SWCE method uses $K = 6$ tapers.

(for instance, EERs of 2.9% and 1.1% were reported for male and female trials). Since i-vector and GMM-JFA share almost the same components—factor analysis on GMM supervectors with eigenvoice adaptation—we expect the results, to a certain extent, to generalize to i-vector classifier as well. In fact, preliminary indication of this was recently given in [51] on the SRE 2010 corpus using an independent implementation.

## VII. CONCLUSION

We have advocated the use of multitaper MFCC features in speech processing. By replacing the windowed DFT with multitaper spectrum estimate, we found systematic improvements in three independently constructed recognition systems (GMM-UBM, GMM-SVM and GMM-JFA). The improvements were consistent on two very different corpora (NIST 2002 and NIST 2008) including telephony, microphone and interview segments with severe cross-channel variabilities. These observations, together with analysis of bias and variance on TIMIT, gives us confidence to recommend using multitapers in speaker verification and possibly other speech processing tasks.

The choice of the multitaper type (Thomson, multipeak, SWCE) was found less important than the choice of the number of tapers, $K$, but even the exact choice of $K$ does not appear to be critical. The best results were obtained, in all cases, for $3 \leq K \leq 8$. We recommend to start with $K = 6$. Mel-warping turned out useful also with multitapers. To help the interested reader in exploring the technique further, we provide a sample implementation in the Appendix.

It would be also interesting to see whether variance reduction would lead to higher gains in short duration recognition tasks (10-second) and in speech and language recognition problems. Finally, we expect further improvements using alternative feature normalization strategies that suit better for low-variance MFCCs.

## APPENDIX

Below is an example of multitaper spectrum estimation in Matlab where `SWCE` function produces the tapers and their weights in the SWCE method [9]. The function `multitaperspectra`, which can be also used with other types of tapers, is used for spectrum estimation. For a more complete package, including Thomson [24] and multipeak [26] implementations, refer to the WWW pages of the first author, http://cs.joensuu.fi/pages/tkinnu/webpage/. To generate Thomson's tapers, you may also use the function `dpss` in Matlab's signal processing toolbox.

```
function [h,s] = SWCE(N, K)

% Sine-weighted cepstrum estimator (SWCE) tapers
% N = frame size (samples), K = #tapers
% The tapers are columns of h, their weights in s.

M = fix(N/K);
for i=1:K
    h(:,i)=sqrt(2/(N+1))*sin((pi*i*[1:N]')/(N+1));
end
s=((cos(2*pi*[0:fix(N/M)-1]'*M/N/2))+1) ...
    ./sum(cos(2*pi*[0:fix(N/M)-1]'*M/N/2)+1);
```

```
function spec = multitaperspectra(frames, tapers,
    weights, NFFT)

% Compute multitaper power spectra.
% frames:   (num_frames x N) matrix of frames.
% tapers:   (N x K) matrix of K tapers.
% weights:  (K x 1) vector of taper weights.
% NFFT:     Number of FFT bins.
% spec:     Multitaper power spectra as columns.
%
% Note: the frames should NOT be windowed using
% Hamming/Hann etc type of windows. Give the raw
% "boxcar"-windowed frames as input instead.

spec = zeros(NFFT, size(frames', 2));
for (taper_nbr = 1:size(tapers,2))
  spec = spec + weights(taper_nbr)*abs(fft((frames')
      .* ...
       repmat(tapers(:,taper_nbr),1,size(frames',2))
           , NFFT)).^2;
end
spec = spec(1:NFFT/2+1, :);
```

## ACKNOWLEDGMENT

The first author would like to thank professor J. Alho at UEF for useful suggestions to Section IV.

## REFERENCES

[1] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–84, Jan. 1978.

[2] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 64, no. 4, pp. 561–580, Apr. 1975.

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

[4] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, New Jersey: Prentice-Hall, 2001.

[5] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[6] X. Xiao, E.-S. Chng, and H. Li, "Temporal structure normalization of speech feature for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 14, no. 7, pp. 500–503, Jul. 2007.

[7] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey: Speaker Recognition Workshop (Odyssey 2001)*, Crete, Greece, Jun. 2001, pp. 213–218.

[8] C.-P. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 257–270, Jan. 2007.

[9] M. Hansson-Sandsten and J. Sandberg, "Optimal cepstrum estimation using multiple windows," in *Proc. ICASSP '09*, Taipei, Taiwan, Apr. 2009, pp. 3077–3080.

[10] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications*. Cambridge, MA: Cambridge Univ. Press, 1993.

[11] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Comm.*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

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

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

[14] N. Brümmer, L. Burget, J. H. Černocký, O. Glembek, F. Grézl, M. Karafiát, D. A. v. Leeuwen, P. Matějka, P. Schwartz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2072–2084, Sep. 2007.

[15] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 980–988, Jul. 2008.

[16] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[17] A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shriberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 1987–1998, Sep. 2007.

[18] C. H. You, K. A. Lee, and H. Li, "GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1300–1312, Aug. 2010.

[19] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[20] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 412–424, Mar. 2006.

[21] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 600–613, Mar. 2010.

[22] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," , vol. 12, no. 1, pp. 59–67, Jan. 2004.

[23] L. P. Ricotti, "Multitapering and a wavelet variant of MFCC in speech recognition," *IEE Proc. Vis., Image Signal Process.*, vol. 152, no. 1, pp. 29–35, Feb 2005.

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

[25] K. S. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation," *IEEE Trans. Signal Process.*, vol. 43, no. 1, pp. 188–195, Jan 1995.
[26] M. Hansson and G. Salomonsson, "A multiple window method for estimation of peaked spectra," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 778–781, Mar. 1997.
[27] J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P. Flandrin, and P. Borgnat, "Multitaper estimation of frequency-warped cepstra with application to speaker verification," *IEEE Signal Process. Lett.*, vol. 17, no. 4, pp. 343–346, Apr. 2010.
[28] A. T. Walden, E. McCoy, and D. B. Percival, "The variance of multitaper spectrum estimates for real gaussian processes," *IEEE Trans. Signal Process.*, vol. 42, no. 2, pp. 479–482, Feb. 1994.
[29] T. P. Bronez, "On the performance advantage of multitaper spectral analysis," *IEEE Trans. on Sign. Proc.*, vol. 40, no. 12, pp. 2941–2946, Dec. 1992.
[30] P. D. Welch, "The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, no. 2, pp. 70–73, Jun. 1967.
[31] C. H. Shadle and G. Ramsay, "Multitaper analysis of fundamental frequency variations during voiced fricatives," in *Proc. 6th Int. Seminar Speech Product.*, Dec. 2003, p. CD-6.
[32] N. Erdol and T. Gunes, "Multitaper covariance estimation and spectral denoising," in *Proc. Conf. Rec. 39th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2005, pp. 1144–1147.
[33] T. Kinnunen, R. Saeidi, J. Sandberg, and M. Hansson-Sandsten, "What else is new than the Hamming window? Robust MFCCs for speaker recognition via multitapering," in *Proc. Interspeech*, Makuhari, Japan, Sep. 2010, pp. 2734–2737.
[34] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2005)*, Philadelphia, PA, Mar. 2005, pp. 629–632.
[35] P. Kenny, Joint factor analysis of speaker and session variability: Theory and algorithms Tech. Rep. CRIM-06/08-14, 2006.
[36] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, May 2007.
[37] D. J. Thomson, "Jackknifing multitaper spectrum estimates," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 20–30, Jul. 2007.
[38] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4165–4174, Nov 2009.
[39] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, Mar. 1978.
[40] T. Schneider and A. Neumaier, "Algorithm 808: ARfit-a Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models," *ACM Trans. Math. Softw.*, vol. 27, pp. 58–65, 2001.
[41] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 42–54, Jan. 2000.
[42] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in text-independent speaker verification," in *Proc. Interspeech '05*, Lisbon, Portugal, Sep. 2005, pp. 3117–3120.
[43] R. Saeidi, H. R. S. Mohammadi, T. Ganchev, and R. D. Rodman, "Particle swarm optimization for sorted adapted Gaussian mixture models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 344–353, Feb. 2009.
[44] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 599–602, Jun. 2010.
[45] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku, "Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions," in *Proc. Interspeech '10*, Makuhari, Japan, Sep. 2010, pp. 1477–1480.
[46] H. Li, B. Ma, K.-A. Lee, H. Sun, D. Zhu, K. C. Sim, C. You, R. Tong, I. Kärkkäinen, C.-L. Huang, V. Pervouchine, W. Guo, Y. Li, L. Dai, M. Nosratighods, T. Tharmarajah, J. Epps, E. Ambikairajah, E.-S. Chng, T. Schultz, and Q. Jin, "The I4U system in NIST 2008 speaker recognition evaluation," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP '09)*, Taipei, Taiwan, Apr. 2009, pp. 4201–4204.
[47] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for speaker verification in additive noise," in *Proc. Odyssey 2010: Speaker Lang. Recogni. Workshop*, Brno, Czech Republic, Jun. 2010.
[48] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.
[49] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. 5th Eur. Conf. Speech Commun. Technol. (Eurospeech '97)*, Rhodos, Greece, Sep. 1997, pp. 1895–1898.
[50] D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten, "NIST and NFI-TNO evaluations of automatic speaker recognition," *Comput. Speech Lang.*, vol. 20, pp. 128–158, Apr.–Jul. 2006.
[51] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper MFCC features for speaker verification using i-vectors," in *Proc. IEEE Autom. Speech Recognit. Understanding (ASRU 2011)*, Dec. 2011, pp. 547–552.

**Tomi Kinnunen** (M'12) received the M.Sc. and Ph.D. degrees in computer science from the University of Joensuu, Joensuu, Finland, in 1999 and 2005, respectively.

He was an Associate Scientist at the Institute for Infocomm Research ($I^2R$), Singapore, and as a Senior Assistant in the Department of Computer Science and Statistics, University of Joensuu. He is currently a Post-Doctoral Researcher at the University of Eastern Finland, Joensuu, and his research is funded by the Academy of Finland. His research areas cover speaker recognition and speech signal processing.

**Rahim Saeidi** (S'09) received the B.Sc. degree in electrical engineering from Azad University-Save branch, Saveh, Iran, in 2002, the M.Sc. degree in telecommunication systems engineering from the Iran University of Science and Technology, Tehran, Iran, in 2005, and the Ph.D. degree from the University of Eastern Finland, Joensuu, in 2011.

He is currently a Postdoctoral Researcher at Radboud University, Nijmegen, The Netherlands. His research interests include speech processing, machine learning, neuroscience, and pattern recognition.

**Filip Sedlák** was born in Brno, Czech Republic, in 1985. He received the B.Sc. degree from the Faculty of Information Technology (FIT), Brno University of Technology (BUT). He is currently pursuing the M.S. degree in the School of Computing, University of Eastern Finland (UEF), Joensuu.

In 2008, he moved to UEF as an exchange student. From February 2010 to July 2010, he was visiting the Institute for Infocomm Research (I2R) and Nanyang Technological University (NTU), Singapore, as an intern.

**Kong Aik Lee** received the B.Eng. (first class honors) degree from the University Technology Malaysia, Johor, in 1999 and the Ph.D. degree from Nanyang Technological University, Singapore, in 2006.

He is currently a Senior Research Fellow with the Human Language Technology Department, Institute for Infocomm Research ($I^2R$), Singapore. His research focuses on statistical methods for speaker and spoken language recognition, adaptive echo and noise control, and subband adaptive filtering. He is the leading author of the book *Subband Adaptive Filtering: Theory and Implementation* (Wiley, 2009).

**Johan Sandberg** was born in Sweden in 1980. He received the M.Sc. degree in engineering physics and the Ph.D. degree in mathematical statistics from the Centre for Mathematical Sciences, Lund University, Lund, Sweden, in 2005 and 2010, respectively.

His research field has included time–frequency analysis of time series, spectrum and cepstrum estimation and speech analysis. He is with Nordea Bank, Copenhagen, Denmark, where his research interest is focused on modeling of currency exchange rates and interest rates.

**Maria Hansson-Sandsten** (S'90–M'96) was born in Sweden in 1966. She received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Lund University, Lund, Sweden, in 1989 and 1996, respectively.

Currently, she is a Professor in mathematical statistics with a speciality towards statistical signal processing at the Centre for Mathematical Sciences, Lund University. Her current research interests include multitaper spectrum analysis and time–frequency analysis of stochastic processes with application areas of electroencephalogram signals, heart rate variability signals, and speech signals.

**Haizhou Li** (M'92–SM'01) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronic engineering from the South China University of Technology (SCUT), Guangzhou, in 1984, 1987, and 1990, respectively.

He is currently the Department Head of the Human Language Technology, and the Director of Baidu-I2R Research Centre at the Institute for Infocomm Research, Singapore. He is also a conjoint Professor at the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia. He has worked on speech and language technology in academia and industry since 1988. He taught in the University of Hong Kong (1988-1990), South China University of Technology (1990-1994), and Nanyang Technological University (2006-). He was a Visiting Professor at CRIN/INRIA in France (1994-1995). He was appointed as Research Manager in the Apple-ISS Research Centre (1996-1998), Research Director in Lernout & Hauspie Asia Pacific (1999-2001), and Vice President in InfoTalk Corp. Ltd. (2001-2003). His research interests include automatic speech recognition, natural language processing, and information retrieval. He has published over 200 technical papers in international journals and conferences.

Dr Li. now serves as an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, *ACM Transactions on Speech and Language Processing*, and *Springer International Journal of Social Robotics*. He is an elected Board Member of the International Speech Communication Association (ISCA, 2009-2013), the President of the Chinese and Oriental Language Information Processing Society (COLIPS, 2011-2013), an Executive Board Member of the Asian Federation of Natural Language Processing (AFNLP, 2006-). He served as the Local Arrangement Chair of SIGIR 2008 and ACL-IJCNLP 2009. He was appointed the General Chair of ACL 2012 and INTERSPEECH 2014. He was the recipient of National Infocomm Award of Singapore in 2001. He was named one of the two Nokia Visiting Professors 2009 by the Nokia Foundation in recognition of his contribution to speaker and language recognition technologies.