What Else is New Than the Hamming Window? Robust MFCCs for Speaker Recognition via Multitapering

Tomi Kinnunen[1], *Rahim Saeidi*[1], *Johan Sandberg*[2] *and Maria Hansson-Sandsten*[2]

[1]School of Computing, University of Eastern Finland, Joensuu, Finland
[2]Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Lund, Sweden
{tomi.kinnunen,rahim.saeidi}@uef.fi, {sandberg, sandsten}@maths.lth.se

Abstract

Usually the mel-frequency cepstral coefficients (MFCCs) are derived via Hamming windowed DFT spectrum. In this paper, we advocate to use a so-called multitaper method instead. Multitaper methods form a spectrum estimate using multiple window functions and frequency-domain averaging. Multitapers provide a robust spectrum estimate but have not received much attention in speech processing. Our speaker recognition experiment on NIST 2002 yields equal error rates (EERs) of 9.66 % (clean data) and 16.41 % (-10 dB SNR) for the conventional Hamming method and 8.13 % (clean data) and 14.63 % (-10 dB SNR) using multitapers. Multitapering is a simple and robust alternative to the Hamming window method.
Index Terms: speaker verification, multiple window method

1. Introduction

Current speech, speaker and language recognition applications perform well under clinical laboratory setting but robust recognition under variable environments, handsets and transmission channels remains a constantly challenging problem. A major source of problems are the spectral front-ends based on either discrete Fourier transform (DFT) or linear prediction (LP). The short-term spectrum is subject to many harmful variations. Due to such variations, complex feature normalization, channel compensation and score normalization are required [1].

In this paper, our focus is on the most popular speech front-end, the *mel-frequency cepstral coefficients* (MFCCs) [2]. MFCC computation begins by multiplying a short-term frame of speech by a tapered window function [3] and computing the DFT of the windowed frame. The DFT magnitude spectrum is then smoothed by using a psychoacoustically motivated filterbank, followed by logarithmic compression and, finally, discrete cosine transform (DCT). The final feature vector is usually appended with the first and second order time derivatives (Δ and Δ^2 features) and further processed by cepstral mean and variance normalization (CMVN) and other feature normalizations. In this paper our goal is to make the first step, computation of the base MFCCs, more robust.

From a statistical point of view, we imagine that, for every short-term speech frame there exists a "true" *random process* which generates that particular frame; an example would be a digital filter driven with random inputs but with fixed filter coefficients. For speech signals, we imagine that there exists a speaker- and phoneme-dependent random process from which the actual speech sounds are generated from. This abstract viewpoint, in the context of automatic speaker recognition, is well-modeled by the Gaussian mixture model (GMM) back-end for cepstral features [4]. The means of GMM represent speaker-dependent information and variances model uncertainty in the observed vectors. In this paper, our goal is to reduce that uncertainty by using better MFCC estimator. For different acoustic realizations of the same phoneme spoken by the same speaker, a good MFCC estimator would produce "similar" MFCC vectors. In statistical terms, we wish to have an MFCC estimator with small *variance*. Naturally, we should also require the estimated cepstrum to be, on average, close to the true cepstrum and therefore have small *bias*. These bias and variance [5] can quantitatively be analyzed, without any model of the speech production mechanism itself, but by imposing a mathematical model of the random process corresponding to a single speech frame (e.g. Gaussian zero-mean stationary process as in [6]).

The bias and variance can intuitively be understood by considering the degree of smoothness in a spectrum estimate. Smooth spectrum, such as the DFT spectrum after MFCC filterbank averaging or an all-pole spectrum [7, 8] with a small number of poles, have a small variance because they produce similar spectra for different instances of the same random process. However, over-smoothing increases the bias because of decreased spectral resolution. A good spectrum (and cepstrum) estimate, therefore, should have low variance to be robust against noise and other nuisance factors but also retain low enough bias to be accurate enough representation for the given classification task.

What are the bias and variance of the typical MFCC estimation procedure and could they be improved? We first note that the Hamming-type of time-domain window reduces the *spectral leakage* resulting from the convolution of the signal and window function spectra. The windowing, therefore, reduces the bias. The variance, unfortunately, remains high [5]. One way to reduce the variance of the MFCC estimator is to replace the Hamming window DFT spectrum estimate by a so-called *multitaper* spectrum estimate [6, 9, 10]. The idea in multitapering, as illustrated in Fig. 1, is to pass the analysis frame through different window functions and form the final spectrum estimate as a weighted average of the individual *sub-spectra*. The window functions or *tapers* are designed so that the estimation errors in the individual sub-spectra are approximately uncorrelated. Averaging these uncorrelated spectra gives a low-variance spectrum estimate and, consequently, low-variance MFCC estimate as well. The multitaper method is similar to the well-known *Welch's method* which forms a *time*-averaged spectrum over multiple frames. Multitapers, however, focus only on *one* frame

The work of T. Kinnunen was supported by the Academy of Finland (project no 132129) and the work of R. Saeidi was supported by a scholarship from the Finnish Foundation for Technology Promotion (TES). Computing services from *CSC - IT Center for Science* were used for the experiments (project no uef4836).

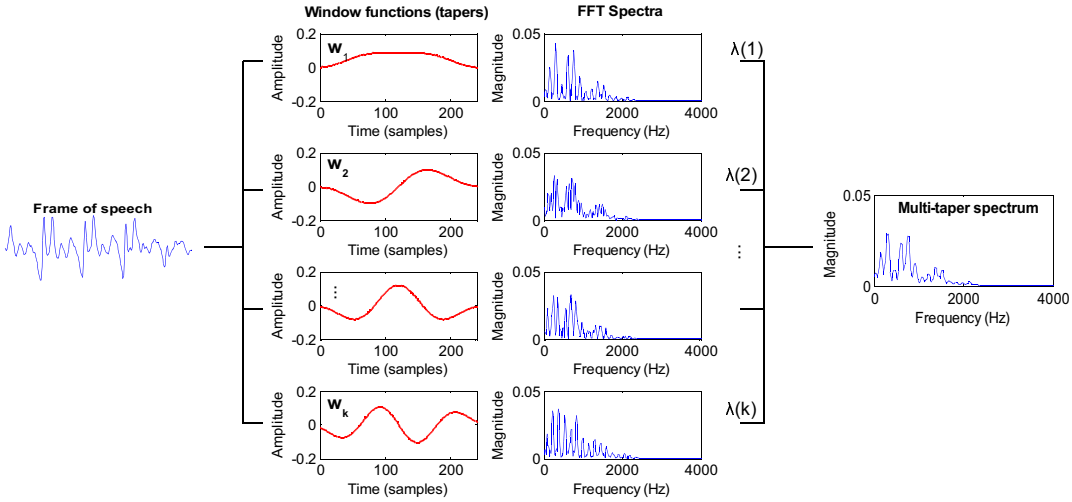

Figure 1: Multiple window method of spectrum estimation analyzes data using independent windows which lead to slightly different magnitude spectra. The final spectrum is formed as a weighted average of the individual spectra. The averaging reduces the *variance* of the spectrum estimate, therefore making the spectrum less sensitive to noise compared to the conventional single-window method.

and therefore make more efficient use of the limited data.

Although multitapering guarantees low variance spectrum estimate, it has not gained much attention in speech processing so far [11]. One reason could be that, since there exists a number of different multitapers to choose from, it may not be clear which suits well for modeling speech signals. It is our goal, therefore, to carry out a comparative evaluation of different multitaper techniques and compare their performance to conventional single-taper technique.

2. MFCC Computation via Multitapering

Let $\mathbf{x} = [x(0) \ \ldots \ x(N-1)]^T$ denote one frame of speech. The most popular spectrum estimate in speech processing, the *windowed periodogram*, 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 $f \in \{0, 1, \ldots, N-1\}$ denotes the discrete frequency index and $\mathbf{w} = [w(0) \ \ldots \ w(N-1)]^{\mathrm{T}}$ is a time-domain window function which usually is symmetric and decreases towards the frame boundaries (e.g. Hamming). From a statistical perspective, the use of a Hamming-type of window reduces the bias of the spectrum, i.e. how much the estimated spectral density value $\hat{S}(f)$ differs from the true value $S(f)$, on average. But the estimated spectrum still has large variance. To reduce the variance, *multitaper* spectrum estimator [5, 9, 12] can be used:

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

Here, k multitapers $\mathbf{w}_j = [w_j(0) \ \ldots \ w_j(N-1)]^T$, $j = 1, \ldots, k$, are used with corresponding weights $\lambda(j)$. The multitaper estimate is therefore obtained as a weighted average of k individual sub-spectra (Fig. 1). The conventional single-window method (1) is obtained as a special case when $k = 1$ and $\lambda = 1$.

A number of different tapers have been proposed in literature for spectrum estimation, such as *Thomson* [9], *sine* [10] and *multipeak* tapers [12]. For cepstrum analysis, the sine tapers are applied with optimal weighting in [13]. 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 general, the tapers are designed so that the estimation errors in the individual subspectra will be approximately uncorrelated, which is the key to variance reduction. The details of finding the optimal tapers for a given process is out of the scope of the current paper but for the interested reader we mention that the solution is obtained from an eigenvalue problem where the eigenvectors and -values correspond to the tapers and their weights, respectively. Additional constraints are often added to the optimization problem to force the designed tapers be robust against violated assumptions of the random process. In practice, many multitapers work well even though designed for another process. For instance, the Thomson window [9], designed for white noise, perform well for any smooth spectrum.

3. Noise Robustness of Multitapering

Figure 2 demonstrates the use of multitaper spectrum estimation for analysis of speech under additive factory noise corruption. The left panel shows spectrum estimate using the conventional single-taper (Hamming) method whereas the right panel shows spectrum estimate using multipeak tapers with $k = 6$ tapers. The upper lines (blue) correspond to clean speech and the lower lines (red) to noisy speech. The single-taper spectrum contains more details and shows large difference between the clean and the noisy frame. The multitaper spectra, in turn, are smooth and look visually more similar between the clean and the noisy version. In short, the multitaper method has smaller variance.

To understand the bias and variance trade-off better, we consider the variance and spectrum resolution of the single- and multi-taper methods. For the windowed periodogram (1), the variance is usually approximated as [5]

$$V[\hat{S}(f)] \approx S^2(f). \qquad (3)$$

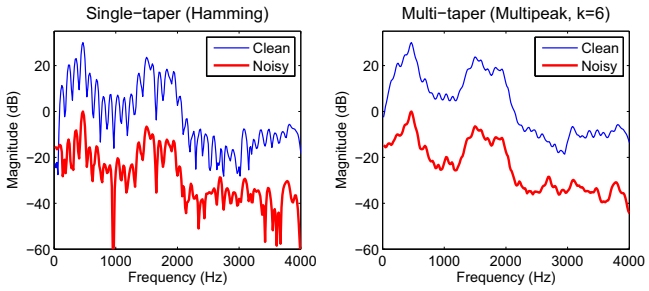

Figure 2: Single- and multitaper methods under additive noise. The spectra in each plot have been shifted for visualization.

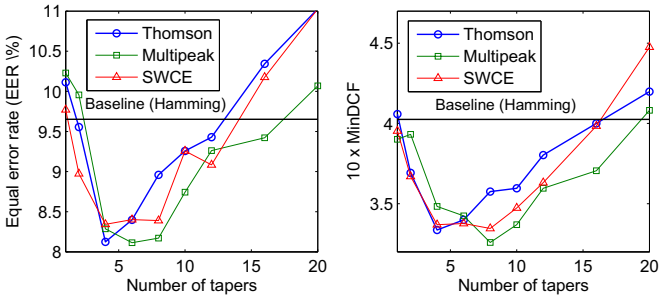

Figure 3: Effect of the number of tapers to EER and MinDCF.

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. This suggests that, even though Hamming window reduces the spectral bias, it has twice as poor spectral resolution as the rectangular window. Note also that (3) does not depend on the window length N and thus, more data will *not* reduce the variance.

For the multiple window spectrum estimator (2), the spectral resolution is approximately $B_w = k/N$ [9] and the variance can be approximated as

$$V[\hat{S}(f)] \approx \frac{1}{k} S^2(f). \quad (4)$$

This result is analogous to the known fact that variance of the mean of sample of size k is inversely proportional to k. The formula is approximately valid also for the Welch's method with 50% overlap between the windows [5]. The formula (4) suggests that, by increasing the number of tapers, we can reduce the variance of the spectrum estimate, hence making the spectrum more robust across random variations. The robustness, however, is traded off with spectrum resolution. We expect an optimal number of tapers to be a compromise between robustness and resolution, as we shall demonstrate by speaker recognition experiments in Section 5.

Note that the formulae (3) and (4) consider variance in spectral and not MFCC domain which are generally different due to the logarithmic compression and cosine transform. Nevertheless, if the estimated spectrum deviates from the true spectrum, so will the resulting MFCC vector deviate from the true MFCC. For a mathematical treatment of MFCC bias and variance, refer to the recent studies [6, 14].

4. Speaker Verification Setup

We use the NIST 2002 speaker recognition evaluation (SRE) corpus for the experiments. It contains 139 males and 191 females and there are 2982 genuine and 36,277 impostor trials. For the baseline *Hamming* method, we compute the MFCCs as it is usually done: Hamming windowing, DFT magnitude spectrum, 27 mel-frequency spaced filters, logarithm and discrete cosine transform. We keep the lowest 18 MFCCs, excluding c_0 as usual. For the *Thomson* [9], *multipeak* [12] and *sine-weighted cepstrum estimator* (SWCE) [13] methods, we proceed similarly but estimate the magnitude spectrum using the multitaper spectrum estimator as described in Section 2. The complete front-end includes RelAtive SpecTrAl (RASTA) filter for the base MFCCs, Δ and Δ^2 coefficients, an energy-based voice activity detector (VAD) and, finally, cepstral mean and variance normalization (CMVN).

We utilize a standard Gaussian mixture model with universal background model (GMM-UBM) [4] as our system backend, with test normalization (Tnorm) [15] applied on the log likelihood ratio scores. We have utilized the same system recently in [8, 16]. The data for UBM training and T-norm models were drawn from the NIST 2001 corpus.

In comparison of the different MFCC estimation methods, we consider both the equal error rate (EER) and the minimum detection cost function value (MinDCF). EER is the error rate for which the miss rate (P_{miss}) and the false alarm rate (P_{fa}) are equal. MinDCF, in turn, is used in the NIST speaker recognition evaluations and defined to be the minimum of the error functional $0.1 \times P_{\text{miss}} + 0.99 \times P_{\text{fa}}$. In addition, we display selected detection error tradeoff (DET) curves for the entire tradeoff of false alarm and miss rates.

For systematic study of the robustness of the four feature sets, we consider their performance under additive *factory noise* degradation (noise drawn from the NOISEX-92 corpus).The UBM and target model training data are kept untouched, but the noises are added to the test files with a given average segmental signal-to-noise ratio (SNR). We consider five SNR levels: clean, 20, 10, 0, and -10 dB, where "clean" refers to the original NIST samples. To focus on differences of the spectrum estimation methods and not to accuracy of the energy VAD (whose accuracy degrades to near-unusable level at SNRs less than 0 dB), we use VAD labels derived from the clean signal in all cases.

5. Speaker Recognition Results

We first study how the number of tapers affects the accuracy on the original NIST data. We compare the EERs and MinDCFs of the three multitaper techniques and also show the conventional single-taper method (Hamming window) as a reference in Fig. 3. Firstly, all the multitaper methods outperform the single-taper method, even when the number of tapers is not set to optimum; the error rates are lower than for the baseline for $4 \leq k \leq 12$. Secondly, the multitaper methods show convex error curves as hypothesized; too low an order increases the variance whereas too high an order increases the bias of the spectrum estimate. The optimum number of tapers, for this dataset, is on the range $4 \leq k \leq 8$.

We next study the performance of the methods under the additive factory noise corruption. Based on Fig. 3, we fix the number of tapers as $k = 4$ for Thomson and $k = 8$ for both multipeak and SWCE. The results are shown in Table 1 and Fig. 4. All methods significantly degrade with decreasing SNR as expected. It is also clear that multitaper methods outperform

Table 1: System performance under factory noise corruption (18 MFCCs). For each row, the best EER and MinDCF are bolded.

Signal-to-noise ratio (dB)	Equal error rate (EER %)				MinDCF			
	Hamming	Thomson	Multipeak	SWCE	Hamming	Thomson	Multipeak	SWCE
clean	9.66	**8.13**	8.23	8.35	4.03	3.33	**3.25**	3.37
20	10.23	8.40	**8.34**	8.74	4.13	**3.43**	3.45	3.47
10	10.45	8.57	8.92	**8.69**	4.15	3.60	**3.53**	3.54
0	11.54	10.29	**10.28**	10.45	5.02	4.46	**4.24**	4.30
-10	16.41	15.49	15.60	**14.63**	7.42	7.11	**6.76**	6.78

the baseline (Hamming) in both EER and MinDCF. From the three multitaper methods Thomson works best on clean data whereas multipeak and SWCE perform better at lower SNRs.

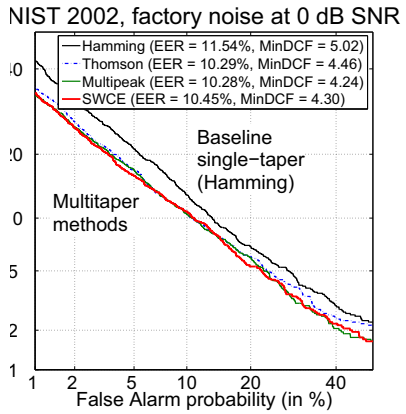

Figure 4: Performance of single- and multitaper methods under additive factory noise (0 dB SNR).

6. Conclusions

In this paper we have promoted to use multitapers for robust MFCC extraction. Our speaker verification results indicate that multitapers, independent of the chosen taper type, clearly outperform conventional single-window technique. We observed this on both clean and noisy data, suggesting insensitivity to both the type and number of tapers (which was optimized on clean data). Due to their (slightly) better performance on the noisier conditions, we recommend to use either the multipeak or the SWCE tapers instead of Thomson. A good choice for the number of tapers is 4 to 8. In conclusion, multitapers are simple and robust alternative for the conventional single-window methods.

7. References

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

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

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

[4] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, January 2000.

[5] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications*. Cambridge University Press, 1993.

[6] J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P. Flandrin, and P. Borgnat, "Multitaper estimation of frequency-warped cepstra with application to speaker verification," *IEEE Signal Processing Letters*, vol. 17, no. 4, pp. 343–346, April 2010.

[7] J. Makhoul, "Linear prediction: a tutorial review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 561–580, April 1975.

[8] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *Sign. Proc. Lett.*, 2010.

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

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

[11] Y. Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE T. Speech, Audio & Lang. Proc.*

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

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

[14] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Trans. on Signal Proc.*, vol. 57, no. 11, pp. 4165–4174, Nov 2009.

[15] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, January 2000.

[16] R. Saeidi, H. Mohammadi, T. Ganchev, and R.D.Rodman, "Particle swarm optimization for sorted adapted gaussian mixture models," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 2, pp. 344–353, February 2009.