

A Study of Low-variance Multi-taper Features for Distributed Speech Recognition

Md Jahangir Alam^{1,2}, Patrick Kenny¹, Douglas O'Shaughnessy²,

¹ CRIM, Montreal, Canada

{Janagir.Alam, Patrick.Kenny}@crim.ca

² INRS-EMT, University of Quebec, Montreal, Canada

dougo@emt.inrs.ca

Abstract. In this paper we study low-variance multi-taper spectrum estimation methods to compute the mel-frequency cepstral coefficient (MFCC) features for robust speech recognition. In speech recognition, MFCC features are usually computed from a Hamming-windowed DFT spectrum. Although windowing helps in reducing the bias of the spectrum, but variance remains high. Multi-taper spectrum estimation methods can be used to correct the shortcomings of single taper (or window) spectrum estimation methods. Experimental results on the AURORA-2 corpus show that the multi-taper methods, specifically the multi-peak multi-taper method, perform better compared to the Hamming-windowed spectrum estimation method.

Keywords: Speech recognition, multi-taper spectrum, AURORA-2.

1 Introduction

Useful information extraction from speech has been a subject of active research for many decades. A feature extractor (or front-end) is the first step in an automatic speech or speaker recognition system, which transforms a raw signal into a compact representation. Since feature extraction is the first step in the chain, the quality of later steps (modelling and pattern matching/classification) strongly depends on it. The MFCC features are the most popular in speech and speaker recognition systems and they demonstrate good performance in speech and speaker recognition. The MFCC representation is an approximation of the structure of the human auditory system [1]. Since MFCC features are computed from an estimated spectrum, it is crucial that this estimate is accurate. Usually, the spectrum is estimated using a windowed periodogram. Despite having low bias, a windowed periodogram has large variance and therefore, MFCC features computed from this estimated spectrum have also high variance. One elegant technique for reducing the variance is to replace a windowed periodogram estimate with a multi-taper spectrum estimate [4-6].

The multi-taper methods reduce the variance of the spectral estimates by using multiple time-domain window functions or tapers rather than a single taper. The multi-taper method has been widely used in geophysical applications and has been shown in multiple cases to outperform the windowed periodogram. It has also been

used in speech enhancement application [2] and, recently, in speaker recognition [3] with promising preliminary results, but not in speech recognition. In this paper, our aim is to compute MFCC features from a multi-taper spectral estimate for speech recognition on the AURORA-2 task [7].

2 Multi-tapering Background

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

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

where $k \in \{0, 1, \dots, K-1\}$ denotes the frequency bin index, N is the frame length, $s(m,j)$ is the time domain speech signal and $w(j)$ denotes the time domain window function called a taper which usually is symmetric and decreases towards the frame boundaries (e.g., Hamming). Eq. (1) is sometimes called a single-taper, modified or windowed periodogram. If $w(j)$ is a boxcar function, eq. (1) is called the periodogram. Windowing reduces the bias, i.e., difference between the estimated spectrum and the actual spectrum, but it does not reduce the variance of the spectral estimate [8] and therefore, the variance of the MFCC features computed from this estimated spectrum is also large. One way to reduce the variance of the MFCC estimator is to replace the windowed periodogram estimate by a so-called multi-taper spectrum estimate [4-6]. The multi-taper spectrum estimator is given by

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

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 $\lambda(p)$ is the weight corresponding to the p -th taper. The tapers $w_p(j)$ are chosen to be orthonormal, i.e.,

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

The multi-taper spectrum estimate is therefore obtained as the weighted average of M individual sub-spectra. Eq. (1) can be obtained as a special case of eq. (2) when $M=1$ and $\lambda(p)=1$. The tapers in the multi-taper method are chosen so that the estimation errors in the individual sub-spectra are uncorrelated. Averaging these uncorrelated spectra gives a low variance spectrum estimate and, consequently low variance MFCC estimate as well. The underlying details of multi-taper method is similar to Welch's modified periodogram [8]; it, however, focus only on one frame rather taking time-averaged spectrum over multiple frames.

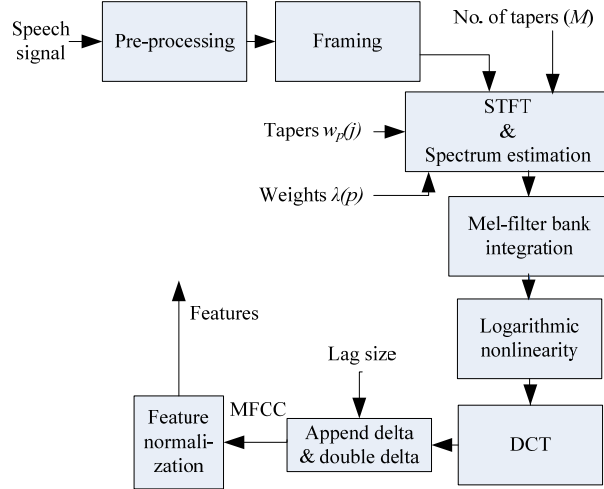


Fig. 1. Generalized block diagram for the single and multi-taper spectrum estimation-based MFCC feature extraction. Lag size used for the delta and double delta computation is 3 and 2, respectively. No. of tapers chosen (experimentally) for all multi-taper methods is 6.

The choice of taper has a significant effect on the resultant spectrum estimate. The objective of the taper is to prevent energy at distant frequencies from biasing the estimate at the frequency of interest. Various tapers have been proposed in the literature for spectrum estimation. A good set of M orthonormal data tapers with good leakage properties are specified from Slepian sequences (also called discrete prolate spheroidal sequences (dpss)) [5]. Another orthogonal family of tapers is the *sine* tapers given by [6]:

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

In [9] the *sine* tapers are applied with optimal weighting for cepstrum analysis and multi-peak tapers are designed for peaked spectra in [10]. As there exists a number of different multi-tapers to choose from, it may not be clear which multi-taper suits well for modeling speech signals. In this paper, our goal is to do a comparative evaluation of various multi-taper methods for MFCC estimation and compare their performance with the conventional Hamming window-based MFCC estimation, in the context of speech recognition on the aurora-2 task [7].

Fig. 1 shows the generalized block diagram of the MFCC feature extraction from the single and multi-taper spectrum estimates. The pre-processing step includes pre-emphasizing, DC removal and signal normalization. As we mentioned above, the Hamming-windowed spectrum estimates can be obtained as a special case of the multi-taper spectrum estimation method. To compute MFCC features from single taper (or window) spectrum estimates we use $M=1$, $\lambda(1)=1$, and $w_1(j)$ is the Hamming window. The analysis frame length is 25 ms with a frame shift of 10 ms.

Delta and double coefficients were calculated using a 3- and 2-frame window, respectively. The MFCC features are normalized using the widely used mean and variance normalized (MVN) technique over the entire utterance.

3 Experiments

3.1 Experimental setup

The AURORA-2 database is used for comparing the performances of the multi-taper methods to the conventional Hamming window technique. 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 clean speech recordings only from 55 male and 55 female adults. The multi-condition training set consists 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 SNRs (20 dB, 15 dB, 10 dB and 5 dB). Test set A is composed of speech with conditions matched to the multi-condition training set, test set B is composed of speech with non-matched background noise (restaurant, street, airport and train-station) and test set C is composed of speech with partly matched background noise and non-matched convolutional noise (MIRS (modified intermediate reference system) filtered subway and street noise). The clean training set constitutes mismatched training/testing conditions whereas the multi-condition training set constitutes much more matched training/testing conditions.

For our experiments, we use 13 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 window, respectively. For all the systems, the MFCC features are normalized using the conventional mean and variance (MVN) normalization technique over the whole utterance. For the recognition task we use the HTK speech recognizer. In the experiments we use a simple HMM-based system with 16 states per word model, 3 Gaussian components per state.

3.2 Results

We use percentage of word accuracy as a performance evaluation measure for comparing the recognition performances of the multi-taper spectrum estimation method to that of the Hamming windowed periodogram technique. Our four speech recognition systems are:

- Hamming: MFCC features are computed from the Hamming windowed 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].

We try various numbers of tapers starting from 4 to 10 and we have found experimentally that MFCC features extracted from multi-taper spectrum estimator with six tapers perform better than the others.

Tables 1 to 3 present the average word accuracy (averaged over 0-20dB SNRs) for test sets A, B and C, respectively, in clean training condition. Tables 4 to 6 present the average word accuracy (averaged over 0-20dB SNRs) for test sets A, B and C, respectively, in multi-condition training. Multi-taper methods perform better than the Hamming window technique in almost all the cases except one case. In multi-condition training and for test set B, Hamming windowed spectrum estimation method provides better results than the multi-taper methods.

Table 1. Average (0-20dB) word accuracy as percentage for test set A in clean training condition. For each column the best result is in boldface.

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

Table 2. Average (0-20dB) word accuracy as percentage for test set B in clean training. For each column the best result is in boldface.

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

Table 3. Average (0-20dB) word accuracy as percentage for test set C in clean training. For each column the best result is in boldface.

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

Table 4. Average (0-20dB) word accuracy as percentage for test set A in multi-condition training. For each column the best result is in boldface.

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

Table 5. Average (0-20dB) word accuracy as percentage for test set B in multi-condition training. For each column the best result is in boldface.

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

Table 6. Average (0-20dB) word accuracy as percentage for test set C in multi-condition training. For each column the best result is in boldface.

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

4 Conclusion

In this paper we have presented three multi-taper spectrum estimation approaches for low variance MFCC computation and compared their performances, in the context of speech recognition on the AURORA-2 corpus. Experimental results on the clean and multi-condition training mode showed that multi-taper methods performed well (specifically in the clean training condition) compared to the single taper (e.g., Hamming) method. Therefore, multi-taper methods (specifically the multi-peak multi-taper method) can be an alternative to the conventional Hamming window technique for the estimation of low variance MFCC features for robust speech 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, In: *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, In: *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, In: *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, France (2000).
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, In: *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).