MULTIPLE WINDOWED SPECTRAL FEATURES FOR EMOTION RECOGNITION

Yazid Attabi[1, 2], Md Jahangir Alam[1, 3], Pierre Dumouchel[2], Patrick Kenny[1], Douglas O'Shaughnessy[3]

[1]Centre de recherche informatique de Montréal, Montréal, Canada
[2]École de technologie supérieure, Montréal, Canada
[3]INRS-EMT, Montréal, Canada

ABSTRACT

MFCC (*Mel Frequency Cepstral Coefficients)* and PLP *(Perceptual linear prediction coefficients)* or RASTA-PLP have demonstrated good results whether when they are used in combination with prosodic features as suprasegmental (long-term) information or when used stand-alone as segmental (short-time) information. MFCC and PLP feature parameterization aims to represent the speech parameters in a way similar to how sound is perceived by humans. However, MFCC and PLP are usually computed from a Hamming-windowed periodogram spectrum estimate that is characterized by large variance. In this paper we study the effect of averaging spectral estimates obtained using a set of orthogonal tapers (windows) on emotion recognition performance. The multitaper MFCC and PLP are examined separately as short-time information vectors modeled using Gaussian mixture models (GMMs). When tested on the FAU AIBO spontaneous emotion corpus, a relative improvement ranging from 2.2% to 3.9% for both MFCC and PLP systems is achieved by multiple windowed spectral features compared to single windowed ones.

Index Terms— Emotion recognition, multitaper spectrum, MFCC, PLP, speech, GMM

1. INTRODUCTION

Research on the extraction of most discriminate feature sets for emotion recognition from speech has been an area of great focus in several studies in last years. Thousands of paralinguistic features are extracted and used in experiments as a whole set or reduced to a subset using feature selection techniques. These features can be classified to one of three categories: Prosodic such as energy, loudness and duration, Voice Quality such as jitter and shimmer, and Spectral such as LPCC (linear prediction cepstral coefficients). The spectral features have been found, when used in combination to other categories of features (or even as a stand-alone feature vector), to improve (or to achieve good) performance [1-4]. Mel Frequency Cepstral Coefficients (MFCCs) [5] and Perceptual Linear Prediction (PLP, with or without RASTA filtering) [6] are examples of spectral features that achieve good results not only on speech processing in general but also on emotion recognition.

The MFCC and PLP parameterization techniques aim to simulate the way how a sound is perceived by a human. Usually, the spectrum is estimated using a windowed periodogram via the discrete Fourier transformation (DFT) algorithm. Despite having low bias, a consequence of the windowing is increased estimator variance. An elegant technique for reducing the spectral variance is to replace a windowed periodogram estimate with a multiple windowed (or multitaper) spectrum estimate [7, 8].

In the multitaper spectral estimation method, a set of orthogonal tapers is applied to the short-time speech signal and the resulting spectral estimates are averaged, which reduces the spectral variance. The multitaper method has been widely used in geophysical applications and more recently, in speech enhancement applications [9] and in speaker and speech recognition [10-12] and it has been shown to outperform the windowed periodogram.

Our main goal in this paper is to study whether the improvements achieved on speech and speaker verification tasks using multitapering, described in section 2, translate to the emotion recognition problem. Multitaper spectrum estimates are applied to the extraction of both MFCC and PLP features (section 3). Multitaper MFCC and PLP are used in experiments on the FAU AIBO corpus, a well-known spontaneous emotion speech corpus (section 4). The extracted features are used as short-term information and modeled using GMM models, which are presented in section 5. The results are presented in section 6 before drawing a conclusion.

2. MULTITAPER SPECTRUM ESTIMATION

In speech processing applications, the power spectrum is often estimated using a windowed direct spectrum estimator. For the m^{th} frame and k^{th} frequency bin an estimate of the windowed periodogram (called also single-taper) can be formulated as:

$$\hat{S}_d(m,k)=\left|\sum_{j=0}^{N-1} w(j)s(m,j)e^{-\frac{i2\pi jk}{N}}\right|^2, \qquad (1)$$

where $k \in \{0,1,\cdots,K-1\}$ 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, also known as taper. The taper, such as *Hamming* window, is usually symmetric and decreases towards the frame boundaries.

Windowing reduces the bias, i.e., expected value of the difference between the estimated spectrum and the actual spectrum, but it does not reduce the variance of the spectral estimate [13]. To reduce the variance of the MFCC or PLP estimator, the multitaper spectrum estimate is used instead of the windowed periodogram estimate [7, 8, 14].

The multitaper spectrum estimator, which uses M orthogonal window functions rather than a single window, can be expressed as

$$\hat{S}_{MT}(m,k)=\sum_{p=1}^{M}\lambda(p)\left|\sum_{j=0}^{N-1} w_p(j)s(m,j)e^{-\frac{i2\pi jk}{N}}\right|^2, \qquad (2)$$

where N is the frame length and w_p is the p^{th} data taper (p = 1, 2,...,.M) used for spectral estimate $\hat{S}_{MT}(\cdot)$. Finally, $\lambda(p)$ is the weight of the p^{th} taper. The tapers $w_p(j)$ are typically chosen to be orthonormal. The multitaper spectrum estimate is therefore obtained as the weighted average of M individual sub-spectra.

The idea behind multitapering is to reduce the variance of the spectral estimates by averaging M direct spectral estimates, each with a different data taper. If all M tapers are pairwise orthogonal and properly designed to prevent leakage, the resulting multitaper estimates outperform the windowed periodogram in terms of reduced variance, specifically, when the spectrum of interest has high dynamic range or rapid variations [15]. Therefore, the variance of the MFCC and PLP features computed via this multitaper spectral estimate will be low as well.

Various tapers have been proposed in the literature for spectrum estimation. A set of M *orthonormal* data tapers with good leakage properties is given by the *Slepian* sequences (also called discrete prolate spheroidal sequences (dpss)), which are a function of a prescribed mainlobe width [7, 9]. The *Slepian* tapers, which underlie the *Thomson* multitaper method [7], are illustrated in Fig. 1 for M = 6 both in time and frequency domains.

The *sine* tapers are another family of tapers, which are very easy to compute and are pairwise orthogonal, and can be formulated as [8]:

$$w_p(j)=\sqrt{\frac{2}{N+1}}\sin\left(\frac{\pi p(j+1)}{N+1}\right),\ j=0,1,....,N-1 \qquad (3)$$

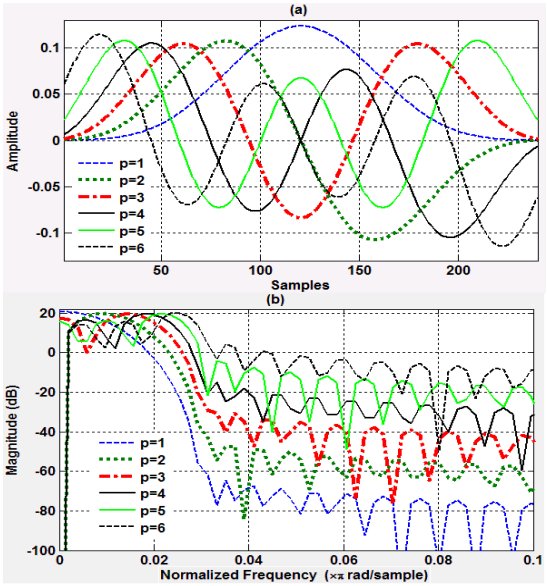

Figure 1. Thomson multitapers for N = 256, M = 6 in (a) time and (b) frequency domains.

The multiplicative constant makes the tapers *orthonormal*. The sine tapers are applied with optimal weighting for cepstrum analysis (called Sinusoidal Weighted Cepstrum Estimator (SWCE)) in [14] and in [16] the multi-peak tapers are designed for peaked spectra. More details about these three tapers can also be found in [10]. In this paper, we use the *Thomson* multitaper [7], the SWCE [14], and the Multi-peak multitaper spectrum estimator [16] to compute the low variance MFCC and PLP features for an emotion recognition system.

3. MULTITAPER MFCC AND PLP EXTRACTION

The feature extraction process of multitaper MFCC and PLP is presented through the block diagram of Figure 2. After pre-processing (DC offset removal and signal spectral pre-emphasis), the speech signal is decomposed into a series of 20-30 ms overlapping frames with a frame shift of 10 ms. Each frame is then multiplied by a single window (when M=1) such as a *Hamming* window or multiple windows, such as a Thomson multitaper, to reduce the effect of discontinuity introduced by the framing process. The power spectrum is estimated by computing the squared magnitude of the discrete Fourier transform (DFT) of the frame. The spectrum of the speech signal is then filtered by a group of triangle bandpass filters that simulate the characteristics of a human's ear called Mel windows.

After these similar processing steps for both features (MFCC and PLP), the MFCC extraction process follows with the natural logarithmic nonlinearity, which aims to approximate the relationship between a human's perception of loudness and the sound intensity. Finally, the DCT (discrete cosine transform) is applied to generate the cepstral coefficients.

For the PLP features, the nonlinearity is based on the power-law proposed by Hermansky [6]. An inverse discrete Fourier transform (IDFT) is then applied to obtain a perceptual autocorrelation sequence following the linear prediction (LP) analysis. Final features are generated from LP coefficients using cepstral recursion [17]. Note that, for the extraction of PLP features, we have followed HTK-based processing [18], that is, for the auditory spectral analysis a Mel filterbank is used instead of a trapezoid-shaped bark filterbank.

Once static MFCC and PLP features are extracted, the log energy of the frame, the delta and double delta features are computed and added to the feature vector characterizing the frame of speech.

4. EMOTION CORPUS

The effectiveness of the low variance multitaper spectrum estimation on an emotion recognition task is tested using the FAU AIBO [3] emotional speech corpus. The dataset consists of spontaneous recordings of German children interacting with a pet robot. The corpus is composed of 9959 chunks for training and 8257 chunks for testing. A chunk is an intermediate unit of analysis between the word and the turn, which is manually defined based on syntactic-prosodic criteria. The chunks are labeled into five emotion categories: *Anger* (**A**), *Emphatic* (**E**), *Neutral* (**N**), *Positive* (**P**, composed of *motherese* and *joyful*) and *Rest* (**R**, consisting of emotions not belonging to the other categories such as *bored*, *helpless*, and so on). The distribution of the five classes is highly unbalanced. For example, the percentage of training data of each class is as follows: **A**(8.8%), **E**(21%), **N**(56.1%), **P**(6.8%), **R**(7.2%).

5. GMM MODELS

Cepstral feature vectors are modeled using a GMM model. GMM is a generative model widely used in the field of speech processing. It is a semi-parametric probabilistic method that offers the advantage of adequately representing speech signal variability. Given a GMM modeling a *D*-dimensional vector, the probability of observing a feature vector given the model $\tilde{\lambda}$ is computed as follows:

$$P\left(\mathbf{x}\,\middle|\,\tilde{\lambda}\right) = \sum_{i=1}^{m} w_i N\left(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \qquad (4)$$

where m, w_i, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ correspond to the number of Gaussians, weight, mean vector and diagonal covariance matrix of the i^{th} Gaussian, respectively.

GMM parameters are estimated using a Maximum Likelihood (ML) approach based on the Expectation Maximization (EM) algorithm [19].

The classification of a test sequence of frames $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ is based on the Bayes decision. Using an equal prior probability for all classes, the classification is achieved by computing the log-likelihood of the test utterance against the GMM of each emotion class. The test recording is classified as the emotion class label that maximizes the log-likelihood value over all class models.

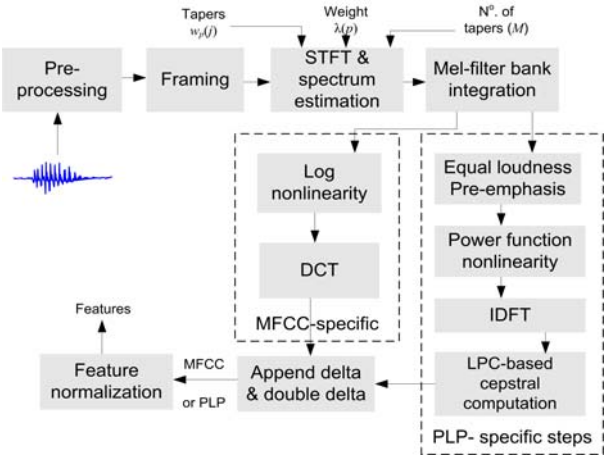

Figure 2 Block diagram illustrating MFCC and PLP feature extraction based on single and multitaper spectrum estimation.

6. EXPERIMENTS

In this section we describe and report the results of the systems designed in order (i) to evaluate the efficiency of multitaper spectrum estimation vs. single-taper, and (ii) to compare between the different methods of multitaper spectrum estimation described in Table 1.

Table 1 Single-taper and multitaper MFCC and PLP feature-based emotion recognition systems.

System	Description
Baseline	MFCC and PLP features are computed from the Hamming windowed direct spectrum estimate.
SWCE	MFCC and PLP features are computed from the sinusoidal weighted (i.e., *sine* tapered) spectrum estimate [14].
Multi-peak	MFCC and PLP features are computed from the multitaper spectrum estimate using multi-peak tapering [16].
Thomson_1	MFCC and PLP features are calculated from the multitaper spectrum estimates with dpss tapering [7]. Eigenvalues are used as the weights instead of uniform weights.
Thomson_2	MFCC and PLP features are calculated using Thomson multitaper method. Adaptive non-uniform weights are used instead of uniform one.

6.1. Experimental setup

The training of GMM models has been made with different numbers of mixtures taken from the set

$\{2,4,8,16,32,64,128,256,512,1024\}$. The best parameter is tuned based on the training data using a 9-fold cross validation protocol. Each fold contains a separate group of speakers to ensure speaker independent evaluation. After optimization, the selected numbers of Gaussians used for test data are as follows. For MFCC based-features we use 256 for the baseline, 128 for SWCE, Multi-peak and Thomson_1 and 32 for Thomson_2 systems. For PLP-based features, we used mixtures of 128 for baseline, SWCE and Multi-peak and mixture of 64 Gaussians for the Thomson_1 and Thomson_2 systems. The results are optimized to maximize the unweighted average recall (UAR) measure and secondly the weighted average recall (WAR) (namely accuracy) given that FAU AIBO emotion classes are highly unbalanced (i.e., one class is disproportionately more represented than the others). Note that a baseline classifier that predicts all the test data of the same class as of the majority one, namely *Neutral*, will achieve 65% of accuracy but only 20% of UAR.

6.2 Multitaper features

For our experiments, 13 static MFCC or PLP features (including the log energy) are extracted with a frame shift of 10 ms. Deltas and double deltas of static features are computed using a 5-frame window and added to static coefficients to compose 39-dimensional MFCC or PLP feature vectors. For this study, we take the number of tapers, *M*, equal to 6 because it is found that this number optimizes the performance for speech recognition [12] and speaker verification problems, as reported in [10, 11].

6.3 Uniform versus non-uniform multitaper weighting

In its simplest formulation, a multitaper estimator is the average of *M* direct spectral estimators and takes the form:

$$\hat{S}_{MT}(m,k) = \frac{1}{M}\sum_{p=1}^{M}\left|\sum_{j=0}^{N-1} w_p(j)\, s(m,j)\, e^{-\frac{i2\pi jk}{N}}\right|^2. \qquad (5)$$

It has been shown in [10], in the context of speaker recognition, that the use of non-uniform weights instead of uniform weights (equal to 1/*M*) to obtain the final multitaper spectral estimate provides better recognition accuracy over the baseline.

In the multitaper systems described in Table 1 (namely SWCE, multi-peak, Thomson_1 and Thomson_2), the final spectrum of the multitaper is estimated by averaging the *M* tapered subspectra using non-uniform weights. The reason is: The central peak at the harmonic frequency is produced by the exclusive contribution of the first taper. The subsequent tapers compensate for the information lost at the extremes of the first taper by producing spectral peaks distributed left and right of the central peak. An equal contribution (uniform weights) of each taper in estimation of the final spectrum yields to a high loss of energy for higher-order tapers. In order to compensate for this increased energy loss, adaptive non-uniform weights are applied as proposed by Thomson in [7]. Multitaper using adaptive non-uniform weights represents the fourth system in experiments in this study and dubbed *Thomson_2* as described in Table 1.

6.4. Evaluation

Table 2 gives the results achieved for each multitaper system. First, we observe that multitapers perform better than the single taper (*Hamming*) for both, MFCC and PLP feature-based, systems. The relative gain with respect to the baseline is ranging from 2.6% to 3.9% for MFCC and substantially the same range for the PLP based-system (from 2.2% to 3.9%). We also note that the adaptive non-uniform multitaper weighting (Thomson_2) slightly outperforms the non-uniform weighting using eigenvalues (Thomson_1). When comparing all multitaper methods together over the baseline, we observe that the SWCE and multi-peak are preferable. Finally, if we compare between MFCC and PLP we observe that PLP preserve its superiority in performance over MFCC with single taper as well as after averaging spectrum estimates.

7. CONCLUSION

In this paper, we have applied multitaper spectral estimates on MFCC and PLP features for an emotion recognition problem. Averaging spectral estimates helps to reduce large variance introduced by single-tapered spectrum estimate. We have shown that multitapering achieves over the baseline system a relative improvement ranging from 2.2% to 3.9% for both MFCC and PLP systems. These results confirm the effectiveness of multitapering in an emotion recognition task as reported in previous studies when applied on speech and speaker recognition. The number of tapers *M* used in this study was based on the optimization made for speech recognition and speaker verification tasks. In future work, we will study if the same value of *M* also optimizes the emotion recognition performance or whether the improvement can be further enhanced for different values of the number of tapers.

Table 2. Emotion recognition results achieved on FAU AIBO test data for the baseline and multitapers systems in terms of UAR and WAR scoring metrics.

	MFCC		**PLP**	
	UAR	WAR	**UAR**	WAR
Hamming	41.7%	45.3%	42.3%	41.5%
SWCE	**43.3%**	40.62%	43.5%	41.8%
Multi-peak	**43.2%**	40.3%	**44.0%**	40.8%
Thomson_1	42.8%	45.3%	42.4%	38.0%
Thomson_2	42.9%	36.7%	43.3%	39.9%

12. REFERENCES

[1] Pao, Tsang-Long, Yu-Te Chen, Jun-Heng Yeh et Wen-Yuan Liao, "Detecting Emotions in Mandarin Speech," *Computational Linguistics and Chinese Language Processing*, vol. 10, p. 347-362, 2005.

[2] Neiberg D, Elenius K, Laskowski K, "Emotion recognition in spontaneous speech using GMMs," *Proc. of INTERSPEECH conference*, pp 809–812, 2006.

[3] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," *Interspeech*, ISCA, Brighton, UK, 2009.

[4] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, N. Boufaden: "Cepstral and long-term features for emotion recognition," *INTERSPEECH* 2009: 344-347

[5] 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.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87 (4), pp. 1738-1752, 1990.

[7] D. J. Thomson, "Spectrum estimation and harmonic analysis," *IEEE proceeding*, vol. 70(9), pp. 1055–1096, 1982.

[8] K. S. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation," IEEE Trans. on Signal Proc., 43(1), 188–195, 1995.

[9] Y. Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," IEEE Trans. On Speech and Audio Proc., 12(1), 59-67, 2004.

[10] Alam, M.J. et al., "Multitaper MFCC and PLP features for Speaker verification using i-vectors," *Speech Communication*, vol. 55, pp. 237-251, 2013.

[11] T. Kinnunen, R. Saeidi, F. Sedlak, K.A. Lee, J. Sandberg, M. Hansson-Sandsten, H. Li, "Low-Variance Multitaper MFCC Features: a Case Study in Robust Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20(7), pp. 1990-2001, 2012.

[12] Alam, J., Kenny, P., and O'Shaughnessy, D., "Low-variance Multitaper Mel-Frequency Cepstral Coefficient Features for Speech and Speaker Recognition Systems," *Springer Cognitive Computation Journal*, to appear. DOI: 10.1007/s12559-012-9197-5.

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

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

[15] McCoy, E. J.; Walden, A. T.; Percival, D. B., "Multitaper Spectral Estimation of Power Law Processes," IEEE Trans. on Signal Processing, 46(3), pp. 655-668, 1998.

[16] 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.

[17] B. G. Gold and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, New York: John Wiley & Sons, Inc, 2000.

[18] S. J. Young et al., "HTK Book, " *Entropic Cambridge Research Laboratory Ltd*., 3.4 editions, 2006.

[19] A. Dempster, N. laird, and. Robin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royel Stastical Society*, vol. B, pp. 1-38, 1997.