# Multi-Taper Spectral Features for Emotion Recognition from Speech

Santosh V. Chapaneri
Dept. Electronics and Telecommunication Engg.
St. Francis Inst. of Technology, Univ. of Mumbai
Mumbai, India
santoshchapaneri@gmail.com

Deepak D. Jayaswal
Dept. Electronics and Telecommunication Engg.
St. Francis Inst. of Technology, Univ. of Mumbai
Mumbai, India
djjayaswal_vcet@yahoo.com

***Abstract*—In this paper, the performance of multi-taper spectral estimate is investigated relative to conventional single taper estimate for the application of emotion recognition from speech signals. Typically, a single taper/window helps in reducing bias of the estimate, but due to its high variance, the resulting spectral features tend to give poor recognition performance. The weighted averages of the multi-tapered uncorrelated eigen-spectra results in more discriminative spectral features, thus increasing the overall performance. We demonstrate that the application of six Multi-peak multi-tapers with support vector machine results in 81% classification accuracy on seven emotions from Berlin emotion database considering only spectral features, compared to 72% using conventional Hamming window method.**

***Keywords— Emotion, Multi-taper, Pattern recognition, SVM***

## I. INTRODUCTION

Human computer interaction (HCI) discipline has seen significant progress in the past decade. An important application is to recognize the emotion from a person's speech, rather than the spoken content. This has various applications ranging from automated call center infrastructure, humanoid robots, car service industry, computer tutorial via avatars, etc. Many commercial applications are being developed recently to serve this purpose. One can imagine an application where the system not only detects jokes made by the caller, but also reacts appropriately by switching to a more casual conversation mode. The performance of such systems depend on the features extracted from the speech emotion signals. If the features are sensitive to noise, then the recognition will be poor leading to various errors in commercial applications. Typically, spectral features such as Mel-frequency cepstral coefficients (MFCC), voice quality features, pitch, and energy are used to determine meaningful discriminative information about the emotion embedded in the speech signal [1]. To reduce the dimensionality of the extracted features, various feature selection techniques such as correlation based feature selection (CFS) [2], greedy feature selection (GFS) [3], etc. are employed. The classifiers used for training and testing the system include k-nearest neighbors [4], Gaussian mixture model [5], support vector machines [6], and hidden Markov models [7], among others.

Generally, due to non-stationarity of speech signals, the spectral information is derived for each frame after multiplying the frame with a single-tapered window, such as Hamming [8], to reduce the spectral leakage and bias of the spectral estimate. However, the variance of the estimate still remains high [9], due to which the resulting spectral features may not be discriminative enough to represent the underlying emotion. Multi-taper concept was introduced by Thomson [10] to reduce the variance by averaging the eigen-spectra resulting from applying multiple orthogonal (in time and frequency) tapers for each frame of the speech signal. It was demonstrated in [11] that the multi-taper spectrum is robust to noise and has less variance relative to conventional spectral estimate. Multi-tapers have been widely used recently for speaker recognition and verification purposes [12-14]. In [15], the authors applied multi-tapers for emotion recognition purpose but using only MFCC and perceptual linear prediction (PLP) features. In this paper, various spectral features are used from both conventional and multi-taper spectral estimates to recognize speech emotions. Though other types of features such as prosodic, voice quality, and linguistic features can be extracted, the purpose of this paper is to determine whether multi-taper spectral estimate features are better for emotion recognition compared to single taper spectral features.

## II. MULTI-TAPER SPECTRAL FEATURES

### A. Single Taper Spectrum

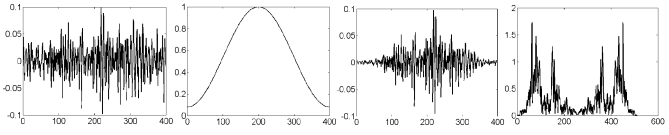



Fig. 1. Illustration of single taper spectrum: (a) Speech frame (25 msec, 16 kHz), (b) Hamming window, (c) Windowed frame, (d) Spectral estimate

Consider a discrete time speech signal sequence $s(n)$, $n = 0, 1, \ldots, N - 1$, with zero mean and power spectral density $S(f)$. Let $w(n)$, $n = 0, 1, \ldots, N - 1$ denote the taper (window) function (such as Hamming or Blackman window). The smoothed periodogram estimator of $S(f)$ is given by

$$\hat{S}(f) = \left| \sum_{n=0}^{N-1} w(n)\, s(n)\, e^{-j2\pi nf} \right|^2, \quad |f| \le \frac{1}{2} \tag{1}$$

with normalized sampling interval. This can be computed efficiently by FFT with $f = n/N$ ($N$ = integer power of 2). Fig. 1 illustrates the spectrum estimate for one speech frame using a single Hamming window. Techniques including Welch's method can be used wherein the speech segment of $N$ samples is split into overlapping segments, Eq. (1) is computed for each segment, and the resulting estimate is the average over these segments. This method reduces the variance to an extent but at the cost of increased bias [16].

### B. Multi-Taper Spectrum

The most important drawback of using single taper estimates is that by using a single window (taper), a significant end-portion of the signal is discarded as observed in Fig. 1 (c). Due to this, the variance of the resulting spectral estimate is much higher [9]. Using multi-tapers, the information lost by one taper is partially recovered by other tapers. The multiple tapers are designed so as to reduce spectral leakage with just few tapers [17]. The resulting multi-taper spectrum is obtained by a weighted sum of the single tapered periodograms. The weights are chosen such that the spectral estimate will have less variance than the single spectral estimate, while also retaining small bias from spectral leakage. Simple statistics tells us that if a random variable $X$ has variance of $\sigma^2$, then the statistical average of $n$ independent samples of $X$ will have variance of $\sigma^2/n$. Since the spectral estimates that result from using multiple orthogonal tapers are somewhat uncorrelated, a weighted average of these will thus have a smaller variance. This is important since most experiments are limited to analyzing a single realization of the speech stochastic process.

Let $\{w_k(n)\}_{n=0}^{N-1}$, $k$ = 1, 2, …, $K$, denote the multiple orthonormal tapers such that

$$\sum_{n=0}^{N-1} w_p(n)\, w_q(n) = \begin{cases} 1, & \text{if } p = q \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

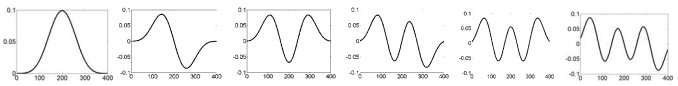
Due to the orthonormality of tapers, we obtain $K$ independent estimators of $S(f)$, and the resulting spectral estimate is the weighted average of $K$ eigen-spectrums, given by Eq. (3).

$$\hat{S}_{MT}(f) = \frac{1}{\sum_{k=1}^{K} \lambda_k} \left( \sum_{k=1}^{K} \lambda_k \left\{ \left| \sum_{n=0}^{N-1} w_k(n)\, s(n)\, e^{-j2\pi nf} \right|^2 \right\} \right), \quad |f| \le \frac{1}{2} \tag{3}$$

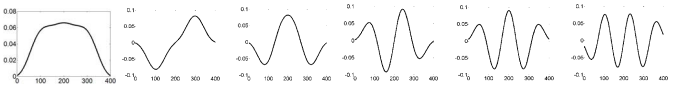
We discuss the three most popularly used multi-tapers, viz. Thomson [10], Multi-peak [18] and Sinusoidal Weighted Cepstrum Estimator (SWCE) [19], and apply them in the context of emotion recognition from speech. In Thomson method, the taper vector $\boldsymbol{w}_k$ is calculated from eigen-equation

$$\mathbf{A}\boldsymbol{w}_k = v_k \boldsymbol{w}_k \tag{4}$$

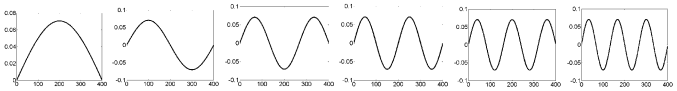
where $v_k$ are the eigen-values ranging from zero to unity, and $\mathbf{A}$ is a real symmetric Toeplitz matrix with each element given by $a_{i,j} = \sin\left[2\pi B(i-j)\right] / \left[\pi(i-j)\right]$, $B$ is the half-frequency bandwidth. To make up for the discarded parts from the first taper, the other orthonormal tapers give increasingly more weight to the ends of the frame. Thus, there is no loss of information of any part of the speech frame. Generally, the bias of the spectral estimate is larger and variance lower when more multi-tapers are used [12]. The number of tapers should be chosen as $K \le \lfloor 2NB \rfloor$ since the sorted low-order eigen-values approaches one and high-order values become insignificant. Typically, about 4 to 8 tapers have been shown to give a good performance in a speaker recognition setup [11]. For weighted averaging, the weights $\lambda_k$ are simply $1/K$, or adaptive weights could also be used from the eigen-values of these discrete prolate spheroidal tapers [10].



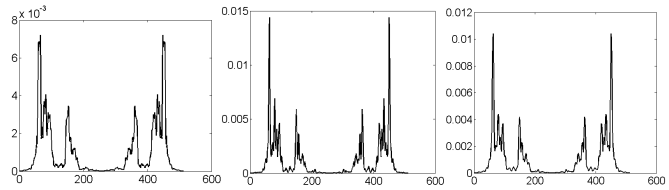
(a) 6 Thomson tapers



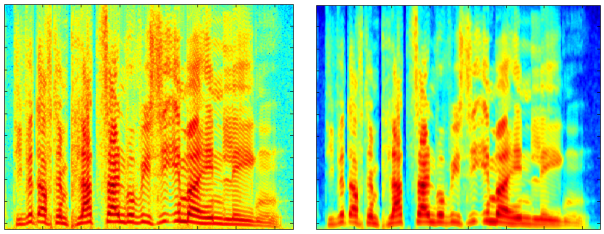
(b) 6 Multi-peak tapers



(c) 6 SWCE tapers



(d) Multi-taper spectral estimates: Thomson, Multi-peak, SWCE (left to right)



(e) Spectrograms of speech with Hamming (left), Multi-peak (right)

Fig. 2. Illustration of multi-taper spectral estimates

Multi-peak tapers were proposed in [18] for peaked spectra with low bias and low variance at the frequency peaks. It is obtained as a solution of the eigen-value problem in Eq. (5)

$$\mathbf{R}_B \boldsymbol{q}_k = v_k \mathbf{R}_Z \boldsymbol{q}_k \tag{5}$$

where $\mathbf{R}_B$ is the ($K$ x $K$) Toeplitz matrix with elements given by $r_B(k) = r_s(k) * B\mathrm{sinc}(Bk)$, $0 \le |k| \le K-1$, $r_s$ is covariance function of signal $s$, and $\mathbf{R}_Z$ is the Toeplitz matrix chosen for suppression of sidelobes with penalty of 30 dB. The weights are obtained from the largest eigen-values by Eq. (6).

$$\lambda_k = v_k \Big/ \sum_{k=1}^{\lfloor N/K \rfloor} v_k \tag{6}$$

A simpler multi-taper exists in the form of SWCE [19] with individual tapers given by Eq. (7) along with the closed form solution of their weights in Eq. (8).

$$w_k(n) = \sqrt{\frac{2}{N+1}} \sin\left(\frac{\pi k(n+1)}{N+1}\right) \quad (7)$$

$$\lambda_k = \frac{\cos\left(\frac{2\pi(p-1)}{M/2}\right)+1}{\sum_{p=1}^{M}\left(\cos\left(\frac{2\pi(p-1)}{M/2}\right)+1\right)}, \; p = 1,2,...,\lfloor N/K \rfloor \quad (8)$$

Fig. 2 (a – c) shows the multiple ($K$ = 6 tapers) obtained via the three techniques, along with their weighted average spectral estimates in Fig. 2 (d). It can be easily seen that the multi-taper spectrums of the speech frame (Fig. 1 (a)) are much smoother than the single taper spectrum of Fig. 1 (d), and thus has less variance, as expected. In general, the variance of multi-taper spectrum is given by $\mathrm{var}\left(\hat{S}_{MT}(f)\right) = \mathrm{var}\left(\hat{S}(f)\right)\Big/K$ [11]. The spectrograms for one speech utterance are shown in Fig. 2 (e) where we observe that the multi-taper spectrogram is much smoother than the single taper one.

### C. Spectral Features

In this work, various spectral features are computed from the spectral estimate (both single tapered and multi-tapered) to characterize the emotion in input speech signal. The signal $s(n)$ is split into overlapping frames with frame size of 25 msec and overlapping of 10 msec. In the following notation, $m$ denotes the frame number ($m$ = 1 to $M$), $f_k(m)$ and $a_k(m)$ denote the frequency and amplitude of frequency bin $k$, using either single or multi-taper spectrum approaches. The normalized STFT amplitude is given by $p_k(m) = a_k(m)\Big/\sum_{k=1}^{F} a_k(m)$, where $F$ is the total number of FFT frequency bins. The following spectral features are considered in this work:

i) *Spectral Energy SE*($m$): It is computed as the sum of squared amplitudes given by Eq. (9). *SE* is usually higher for emotions having high excitation, viz. anger, happiness, and lower for emotions such as fear, boredom.

$$SE(m) = \sum_{k=1}^{F} a_k^2(m) \quad (9)$$

ii) *Spectral Centroid SC*($m$): It is computed as the first moment of spectral estimate given by Eq. (10). It represents the spectral center of gravity.

$$SC(m) = \sum_{k=1}^{F} f_k(m) \cdot p_k(m) \quad (10)$$

iii) *Spectral Spread SS*($m$): It represents the spectral standard deviation, i.e. spread of spectral estimate around its mean value, and is given by Eq. (11).

$$SS(m) = \sqrt{\left(\sum_{k=1}^{F}\left(f_k(m) - SC(m)\right)^2 \cdot p_k(m)\right)} \quad (11)$$

iv) *Spectral Skewness SSK*($m$): It represents the measure of symmetry of spectral estimate around its mean value, and is given by Eq. (12). *SSK* > 0 implies a spectral tilt with higher concentration of energy in higher frequency range.

$$SSK(m) = \left(\sum_{k=1}^{F}\left(f_k(m) - SC(m)\right)^3 \cdot p_k(m)\right)\Big/SS^3(m) \quad (12)$$

v) *Spectral Kurtosis SK*($m$): It represents the measure of flatness of spectrum around its mean value, given by Eq. (13). If *SK* > 3, it implies that the spectral estimate has a peakier or compact distribution, and *SK* < 3 implies flat distribution.

$$SK(m) = \left(\sum_{k=1}^{F}\left(f_k(m) - SC(m)\right)^4 \cdot p_k(m)\right)\Big/SS^4(m) \quad (13)$$

vi) *Spectral Rolloff SR*($m$): It represents the frequency below which 95% of the total spectral energy is accounted for, given by Eq. (14) where $fs/2$ is the Nyquist frequency.

$$\sum_{f=0}^{SR(m)} a_f^2(m) = 0.95 \sum_{f=0}^{fs/2} a_f^2(m) \quad (14)$$

vii) *Spectral Decrease SD*($m$): It represents the degree to which there is more low frequency sound than high frequency sound, given by Eq. (15).

$$SD(m) = \frac{\sum_{k=2}^{F} \frac{a_k(m) - a_1(m)}{k-1}}{\sum_{k=2}^{F} a_k(m)} \quad (15)$$

viii) *Spectral Slope SSL*($m$): It is a measure of voice quality found using linear regression given by Eq. (16), and it represents the amount of decrease of spectral amplitude based on human perception.

$$SSL(m) = \frac{1}{\sum_{k=1}^{F} a_k(m)} \cdot \frac{F\sum_{k=1}^{F} f_k(m)a_k(m) - \sum_{k=1}^{F} f_k(m)\sum_{k=1}^{F} a_k(m)}{F\sum_{k=1}^{F} f_k^2(m) - \left(\sum_{k=1}^{F} f_k(m)\right)^2} \quad (16)$$

ix) *Spectral Variation SV*($m$): It represents the flux/variation of spectral estimate over time and is computed from the normalized cross-correlation between successive spectra $a_k$, given by Eq. (17). *SV* close to 1 implies that the successive spectra are highly dissimilar to each other.

$$SV(m) = 1 - \frac{\sum_{k=1}^{F} a_k(m-1)a_k(m)}{\sqrt{\sum_{k=1}^{F} a_k^2(m-1)}\sqrt{\sum_{k=1}^{F} a_k^2(m)}} \quad (17)$$

x) *Spectral Flatness SF*($m$): It represents the measure of noisiness of a spectral estimate, computed by the ratio of geometric mean to the arithmetic mean of spectral values, given by Eq. (18). For example, the spectrum of white noise is essentially "flat", since it represents equally all frequencies in the spectral domain.

$$SF(m) = \frac{\left(\prod_{k=1}^{F} a_k(m)\right)^{1/F}}{\frac{1}{F}\sum_{k=1}^{F} a_k(m)} \quad (18)$$

xi) *Spectral Crest SCR*($m$): It is computed as ratio of the maximum value with the arithmetic mean of the spectral estimate, given by Eq. (19).

$$SCR(m) = \frac{\max_k a_k(m)}{\frac{1}{F}\sum_{k=1}^{F} a_k(m)} \quad (19)$$

xii) *Spectral Entropy SEP*($m$): For the given spectral estimate $a_k(m)$, normalized power spectral density $d_k(m)$ is computed by Eq. (20), following which, Eq. (21) gives a measure of spectral entropy per frame.

$$d_k(m) = \frac{|a_k(m)|^2}{\sum_{k=0}^{F/2} |a_k(m)|^2} \quad (20)$$

$$SE(m) = -\sum_{k=0}^{F/2} d_k(m) \cdot \log_2\left(d_k(m)\right) \quad (21)$$

To avoid the non-speech portion of the signal, heuristics mentioned in Eq. (22) are applied. The first heuristic limits the frequency range of the signal to perceptual region of telephonic speech, and the second one avoids extreme strong tones.

$$\begin{aligned} |a_k(m)|^2 &= 0,\ k < 300 \text{ Hz or } k > 3500 \text{ Hz}, \\ d_k(m) &= 0 \text{ if } d_k(m) \geq 0.9 \end{aligned} \quad (22)$$

xiii) *Mel Frequency Cepstral Coefficients* (*MFCC*): It represents the shape of the spectral estimate with very few coefficients. It is defined as the Discrete Cosine Transform (DCT) of the logarithm of the Mel-filtered spectral estimate, given by Eq. (23). The use of Mel frequency scale adopts the human auditory system, which is linear till 1 kHz and non-linear above this frequency, to better take into account the mid-frequencies part of the signal. Excluding the first coefficient, typically 13 coefficients are stored for each frame $m$.

$$MFCC(m,:) = \text{dct}\left\{\log\left(\text{abs}\left(a_k(m)\right) \cdot \text{MelFilterBank}\right)\right\} \quad (23)$$

xiv) *Spectral Contrast* (*SCT*, *SCV*): The MFCC averages the spectral distribution in each Mel sub band due to which the relative spectral information is lost. For a better discrimination, octave based spectral contrast features were proposed in [20], using sub bands based on octave scale filters. The values of spectral contrast and spectral valley are computed for each sub band per frame $m$. The spectrum of each frame $a_k(m)$ is divided into six octave-scaled sub bands (with boundaries at 0, 200Hz, 400Hz, 800Hz, 1.6kHz, 3.2kHz, and 8kHz, for sampling frequency of 16kHz), resulting in $\{a_{b,1}(m), a_{b,2}(m), \ldots, a_{b,N}(m)\}$ where $N$ is the number of FFT frequency bins in the $b^{th}$ sub band. This vector is sorted in a non-ascending order represented as $\{a'_{b,1}(m), a'_{b,2}(m), \ldots, a'_{b,N}(m)\}$. The spectral peak (*SCP*), spectral valley (*SCV*), and spectral contrast (*SCT*) in the $b^{th}$ sub band is computed using Eqs. (24) – (26), where $\alpha$ is the neighborhood factor deciding the part of all bins ($k$) in the band to average over (empirically, $\alpha$ = 0.2). The spectral contrast features are given by ($SCT_b$, $SCV_b$) with $b \in [1, 6]$.

$$SCP_b(m) = \log\left\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} a'_{b,i}(m)\right\} \quad (24)$$

$$SCV_b(m) = \log\left\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} a'_{b,N-i+1}(m)\right\} \quad (25)$$

$$SCT_b(m) = SCP_b(m) - SCV_b(m) \quad (26)$$

In this work, we compute the above mentioned spectral features for each frame of the speech signal leading to a total of 12 (Statistical Spectral features) + 13 (MFCC coefficients) + 12 (Spectral Contrast features) per frame. To reduce the dimensionality, we determine the statistics of each spectral feature contour using mean, standard deviation, median and inter-quartile range. This results in a total of (12 + 13 + 12) x 4 = 148 spectral features per emotion speech signal.

## III. Feature Classification

Support Vector Machines (SVM) [21], which is a widely used supervised learning classification algorithm, is used in this work for classifying the spectral features into appropriate emotion classes. It determines an optimal separating hyperplane for binary classification task. Given training vectors (in this work, spectral features) $\mathbf{x}_i \in R^d$, the model is estimated as given in Eq. (27), where $L^2$-norm is used, $b$ indicates the distance from hyperplane to the origin, $\xi_i$ are slack variables (for non-separable cases), $y_i \in \{+1, -1\}$ are the output labels and $C$ is the cost penalty of the classification error.

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{L} \xi_i \\ &\text{s.t. } y_i\left(\mathbf{w}^T \varphi(\mathbf{x}_i) + b\right) \geq 1 - \xi_i,\ \xi_i \geq 0 \end{aligned} \quad (27)$$

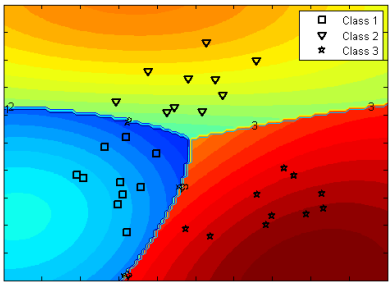



Fig. 3. Multi-class capability of SVM classifier

$$K(\mathbf{x}_i, \mathbf{x}_j) = \varphi^T(\mathbf{x}_i)\varphi(\mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \quad (28)$$

The non-linear function $\varphi$ converts the non-linear data in lower dimension to liner data in higher dimensional space using kernels that satisfy Mercer's conditions. In this work, radial basis function (RBF) kernels are used given by Eq. (28). The optimal parameters $\gamma$ and $C$ are determined by grid-search

algorithm with 10-fold cross-validation. Further, for multi-class instances of emotion in this work, pairwise classification is done using one-against-one approach and the final class decision is performed by majority voting. An illustration of the multi-class capability of SVM is shown in Fig. 3 for 3 classes.

## IV. Experimental Results

The performance of conventional spectral and multi-taper spectral features for the purpose of emotion recognition is evaluated using the Berlin emotion database (EMO-DB) [22]. It includes 535 simulated speech utterances of 6 emotions ranging from anger, boredom, disgust, fear, happy, and sad, along with neutral. 70% of the dataset is used for training the multi-class SVM classifier with 10-fold cross-validation for parameter optimization. Fig. 4 shows the normalized sample means and standard deviations of statistical spectral and MFCC features for one speech utterance of happiness emotion. The sample means are almost same for both conventional as well as multi-taper features, thus bias is not affected. However, the standard deviations are significantly reduced for the (Multi-peak) multi-taper features, thus the variance is reduced owing to the use of multiple windowed tapers.

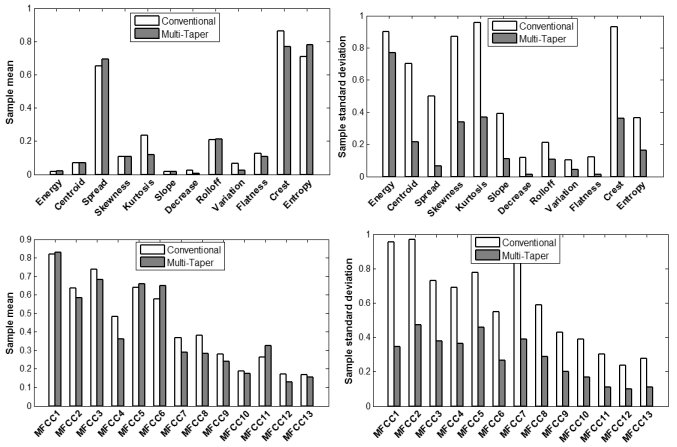



Fig. 4. Variance reduction due to multi-peak multi-tapers in statistical spectral features (top panel), and MFCC features (bottom panel)

Table I shows the accuracy of the recognition system for different number of emotions using single windowed spectral and multi-taper spectral features with different types of tapers. The emotions used from the database include anger (A), boredom (B), disgust (D), fear (F), happiness (H), and sadness (S), along with a neutral (N) state. By increasing the number of emotions to recognize, the overall accuracy decreases, reaching to a maximum of 81.08% achieved by the use of Multi-peak multi-tapers, giving an improvement of 8.5% over the conventional single-windowed Hamming spectral estimation technique. Further, these emotions are categorized into activation, valence and emotion-against-neutral binary classes [23] as follows: Activation = {(A, D, F, H), (B, S)}; Valence = {(H), (A, B, D, F, S)}; Emotion/No-Emotion = {(A, B, D, F, H, S), (N)}, where the first sub-set refers to positive class and the second sub-set refers to negative class. For these binary classification problems as well, the multi-taper approach out-performs the conventional one as seen from Table I.

TABLE I. Accuracy of Single and Multi-taper Features

| Emotions | Hamming (1 taper) | Thomson (6 tapers) | Multi-peak (6 tapers) | SWCE (6 tapers) |
|---|---|---|---|---|
| **Two** | 100% | 100% | **100%** | 100% |
| **Three** | 90.32% | 91.43% | **93.45%** | 92.33% |
| **Four** | 82.5% | 83.55% | **85.32%** | 82.80% |
| **Five** | 82.17% | 84% | **88.38%** | 85.53% |
| **Six** | 75.65% | 76.74% | **81.43%** | 79.83% |
| **Seven** | 72.57% | 73.53% | **81.08%** | 77.49% |
| **+ / - Activation** | 93.28% | 95.34% | **96.94%** | 94.67% |
| **+ / - Valence** | 92.11% | 94.50% | 95.52% | **95.92%** |
| **+ / - Emotion** | 95.71% | 96.35% | **98.21%** | 97.25% |

TABLE II. Multi-class Confusion Matrix (Multi-peak Tapers)

| | A | B | D | F | H | N | S | Recall |
|---|---|---|---|---|---|---|---|---|
| **A** | **16** | | 1 | | 2 | | | 84.21 |
| **B** | | **9** | | | | | 2 | 81.82 |
| **D** | | 1 | **14** | 1 | 1 | | 1 | 77.78 |
| **F** | | | 1 | **10** | 2 | 1 | 1 | 66.67 |
| **H** | 1 | | | 1 | **17** | | | 89.47 |
| **N** | | 3 | | | | **11** | | 78.57 |
| **S** | | 1 | | 1 | | 0 | **13** | 86.67 |
| **Prec.** | 94.12 | 64.29 | 87.50 | 76.92 | 77.27 | 91.67 | 76.47 | |

TABLE III. Confusion Matrices: Activation, Valence, Emotion (Multi-peak Tapers)

| | + Activ. | - Activ. | Recall |
|---|---|---|---|
| **+ Activ.** | **70** | 2 | 97.22 |
| **- Activ.** | 1 | **25** | 96.15 |
| **Prec.** | 98.59 | 92.59 | |

| | + Val. | - Val. | Recall |
|---|---|---|---|
| **+ Val.** | **16** | 3 | 84.21 |
| **- Val.** | 1 | **78** | 98.73 |
| **Prec.** | 94.12 | 96.3 | |

| | Emotion | Neutral | Recall |
|---|---|---|---|
| **Emotion** | **97** | 1 | 98.98 |
| **Neutral** | 1 | **13** | 92.86 |
| **Prec.** | 98.98 | 92.86 | |

Table II shows the confusion matrix for recognizing seven emotion states using Multi-peak multi-tapers, along with precision and recall values for each emotion class, with overall accuracy of 81.08%. Table III shows the confusion matrices for binary emotion classification problems. Here, precision, recall, and accuracy is calculated according to Eq. (29), where $N_C$ is the number of accurately predicted emotions, $N_F$ is the number of falsely predicted emotions, $N_M$ is the number of missed emotions, $total_C$ is the number of all accurately predicted emotions, and $total_M$ is the number of all emotion samples. The evaluations of precision and recall are shown in the last row and column of confusion matrices.

$$\text{Precision} = \frac{N_C}{N_C + N_F};\ \text{Recall} = \frac{N_C}{N_C + N_M};\ \text{Accuracy} = \frac{total_C}{total_M} \quad (29)$$

Fig. 5 illustrates the impact of changing the number of tapers on the performance of recognizing seven emotions, where we observe a sharp local maxima for Multi-peak multi-tapers. The optimum range of number of tapers is $3 \leq K \leq 10$, since with less tapers, the initial taper shapes do not completely capture the features of each frame, and with more tapers, the bias of the estimate increases, thus degrading the performance.

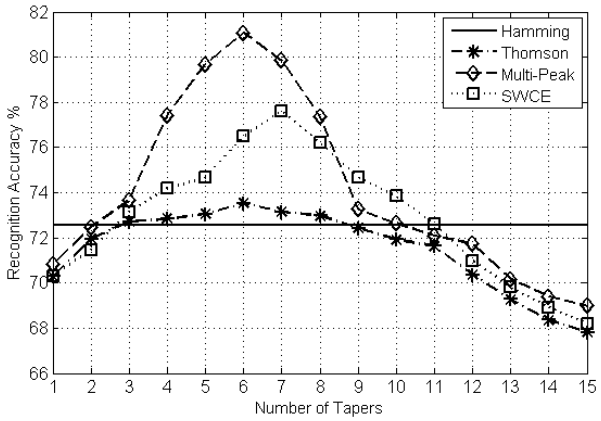



Fig. 5. Impact of number of tapers for recognition of seven emotions

## V. CONCLUSIONS

This work is motivated by the goal to make human computer interaction more natural due to advancement in smartphone and related technology. The focus in the area of emotion recognition from speech must shift from conventional single window spectral estimates to multi-taper methodology, since the resulting multi-taper spectral features have less variance and are more discriminative for better classification of emotions. The number of tapers should be chosen in the range of three to ten for effective recognition. From the various types of multi-tapers, Multi-peak shows an improved performance with six tapers applied on each frame. It is worth mentioning that the extra computations involved in computing multi-taper spectral estimates are negligible as demonstrated in [24]. For enhanced accuracy, the multi-taper spectral features should be appended with other relevant features such as pitch, jitter, shimmer, attack time, etc.

## REFERENCES


[1] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition", *Speech Communication*, vol. 52, no. 8, pp. 613-625, Aug 2010.

[2] T. Vogt, and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition", *IEEE Intl. Conf. Multimedia and Expo*, pp. 474-477, Jul 2005.

[3] P. Giannoulis, and G. Potamianos, "A hierarchical approach with feature selection for emotion recognition from speech", *Proc. Eighth Intl. Conf. Language Resources and Evaluation*, Turkey, pp. 1203-1206, May 2012.

[4] M. Feraru, and M. Zbancioc, "Speech emotion recognition for SROL database using weighted KNN algorithm", *IEEE Intl. Conf. Electronics, Computers and Artificial Intelligence*, pp. 1-4, Jun 2013.

[5] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs", *Intl. Conf. Spoken Language Processing* (*INTERSPEECH*), pp. 809-812, Sep 2006.

[6] N. Yang, R. Muraleedharan, J. Kohl, I. Demirkol, W. Heinzelman, and M. Sturge-Apple, "Speech-based emotion classification using multiclass SVM with hybrid kernel and thresholding fusion", *Proc. 4th IEEE Workshop on Spoken Language Technology*, pp. 455-460, Dec 2012.

[7] T. Nwe, S. Foo, and L. De Silva, "Speech emotion recognition using hidden Markov models", *Speech Communication*, vol. 41, no. 4, pp. 603-623, Nov 2003.

[8] Q. Jin, and T. Zheng, "Overview of front-end features for robust speaker recognition," *Proc. Asia-Pacific Signal and Information Processing Association* (*APSIPA*), Oct 2011.

[9] B. Babadi, and E. Brown, "A review of multi-taper spectral analysis", *IEEE Tran. Biomedical Engg.*, vol. 61, no. 5, pp. 1555-1564, May 2014.

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

[11] T. Kinnunen, R. Saeidi, J. Sandberg, and M. Sandsten, "What else is new than the Hamming window? Robust MFCCs for speaker recognition via multi-tapering", *Proc. Intl. Conf. Speech Communication Association* (*INTERSPEECH*), pp. 2734-2737, Sep 2010.

[12] T. Kinnunen, R. Saeidi, F. Sedlak, K. Lee, J. Sandberg, M. Sandsten, and H. Li, "Low-variance multi-taper MFCC features: a case study in robust speaker verification", *IEEE Tran. Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1990-2001, Sep 2012.

[13] M. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors", *Speech Communication*, vol. 55, no. 2, pp. 237-251, Feb 2013.

[14] M. Alam, V. Gupta, P. Kenny, and P. Dumouchel, "Use of multiple front-ends and i-vector based speaker adaptation for robust speech recognition", *Reverb Workshop IEEE Intl. Conf. Acoustics, Speech, and Signal Proc.* (*ICASSP*), pp. 1-8, May 2014.

[15] Y. Attabi, M. Alam, P. Dumouchel, P. Kenny, and D. O'Shaughnessy, "Multiple windowed spectral features for emotion recognition", *IEEE Intl. Conf. Acoustics, Speech, and Signal Proc.* (*ICASSP*), pp. 7527-7531, May 2013.

[16] P. Abry, P. Gonçalves, and P. Flandrin, "Wavelets, spectrum analysis and 1/*f* processes", *Lecture Notes in Statistics: Wavelets and Statistics*, vol. 103, pp. 15-29, Aug 1995.

[17] S.. Haykin, D. Thomson, and H. Reed, "Spectrum sensing for cognitive radio", *Proc. IEEE*, vol. 97, no. 5, pp. 849-877, May 2009.

[18] M. Hansson, and G. Salomonsson, "A multiple window method for estimation of peaked spectra", *IEEE Trans. Signal Proc.*, vol. 45, no. 3, pp. 778-781, Mar 1997.

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

[20] D. Jiang, L. Lu, H. Zhang, J. Tao, and L. Cai, "Music type classification by spectral contrast feature", *IEEE Intl. Conf. Multimedia and Expo*, vol. 1, pp. 113-116, Aug 2002.

[21] A. Hassan, and R. Damper, "Multi-class and hierarchical SVMs for emotion recognition", *Proc. Intl. Conf. Speech Communication Association* (*INTERSPEECH*), pp. 2354-2357, Sep 2010.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech", *Proc. 9th European Conf. Speech Comm. and Tech.* (*INTERSPEECH*), pp. 1-4, Jan 2005.

[23] E. Marchi, B. Schuller, A. Batliner, S. Fridenzon, S. Tal, and O. Golan, "Emotion in the speech of children with autism spectrum conditions: Prosody and everything else", *Proc. Workshop on Child, Computer and Interaction* (*WOCCI*), Sep 2012.

[24] D. Huynh, "Frequency Estimation of Musical Signals using STFT and Multi-tapers", *IEEE Intl. Symp. Image and Signal Processing and Analysis*, pp. 34-39, Sep 2009.