# Long-Term Spectro-Temporal Information for Improved Automatic Speech Emotion Classification

*Siqing Wu, Tiago H. Falk, and Wai-Yip Chan*

Department of Electrical and Computer Engineering, Queen's University, Canada
siqing.wu@ece.queensu.ca, {falkt, chan}@ee.queensu.ca

## Abstract

This paper investigates the contribution of features which convey long-term spectro-temporal (ST) information for the purpose of automatic emotional speech classification. The ST representation is obtained by means of a modulation filterbank decomposition of long-term temporal envelopes of the outputs of a gammatone filterbank. The two-dimensional discrete cosine transform is used to reduce the dimensionality of the representation; candidate features are then derived from statistics computed from the DCT coefficients. Sequential forward feature selection is used to select the most salient features. Two types of experiments are described which use the Berlin emotional speech database to test the performance of the ST features alone and in combination with prosodic features. In a multi-class experiment, simulation results with a support vector classifier show that a 44% reduction in classification error is attained once prosodic features are combined with the proposed ST features. Additionally, in a one-against-all experiment, an average increase in F-score of 33% is attained when the proposed ST features are included.

**Index Terms**: speech emotion recognition, spectro-temporal features, modulation spectrum, affective computing.

## 1. Introduction

Emotion is an essential element of human communication, thus automatic recognition of emotions in speech has become an active research area as wide applications exist, e.g. to add "emotional intelligence" to human-computer interaction, to design lively humanoid robots, and to improve the performance of current speech recognition and synthesis systems. Features employed for emotion recognition can be broadly classified as containing prosodic or spectral information. A brief review of recent studies [1, 2, 3, 4, 5] reveals that, while statistical measures of pitch trajectories and intensity contours are commonly used as prosodic features, features obtained from short-time ($\sim$20 ms) speech segments, e.g., mel-frequency cepstral coefficients (MFCC), are widely used for spectral content characterization. Moreover, short-term ($\sim$50 ms) temporal information is commonly used and incorporated in the form of delta or delta-delta features [1, 4, 5].

However, recent psychoacoustical and neurophysiological evidence indicates the existence of spectro-temporal receptive fields in mammalian auditory cortex which can extend up to temporal spans of hundreds of milliseconds [6, 7, 8]. These studies reveal the limitation of features derived from short-time spans as the long-time temporal information used by human listeners is discarded. In this paper, we exploit long-term spectro-temporal (ST) information to derive a novel feature set which is shown to improve the performance of automatic speech emotion recognition. The proposed features are derived from an auditory spectro-temporal representation of speech. The representation is obtained by first filtering the speech signal with a critical-band gammatone filterbank, to emulate the processing performed by the cochlea. Temporal envelopes are then extracted, by means of the Hilbert transform, from the output of each critical-band filter. Lastly, a modulation filterbank is applied to each temporal envelope. The resulting representation is henceforth referred to as the modulation spectrum.

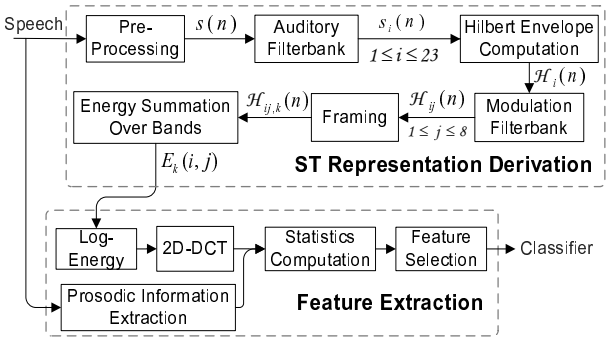


Figure 1: *Flowchart of the proposed algorithm.*

Deployment of the modulation filterbank allows for characterization of the rate of change of the signal temporal envelope. Since the rate at which people's vocal articulators (e.g. lips, jaw, tongue) move varies as their emotional states change [9], we expect features derived from this signal representation to be useful for emotion recognition. To validate this assumption, two types of experiments with the Berlin emotional speech database are conducted to test the proposed features alone and in combination with the commonly used prosodic features. In a multi-class experiment, by combining the proposed features with prosodic features, a recognition accuracy improvement from 62% to 78.7% is observed (i.e., an approximately 44% reduction in classification error rate). In a one-against-all (binary) experiment, an average increase in F-score of 33% is attained once the proposed ST features are included. Improved performance indicates that long-term ST information is useful for speech emotion recognition.

## 2. ST Representation of Speech

The auditory ST representation is obtained via the steps depicted within the top-most dashed box in Fig. 1, denoted by "ST Representation Derivation". First, the original speech signal $x(n)$ is normalized and resampled to 8kHz before processing; inactive speech segments are discarded. The preprocessed

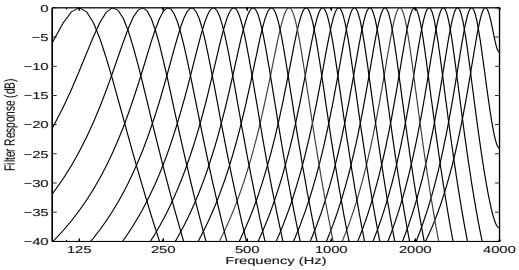


Figure 2: *Frequency response of a 23-band gammatone filterbank with center frequencies ranging from 125Hz to 3.6kHz.*

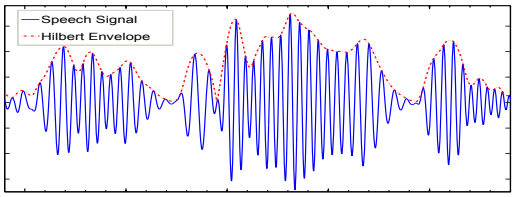


Figure 3: *One example of Hilbert envelope: the solid and dotted lines trace the speech signal and its Hilbert envelope, respectively.*

signal $s(n)$ is then filtered by a critical-band gammatone filterbank, with 23 filters, to emulate the auditory processing of acoustic signals by the human cochlea. The bandwidths of these filters are proportional to their center frequencies, as characterized by the equivalent rectangular bandwidth (ERB) [10]:

$$ERB_k = \frac{F_k}{Q_{ear}} + B_{min} \ , \tag{1}$$

where $F_k$ is the center frequency (in Hz) of the $k$th critical-band filter. $Q_{ear}$ and $B_{min}$ are set to 9.26449 and 24.7, respectively. The frequency response of the auditory filterbank employed in our work is depicted in Fig. 2. The first filter centered at 125Hz has a bandwidth of 38Hz whereas the last filter centered at 3568Hz has a bandwidth of 410Hz.

Using the critical-band filterbank for signal decomposition, 23 outputs $s_i(n)$ are obtained ($1 \leq i \leq 23$). The Hilbert envelope $\mathcal{H}_i(n)$ is then computed for each $s_i(n)$. In the time domain, the Hilbert envelope $\mathcal{H}(n)$ of a real-valued signal $\mathrm{x}(n)$ is the magnitude of the analytic representation of $\mathrm{x}(n)$, i.e.,

$$\mathcal{H}(n) = \sqrt{\mathrm{x}^2(n) + \mathrm{H}^2\{\mathrm{x}(n)\}}. \tag{2}$$

where $\mathrm{H}\{\cdot\}$ denotes the Hilbert transform. Fig. 3 illustrates a representative Hilbert envelope (dotted lines) for a short-term speech segment.

An eight-band modulation filterbank, as shown in Fig. 4, is then applied to each $\mathcal{H}_i(n)$ to generate outputs $\mathcal{H}_{ij}(n)$, where $1 \leq j \leq 8$, with $i$ and $j$ being indexes for auditory and modulation channels, respectively. $\mathcal{H}_{ij}(n)$ is then framed by a rectangular window of 250 ms length with 30 ms frame shift. This relatively long temporal extent is determined according to the physiological studies mentioned earlier and helps to capture long-term temporal dynamics information.

Denote the output of the $i$th auditory channel and $j$th modulation band at the $k$th frame as $\mathcal{H}_{ij,k}(n)$. Thus, the ST energy representation of frame $k$, termed $E_k(i,j)$, is given by

$$E_k(i,j) = \sum_{n=1}^{N} \mathcal{H}_{ij,k}^2(n) \ , \tag{3}$$

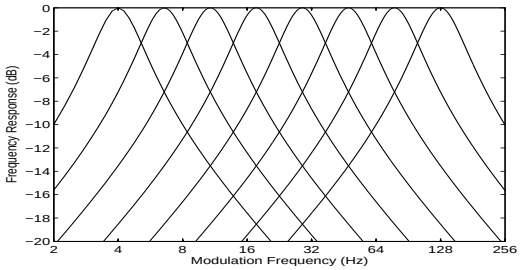


Figure 4: *Frequency response of an 8-band modulation filterbank with center frequencies ranging from 4Hz to 128Hz.*

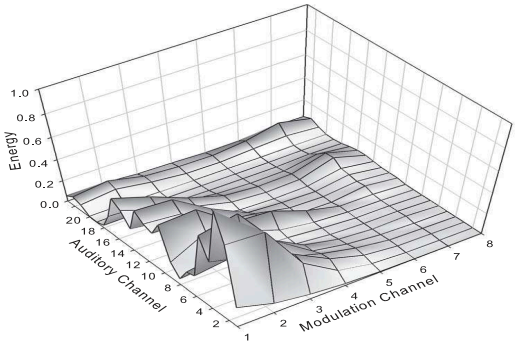


Figure 5: $E(i,j)$ *for one frame of a* $neutral$ *speech file: low channel index indicates low frequency; energy is normalized by the maximum value for visual purpose, sic passim.*

where $1 \leq k \leq T$, with $N$ and $T$ representing the number of samples in one frame and the number of active speech frames, respectively. Therefore, each frame results in a $23 \times 8$ modulation energy matrix. Fig. 5 depicts $E(i,j)$ of one frame from a $neutral$ speech file. With respect to the physical limitation and inertia of vocal organs, most of the dynamic details are concentrated within low modulation frequencies, as shown in Fig. 5.

## 3. Feature Extraction

In this section, a description of the feature extraction module, as depicted within the bottom dashed box in Fig. 1, is presented. Prosodic features considered in our experiments are also described.

### 3.1. 2D-DCT and feature computation

The two-dimensional Discrete Cosine Transform (2D-DCT) is applied to the log of the energy term described by (3), $\log(E_k(i,j))$, on a frame-by-frame basis in order to perform dimensionality reduction. DCT coefficients $C_k(p,q)$, calculated for the $k$th frame, are used to compute features for speech emotion recognition. Coefficients are computed as:

$$C_k(p,q) = \alpha(p)\beta(q) \sum_{i=0}^{L-1} \sum_{j=0}^{M-1} \log(E_k(i+1,j+1))\Phi(p,q) \ , \tag{4}$$

where $0 \leq p \leq L-1$, $0 \leq q \leq M-1$, and

$$\Phi(p,q) = \cos(\frac{\pi(2i+1)p}{2L}) \cos(\frac{\pi(2j+1)q}{2M}) \ . \tag{5}$$

Here, $L = 23$, $M = 8$, and

$$\alpha(p) = \begin{cases} 1/\sqrt{L}, p = 0 \\ \sqrt{2/L}, p \neq 0 \end{cases} \ , \ \beta(q) = \begin{cases} 1/\sqrt{M}, q = 0 \\ \sqrt{2/M}, q \neq 0 \end{cases} \ .$$

In our implementation, the 25 DCT coefficients with the 25 largest variances are selected for feature computation. The choice for the number of selected coefficients is a compromise between feature candidate pool size and recognition accuracy. Lastly, eight statistics are computed from each of the 25 selected coefficients: mean, standard deviation, maximum, minimum, range, median, 25% and 75% quartiles. In total, 200 ($25 \times 8$) scalar features are calculated from the ST representation.

### 3.2. Prosodic features

Prosodic features have been proven useful for speech emotion classification and are widely used in previous studies [1, 2, 3, 4]. In order to explore the contributions of the proposed features to prosodic features, pitch and intensity trajectories are extracted from the speech signal. To capture local temporal dynamics information, the first and second derivatives of such trajectories are also calculated. The eight aforementioned statistics are computed for the six trajectories, resulting in 48 prosodic features. In total, 248 features (48 prosodic and 200 proposed) are extracted from each speech signal. Sequential forward feature selection (SFS) [11] is used to select the most salient features. It is observed in our experiments that selecting the top 15 features suffices, as adding more features did not improve performance.

## 4. Experimental Evaluation

In this section, a description of the database used in our experiments is given. Experimental results are also presented and discussed.

### 4.1. Berlin emotional speech database

The Berlin emotional speech database employed in our experiment describes emotions as discrete emotional states. The database was recorded with 16 bit precision and 16kHz sampling rate for the purpose of studying acoustical features of emotional speech. Five female and five male actors each uttered ten sentences (5 short and 5 longer, generally between 1.5 and 4 seconds) in German to simulate seven emotions. Utterances with a subjective recognition rate better than 80% were chosen, as evaluated by a subjective listening test. The final number of speech files in the presented database is 535 divided (not uniformly) among seven emotions: $anger$ (127), $boredom$ (81), $disgust$ (46), $fear$ (69), $joy$ (71), $neutral$ (79) and $sadness$ (62). More details can be found in [12].

### 4.2. Results and discussion

A radial basis function support vector classifier (SVC) is employed for classification [13]. Unless otherwise specified, all results to follow are obtained using 10-fold cross-validation. Fifty utterances are randomly selected from each emotion. In our experiments, the emotion "disgust" is excluded as it contains a limited number of files, thus only six emotion classes are used. In total, each cross validation trial uses 270 speech files for training and 30 files for testing.

Three multi-class classification experiments are performed, using: ($i$) only prosodic features, ($ii$) only the proposed features, ($iii$) combined prosodic-proposed features. Tables 1 and 2 list classification results using only prosodic and the proposed features, respectively; an average classification accuracy of 62% and 69.3% is attained. Further analysis of the confusion matrices shows that the emotion $joy$ attains the poorest classification accuracy and is usually misclassified as $anger$. Moreover, the proposed features exhibit a significantly stronger capability of classifying $sadness$. Fig. 6 compares the ST representation $E(i, j)$ of $sadness$ with $neutral$; in order to gain some insight into the statistical trend, $E(i, j)$ is averaged over all frames of all "sad" or "neutral" speech files. The difference in the figure is notable - for $sadness$, most of the energy is concentrated within very low auditory and modulation channels. Such behavior of "sad" is not witnessed for other emotion classes. For sad speech, it is known that vocal organs move slower, e.g., in [9], it is reported that emotion $sadness$ is associated with the relatively smaller average tongue tip movement velocity of vowels in comparison with other emotions. Therefore, more pronounced ST energy in lower modulation frequencies is expected for $sadness$. On the contrary, ST energy is observed in higher modulation frequency channels (e.g. the last 4 channels of the modulation filterbank) for emotions such as $anger$ and $joy$.

Table 1: *Classification accuracy with **prosodic** features (the left-most column refers to the true emotion, sic passim).*

| Emotion | Anger | Boredom | Fear | Joy | Neutral | Sadness |
|---|---|---|---|---|---|---|
| Anger | **68%** | 4% | 0% | 24% | 4% | 0% |
| Boredom | 0% | **76%** | 0% | 0% | 16% | 8% |
| Fear | 8% | 4% | **68%** | 4% | 8% | 8% |
| Joy | 48% | 0% | 4% | **44%** | 4% | 0% |
| Neutral | 4% | 16% | 4% | 0% | **52%** | 24% |
| Sadness | 0% | 12% | 0% | 0% | 24% | **64%** |

Table 2: *Classification accuracy with the **proposed** features*

| Emotion | Anger | Boredom | Fear | Joy | Neutral | Sadness |
|---|---|---|---|---|---|---|
| Anger | **68%** | 0% | 16% | 16% | 0% | 0% |
| Boredom | 0% | **72%** | 4% | 0% | 16% | 8% |
| Fear | 0% | 4% | **64%** | 12% | 16% | 4% |
| Joy | 20% | 8% | 4% | **60%** | 8% | 0% |
| Neutral | 4% | 16% | 12% | 8% | **60%** | 0% |
| Sadness | 0% | 8% | 0% | 0% | 0% | **92%** |

Table 3: *Classification accuracy with **combined** features*

| Emotion | Anger | Boredom | Fear | Joy | Neutral | Sadness |
|---|---|---|---|---|---|---|
| Anger | **80%** | 0% | 0% | 20% | 0% | 0% |
| Boredom | 0% | **80%** | 0% | 0% | 12% | 8% |
| Fear | 8% | 0% | **80%** | 8% | 4% | 0% |
| Joy | 28% | 0% | 4% | **68%** | 0% | 0% |
| Neutral | 0% | 16% | 0% | 0% | **72%** | 12% |
| Sadness | 0% | 4% | 4% | 0% | 0% | **92%** |

Table 3 reports classification accuracy for the combined feature set. As observed from Tables 1 and 3, an average 16 percentage-point improvement is attained once ST features are considered (from 62% with only prosodic features to 78.7% for the combined features). This improvement translates into an approximately 44% reduction in classification error rate. This improvement in performance suggests that the proposed ST features are advantageous additions to the widely used prosodic features.

In order to further explore the gains obtained with the proposed features, one-against-all (i.e. binary) classification experiments are performed. In this experiment, six binary SVCs are trained, each of which is used to recognize one of the six emotions, with the remaining five emotions treated as an "unwanted" class. The F-score [14] is calculated as the performance measure as it is particularly useful when data is not equally distributed among classes. The F-score $F$ is given by:

$$F = \frac{2 \times P \times R}{P + R}\,, \tag{6}$$

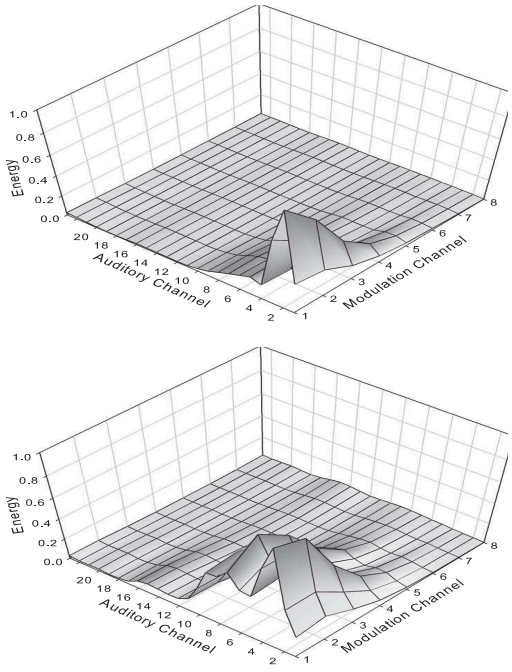


Figure 6: *Top and bottom panels depict average $E(i,j)$ for emotion $sadness$ and $neutral$, respectively.*

where $P$ and $R$ refer to $Precision$ and $Recall$, respectively.[1] Fig. 7 compares the classification results for the three types of features. It is evident that the proposed features improve the overall classification accuracy. Compared with the F-score computed for prosodic features, the improvement achieved by including the proposed features is 33% on average (the minimum improvement is 5% for $boredom$ and the maximum 83% for $sadness$).

Furthermore, it is also evident from the confusion matrices shown in Tables 1-3 that the exhibited tendency of misclassification is similar – if we consider $anger, fear, joy$ as high activation emotions and $boredom, neutral, sadness$ as low activation emotions, confusions within the same activation class are more pronounced. Such misclassification behavior can also be found in [1, 3, 5]. A further experiment with the combined features shows that a 96.3% accuracy is achieved when only classifying the two activation levels. Hence improved emotion recognition performance may be attained via cascaded classification as described in [3].

## 5. Conclusions

We have introduced a novel feature set for speech emotion recognition which is based on an auditory spectro-temporal (ST) representation of speech. The ST representation is derived to simulate the stimuli perception process in human cortical receptive fields. By modeling such characteristics, the proposed features are shown to improve the performance of speech emotion recognition. Experimental results show that the proposed features outperform conventional prosodic features when tested

[1]This type of F-score is also known as F1 measure, because $Precision$ and $Recall$ are evenly weighted. $Precision$ for a class is the number of true positives for the class divided by the total number of test samples classified as belonging to the class. $Recall$ for a class is the number of true positives for the class divided by the total number of test samples that actually belong to the class. $F$ reaches its best value at 1 and worst value at 0 by definition.

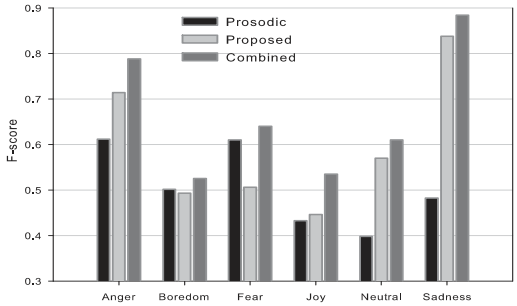


Figure 7: *F-score of one-against-all classification tests.*

alone and render a substantial improvement in recognition accuracy when combined with prosodic features. Future work will investigate the application of the proposed features to address the problem of dealing with real-world emotional speech, which mainly involves the following issues: spontaneous (vs. acted) speech, continuous (vs. discrete) emotions, and noisy (vs. noise-free) conditions [2, 4].

## 6. References

[1] M. Shami and W. Verhelst, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech," *Speech Communication*, vol.49, pp. 201-212, 2007.

[2] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol.49, pp. 787-800, 2007.

[3] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," In *Proc. ICASSP*, Apr. 2007, vol.4, pp. 17-20.

[4] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," In *Proc. ICASSP*, Apr. 2007, vol.4, pp. 941-944.

[5] M. Ayadi, M. Kamel, and F. Karray, "Speech emotion recognition using gaussian mixture vector autoregressive models," In *Proc. ICASSP*, Apr. 2007, vol.4, pp. 957-960.

[6] T. Chih, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, pp. 887-906, 2005.

[7] S. Shamma, H. Versnel, and N. Kowalski, "Ripple analysis in ferret primary auditory cortex: I. response characteristics of single units to sinusoidally rippled spectra,", *Aud. Neurosci.*, vol. 1, 1995.

[8] D. Klein, D. Depireux, J. Simon, and S. Shamma, "Robust spectro-temporal reverse correlation for the auditory system: Optimizing stimulus design," *J. Comput. Neurosci.*, vol. 9, 2000.

[9] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," In *Proc. Interspeech*, 2005, pp. 497-500.

[10] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, 1990.

[11] J. Kittler, "Feature set search algorithms," *Pattern Recognition and Signal Processing*, pp. 41-60, 1978.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," In *Proc. Interspeech*, 2005, pp. 1517-1520.

[13] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[14] C. Rijsbergen, *Information Retrieval*, Butterworths, London, 1979.