

Spectro-Temporal Modulations for Robust Speech Emotion Recognition

Lan-Ying Yeh, and Tai-Shih Chi

Department of Electrical Engineering
National Chiao Tung University, Hsinchu, Taiwan 300, R.O.C.

sidneyyeh.cm97g@nctu.edu.tw, tschi@mail.nctu.edu.tw

Abstract

Speech emotion recognition is mostly considered in clean speech. In this paper, joint spectro-temporal features (RS features) are extracted from an auditory model and are applied to detect the emotion status of noisy speech. The noisy speech is derived from the Berlin Emotional Speech database with added white and babble noises under various SNR levels. The clean train/noisy test scenario is investigated to simulate conditions with unknown noisy sources. The sequential forward floating selection (SFFS) method is adopted to demonstrate the redundancy of RS features and further dimensionality reduction is conducted. Compared to conventional MFCCs plus prosodic features, RS features show higher recognition rates especially in low SNR conditions.

Index Term: Emotion recognition, robust, spectro-temporal modulations

1. Introduction

Speech emotion recognition has been a popular research topic over the last decade. Knowing the emotion status of the speaker is important for human-machine interfaces with better interaction experiences. Many modern applications, such as interactive robots, infant or elder caring systems and speech-recognition based customer service lines, can use such information. Researchers have been devoted to searching novel features and designing powerful classifier to improve the recognition rate [1, 2]. The best feature sets have been discussed over years, and it is well acknowledged that pitch, energy, and duration contribute the most to emotion recognition [3, 4]. Spectral information or formants are also discussed frequently. However, early studies are often launched on “perfect” conditions, i.e., clean speech with acting emotions, which is far from real-world applications.

Recently, more and more studies pay attention to natural emotion or real environments with noises [5, 6]. However, these studies often went forward to find thousands of features in order to obtain an optimal set of features with the highest recognition rate for any particular testing environment. These brute-force

methods seem working, but these studies only evaluate their performance under the matched condition, where the testing data is under the same noise level as the training data. Undoubtedly, degraded performance is expected with changes of testing environments.

In this work, we intend to find a robust feature set from a spectro-temporal auditory perceptual model [7] for speech emotion recognition. This model consists of two computational modules, where the first module is to model functions of the cochlea and the second module is to model functions of the auditory cortex (A1). The Berlin emotional speech database with additive noises is utilized to test the robustness of proposed spectro-temporal auditory features. A linear-kernel SVM [8] is used as the emotion classifier. Recognition rates of our spectro-temporal auditory RS features are evaluated and compared to conventional spectral features (MFCCs) plus additional prosodic features under additive white and babble noises. Furthermore, the dimensionality reduction of our RS features is conducted and corresponding performance is investigated.

This paper is organized as follows. In section 2, a brief review of the two-module spectro-temporal auditory model is given. Speech database and three sets of features used in this study are introduced in section 3. In section 4, recognition results with and without the dimensionality reduction are demonstrated. We end in section 5 with conclusions and discussions.

2. Auditory Model

The auditory features adopted in this study are extracted from stages of a physiological based auditory model, which consists of an early cochlear (ear) and a central cortical (A1) module.

2.1. Cochlear Module

The cochlear module models functions of the peripheral auditory system. As shown in Figure 1, it first consists of a bank of 128 overlapping asymmetric constant-Q bandpass filters ($Q_{3dB} \approx 4$) which mimic the frequency selectivity of the cochlea. These filters distribute evenly over 5.3 octaves with

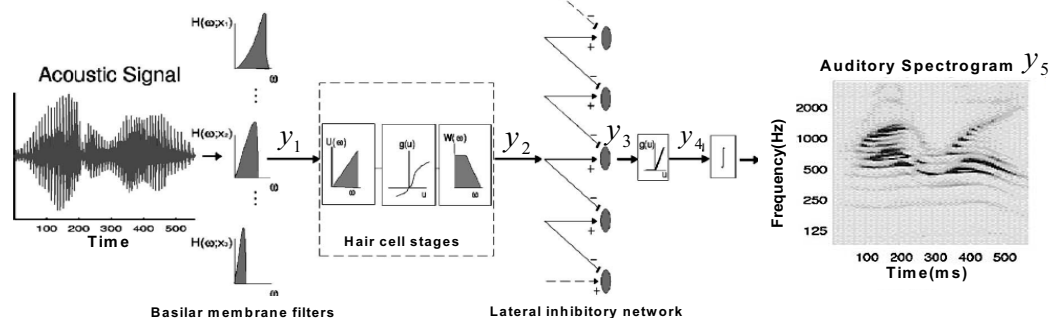


Figure 1: Stages of the early cochlear module (adopted from [7])

24 filters/octave frequency resolution. The output of each filter is fed into a non-linear compression stage and a lateral inhibitory network (LIN), and then processed by an envelope extractor (a half-wave rectifier followed by a low-pass filter). The non-linear compression models the saturation of the inner hair cells, and the LIN models the frequency masking effect. In this study, a simplified linear version of this module without the hair cell stage is used. All tested speech signals are normalized in advance to avoid the high-gain compression done by hair cells. Outputs of different stages of this module can then be written as:

$$y_1(t, \omega) = s(t) * h(t; \omega) \quad (1)$$

$$y_3(t, \omega) = \partial_{\omega} y_1(t, \omega) \quad (2)$$

$$y_4(t, \omega) = \max(y_3(t, \omega), 0) \quad (3)$$

$$y_5(t, \omega) = y_4(t, \omega) * \mu(t; \tau) \quad (4)$$

where $s(t)$ is the input speech; $h(t; \omega)$ is the impulse response of the constant-Q cochlear filter with center frequency ω ; $*_t$ depicts the convolution in time; ∂_{ω} is the partial derivative of variable ω ; the integration window $\mu(t; \tau) = e^{-t/\tau} \cdot u(t)$ with the time constant τ models the current leakage along the neural pathway to the midbrain; and $u(t)$ is the unit step function. More detailed descriptions of related auditory process and mathematic formulations of this cochlear module can be found in [7].

The output $y_5(t, \omega)$ is referred to as an auditory spectrogram, which represents neuron activities along the time and log-frequency axis. Intuitively, it is similar to the magnitude response of a mel-scaled FFT based spectrogram, where our constant-Q criterion approximates the mel-scale and our local envelope approximates the magnitude of a FFT based spectrogram.

2.2. Cortical Module and Rate-Scale Representation

The second module models the spectro-temporal selectivity of neurons of the auditory cortex (A1). Briefly speaking, the auditory spectrogram $y_5(t, \omega)$ is further analyzed by cortical neurons which are modeled by two-dimensional filters tuned to different spectro-temporal modulation parameters [7]. The rate (or velocity) parameter in Hz reflects how fast the local spectro-temporal envelope varies along the temporal axis. The scale (or density) parameter in cycle/octave characterizes how broad the signal's local spectro-temporal envelope distributed along the log-frequency axis.

In addition to the rate and scale, cortical neurons are also found to be sensitive to the direction of the FM sweep. This directionality is characterized in this module by the sign of the rate (negative for upward sweeping; positive for downward sweeping). From functional point of view, this module models cortical neurons as performing a joint spectro-temporal multi-resolution analysis (due to various rate-scale combinations) on the input auditory spectrogram. The excitation pattern of cortical neurons pertaining to a single t-f unit in the input spectrogram is referred to as the rate-scale representation of that particular t-f unit. Each rate-scale representation is labeled by neurons' tuning characteristic of rate, scale, and directionality.

Two averaged rate-scale plots over the frequency axis around 200 and 550 ms are given in Figure 2. Two aspects are clearly shown in each rate-scale plot: (1) spectro-temporal

modulations of envelopes and (2) resolved pitch below 512 Hz. Take the 550 ms frame as an example. The resolved pitch around 230 Hz excites {high rate, fine scale} neurons, thus produces the corresponding rate-scale representation. On the other hand, envelopes of the almost flat harmonic structure shown at 230, 460 and 1150 Hz excite neurons tuned to {low rate (due to the flatness), low scale (2 cycles within 2.32 octave)} and produce strong rate-scale responses at regions less than 8 Hz and less than 1 cycle/octave. Since flat envelopes do not favor any sweeping directions, symmetric responses to rate are clearly shown in the {low rate, low scale} region. More detailed description and mathematic formulation of this cortical module can be found in [7].

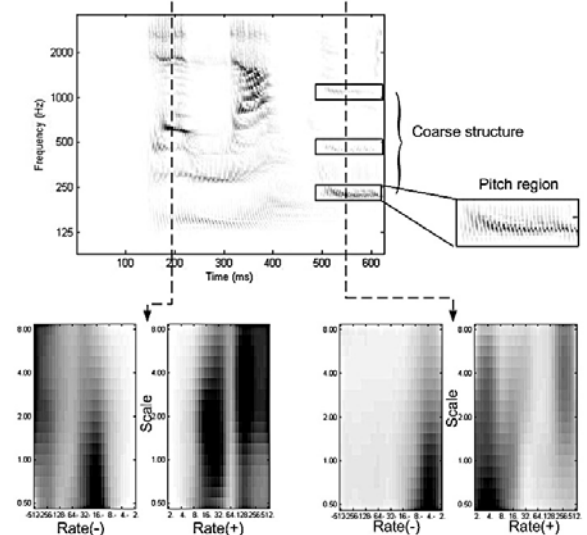


Figure 2: Rate-scale representation produced by the cortical module.

3. Speech Sample and Feature Extraction

The flowchart of our proposed method is shown in Figure 3. First, the clean speech is distorted by additive noises. Secondly, features are extracted from both clean and noisy speech, and then are used to train and test a linear-kernel SVM classifier, respectively. Rate-Scale (RS) representations from the auditory model and conventional MFCCs plus prosodic features are extracted for a head-to-head contest in a speech emotion recognition task.

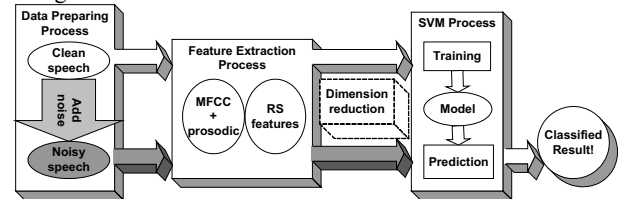


Figure 3: Overall flowchart of proposed method.

3.1. Speech Samples

The popular Berlin Emotional Speech Database (EMO-DB) [9] is used in all simulations in this study. Clean speech samples are uttered by five female and five male actors. Each actor speaks ten sentences in German, containing emotions of anger, happiness, sadness, fear, disgust, boredom, and neutral. Only those utterances scoring higher than 80% emotion recognition rate in a subjective listening test are included in the EMO-DB database [9]. Hence, there are 526 sentences in total with seven classes of emotions. Original speech samples are recorded with 16 kHz sampling frequency under studio condition, and are

downsampled to 8 kHz to cover the fundamental frequencies of male speakers when analyzed by our 5.3-octave frequency coverage cochlear filterbank in our auditory model (see section 2.1). White noise and babble noise are then extracted from the NOISEX-92 database [10] and added to clean speech to simulate various SNR conditions. A simple energy-based VAD is first applied to each clean utterance to determine its active regions. Only durations of active regions are considered in calculating SNR.

3.2. Rate-Scale (RS) Features

As mentioned in section 2.2, rate-scale plots reveal joint spectro-temporal modulations of the speech. The slow modulations, which are related to the speaking rate (i.e., the changing rate of the vocal track), are shown in low rate regions. On the other hand, the energy of resolved pitch is captured in high rate regions. In this study, we consider rates at $\pm 2^{1 \dots 9}$ Hz to cover the complete temporal structures (speaking rate and pitch) of the speech. As for the scale region, we emphasize on the $2^{-1 \dots 3}$ cycle/octave to cover complete frequency structures, from formants (captured by low scales) to harmonics (captured by high scales). Therefore, 90 rate-scale features (9 rates, 5 scales and both directions) are extracted per frame. The mean and standard deviation of these 90 RS features are then calculated over the entire utterance. Finally, 180 RS features per utterance are preserved for emotion recognition.

3.3. MFCC Features

The mel-frequency cepstral coefficients (MFCCs) are widely used in the speech analysis field. Here, the first 13 MFCCs (including the zero-order coefficient) are extracted from 25 ms Hamming-windowed frame every 10 ms with the pre-emphasis coefficient 0.97. The mean, standard deviation, skewness, and kurtosis of these 13 MFCCs, their deltas, and double-deltas are computed as 156 features per utterance.

3.4. Prosodic Features

The 180 RS features mentioned above contain pitch and timbre (i.e., the formant structure) information, however, conventional MFCCs only carry timbre information. To make a fair comparison, prosodic features (pitch, energy and duration) are extracted and combined with MFCC features.

The fundamental frequency (F0) contour is extracted by STRAIGHT [11]. The algorithm estimates the aperiodic power (AP) of each frame. Frames with high AP are assumed unvoiced with zero F0. Only low-AP frames are treated as voiced frames and return valid F0 estimate. The energy contour is extracted every 10 ms with a 25 ms Hamming window. Duration related features are derived from the voiced/unvoiced discrepancy obtained in F0 estimation.

Statistics of these prosodic features used in this study are similar to those used by other researchers [1, 2]. However, not to form a huge feature set with 1000 ~ 4000 parameters, a reasonably small-sized feature set is constructed. As a result, some features are omitted or replaced. For example, the mean of the positive and the negative dF0 are calculated separately to represent the upward and the downward trend, respectively, instead of the mean of all dF0. As for the energy, the minimum value of energy must be close to zero such that the min value, relative position of min, and range would not provide crucial information and hence are dropped from our feature list. Finally, 30 prosodic features are extracted and referred to as the PRO30 feature set. The description of this feature set is given in Table 1.

Table 1. 30 prosodic features

F0 (8 features)	mean, std, max value, relative position of max, min value, relative position of min, range, number of local max point
dF0 (8 features)	mean of positive, mean of negative, std, max value, relative position of max, min value, relative position of min, ratio of positive
logE (3 features)	std, max value, relative position of max
dlogE (8 features)	mean of positive, mean of negative, std, max value, relative position of max, min value, relative position of min, ratio of positive
Duration (3 features)	speaking rate (1/average duration), std of voiced duration, mean pause time

4. Simulation Results

The implemental Support Vector Machine (SVM) algorithm, libsvm [11], is used in this study as the emotion classifier. SVM algorithms are very popular due to their remarkable performance. Many kinds of kernels are available for the SVM to map problems onto higher dimensional spaces. Although the radial basis function (RBF) kernel is suggested to use the first, different choices of parameters would affect results radically [8]. These parameters need to be fine tuned by the grid search for each training condition. Therefore, a simpler linear kernel is adopted in this study only to investigate the robustness of features. Before building the SVM, features extracted from each speaker are processed by the mean subtraction and variance normalization. In addition, all training and testing features are linearly scaled to [0, 1].

To evaluate the robustness of RS features in unknown environments, mismatched tests (clean data for training while noisy data for testing) are performed under various SNR conditions. The 10-fold cross-validation procedures are adopted in our test. Speech samples are randomly divided into 10 subsets. In each trial, one subset is used for testing while the other nine subsets are used for training the SVM recognizer. Final recognition rates are obtained by averaging over 10 trials.

Table 2. Recognition rates (%) in additive white noises

white noise	clean	20dB	15dB	10dB	5dB	0dB
RS180	74.32	73.57	72.98	72.60	72.22	62.91
MFCC156+PRO30	83.09	69.03	65.99	61.42	53.62	52.44

Table 3. Recognition rate (%) in additive babble noises

babble noise	clean	20dB	15dB	10dB	5dB	0dB
RS180	75.48	75.49	74.73	72.25	64.62	49.46
MFCC156+PRO30	82.54	73.50	66.55	61.38	56.78	37.53

Table 2 and 3 show recognition rates of using RS180 and MFCC156+PRO30 features in additive white and babble noises, respectively. RS180 outperforms MFCC156+PRO30 in all SNR conditions (20dB~0dB), except in the clean condition. The MFCC156+PRO30 features from training samples depict magnitude spectra and pitch values with high precision. Such precise representations would produce good matches in clean condition, but are also prone to degradations by noises. On the other hand, RS features only carry the information of

spectro-temporal amplitude modulations, which is equivalent to the spectro-temporal envelopes without carriers' fine structure (phase) information. While not providing accurate matches in the clean condition, RS features are more resistant to deteriorations from spectro-temporal envelopes of noises.

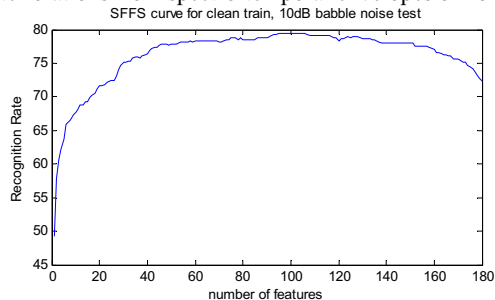


Figure 4: Recognition rate (%) of RS180 by SFFS method

A feature selection method, sequential forward floating selection (SFFS) [12], is used to examine contributions within RS180 features. It starts from an empty feature set and sequentially includes (or excludes) a feature into the selected set, then evaluates the performance of newly constructed feature set. As shown in Figure 4, the performance peaks around using 100 features and does not vary a lot from using 60 to 140 features. Tests on other SNR conditions have the similar trend. These results simply imply our RS features are highly redundant, which is not unexpected due to the highly overlapped two-dimensional filters in the cortical module [7]. Therefore, RS180 can be further downsampled to RS92 by choosing rate-scale combinations of gray spots in Figure 5. Note, only downward direction (positive rate) is shown in the figure.

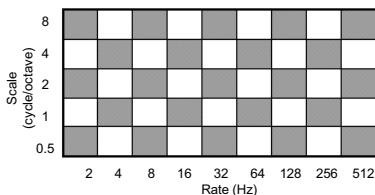


Figure 5: Rate-scale selections (gray areas) of RS92

Two subsets of MFCC156 are selected to compare with our reduced RS92 features. The first subset (MFCC78) contains the mean and standard deviation of 13 MFCCs, 13 Δ MFCCs and 13 $\Delta\Delta$ MFCCs. The second subset (MFCC52) contains the mean, standard deviation, skewness, and kurtosis of 13 MFCCs. Both subsets are then combined with PRO30 features to have comparable feature numbers as RS92. Recognition rates are shown in Table 4 and 5. Results show that RS92 has almost the same performance as RS180 in white noise while performs slightly worse in high-SNR (20, 15 dB) babble noise. The reason for that is white noise has vastly different spectro-temporal modulations from speech, while the babble noise has similar modulations to speech. Hence, a higher resolution in the RS domain is preferred for babble noise. Nevertheless, RS92 outperforms MFCC78+PRO30 and MFCC52+PRO30 in almost all SNR conditions (except in the 20 dB babble noise), especially in low SNR conditions.

Table 4. Recognition rates (%) in additive white noises

white noise	clean	20dB	15dB	10dB	5dB	0dB
RS92	72.79	73.18	72.8	71.66	71.06	64.63
MFCC78+PRO30	84.81	71.27	68.83	62.36	54.94	51.93
MFCC52+PRO30	82.32	72.88	69.59	68.98	60.65	58.56

Table 5. Recognition rate (%) in additive babble noises

babble noise	clean	20dB	15dB	10dB	5dB	0dB
RS92	72.44	72.26	72.07	71.3	64.8	50.01
MFCC78+PRO30	83.11	73.57	71.29	62.73	55.54	42.58
MFCC52+PRO30	82.52	72.44	70	64.45	61.01	42.58

5. Conclusions and Future Work

In this paper, features from spectro-temporal modulations are shown more robust to additive white and babble noises than conventional MFCCs plus prosodic features in mismatched emotion recognition simulations, especially in low SNR (≤ 10 dB) conditions. However, the Berlin database we used only contains acting emotions. Currently, we are extending our works to larger databases with natural emotions, such as the FAU Aibo Emotion Corpus [13].

6. Acknowledgements

This research is partially supported by the National Science Council, Taiwan with Grant No. NSC 98-2221-E-009-092.

7. Reference

- [1] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov Model-Based Speech Emotion Recognition," *Proc. ICASSP*, 2003, vol. 2, pp. 1-4.
- [2] Dan-Ning Jiang, and Lian-Hong Cai, "Speech Emotion Classification with the Combination of Statistic Features and Temporal Features", *ICME, 2004*, pp. 1967-1970.
- [3] V. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Comm.*, vol. 48, no. 9, pp. 1162-1181, September 2006.
- [4] Z. Zeng, M. Pantic, G. I. Rosiman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39 - 58, 2009.
- [5] B. Schuller, D. Arsić, F. Wallhoff, and G. Rigoll, "Emotion Recognition in the Noise Applying Large Acoustic Feature Sets," in *Proc. Speech Prosody*, 2006.
- [6] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards More Reality in the Recognition of Emotional Speech," *Proc. ICASSP*, 2007, Vol. IV, pp. 941-944.
- [7] T. Chi, P. Ru, and S.A. Shamma, "Multi-resolution spectro-temporal analysis of complex sounds," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887-906, 2005.
- [8] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [9] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter Sendlmeier und Benjamin Weiss, "A Database of German Emotional Speech", *Proc. Interspeech*, Lissabon, Portugal, 2005, pp. 489-492.
- [10] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Comm.*, vol.12(3), pp. 247-251, 1993.
- [11] H. Kawahara, Alain de Cheveign'e, H. Banno, T. Takahashi and T. Irino, "Nearly Defect-free F0 Trajectory Extraction for Expressive Speech Modifications based on STRAIGHT," *Proc. Interspeech*, 2005, pp. 537-540.
- [12] P. Pudil, F.J. Ferri, J. Novovicova, J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," *Proc. international Conference on Computer Vision & Image Processing*, pp. 279-283, 1994.
- [13] S. Steidl, Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech, Logos Verlag, Berlin, 2009.