MSF-Based Speaker Automatic Emotional Recognition in Continuous Chinese Mandarin

Yuqiang Qin a, b, * , Xueying Zhang a

*Taiyuan University of Technology, Taiyuan, 030024, China
bTaiyuan University of Science and Technology, Taiyuan, 030024, China

Abstract

In the paper, modulation spectral features (MSFs) are proposed for the automatic emotional recognition for speech signal. The features are extracted from an auditory-inspired long-term spectro-temporal (ST) representation. On an experiment assessing classification of 4 emotion categories, the MSFs show promising performance in comparison with features that are based on mel-frequency cepstral coefficients and perceptual linear prediction coefficients, two commonly used short-term spectral representations. The MSFs further express a substantial improvement in recognition performance when used to augment prosodic features, which have been extensively used for speech emotion recognition. Using both types of features, an overall recognition rate of 91.55 % is obtained for classifying 4 emotion categories.

Keywords: Speech emotional recognition; Modulation spectral features (MSFs); Spectro-temporal (ST); Automatic emotional computing; Speech emotion analysis

Introduction

Affective computing, an active interdisciplinary research field, is concerned with the automatic recognition, interpretation, and synthesis of human emotions[1]. Within its areas of interest, speech emotion recognition (SER) aims at recognizing the underlying emotional state of a speaker from the speech signal[2]. The paralinguistic information conveyed by speech emotions has been found to be useful in multiple ways in speech processing, especially serving as an important ingredient of “emotional
intelligence” of machines and contributing to human–machine interaction[3]. The limitations of short-term spectral features for speech recognition, however, are considerable. On the other hand, advances in neuroscience suggest the existence of spectro-temporal(ST) receptive fields in mammalian auditory cortex which can extend up to temporal spans of hundreds of milliseconds and respond to modulations in the time-frequency domain. In line with these findings, long-term modulation spectral features (MSFs) are proposed in this paper for emotion recognition[4]. These features are based on frequency analysis of the temporal envelopes (amplitude modulations) of multiple acoustic frequency bins, thus capturing both spectral and temporal properties of the speech signal[5].

1. ST representation of emotional speech

The auditory-inspired spectro-temporal (ST) representation of speech is obtained via the steps depicted in Figure 1[6]. The initial pre-processing module resamples the speech signal to 8 kHz and normalizes its active speech level to-26 dBov using the P.56 speech voltmeter. Since emotions can be reliably conveyed through band-limited telephone speech, we consider the 8 kHz sampling rate adequate for SER. Emotional speech frames (without overlap) are labeled as active or inactive by the G.729 voice activity detection (VAD) algorithm described in and only active speech frames are retained. The preprocessed speech signal $s(n)$ is framed into long-term segments $s_k(n)$ by multiplying a 256 ms Hamming window with 64 ms frame shift, where $k$ denotes the frame index. Because the first subband filter in the modulation filterbank (described below) analyzes frequency content around 4 Hz, this relatively long temporal span is necessary for such low modulation frequencies[7].

![Flowchart for deriving the ST representation](image)

It is well-known that the human auditory system can be modeled as a series of over-lapping band-pass frequency channels, namely auditory filters with critical bandwidths that increase with filter center frequencies. The output signal of the $i$th critical-band filter at frame $k$ is given by:

$$s_k(i,n) = s_k(n) * h(i,n)$$

where $h(i,n)$ denotes the impulse response of the $i$th channel, and $*$ denotes convolution. Here, a critical-band gammatone filterbank with $N$ subband filters is employed. The implementation in is used. The center frequencies of these filters (namely acoustic frequency, to distinguish from modulation frequency of the modulation filterbank) are proportional to their bandwidths, which in turn, are characterized by the equivalent rectangular bandwidth:

$$ERB_i = \frac{F_i}{Q_{acq}} + B_{min}$$

where $F_i$ is the center frequency (in Hz) of the $i$th criticalband filter, and $Q_{acq}$ and $B_{min}$ are constants set to 9.26449 and 24.7, respectively. In our simulations, a gammatone filterbank with 19 filters is used, where the first and the last filters are centered at 125 Hz and 3.5 kHz, with bandwidths of 38 and 400 Hz, respectively. The temporal envelope, or more specifically, the Hilbert envelope $H_k(i,n)$, is
then computed from $s_k(i, n)$ as the magnitude of the complex analytic signal $\hat{s}_k(i, n) = s_k(i, n) + jH\{s_k(i, n)\}$, where $H\{\}$ denotes the Hilbert transform. Hence,

$$H_k(i, n) = |\hat{s}_k(i, n)| = \sqrt{s_k^2(i, n) + H^2\{s_k(i, n)\}}$$ (3)

The auditory spectral decomposition modeled by the critical-band filterbank, however, only comprises the first stage of the signal transformation performed in the human auditory system. The output of this early processing is further interpreted by the auditory cortex to extract spectro-temporal modulation patterns. An $M$-band modulation filterbank is employed in addition to the gammatone filterbank to model such functionality of the auditory cortex. By applying the modulation filterbank to each $H_k(i, n)$; $M$ outputs $H_k(i, j, n)$ are generated where $j$ denotes the $j$th modulation filter, $1 \leq j \leq M$. The filters in the modulation filterbank are second-order bandpass with quality factor set to 2. In this work we use an $M = 5$ filterbank whose filter center frequencies are equally spaced on logarithm scale from 4 to 64 Hz. The filterbank was shown in preliminary experiments to strike a good balance between performance and model complexity.

Lastly, the ST representation $E_k(i, j)$ of the $K$th frame is obtained by measuring the energy of $H_k(i, j, n)$, given by:

$$E_k(i, j) = \sum_{k=1}^{L} |H_k(i, j, n)|^2$$ (4)

where $1 \leq k \leq T$ with $L$ and $T$ representing the number of samples in one frame and the total number of frames, respectively. For a fixed $j = j^*$, $E_k(i, j^*)$ relates the auditory spectral samples of modulation channel $j^*$ after criticalband grouping. By incorporating the auditory filterbank and the modulation filterbank, a richer two-dimensional frequency representation is produced and allows for analysis of modulation frequency content across different acoustic frequency channels[8].

2. ST-based modulation spectral features extraction

Two types of MSFs are calculated from the ST representation, by means of spectral measures and linear prediction parameters. For each frame $k$, the ST representation $E_k(i, j)$ is scaled to unit energy before further computation, i.e.

$$\sum_{k=1}^{N} E_k(i, j) = 1$$

Six spectral measures $\Phi_1 - \Phi_6$ are then calculated on a per-frame basis. For frame $k$, $\Phi_{1,k}(j)$ is defined as the mean of the energy samples belonging to the $j$th modulation channel ($1 \leq j \leq 5$):

$$\Phi_{1,k}(j) = \frac{\sum_{i=1}^{N} E_k(i, j)}{N}$$ (5)

Parameter $\Phi_1$ characterizes the energy distribution of speech along the modulation frequency. The second spectral measure is the spectral flatness which is defined as the ratio of the geometric mean of a spectral energy measure to the arithmetic mean. In our calculation, $E_k(i, j)$ is used as the spectral energy measure at frame $k$ for modulation band $j$ and $\Phi_2$ is thus defined as:

$$\Phi_{2,k}(j) = \sqrt[N]{\prod_{i=1}^{N} E_k(i, j)} / \Phi_{1,k}(j)$$ (6)

A spectral flatness value close to 1 indicates a flat spectrum, while a value close to 0 suggests a spectrum with widely different spectral amplitudes. The third measure employed is the spectral centroid
which provides a measure of the “center of mass” of the spectrum in each modulation channel. Parameter $\Phi_3$ for the $j$-th modulation channel is computed as:

$$\Phi_{3,j}(j) = \frac{\sum_{i=1}^{N} f(i) E_k(i,j)}{\sum_{i=1}^{N} E_k(i,j)}$$

(7)

Two types of frequency measure $f(i)$ have been experimented: (1) $f(i)$ being the center frequency (in Hz) of the $i$-th critical-band filter of the auditory filterbank and (2) $f(i)$ being the index of the $i$-th criticalband filter, i.e., $f(i) = i$. No remarkable difference in performance is observed between the two measures, thus the latter is chosen for simplicity. Moreover, given the observation that adjacent modulation channels usually have considerable correlation, the spectral flatness and the centroid parameters of adjacent modulation channels also exhibit high correlation. In order to alleviate such information redundancy, $\Phi_{2,k}(j)$ and $\Phi_{3,k}(j)$ are only computed for $j \in \{1, 3, 5\}$.

3. Experiment and analysis of results

3.1. MSF-based speech spectro-temporal (ST)

Figure 2 shows the ST representation $E(i,j)$ for the 4 emotions in the speech corpus database, where every $E(i,j)$ shown is the average over all the frames and speakers available in the database for an emotion. As illustrated in the figure, the average ST energy distribution over the joint acoustic-modulation frequency plane is similar for some emotions (e.g. anger vs. joy), suggesting they could become confusion pairs, while very distinct for some others (e.g. anger vs. sadness), suggesting they could be well discriminated from each other. As reasonably expected, the less expressive emotions such as neutral and sadness have significantly more low acoustic frequency energy than anger and joy.

Fig.2. Average E(i,j) for 4 emotion categories

3.2. Evaluation

Confusion matrix are shown in Tables 1 (left-most column being the true emotions), for the best recognition performance achieved by prosodic features alone and combined prosodic and proposed features (LDA + SN), respectively. We can see from the confusion matrix that adding MSFs contributes to improving the recognition and precision rates of all emotion categories. It is also shown that most emotions can be correctly recognized with above 89% accuracy, with the exception of joy, which forms
the most notable confusion pair with anger, though they are of opposite valence in the activation–valence emotion space.

Table 1. MSF-base confusion matrix for automatic emotional recognition

| Emotion | Anger | Joy | Sadness | Neutral | Rate(%) |
|---------|-------|-----|---------|---------|---------|
| Anger   | 119   | 8   | 0       | 0       | 93.7    |
| Joy     | 12    | 52  | 0       | 3       | 77.6    |
| Sadness | 0     | 0   | 62      | 0       | 100     |
| Neutral | 2     | 2   | 0       | 75      | 94.9    |
| Precision (%) | 89.5 | 83.9 | 100 | 96.2 |

4. Conclusion

This work presents novel MSFs for the recognition of human emotions in speech. The MSFs are evaluated first on the database to emotional corpus database to classify 4 discrete emotions. Simulation results show that the MSFs serve as powerful additions to prosodic features, as substantial improvement in recognition accuracy is achieved once prosodic features are combined with the MSFs, with up to 91.6% overall recognition accuracy attained.

Acknowledgements

This work is sponsored by Research of Emotional Speech Recognition and HMM Synthesis Algorithm (China-Belgium cooperation) (Shanxi International Cooperation Foundation of Science and Technology of China, NO.2011081047), Research of HMM-Based Emotional Speech Synthesis Algorithm (Shanxi Natural Science Foundation of China, NO. 2010011020-1).

References

[1] Ishi, C., Ishiguro, H., Hagita, N. Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech. EURASIP J. Audio Speech Music Process, article ID 528193, 2010,12 pages.
[2] Falk, T.H., Chan, W.-Y. A non-intrusive quality measure of dereverberated speech. In: Proc. Internat. Workshop for Acoustic Echo and Noise Control, 2008, 155-189.
[3] Falk, T.H., Chan, W.-Y. Modulation spectral features for robust far-field speaker identification. IEEE Trans. Audio Speech Language Process. 18, 2010, 90–100.
[4] Falk, T.H., Chan, W.-Y. Temporal dynamics for blind measurement of room acoustical parameters. IEEE Trans. Instrum. Meas. 59, 2010b, 978–989.
[5] Giannakopoulos, T., Pikrakis, A., Theodoridis, S. A dimensional approach to emotion recognition of speech from movies. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, 2009, 65–68.
[6] Wu, S., Falk, T., Chan, W.-Y. Automatic recognition of speech emotion using long-term spectro-temporal features. In: Proc. Internat. Conf. on Digital Signal Processing, 2009, 1–6.
[7] Wollmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., Cowie, R. Abandoning emotion classes – Towards continuous emotion recognition with modelling of long-range dependencies. In: Proc. Interspeech, 2008, 597–600.
[8] Sun, R., E., M., Torres, J. Investigating glottal parameters for differentiating emotional categories with similar prosodics. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, 2009, 4509–4512.