Convolutional Neural Network with Spectrogram and Perceptual Features for Speech Emotion Recognition

Linjuan Zhang[1], Longbiao Wang[1(✉)], Jianwu Dang[1,2(✉)], Lili Guo[1], and Haotian Guan[3]

[1] Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
{linjuanzhang,longbiao_wang,liliguo}@tju.edu.cn
[2] Japan Advanced Institute of Science and Technology, Ishikawa, Japan
jdang@jaist.ac.jp
[3] Intelligent Spoken Language Technology (Tianjin) Co., Ltd., Tianjin, China
htguan@huiyan-tech.com

Abstract. Convolutional neural network (CNN) has demonstrated a great power at mining deep information from spectrogram for speech emotion recognition. However, perceptual features such as low-level descriptors (LLDs) and their statistical values were not utilized sufficiently in CNN-based emotion recognition. To solve this problem, we propose novel features to combine spectrogram and perceptual features in different levels. Firstly, frame-level LLDs are arranged as time-sequence LLDs. Then, spectrogram and time-sequence LLDs are fused as compositional spectrographic features (CSF). To fully utilize perceptual features and global information, statistical values of LLDs are added in CSF to generate rich-compositional spectrographic features (RSF). Finally, the proposed features are individually fed to CNN to extract deep features for emotion recognition. Bi-directional long short-term memory was employed to identify emotions and the experiments were conducted on EmoDB. Compared with spectrogram, CSF and RSF improve the unweighted accuracy by a relative error reduction of 32.04% and 36.91%, respectively.

Keywords: Speech emotion recognition · Spectrogram
Perceptual features · Convolutional neural network
Bi-directional long short-term memory

1 Introduction

The field of man-machine communication has witnessed a tremendous improvement in recent years. We still have difficulties in communicating with machines naturally. It is believed that speech emotion is particularly useful in human-computer interface, because the emotion carries the essential semantics and helps

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11304, pp. 62–71, 2018.
https://doi.org/10.1007/978-3-030-04212-7_6

machines better understand human speech [1]. However, speech emotion recognition is technically challenging because it is not clear what kinds of speech features are salient to efficiently characterize different emotions [2,3]. The aim of this study is to find new affect-salient features for speech emotion recognition.

Conventional speech emotion recognition approaches rely mostly on feature selection. Perceptual features have been intensively selected to estimate the emotion of speakers [2,4,5]. Perceptual features [6] consist of LLDs and statistical features, which are described in Table 1. LLDs are zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, fundamental frequency (F0), harmonics-to-noise ratio (HNR) by autocorrelation function, and mel-frequency cepstral coefficients (MFCC) 1–12. To each of these, the delta coefficients are additionally computed. Statistical features are statistical values of LLDs. Utilizing deep neural networks (DNN) to learn deep features from perceptual features is common in speech emotion recognition task. For example, DNN is utilized to obtain probability distribution of emotional state from perceptual features, and extreme learning machine (ELM) is used for classification [7]. Some studies proposed that the combination of bi-directional long short-term memory recurrent neural network (BLSTM-RNN) and full-connected neural network with a model of attention performs well when using perceptual features [8,9]. These studies highlight perceptual features based on deep networks for speech emotion recognition.

Table 1. The composition of perceptual features

LLDs (16*2)	Statistical Values
(Δ)ZCR (Δ)RMS energy	Mean, standard deviation, kurtosis, skewness,
(Δ)F0 (Δ)HNR	Extremes: value, rel. position, range
(Δ)MFCC(1-12)	Linear regression: offset, slope, MSE

Perceptual features are chosen by experience and not comprehensive. Thus it is uncertain whether the features selected from our prior knowledge are adequate for good performance in all the situations. Compare to perceptual features, using spectrogram for speech recognition proved to be successful [10–12]. It is recognized that emotional contents of utterances influence spectral energy in the frequency domain [13]. Deep networks with different structures based on raw spectrogram have shown significantly improvements of speech emotion recognition. In [14–16], they extracted deep features from raw spectrogram with CNN to find good results. These studies indicated that CNN can process spectrogram more effective and help identify emotions.

Comprehensive spectrogram can be obtained from the speech directly not by experience from the prior knowledge. If applying CNN on spectrogram alone, it is difficult to sufficiently learn the prior knowledge of perceptual features for automatic speech emotion recognition. In order to overcome this problem,

we propose novel features to utilize the prior knowledge and comprehensive spectrographic information simultaneously. First, frame-level LLDs are arranged in timeline to make time-series LLDs. Then, segmental spectrogram and time-series LLDs are fused as CSF based on timeline. In [19], global features are thought to be important for speech emotion recognition. However, LLDs in CSF contain a wealth of local information. To solve this problem and fully utilize perceptual features, statistical features that are statistical values of LLDs are added in CSF manually to generate RSF. Finally, CNN is employed to extract deep features from our proposed features. It is the first work to combine perceptual features and spectrogram before feature extraction and treat them as 2-D images to be fed into the CNN model so as to extract deep features for emotional classification task.

The outline of this paper is as follows. The baseline system is described in Sect. 2. Section 3 introduces the proposed features and fusion methods. Sections 4 and 5 cover the experiments and conclusions.

2 Baseline System

In this section, our baseline system is described in Fig. 1 according to previous works [15–18]. First, speech signals are split into segments with a fixed length. Secondly, short-time Fourier transform (STFT) are used to transform segmental signals into amplitude spectrogram. When doing STFT, the FFT points are 256. Then, segmental spectrogram are fed to CNN to extract deep spectrogram. Finally, BLSTM is used to identify utterance-level emotions. The baseline system does not make use of the prior knowledge (e.g. F0).

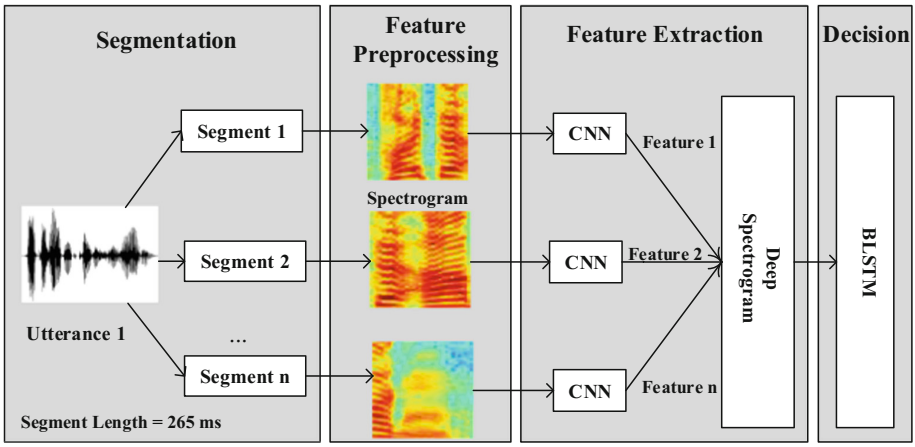


Fig. 1. Structure of the baseline system

The reasons why we primarily focus on CNN-BLSTM are:

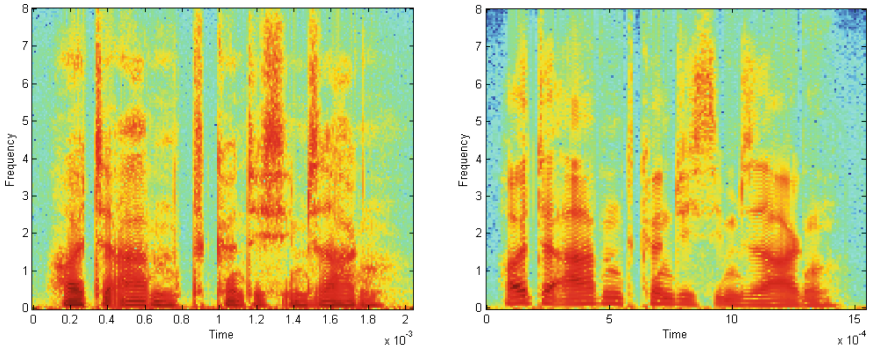
(1) Since CNN models temporal and spectral local correlations [20], it is chosen first to extract deep features from the 2-D representations of our proposed features.

(2) Emotion is manifested in speech through a variable range of temporal dependencies. BLSTM is used to recognize the sequential dynamics in an utterance [21]. It is expected that BLSTM network captures long short-term dependent temporal details of the CNN-based features in a consecutive utterance.

3 CNN Based on Spectrogram and Perceptual Features

3.1 Motivation for Fusing Spectrogram and Perceptual Features

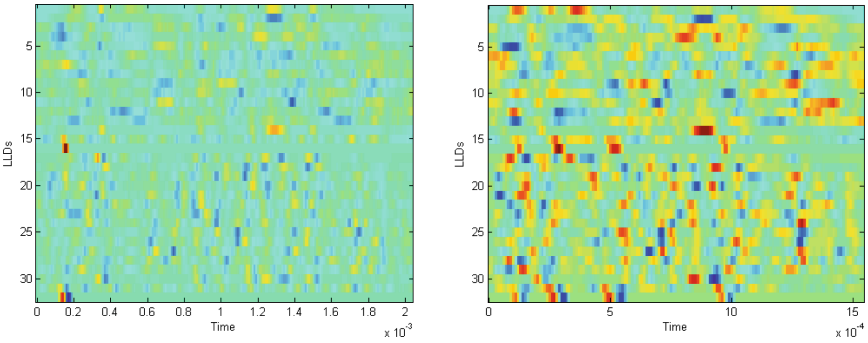
In this section, the motivation of the feature fusion is described from visual and theoretical perspective. Figures 2 and 3 show utterance-level spectrogram and time-sequence LLDs on different emotions with the same contents. Figure 2(a) describes the spectrogram of sadness emotion, and Fig. 2(b) depicts neutral emotion. The depth of reddish color implies the level of frequency energy. It is clear from Fig. 2 that the spectrogram of sadness and neutral standing for low-arousal emotions have similar patterns. In order to classify emotions with similar arousal, utilizing LLDs may be useful.



(a) Sadness (b) Neutral

Fig. 2. Visualization of spectrogram

In Fig. 3, the horizontal axis represents time-domain of utterance. The vertical axis represents the 32-dimensional LLDs. Figure 3(a) and (b) are obviously different, which means the selected LLDs are easier to distinguish similar arousal emotions than spectrogram in this situation. However, it is unclear which kind of features are more effective to distinguish emotions in different cases. In order to adapt the features to various situations, our attempt is to fuse spectrogram and perceptual features for speech emotion recognition.



(a) Sadness (b) Neutral

Fig. 3. Visualization of time-sequence LLDs

From a theoretical perspective, CNN is excellent at mining deep information from raw spectrogram. In the use of wide-band spectrogram, formants are emphasized but F0 is not, whereas F0 is known to compose the main vocal cue for emotion recognition [22]. Perceptual features can provide lots of prior knowledge (e.g. F0) that is useful for emotion recognition. In order to make use of spectrogram and perceptual features simultaneously, we propose novel features such as CSF and RSF for speech emotion recognition.

3.2 Fusion Strategy of CSF and RSF

In this study, the fusion Strategy of CSF and RSF consists of three steps.

The first step is to calculate segmental spectrogram. After the STFT, raw spectrographic matrix is obtained with the size of 25×129, where 25 is the number of time points, and 129 depends on the selected region and frequency resolution.

The next step is to utilize the openSMILE toolkit to get frame-level LLDs and segment-level statistical features that have been described in Table 1. LLDs are organized in time series. Each 25 frame-level LLDs constitute a segmental time-sequence LLDs. After normalization, the matrix of time-sequence LLDs is obtained with the size of 25×32, where 25 represents the number of frames in a segment, and 32 is the dimension of LLDs.

The third step is the feature fusion. Based on the timeline, segmental spectrogram and time-sequence LLDs are spliced together as CSF, where the size of CSF is 25×161. CSF vector of the j-th segment in the i-th utterance can be formulated as:

$$CSF_{ij} = [S_{ij}, L_{ij}], \tag{1}$$

where the S_{ij}, L_{ij} correspond to spectrogram vector and time-sequence LLDs vector of the j-th segment in the i-th utterance, respectively.

In order to splice spectrogram, time-sequence LLDs and statistical features, 384-dimensional statistical features are firstly reduced to 375 dimensions using PCA. Then, the resized statistical features are reshaped as 25×15. Finally, segmental spectrogram, time-sequence LLDs and statistical features are spliced together as RSF, where the size of RSF is 25×176. RSF vector of the j-th segment in the i-th utterance can be formulated as:

$$RSF_{ij} = [S_{ij}, L_{ij}, C_{ij}], \tag{2}$$

where the C_{ij} represents statistical features vector of the j-th segment in the i-th utterance. Figure 4 depicts the detailed feature extraction of RSF.

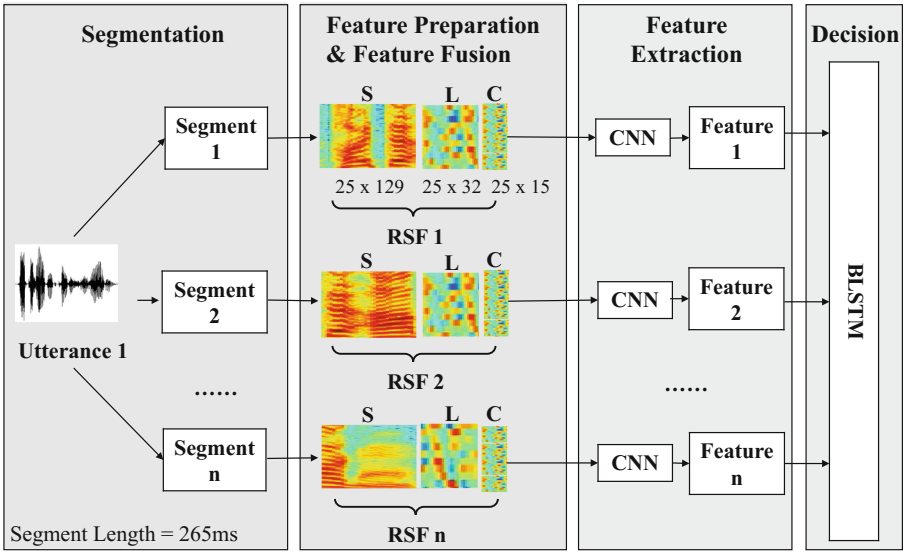


Fig. 4. Extraction of RSF. S represents spectrogram. L represents time-sequence LLDs. C represents statistical features.

4 Experiment

4.1 Experimental Setup

We choose speech materials from the EmoDB [23], which has seven categorical emotion types including disgust, sadness, fear, happiness, neutral, boredom and angry, where the number of utterances in each category are 46, 62, 69, 71, 79, 81 and 127, respectively. There are 535 simulated emotional utterances in German. All the utterances of approximately 2–3 s are sampled at 16000 Hz. The arousal is a descriptor of the intensity of the emotion. In terms of the arousal space [24], angry, fear, disgust and happiness belong to the high-arousal emotion, while, sadness, boredom and neutral belong to the low-arousal emotion.

According to [25], a speech segment contains sufficient emotional contents longer than 250 ms. In our experiment, the utterances are split into segments with a 265-ms window size. Each segment is divided into 25 frames using a 25-ms window, shifting 10 ms each time. About 50,000 segments are collected in this way. Table 2 depicts the detail of the network. Other parameters are also tested for experiments, and the configuration of Table 2 resulted the best performance.

Table 2. Parameters of the CNN-BLSTM network

Layers	Parameters
Convolution 1	32 filters of 5×5
Max-Pooling	2×2
Convolution 2	64 filters of 5×5
Max-Pooling	2×2
Dense layer	Length 1024
LSTM	Bi-directional, 200
LSTM	Bi-directional, 200
Dense layer	Length 7, softmax

Due to the limited size of the Berlin Emotion Database, we run a 10-fold cross validation. The weighted accuracy (WA), unweighted accuracy (UA), F1 and relative error reduction are used to evaluate the results. WA is the accuracy of all the test utterances. UA is defined as average of per emotional category recall. F1 is the harmonic average of precision and recall. Relative error reduction is the ratio of error reduction to original error.

4.2 Experimental Results

This section shows the classification results of our proposed features. From Table 3, we conclude: (1) Compared with spectrogram, the proposed time-sequence LLDs improve the WA and UA by a relative error reduction of 11.23% and 10.29%, respectively. One of the reasons is that time series information of LLDs is used more adequately by BLSTM. Another reason is that selected LLDs perform better than raw spectrogram on a small amount of training data. (2) CSF outperforms spectrogram with 33.76% and 32.04% relative error reduction in terms of WA and UA, respectively. RSF outperforms spectrogram with 38.06% and 36.91% relative error reduction in terms of WA and UA, respectively. The results reveal that spectrogram and perceptual features are complementary. Moreover, our proposed features are significantly effective. (3) RSF performs better than CSF. The results indicate that it is useful to add additional statistical features into CSF. And it is effective to reshape the statistical features and treat them as an additional graph for CNN to learn.

Table 3. WA and UA of different features with CNN-BLSTM

ID	Features	Components	Size	WA (%)	UA (%)
1	Spectrogram (baseline)	Spectrogram	25 × 129	86.73	86.40
2	Time-sequence LLDs (proposed)	LLDs	25 × 32	88.22	87.80
3	CSF (proposed)	Spectrogram + LLDs	25 × 161	91.21	90.76
4	RSF (proposed)	Spectrogram + LLDs + Statistical features	25 × 176	**91.78**	**91.42**

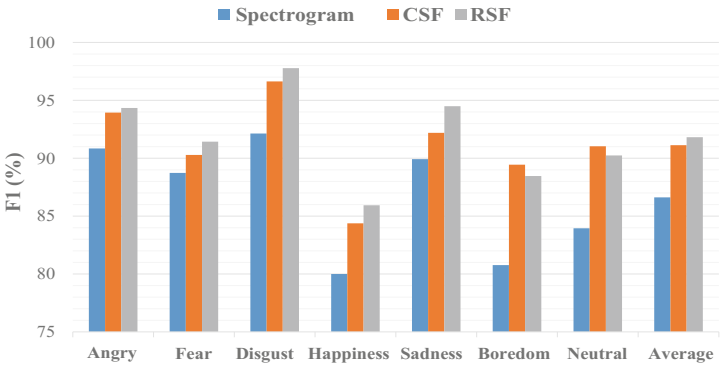


Fig. 5. The F1(%) of different features on different emotions

Figure 5 shows the contribution of proposed features on classifying different kinds of emotion in comparison to spectrogram (baseline features). (1) The results of CSF and RSF are both better than spectrogram on all kinds of emotions, especially on boredom emotion. (2) RSF performs better than CSF on most kinds of emotions. However, when classifying boredom and neutral, CSF performs better than RSF. We assume that there is no noticeable changes on LLDs in both boredom and neutral utterances. Therefore, it is unnecessary to add extra statistical features in this situation. (3) On average F1 of seven emotions, CSF and RSF significantly outperform spectrogram by relative error reduction of 33.68% and 38.80%, respectively. Overall, both CSF and RSF are effective for identifying different categories of emotions.

5 Conclusions and Future Works

In this paper, we first proposed time-sequence LLDs, CSF and RSF for speech emotion recognition. Then, the proposed features were individually fed into the CNN model to extract deep features. Finally, the BLSTM was employed to do

final classification. It is the first work to combine spectrogram and perceptual features simultaneously for speech emotion recognition. Our results indicated that spectrogram and perceptual features were complementary and our proposed features were effective.

For future work, we will evaluate our proposed features on other large datasets and consider integrating speaker, gender and linguistic features in our experiment.

Acknowledgments. The research was supported by the National Natural Science Foundation of China (No. 61771333 and No. U1736219) and JSPS KAKENHI Grant (16K00297).

References

1. Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., Wróbel, M.R.: Emotion recognition and its applications. In: Hippe, Z.S., Kulikowski, J.L., Mroczek, T., Wtorek, J. (eds.) Human-Computer Systems Interaction: Backgrounds and Applications 3. AISC, vol. 300, pp. 51–62. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08491-6_5
2. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. **44**(3), 572–587 (2011). https://doi.org/10.1016/j.patcog.2010.09.020
3. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. **53**(9–10), 1062–1087 (2011). https://doi.org/10.1016/j.specom.2011.01.011
4. Ringeval, F., et al.: Av+ ec 2015: the first affect recognition challenge bridging across audio, video, and physiological data. In: 5th International Workshop on Audio/Visual Emotion Challenge, pp. 3–8. ACM (2015). https://doi.org/10.1145/2808196.2811642
5. Valstar, M., et al.: Avec 2016: depression, mood, and emotion recognition workshop and challenge. In: 6th International Workshop on Audio/Visual Emotion Challenge, pp. 3–10. ACM (2016). https://doi.org/10.1145/2964284.2980532
6. Schuller, B., Steidl, S., Batliner, A.: The Interspeech 2009 emotion challenge. In: Tenth Annual Conference of the International Speech Communication Association (2009)
7. Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. In: INTERSPEECH, pp. 223–227 (2014). https://www.microsoft.com/en-us/research/publication/speech-emotion-recognition-using-deep-neural-network-and-extreme-learning-machine/
8. Huang, C. W., Narayanan, S. S.: Attention assisted discovery of sub-utterance structure in speech emotion recognition. In: INTERSPEECH, pp. 1387–1391 (2016). https://doi.org/10.21437/interspeech.2016-448
9. Mirsamadi, S., Barsoum, E., Zhang, C.: Automatic speech emotion recognition using recurrent neural networks with local attention. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227–2231. IEEE (2017). https://doi.org/10.1109/icassp.2017.7952552
10. Variani, E., Lei, X., McDermott, E., Moreno, I. L., Gonzalez-Dominguez, J.: Deep neural networks for small footprint text-dependent speaker verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4052–4056. IEEE (2014). https://doi.org/10.1109/icassp.2014.6854363

11. Hannun, A., et al.: Deep Speech: Scaling up End-to-end Speech Recognition (2014). http://arxiv.org/abs/1412.5567
12. Amodei, D., et al.: Deep Speech 2: end-to-end speech recognition in English and Mandarin. In: International Conference on Machine Learning, pp. 173–182 (2016). http://dl.acm.org/citation.cfm?id=3045390.3045410
13. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden markov models. Speech Commun. **41**(4), 603–623 (2003). https://doi.org/10.1016/S0167-6393(03)00099-2
14. Huang, Z., Dong, M., Mao, Q., Zhan, Y.: Speech emotion recognition using CNN. In: 22nd ACM international conference on Multimedia, pp. 801–804. ACM (2014). http://doi.acm.org/10.1145/2647868.2654984
15. Lim, W., Jang, D., Lee, T.: Speech emotion recognition using convolutional and recurrent neural networks. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1–4. IEEE, Asia-Pacific (2016). https://doi.org/10.1109/apsipa.2016.7820699
16. Satt, A., Rozenberg, S., Hoory, R.: Efficient emotion recognition from speech using deep learning on spectrograms. In: INTERSPEECH, pp. 1089–1093 (2017). https://doi.org/10.21437/interspeech.2017-200
17. Guo, L., Wang, L., Dang, J., Zhang, L., Guan, H.: A feature fusion method based on extreme learning machine for speech emotion recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2666–2670 (2018). https://doi.org/10.1109/icassp.2018.8462219
18. Guo, L., Wang, L., Dang, J., Zhang, L., Guan, H., Li, X.: Speech emotion recognition by combining amplitude and phase information using convolutional neural network. In: INTERSPEECH, pp. 1611–1615 (2018). https://doi.org/10.21437/interspeech.2018-2156
19. Hu, H., Xu, M.X., Wu, W.: Fusion of global statistical and segmental spectral features for speech emotion recognition. In: INTERSPEECH, pp. 2269–2272 (2007)
20. Yu, D., et al..: Deep convolutional neural networks with layer-wise context expansion and attention. In: INTERSPEECH, pp. 17–21 (2016). https://doi.org/10.21437/interspeech.2016-251
21. Lee, J., Tashev, I.: High-level feature representation using recurrent neural network for speech emotion recognition. In: Sixteenth Annual Conference of the International Speech Communication Association (2015). https://www.microsoft.com/en-us/research/publication/high-level-feature-representation-using-recurrent-neural-network-for-speech-emotion-recognition/
22. Petrushin, V. A.: Emotion recognition in speech signal: experimental study, development, and application. In: Sixth International Conference on Spoken Language Processing, pp. 222–225 (2000)
23. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B.: A Database of German Emotional Speech. In: Ninth European Conference on Speech Communication and Technology, pp. 1517–1520 (2005)
24. Xie, B.: Research on Key Issues of Mandarin Speech Emotion Recognition [Ph.D. Thesis]. Hangzhou: Zhejiang University (2006)
25. Provost, E. M.: Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3682–3686. IEEE (2013). https://doi.org/10.1109/icassp.2013.6638345