Research Progress of Speech Emotion Recognition Based on Discrete Emotion Model

Chunfeng Yang¹, Jiajia Lu¹*, Qiang Wu¹ and Huiyu Chen¹
¹ School of computer science and Engineering, Northwest Normal University, Lanzhou, GanSu Province, 730000, China
*Corresponding author’s e-mail: 2020221935@nwnu.edu.cn

Abstract. Speech emotion recognition is one of the important technologies of human-computer interaction, and neural networks have made great contributions in it. In this survey, the commonly used discrete emotion databases and speech emotion feature parameters are firstly introduced. Then the feature extraction methods and emotion recognition models in recent three years’ research in China are summarized, and the former one can be divided into two directions, selecting the feature subset from the existing features and using neural network to extract new features. Finally, the challenges encountered in speech emotion recognition are described, and the future research and development of speech emotion recognition based on discrete emotion models are prospected.

1. Introduction

Speech is one of the main means of human communication, whose acoustic signal contains rich speaker information, semantic information and emotional information. Different information components promote the development of phonetics in different directions: speech recognition, voiceprint recognition and emotion recognition [1], and speech emotion recognition can be described as a technology to deduce the emotion type of speech from features of emotion signal through computer processing [2]. Therefore, it can be divided into two parts: feature extraction and classification, from which this paper analyzed.

There are two kinds of emotion theory models commonly used in speech emotion recognition technology, discrete emotion model or dimensional emotion model [3]. The former defines "basic emotion" and thinks that other emotions are modified and combined by "basic emotion", which is concise, easy to study, but weak in expressing complex emotions. The latter regards emotion as a gradual and smooth transition and describes the emotional state in the multi-dimensional emotional space, which is easy to cover all the emotional states, but difficult to convert the qualitative emotional state to the quantitative spatial coordinates[4]. Therefore, the research based on discrete emotion model is more extensive, and this paper mainly analyzes the relevant research on discrete speech emotion recognition in recent three years in China.

The most commonly used discrete speech emotion data sets in China are Berlin EMO-DB German emotion database[5] and CASIA Chinese emotion database. The former was recorded by Technische Universität Berlin through inviting 5 actors and 5 actresses to simulate 7 emotions respectively, which contains 535 voice sentences; the latter was obtained by Institute of Automation, Chinese Academy of Sciences through recording the performance of 5 different emotions by 2 actors and 2 actresses, which contains 9600 voice sentences. In addition, some researchers used Belfast English emotion database or King Saud University Emotions(KSU), which is an Arabic emotion corpus. Besides, voice signals
extracted from video databases, such as eNTERFACE’05 data set and IEMOCAP data set [6], have also been used.

2. Feature selection and its improvement

2.1 Emotional features of speech

There is no acknowledged speech emotion features at home and abroad so far, and a commonly used classification method is dividing it into acoustic emotional features and semantic emotional features. Acoustic emotional is shown in Table 1, and semantic emotional features contains linguistic feature, which can be obtained by extracting semantic information, and contextual feature, which describes differences in gender and cultural background [7]. Moreover, the features of speech emotion recognition can also be obtained by feature extraction of speech signals or spectrograms through neural networks.

| Table 1 Traditional Speech Emotion Feature Classification |
|-----------------------------------------------|
| feature | description | specific features |
|----------|-------------|-------------------|
| Rhythmic Features | Reflect changes in speech, including intonation, pitch, and rhythm, etc. | Fundamental frequency, speech energy and duration, etc. |
| Spectral Correlating Features | Describe the correlation of vocal cord movement and changes in vocal tract shape. | Linear spectrum includes linear prediction coefficients, linear prediction cepstrum coefficients, etc. The most famous cepstrum is the Mel frequency cepstrum coefficient. |
| Quality Features | Reflect whether the speech is pure, clear, smooth and so on. | Formant peak, bandwidth, frequency perturbation and amplitude perturbation, etc. |
| Non-linear Features | The geometric features which obtained by using the phase space reconstruction technique. | Hurst exponent, Lyapunov exponent, etc. |

2.2 Research on feature selection

In recent research, the features used can be roughly divided into following two types: selecting the feature subset with better distinguishing ability from the existing features, and using neural network to explore new features with stronger emotional expression ability.

There are usually two ways to select the feature subset from the existing acoustic features, one is manual selection, and the other is neural network selection.

Duan Junyi et al. input speech signals, Mel frequency cepstral coefficients(MFCC) and linear frequency cepstral coefficients(LFCC) into the convolutional neural network with corresponding dimensional convolution kernel, and found that the accuracy using MFCC and LFCC are both higher than the one which using the original speech. The accuracy of emotion recognition of MFCC and LFCC is similar, but performances are different on different data sets: the accuracy of LFCCs on the Emo-DB data set is higher, reaching 92.1%; but the opposite is true on the SAVEE data set, MFCC have a higher accuracy rate of 88.3%[8]. Liu Mingzhu et al. selected three main features, pitch frequency, amplitude energy and formant [9] into the later classification model; Zhang Yusha et al. used the Mel frequency cepstral coefficient(MFCC) and the Mel energy spectrum dynamic coefficient(MEDC) obtained after speech preprocessing to classify speech[10]; Zhang Huiyun et al. used normalized 13-dimensional MFCC, zero crossing rate, spectral center of gravity, harmonic noise ratio and pitch [11]; Yu Hua selected 34 different acoustic features, including 3 time domain features, 5 frequency domain features, 13 MFCC features and 13 chromaticity features [12]; Jiang Pengxu et al. used OpenSMILE software to extract 384-dimensional emotional features from commonly used prosodic features, sound quality features and spectral features [13].
It is also a common feature screening method that filtering traditional acoustic features through neural networks to reduce the feature dimension. After training 13 acoustic features on SVM, BP and RF networks, Chu Yu fused three acoustic features, MFCC, pitch frequency and formant, which performed well on multiple data sets, and compared it with single feature and random combination of features, the results show that the fusion features extracted in that paper can be well recognized on different data sets than others[14]. According to the attention parameters corresponding to each feature in the attention matrix obtained by training, Hu Tingting et al. selected the top 51 ones among 88 acoustic features to conduct experiments, and the results show that the accurate rate has increased slightly while reducing the dimensionality, and the performance has remained stable on other data sets[15].

There are several ways to use neural network to complete feature selection. One is to extract the features of original speech signals [16], spectrum [17-20], or various acoustic features and their combinations[21, 22] through the traditional convolutional neural network or recurrent neural network and their variants. The neural networks are trained with the features extracted from the spectrogram by PCANET [23] and the traditional acoustic features separately, and then the voting mechanism is used to complete the classification [24], and the process is shown in Figure 1. The other one is to use DBM to extract features from combination of prosodic features, Mel, nonlinear attributes and nonlinear geometric features[25, 26]. The multi-layer DBM feature extraction networks used are shown in Fig. 2 and Fig. 3 respectively. In addition, some researchers used hybrid neural network models for feature extraction, for example, the CGRU model proposed by Zheng Yan, which input 40-dimensional MFCC features into one-dimensional convolution (CNN) at first, and then input the results to gated recurrent unit(GRU) along time steps to obtain global emotional features, finally, filtered all the emotional features acquired through random forest [27]; Hu Desheng input the acoustic features and its statistical features in the segment into the BLSTM-Attention network as the main network to extract the depth segment features, input the Mel spectrogram into the CNN-GAP network as an auxiliary network to extract deep Mel spectrogram features, and then used the feature fusion as the input of the classification model[28].

![Figure 1 Flow chart of fused speech emotion recognition algorithm][24]

![Figure 2 multi-layer DBM feature extraction networks][25]

![Figure 3 multi-layer DBM feature extraction networks][26]
Moreover, some researchers used multi-modal information to assist emotion recognition, which includes expression, posture, text and etc. Yao Yiqin extracted the features in the spectrogram and text through the convolution-LSTM recurrent neural network with attention mechanism and the BERT model respectively, and then used the key feature selection network GATAS A (global acoustic-to-text and acoustic-to-self-acoustic to text) to obtain deep features, which composed of GATA (global acoustic text acoustic) and ASATA (acoustic-to-self-acoustic to text)[29].

3. Classification model and its improvement

Researchers mainly made improvements on the basic models of LSTM, SVM, and CNN, for example, Wang Chuanyu et al. input the fusion features into the three-layer LSTM network, replacing the fully connected layer with a global average pooling layer [25]; Gao Fan et al. input the features selected by the DBM into the LSTM, using cost function based on mean square error and cross entropy[26]; Zhang Yusha et al. used SVM based on radial basis function (RBF) kernel to train input features[10].

A more extensive operation is to add the attention mechanism into the LSTM. For example, Liu Tianbao, input the hidden state and unit state of the stacked LSTM to the self-attention mechanism module [17], which is shown in Figure 4; Zhang Huiyun used the Attention Advanced LSTM(AA-LSTM) network and multi-task learning method [11], to enhance the performance of the network, whose network structure is shown in Figure 5. Feng Tianyi also used multi-task learning to assist speech emotion recognition [22], for the existing literature shows that emotional state can be considered gender-dependent, and there is a certain connection between speaker identity and emotion.

A summary of the above-mentioned representative speech emotion recognition research results is shown in Table 2. As the researches adopted discrete emotion models and conducted research on discrete emotion data sets, final task was classification and the evaluation indicators were accuracy or a combination of weighted accuracy (WA) and unweighted accuracy (UA). In view of the limited sample size of a single data set and the differences in emotional expression of different languages, even if the same methods were used in different data sets, the accuracy rate were quite different. Comparing experiments on CNN, SVM, and LSTM on IEMOCAP, Emo-DB, and CASIA, the overall accuracy of Emo-DB is "over 10%" [30]. Similar results obtained in the test of BP and RF[14], so when different models are tested on different data sets, the accuracy shown in the literature is not enough to comprehensively compare the performance of different models.

| Literature | Feature Selection | Classification Model | Data Set | Evaluating Indicator | Data |
|------------|------------------|---------------------|---------|----------------------|------|
| [8](2020)  | MFCC, LFCC       | CNN                 | EMO-DB, SAVEE | Accuracy            | 92.1%, 86.5% |
| [10](2020) | MFCC, MEDC       | Improved SVM        | EMO-DB,   | Accuracy             | 82%  |

Figure 4 Schematic diagram of LSTM model with attention mechanism

Figure 5 Schematic diagram of AA-LSTM

Table 2 Statistical table of discrete speech emotion recognition research
| Reference | Feature Type | Model | Dataset | Accuracy |
|-----------|--------------|-------|---------|----------|
| [11] (2021) | Normalized acoustic feature | AA-LSTM | EMO-DB | 70.09% |
| [13] (2019) | Acoustic feature | Improved Lenet-5 model | CASIA, EMO-DB | 85.8%, 85.7% |
| [15] (2019) | Sub feature collection screening by attention coefficient | LSTM | IEMOCAP, eNTERFACE’05 | WA, UA | 64.0%, 63.9% |
| [16] (2021) | Features from speech signal extracted by CNN | CRNN model with improved BLSTM output mode | IEMOCAP | WA, UA | 71.39%, 61.06% |
| [17] (2021) | Features of speech signal extracted by CNN | LSTM with interlayer attention mechanism | RML, AFEW6.0, eNTERFACE’05 | Accuracy | 90.11%, 54.73%, 59.32% |
| [18] (2019) | Features of spectrum extracted by CNN | The multi kernel function is applied to SVM | EMO-DB | Accuracy | 85% |
| [21] (2019) | MFCC | MTL-RNN | CASIA, KSU | Accuracy | 90.87%, 91.86% |
| [22] (2019) | Features from spectrum extracted by CNN | LSTM | EMO-DB | Accuracy | 91.74% |
| [24] (2020) | Artificial features, features extracted from spectrogram by PCANET | SVM of differential voting mechanism | Self made data sets, EMO-DB | Accuracy | 76.6%, 86.57% |
| [25] | Features extracted from acoustic emotional features by DMB | LSTM, GAP, softmax | Dataset cut from Cheavd2.0 | Accuracy | 89.1% |
| [26] (2020) | Features extracted from acoustic emotional features by DMB | LSTM | EMO-DB | Accuracy | 99% |
| [28] | Depth segment feature, Deep Mel spectrogram features | softmax | IEMOCAP | WA, UA | 74.45%, 72.50% |
| [29] (2021) | Fusion features of the sonogram features and text features again by GATASA | FNN+ softmax | IEMOCAP | WA, UA | 77.0%, 77.7% |

**4. Summary and outlook**

At present, the data set of speech emotion recognition is not rich enough and the sample size is limited. Besides, different culture and language lead to different expression of emotion. All this limited the features extracted from the data set, resulting in poor recognition accuracy and generalization ability. Furthermore, the reliability of the data set is limited: noise, acoustic variation caused by different
acquisition methods, and the ability of the recorder to control emotions affect the accuracy. Therefore, building a more reliable and rich open corpus is of great significance, and the application of few-shot learning or transfer learning in speech emotion recognition is expected to solve this problem, two.

There are no acknowledged effective features in discrete speech emotion recognition. Two main ideas are selecting the feature subset with the best distinguishing ability from the existing acoustic features, or using neural network to explore new effective features. It is certain that extracting effective features will still be an important topic of speech emotion recognition. Under the background of interdisciplinary integration, perhaps the research of brain science and psychology can be expected to promote it.

As for classification model, researchers added new modules in the traditional models, attention mechanism or kernel function, or adopt hybrid model to improve the recognition accuracy. However, due to the differences in emotional features of different data sets, it is difficult to accurately compare the performance of different models with the accuracy only. The performance in different language datasets of the model trained with multiple datasets are also different, which shows that the generalization ability is not strong enough. How to apply transfer learning to speech emotion recognition to improve the generalization ability in different datasets is also an important direction.

References

[1] Zheng, C.J., Wang, C.L., Jia, N. (2020) Survey of Acoustic Feature Extraction in Speech Tasks. Computer Science, 47:110–9 (in Chinese).
[2] Sun, X.H., Li, H.J. (2020) Overview of speech emotion recognition. Computer Engineering and Applications, 56(11): 1-9 (in Chinese).
[3] Han, W.J., Li, H.F., Ruan, H.B., Ma, L. (2014) Review on speech emotion recognition. Ruan Jian Xue Bao/Journal of Software, 25(1): 37–50 (in Chinese).
[4] Li, H.F., Chen, J., Ma, L., Bo, H.J., Xu, C, Li, H.W. (2020) Dimensional speech emotion recognition review. Ruan Jian Xue Bao/Journal of Software, 31(8):2465–91 (in Chinese).
[5] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeyer, W. F. and Weiss, B., 2005. A database of German emotional speech. In: INTERSPEECH. Lisbon, Portugal:1517–20.
[6] Busso, C., Bulut, M., Lee C.-C., et al. (2009) IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42: 335-9.
[7] Gao, Q.J, Zhao, Z.H., Xu, D., et al. (2020) Review on speech emotion recognition research. CAAI transactions on intelligent systems, 15(1): 1–13 (in Chinese).
[8] Duan, J.Y., Zhao, J.F. (2021). Analysis and Comparison of Speech Emotion Recognition in Time-Frequency Domain Based on CNN. https://kns.cnki.net/kcms/detail/15.1049.N.20210610.1803.002.html (in Chinese).
[9] Liu, M.Z., Li, X.Q., Chen, H.H. (2019) Research on Speech and Emotional Recognition Algorithm Based on Support Vector Machine. Journal of Harbin University of Science and Technology, 24:118–26 (in Chinese).
[10] Zhang, Y.S., Jiang, S.Y. (2020) SPEECH EMOTION DATA MINING CLASSIFICATION AND RECOGNITION METHOD BASED ON MFCC FEATURE EXTRACTION AND IMPROVED SVM. Computer Applications and Software, 37: 160–5+212 (in Chinese).
[11] Zhang, H.Y., Huang, H.M., Huang, Z.D. (2021) AA-LSTM Network- Based Speech Emotion Recognition. Computer Simulation, 38: 211–6 (in Chinese).
[12] Yu, H., Yan, B.C. (2020) Speech Emotion Recognition Based on CTC-RNN. Chinese Journal of Electron Devices, 43: 934–7 (in Chinese).
[13] Jiang, P.X., Fu, H.L., Tao, W.H., Lei, P.Z., (2019) Feature Characterization Based on Convolution Neural Networks for Speech Emotion Recognition. Chinese Journal of Electron Devices, 42:998–1001 (in Chinese).
[14] Chu, Y., Li, T.G., Ye, S., Ye, G.M. (2020) Research on feature selection method in speech emotion recognition. Journal of Applied Acoustics, 38: 216–222(in Chinese).
[15] Hu, T.T., Feng, Y.Q., Shen, L.J., Wang, W. (2019) The salient feature selection by attention mechanism based LSTM in speech emotion recognition. Technical Acoustics, 38: 414–21.
[16] Lô, H.L., Hu, W.P. (2021) Research on speech emotion recognition based on end-to-end deep neural network. Journal of Guangxi Normal University (Natural Science Edition), 39(3): 20–6(in Chinese).

[17] Liu, T.B, Zhang, L.T., Yu, W.T., et al. (2021) Hierarchical LSTM-Based Audio and Video Emotion Recognition with Embedded Attention Mechanism. Laser & Optoelectronics Progress, 58: 183–190.

[18] Wang, Z.M., Liu, G., Song, H. (2019) Speech emotion recognition method based on multiple kernel learning feature fusion. Computer Engineering, 45(8): 248–254(in Chinese).

[19] Miao, Y.Q., Zou, W., Liu, T.L., et al. (2019) Speech emotion recognition model based on parameter transfer and convolutional recurrent neural network. Computer Engineering and Applications, 55(10): 135–140(in Chinese).

[20] Qiao, D., Chen, Z.J, Deng, L., Tu, C.L., 2021. Speech Emotion Recognition Based on Improved Speech Processing and Convolution Neural Network. https://kns.cnki.net/kcms/detail/31.1289.tp.20210207.1830.014.html

[21] Yang, M.J, Zhang, J.B. (2019) Method of speech emotion recognition method based on deep neural networks. Science Technology and Engineering, 19(8): 127–131(in Chinese).

[22] Feng, T.Y., Yang, Z. (2019) Speech Emotion Recognition Algorithm Based on Multi Task Learning and Recurrent Neural Network. Journal of Signal Processing, 35(7): 1133–40(in Chinese).

[23] Chan, T.H., Jia, K., Gao, S., et al. (2015) PCANet: A Simple Deep Learning Baseline for Image Classification?. IEEE Transactions on Image Processing, 24(12):5017–32(in Chinese).

[24] Zhao, X.L., Xu, X.B. (2020) SPEECH EMOTION RECOGNITION BASED ON SHALLOW LEARNING AND DEEP LEARNING MODEL. Computer Applications and Software, 37: 108–12+76(in Chinese).

[25] Wang, C.Y., Li, W.X, Chen, Z.H.,2021. Research of Multi-modal Emotion Recognition Based on Voice and Video Images. https://kns.cnki.net/kcms/detail/11.2127.TP.20210610.0922.004.html.

[26] Gao, F., Zhang, X.Y., Huang, L.X., Li, B.Y.(2020) Multi-feature speech emotion recognition based on DBM-LSTM. Computer Engineering and Design, 41: 465–4706(in Chinese).

[27] Zhen, Y., Chen, J.N., Wu, F., Fu, B. (2020) Research and Implementation of Speech Emotion Recognition Based on CGRU Model. Journal of Northeastern University (Natural Science), 41: 1680–85(in Chinese).

[28] Hu, D.S., Zhang, X.Y., Zhang, J., Li, B.Y., (2021).Feature Fusion Based on Main-Auxiliary Network for Speech Emotion Recognition. https://kns.cnki.net/kcms/detail/14.1220.N.20210401.1511.004.html.

[29] Yao, Y.Q., Guo, W. (2021) Multimodal emotion recognition algorithm based on interactive attention mechanism. Application Research of Computers, 38: 1689-93(in Chinese).

[30] Zhong, Q., Feng, Y.Q., Wang, W. (2019) Comparison of speech emotion recognition in cross language corpus. Journal of Nanjing University(Natural Science), 55: 765–73(in Chinese).