Speech Emotion Recognition algorithm based on deep learning algorithm fusion of temporal and spatial features

Speech Emotion Recognition algorithm based on deep learning algorithm fusion of temporal and spatial features

XuDong An[1]*, Zhou Ruan[1]

[1.]China Telecommunication Technology Labs, China Academy of Information and Communications Technology, Beijing 100000, China

*correspondence author, anxudong@caict.ac.cn

Abstract. In recent years, human-computer interaction systems are gradually entering our lives. As one of the key technologies in human-computer interaction systems, Speech Emotion Recognition(SER) technology can accurately identify emotions and help machines better understand users' intentions to improve the quality of human-computer interaction, which has received a lot of attention from researchers at home and abroad. With the successful application of deep learning in the fields of image recognition and speech recognition, scholars have started to try to use it in SER and have proposed many deep learning-based SER algorithms. In this paper, we conducted an in-depth study of these algorithms and found that they have problems such as too simple feature extraction methods, low utilization of human-designed features, high model complexity, and low accuracy of recognizing specific emotions. For the data processing, we quadrupled the RAVDESS dataset using additive Gaussian white noise (AWGN) for a total of 5760 audio samples. For the network structure, we build two parallel convolutional neural networks (CNN) to extract spatial features and a transformer encoder network to extract temporal features, classifying emotions from one of 8 classes. Taking advantage of CNN's advantages in spatial feature representation and sequence encoding conversion, I obtained an accuracy of 80.46% on the hold-out test set of the RAVDESS data set.

1. Introduction

With the progress and development of technology, human-computer interaction is widely used in today's society. Speech recognition, as one of the media of human-computer interaction, has gradually become the key to realize natural human-computer interaction [1]. Traditional speech information processing systems, including speech understanding and speech conversation models, focus on the correctness of the extraction of the expressed vocabulary in the speech signal and the readability of the generated text in the speech signal. But the speech signal contains not only the information communicated and the words expressed, but also the emotional state of the vocalist is implied. SER of computers, which reflects the corresponding human emotions by extracting the acoustic features of speech, is the basis for achieving more harmonious and efficient human-computer interaction, and is of great research significance and application value [2].

The traditional SER method mainly includes 3 steps: speech signal pre-processing, speech emotion feature extraction and speech emotion classification recognition. Among them, the extraction of emotion features and the model of emotion recognition are the keys of speech signal processing, which directly affect the accuracy of SER. Nwe et al [3] proposed a method based on Hidden Markov model (HMM) based method for SER, which establishes a Hidden Markov model for each emotion state and

calculates the probability function of the output using a mixed Gaussian distribution, which can maintain the overall non-stationarity and local smoothness of the speech signal more accurately. However, due to the inclusion of multiple emotional states in speech, the method requires multiple HMM, which results in a large training computation, difficult classification of emotional states, and a low overall recognition rate. Hervé et al [4] proposed a method for estimating posterior probabilities using Artificial neural network (ANN) to solve the problem of large computation and insufficient accuracy of posterior probabilities of Hidden Markov models, which improves the accuracy of posterior probabilities by using transformed local conditional posterior probabilities to estimate the overall posterior probabilities. Kipyatkova et al [5] proposed the Long-short-term memory network (LSTM) through experiments is more effective for large scale acoustic modeling, modeling speech sequences in each layer of the network for the long-term dependency characteristics of speech sequences, which can converge more rapidly compared to DNNs, with more comprehensive extraction of contextual information and higher overall recognition performance. Zhang et al [6] proposed an attention-based mechanism with a convolutional neural network (CNN) for speech. The method calculates the relevance weights of each temporal domain and emotional features in speech signals through the attention mechanism. The method uses the attention mechanism to calculate the relevance weights of each time domain in the speech signal and the emotional features, then compares the relevance weights of different time domains in the speech signal, and selects the time domain signal with larger weight for recognition, so that the network can focus more on the emotional salience of the speech and ensure that the key information will not be lost. The recognition accuracy of this method is better than that of the basic convolutional neural network. The disadvantages of this method are that it requires large-scale training to calculate the sentiment features through spectrograms and it is more difficult to select parameters for pre-processing.

To address the problems of low accuracy of the above methods in SER and the lack of information in the extracted speech signal features, this paper proposes a deep learning model that combines temporal and spatial features. For the data processing, we quadrupled the RAVDESS dataset using additive Gaussian white noise (AWGN) [7] for a total of 5760 audio samples. For the network structure, we use Mel spectrogram as speech emotion feature input. We use the image classification and spatial feature representation capabilities of CNN to treat the Mel spectrogram as a grayscale image with the width as the time scale and the height as the frequency scale. the value of each pixel in the Mel spectrogram is the intensity of the audio signal at a specific Mel frequency at a time step. building two parallel convolutional neural networks (CNN) to extract spatial features and a transformer encoder network to extract temporal features, classifying emotions from one of 8 classes. The expansion of CNN filter channel size and the reduction of feature mapping will provide the most expressive feature representation at the lowest computational cost, while the use of transformer encoders assumes that the network will learn to predict the frequency distribution of different emotions based on the global structure of the Mel spectrogram for each emotion. Taking advantage of CNN's advantages in spatial feature representation and sequence encoding conversion, I obtained an accuracy of 80.46% on the hold-out test set of the RAVDESS dataset [8].

2. Methods

2.1. Emotion Database

RAVDESS is a validated multimodal database of emotional language and songs. The database consists of 24 professional actors vocalizing lexically matched utterances in a neutral North American accent with gender balance. Voices include expressions of calm, happiness, sadness, anger, fear, surprise, and disgust, and songs contain emotions of calm, happiness, sadness, anger, and fear. Each expression produced two levels of emotional intensity, plus a neutral expression. All conditions are available in face and voice, face and voice only formats. The set of 7356 recordings was rated 10 times each for emotional validity, intensity and authenticity. Ratings were provided by 247 untrained study

participants from North America. An additional group of 72 participants provided test-retest data. High levels of affective validity and test-retest internal reliability were reported.

2.2. Spatial and Temporal Feature-based SER model

Spatial and Temporal Feature-based SER model proposed in this paper mainly includes: speech data pre-processing(AWGN) and MFCC [9] extraction, CNN emotion feature extraction, attention mechanism weight calculation and emotion classification output(Fig.1).

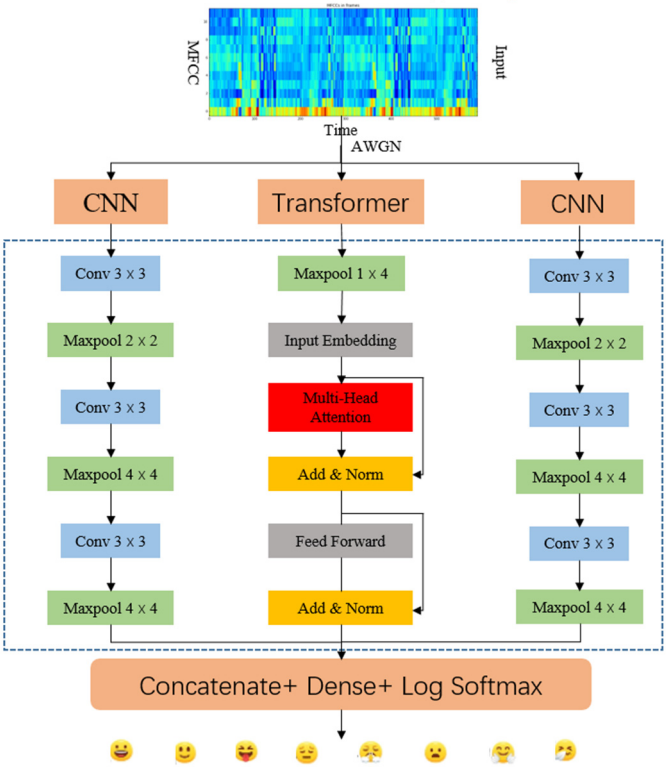

Fig.1 Spatial and Temporal Feature-based SER model

2.3. Data pre-processing

Due to the small sample size of the dataset, overfitting- can easily occur in the highly parameterized neural network model in this paper. Therefore, we design to use the data augmentation method. However, generating more real samples would be incredibly difficult. We need a good balance of noise - too little noise will be useless, and too much noise will make it too difficult for the network to learn from the training data. Additive white Gaussian noise (AWGN) is a channel model in which the only impairment to communication is the linear addition of broadband or white noise with constant spectral density (expressed in watts per hertz bandwidth) and amplitude Gaussian distribution. This model does not take into account fading, frequency selectivity, interference, nonlinearity or dispersion. We will use Additive White Gaussian Noise (AWGN). It is Additive because we are adding it to the source audio signal, it is Gaussian because the noise vector will be sampled from a normal distribution and have a zero-time average (zero mean), and it is white because after the whitening transformation the noise will add power to the audio signal uniformly across the frequency distribution.

2.4. MFCC extraction

Traditional phonological features are divided into acoustic features, rhythmic features and phonological features. Acoustic features are mainly divided into vowels and consonants in speech; rhythmic features are divided into energy, speech rate and resonance peaks; and phonological features represent the clarity of the speech signal [10]. Due to the complexity of speech emotion, many

emotions are difficult to be recognized effectively by rhyme and sound quality alone, resulting in poor differentiation of a single original feature, which requires the confusion of different speech features for speech recognition. Moreover, the speech signal is a non-smooth random process with strong time-variability, so to increase the practicality of feature parameters and reduce the complexity of feature extraction, MFCC is selected as the speech emotion feature in this paper.

MFCC is a speech emotion feature parameter that is an inverse spectral coefficient extracted in the frequency domain of the Mel scale, a feature widely used in automatic speech and speaker recognition. The Mel scale is very accurate in describing the nonlinear characteristics of human ear frequency. The computational relationship between it and the frequency can be expressed in equation (1):

$$\mathrm{M}(f) = 2595 \times \log_{10}(1 + \frac{f}{700}) \tag{1}$$

where: M denotes the Meier frequency function and f denotes the linear frequency. equation (1) represents the relationship between the Mayer frequency and the linear frequency. In the Meier frequency domain, the human perception of audio is linear, and the MFCC are constructed by simulating the human ear characteristics to construct the characteristic parameters through human auditory features.

2.5. Network Structure

We have 5760 MFCC maps(2880 native and 2880 noise-enhanced), each MFCC map has a shape of 40 x 282, where 40 MFC coefficients represent different melodic pitch ranges and each MFC coefficient has 282 time steps. We can imagine the MFCC map as a black and white image with 1 signal strength channel. Thus, our MFCC input feature tensor will be of the shape (5760, 1, 40, 282) before splitting training. The network structure can be known from Fig.1. The expansion of CNN filter channel size and the reduction of feature mapping will provide the most expressive feature representation at the lowest computational cost. we used the Transformer-Encoder layer introduced in [11], hoping that the network would learn to predict the frequency distribution of different emotions based on the global structure of the MFCCs for each emotion.

We train every architecture for up to 200 epochs, optimizing the cross-entropy loss using stochastic gradient descent with a momentum of 0.9. The initial learning rate=0.01, batch size=16, and weight decay=0.0001. All experiments are performed on NVIDIA Tesla V100. The experiments in this paper were done on the Pytorch framework.

3. Test Results and Discussions

In this paper, the obtained models are tested in the test set following the experimental procedure. Table 1 shows the normalized confusion matrix for SER. The recognition accuracy of expression recognition achieved by the proposed method on the RAVDESS dataset is better than existing studies on SER. The average classification accuracy on RAVDESS is listed in Table 2

Table 1 The Normalized confusion matrix for SER

	surprised	neural	calm	happy	sad	angry	fearful	disgust
surprised	0.88	0.02	0	0.04	0.06	0	0	0
neural	0.01	0.51	0.43	0.05	0	0	0	0
calm	0	0	0.96	0	0.04	0	0	0
happy	0.05	0.03	0	0.92	0	0	0	0
sad	0	0.15	0.05	0	0.75	0	0	0.05
angry	0.07	0	0	0.03	0.05	0.85	0	0
fearful	0.07	0	0	0.03	0.1	0	0.77	0.1
disgust	0	0	0	0	0	0.18	0	0.82

Table 2 Classification accuracy on RAVDESS

Method	Accuracy (%)
DCNN[12]	71.61
The proposed algorithm	80.46

Some discriminations of 'sad' for ' neural ' fail. Perhaps surprisingly, the frequency of confusion between 'fear' and 'happiness' is as high as the frequency of confusion between 'sadness' or 'disgust'-perhaps this is because fear is a true multi-faceted emotion. Based on this, we will compare the characteristics of confounding emotions more carefully to see if there are any differences and how to capture them. For example, translating spoken words into text and training the network for multimodal prediction, and combining semantics for sentiment recognition.

4. Conclusion

In recent years, SER technology as one of the key technologies in human-computer interaction systems, has received a lot of attention from researchers at home and abroad for its ability to accurately recognize emotions and thus improve the quality of human-computer interaction. In this paper, we propose a deep learning algorithm with fused features for SER. in terms of data processing, we quadruple-processed the RAVDESS dataset with 5760 audio samples using additive Gaussian white noise (AWGN). For the network structure, we constructed two parallel convolutional neural networks (CNNs) to extract spatial features and a transform encoder network to extract temporal features to classify emotions from one of the eight categories. Taking advantage of CNNs in spatial feature representation and sequence coding transformation, we obtained an accuracy of 80.46% on the holdout test set of the RAVDESS dataset. Based on the analysis of the results, the recognition of emotions by converting speech into text combined with semantics is considered.

References

[1] Wu S, Li F, Zhang P. Weighted Feature Fusion Based Emotional Recognition for Variable-length Speech using DNN[C]//2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, 2019: 674-679. (in USA).

[2] Anagnostopoulos C N, Iliou T, Giannoukos I. Features and classifiers for emotion recognition from speech: a survey from 2000 to 201 1[J]. Artificial Intelligence Review, 2015, 43(2): 155-177. (in NETHERLANDS).

[3] Nwe T L, Foo S W, De Silva L C. Speech emotion recognition using hidden Markov models[J]. Speech communication, 2003, 41(4): 603-623. (in NETHERLANDS).

[4] Bourlard H，Konig Y, Morgan N, et al. A new training algorithm for hybrid HMM/ ANN speech recognition systems[C]/ / 1996 8th European Signal Processing Conference. Trieste, Italy: IEEE, 1996:1-4. (in Italy).

[5] Kipyatkova I. LSTM-based language models for very large vocabulary continuous Russian speech recogni tion [M]/ /Speech and Computer. Cham: Springer International Publishing, 2019: 219-226. (in Germany).

[6] Zhang Y Y，Du J，Wang Z R，et al. Attention based recognition [C]//2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Honolulu, USA: IEEE, 2018: 1771-1775. (in USA).

[7] Bystrom M, Modestino J W. Combined source-channel coding schemes for video transmission over an additive white Gaussian noise channel[J]. IEEE Journal on Selected Areas in Communications, 2000, 18(6): 880-890.(in USA).

[8] Livingstone S R, Russo F A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English[J]. PloS one, 2018, 13(5): e0196391. (in USA).

[9] Likitha M S, Gupta S R R, Hasitha K, et al. Speech based human emotion recognition using MFCC[C]//2017 international conference on wireless communications, signal processing and networking (WiSPNET). IEEE, 2017: 2257-2260. (in USA).

[10] Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction[J]. IEEE Signal processing magazine, 2001, 18(1): 32-80. (in USA).
[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017. (in USA).
[12] Issa D, Demirci M F, Yazici A. Speech emotion recognition with deep convolutional neural networks[J]. Biomedical Signal Processing and Control, 2020, 59: 101894.(in ENGLAND).