Research on speech emotion recognition based on multi-feature fusion

Zhiqiang Huang[1a], Mingchao Liao[1b]

2863511975@qq.com[a], 411781160@qq.com[b]

School of Mathematics and Computer Science, Wuhan Polytechnic University, Hubei, Wuhan 430048

Abstract: In the field of emotion recognition, speech emotion recognition research is very prevalent now, and the information contained in a single feature has a limit. To address the issue of low identification accuracy caused by a single feature, we provide an audio multi-feature fusion approach that fully fuses feature information and extracts information for recognition in several dimensions in this study. Initially, pre-processing and feature extraction are conducted on audio files, and the Mel-spectrogram and multi-F feature sets are extracted. The Mel-spectrogram feature map is then fed into the Convolutional Block Attention Module (CBAM) for higher dimensional feature mapping, and the output results are cascaded with multi-F before being fed into the fully connected layer and SoftMax layer to complete the classification with an accuracy of 81.5% on IEMOCAP datasets.

Key words: Emotion recognition; Multi-feature; CBAM

1 Introduction

The first research on speech emotion recognition was conducted by Williams [1], who experimentally found that when a person is angry, scared, and sad, there are differences between speech signals, mainly in intelligibility, fundamental contour, and power spectrum. Some researchers subsequently discovered more acoustic features related to emotion, such as FO mean, resonance peaks, etc. [2]. In the 1990s, the concept of affective computing was introduced by the MIT Multimedia Laboratory [3]. In 1999, the first speech interface for image capture systems was developed by Moriyama et al. [4] which enabled the first commercial implementation of SER (Speech Expression Recognition). With the development of big data technology, cloud computing and deep learning, more and more researchers have adopted CNN and RNN networks as models for the acquisition and classification of speech emotion information, such as Lee [5], who used RNNs to learn the temporal features of speech signals. Some other researchers directly use one-dimensional convolution to extract features from discrete-time waveforms and use LSTM networks to model the temporal structure of speech to achieve speech emotion prediction [11].

Several emotional features of speech are derived when measuring the concrete emotion in human speech. The speech emotion characteristic parameters provide the foundation for computer machine learning. The correlation degree of various emotions varies greatly for each feature parameter, and the phonetic emotion feature parameters may be classified into prosodic features, sound quality features, and spectral characteristics.

EMIS 2023, February 24-26, Chengdu, People's Republic of China

DOI 10.4108/eai.24-2-2023.2330662

The duration feature and the energy feature are the two basic analytical indices of prosodic characteristics. The duration feature is a characteristic parameter that affects the speed of speech. It consists primarily of the following parameters: speech speed, the number of voice segment frames, the number of silent segment frames, the size of the voice segment area, the size of the silent segment area, and the size of the relative pronunciation area. As people talk, their speaking pace and pause vary according to their emotions. For example, when they are pleased, they speak quicker, and when they are sad, therefore the duration is highly correlated with mood. [10, 12] The volume of a person's speech is a generalization of energy; the higher the energy, the bigger the signal amplitude; the lower the volume, the lower the power, the smaller the signal amplitude. There was also a considerable association between individuals's emotional and energetic features, with furious people speaking with more energy than melancholy people.

Short-time energy or short-time amplitude is a method for converting the energy characteristics of speech signals into standardized expressions, and short-time amplitude prevents the difference between tiny and big sample values from becoming excessively great owing to square in computation. Spectral features are often short-time representations of speech signals, whereas prosodic characteristics are continuous representations. Because the creation of speech signals is linked to various organs in the thoracic cavity and mouth, the short-time representation is utilized. Because the vocal organs rely on muscular control, they do not alter dramatically in a short period of time. As a result, the properties are very steady, and the spectral characteristics do not vary significantly.

Short-time Fourier transform (STFT), Linear Predictor Coefficients (LPC) (Williams CE, 1972), Mel Frequency Cepstrum Coefficients (MFCC)(Schuller B, 2003), Line Spectrum Pair parameters (LSP), Perceptual Linear Predictive Cepstral Coefficients (PLP)[10], Short Time Coherence (SMC), and others are examples of classical spectral features.

Even though there are several speech emotion characteristic characteristics, this research focuses on four that have the most impact: pitch frequency, formant, short-term energy, and MFCC cepstrum coefficient.

2 FEATURE EXTRACTION

2.1 Mel spectrogram

The initial stage in tackling classification issues using machine learning is feature extraction. The signal can be turned into a Spectrogram in the field of audio and speech signal study, where the horizontal coordinate x refers to time and the vertical coordinate y is frequency, and the value of the (x,y) coordinate reflects the magnitude of the y frequency at the x time point. Color distinguishes the magnitude of the amplitude in the spectrogram. The frequencies in the spectrum are linearly distributed, but because human hearing is more sensitive to low frequency sounds and it is difficult to detect changes in high frequencies, a spectrum with a linear distribution in the frequency dimension is insufficient to fit the changes in emotion in speech. The Mel spectrogram is based on a logarithmic modification of the speech spectrogram, which is more congruent with human ear recognition features.

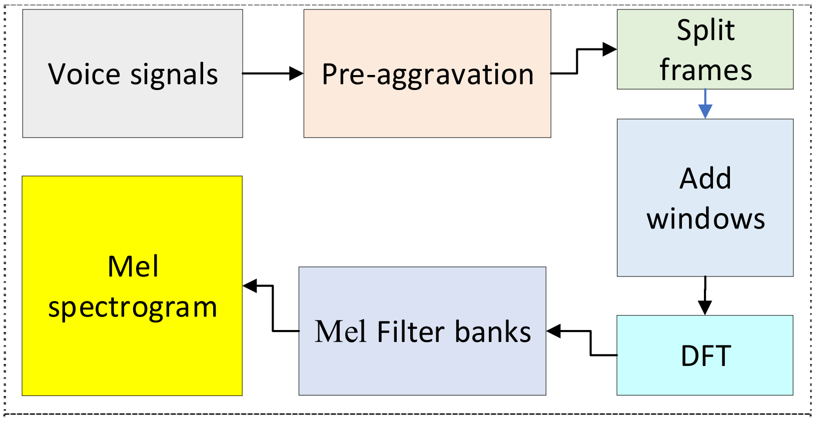

Fig. 1 step diagram of Mel spectrogram extraction

Step1: Pre-emphasis. Pre-emphasis operation can effectively reduce the impact of high-frequency noise on the experimental results, improve audio quality, and keep the spectrum of the signal in a stable frequency band. Make the audio smoother. The output signal is as in Eq. (1).

$$H(Z) = 1 - \alpha z^{-1} \tag{1}$$

$$y(n) = x(n) - \alpha x(n-1) \tag{2}$$

where: $x(n)$ is the original signal, $y(n)$ is the output signal value after processing, and α is taken as 0.97 in the experiments of this paper.

The second step is to frame the audio. The audio data is a continuously fluctuating temporal signal that can be considered steady for a limited length of time. The signal is separated into 20ms frames in this article, which is useful for subsequent processing. There is a slight overlap between frames to decrease the loss of boundary information between frames, and the frame shift is 10ms.

Audio windowing is the third step. The impact of windowing on the signal is that the amplitude of each frame tapers to zero at both ends, making the Fourier transform easier.

$$c(n) = \omega(n) * s(n) \tag{3}$$

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos\left[\dfrac{2\pi n}{(N-1)}\right], 0 \le n \le N-1 \\ 0 \qquad\qquad\qquad\qquad\qquad ,\text{others} \end{cases} \tag{4}$$

Step4: Discrete Fourier transform (DFT). After adding the window, the speech signal needs to be shifted from time domain analysis to frequency domain analysis for energy observation, which is achieved by DFT, and the expression is shown in (5).

$$C(k)=\sum_{n=0}^{N-1}c(n)e^{-j\frac{2\pi}{N}kn},0\le k\le N \tag{5}$$

Step5: Mel filter set. The spectral energy of the speech signal can be obtained after the mode-square operation of the spectrum of the speech signal. Assume that the total number of triangular filters in the system is composed of triangular filters, and each of them is Mel standard, and then make the energy spectrum pass in the middle. The center frequency is denoted by $f(m)$ and the value of m is taken as $[1,\ M]$, and the spectral energy is expressed as $|X(k)|^2$. After passing through the triangular filter, the response expression of the frequency is shown in Eq. (6), and the weighted summation expression is shown in Eq. (7).

$$H_m(k)=\begin{cases}0 & ,k<f(m-1)\\ \dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & ,f(m-1)\le k\le f(m)\\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & ,f(m)\le k\le f(m+1)\\ 0 & ,f(m+1)\le k\end{cases} \tag{6}$$

$$MelSpec(m)=\sum_{k=f(m-1)}^{f(m+1)}\log\left(H_m(k)|X(k)|^2\right) \tag{7}$$

2.2 Multi-F

We fuse the Zero Crossing Rate, amplitude envelope, and root-mean-square energy into multi-F features by a weighting method. In order to efficiently fuse the three features, this paper first conducts classification experiments using neural networks and obtains their classification results. In this paper, we first use the neural network to perform classification experiments and obtain the classification results. The excess-zero rate $Z(z_1,z_2...z_n)$, amplitude envelope $A(a_1,a_2...a_k)$ and root-mean-square energy $R(r_1,r_2...r_t)$ are all one-dimensional feature vectors, and the features can be simply cascaded and directly stitched together using the Conca function, and then form a feature vector of length $n+k+t$, which is finally used for speech emotion recognition. However, different features have large differences in the magnitude of their values due to different measurements and different units, and the amount of information and accuracy of different features themselves also vary. The objective function becomes flat when the neural network is trained with data that are not normalized, and the gradient direction of the fitted function does not go in the direction of the minimum value, but deviates. Therefore, this chapter normalizes the audio feature set before fusion to address both the requirements of neural network training and to weaken the problem of inconsistent metrics between different features. After the normalization is completed, this paper uses the calculated weights to weight the features.

3 Method in speech emotion recognition

To create a worldwide attention effect, we employ a convolutional attention module to discriminate the relevance of different channels and give different weights to different areas of the same channel. SENet is a channel attention technique that arose to overcome the problem of information loss caused by treating each channel in the convolution process as having the same significance. Previous studies assumed that the value of each channel was equal by default; but, in real-world scenarios, the information contained in each channel differs, hence their relevance should be distinguished as well. Similarly, the spatial attention mechanism is intended to enable the model to recognize that the significance of information fluctuates across various sections of the same channel. In this paper, CBAM is applied to attach weights to channels and spaces, and its model structure is shown in Figure 2.

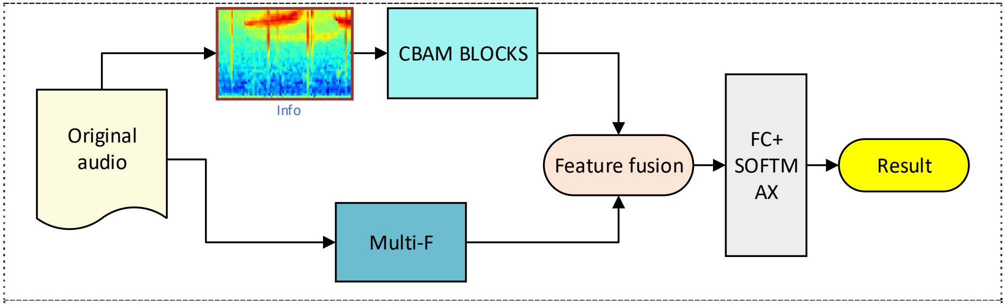

Fig. 2 model structure of the proposed method

3.1 Channel attention

The input channel attention module's data is a feature map of size that is first subjected to global average pooling with a pooling size of and the output size acquired by this operation is. The CBAM employed in this work varies from SENet in that it conducts Global Max Pooling on the input features after extracting the Global Average Pooling. By the same token, maximal pooling yields a feature map of size. $(1 \times 1 \times c)$. The pooling layer itself is designed to extract higher dimensional information, and applying two different poolings also means that the model extracts more feature information. After the two pooling methods, the model outputs two one-dimensional feature vectors, after which the two one-dimensional features are connected to two fully connected layers. The purpose of the fully connected layer is to increase the processing of nonlinear relationships between channels, which can better match the complex and actual inter-channel correlations. The number of neurons in the full-connected layer of dimensionality reduction is generally taken as $c/16$, which is rounded upward in this paper as the first layer number. The second layer needs to be restored to the number of channels, so the number of neurons in the second fully connected layer is equal to the number of channels. The final channel weights are computed by the sigmoid function after summing $F_{avg}(f_{a1}, f_{a2} \cdots f_{ac})$ and $F_{\max}(f_{m1}, f_{m2} \cdots f_{mc})$ the fully connected and the final channel weights, which are calculated as in Equation 3.1

$$W\left[w_1, w_2 \cdots w_c\right] = Sigmoid\left[(f_{a1} + f_{m1}),(f_{a2} + f_{m2})\cdots(f_{ac} + f_{mc})\right] \quad (8)$$

where: $W\left[w_1, w_2 \cdots w_c\right]$ is the output weight vector of each channel.

3.2 Spatial attention

The spatial attention mechanism has an input feature map of size $(h \times w \times c)$, firstly, the global average pooling and global maximum pooling are performed in the same way to obtain two feature maps of size equal to , and then the two feature maps are connected based on the features to obtain the output of size $(h \times w \times 2)$. The number of channels of the feature map is reduced back to 1 by a 7×7 convolution operation. The spatial attention matrix is generated by the Sigmoid function, and the output is obtained by full multiplication of the spatial attention matrix with the input feature map.

4 Experiment and Result analysis

The confusion matrix of the experimental findings (Table 1) shows that the speech emotion detection algorithm described in this chapter, based on Mel-spec and multi-F feature fusion, performs well on the IEMOCAP dataset. The best accurate recognition of the emotion "happy" is 81%. The categorization method is less accurate for "fear" emotion speech, with a true recognition rate of 79%. For the IEMOCAP dataset, the fusion method attained an average accuracy of 81.5%, which is a high accuracy rate.

Tab1 Confusion matrix of speech multi-feature fusion experimental results

	Happy	Sad	Fear	neutral
Happy	0.81	0	0.10	0.09
Sad	0	0.84	0.10	0.06
Fear	0	0.06	0.79	0.15
neutral	0.02	0.10	0.06	0.82

In order to further verify the effectiveness of the Mel-Spec and Multi-F feature fusion sentiment classification algorithms, this paper then designs a set of control experiments to compare the performance of sentiment classification with the single-feature algorithm using the fusion algorithm designed in this chapter, and the comparison results are shown in Table 2.

Tab2 Comparison of experimental results for each audio feature

Emotion features	Happy	Sad	Fear	neutral	Mean accuracy
Mel-Spec	78	84	67	83	78
Spectrogram	75	78	68	78	74.75
ZCR+AE+RMSE	51	56	48	54	52.25
Multi-F	71	64	52	69	64
Mel-Spec、Multi-F	81	84	79	82	81.5

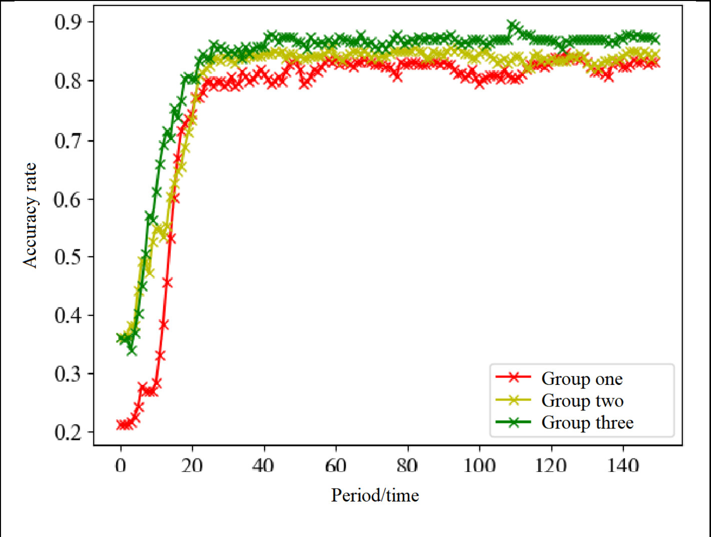

Fig. 3 Model Training Diagram

The test curve was assessed after 150 batches of 64 voice data were trained. This strategy is contrasted with two other approaches. To begin with, as compared to Group three networks, Group two networks perform better on test sets and can achieve higher learning states, and the accuracy of model identification has improved to some extent when compared to Group one networks. Second, the Group three network achieves a more consistent and greater recognition accuracy on the test set than the Group two network. The model structure analysis reveals that the inclusion of the multi-feature features presented in this technique causes the model to pay greater attention to the interaction between the feature graph and its sub-feature graph. Richer and deeper time sequence information may be obtained to sharpen the quiet and noisy point information and emphasize essential time sequence information, making identification and classification easier. As a result, the strategy suggested in this research is also very important for model recognition performance under the same input.

5 Conclusion

In this research, we focus on the speech modality identification method in multi-feature fusion emotion recognition. The book initially offers the feature fusion framework diagram, then covers the process and procedures of obtaining the Mel Spectrogram features, develops a new feature set multi-F features, and provides the unique feature fusion approach. This article discusses the exact experimental contents, data, and analyses the experimental outcomes in the experimental section. Experiments show that the feature fusion technique suggested in this chapter outperforms the classic single-feature classification model of this approach in sentiment classification.

Reference

[1] B S Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification[J]. The Journal of the Acoustical Society of America, 1974, 55(6): 1304-1322.

[2] B. Yang, M.Lugger. Emotion recognitio from speech signals using new harmony features[J]. Signal Processing, 2011, 90(5):1415-1223.

[3] Cahn J E. The generation of affect in synthesized speech[J]. Journal of the American Voice I/O Society, 1990, 8(1); 1-1.

[4] Han Fang, Zheng Jingjing. An improved formant detection algorithm based on LPC [J]. Electronic Design Engineering, 2017, 25(17): 85-89.

[5] Kwon O W, Chan K Hao J, Lee T W. "Emotion Recognition by Speech Signals", [C]. Proceedings of EUROSPEECH-2003, pp.125-128.

[6] Lee J, Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition[C] Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[7] Moriyama T, Ozawa S. Emotion recognition and synthesis system on speech[C] Proceedings IEEE International Conference on Multimedia Computing and Systems. IEEE,1999, l: 840-844.

[8] R.Le Bouquin. Enhancement of noisy speech signals: application to mobile radio communications[J]. Speech Communication, 1996, 18(1):3-19.

[9] Schuller B, Rigoll G, Lang M. Hidden Markov model-based speech emotion recognition", [C]. Proceedings of the ICASSP 2003, IEEE, pp.1-4.

[10] Siqing Wu, Tiago H. Falk, Wai-Yip Chan. Automatic speech emotion recognition using modulation spectral features[J]. Speech Communication, 2011, 24(7):768-785.

[11] Trigeorgis G, Ringeval F, Brueckner R, et al.. Adieu features End-to-end speech emotion recognition using a deep convolutional recurrent network[C]2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016:5200-5204.

[12] Van Bezooijen R, Otto SA, Heenan TA. Recognition of vocal expressions of emotion: Athree nation study to identify universal characteristics[J], Journal of Cross- Cultural Psychology, 1983,14(4): 387-406.

[13] Williams CE, Stevens KN. Emotions and Speech: Some Acoustical Correlates[J]. Journal of the Acoustical Society of America, 1972, 52(4); 1238-1250.

[14] Yun Jin, Peng Song, Wenming Zheng, Li Zhao. Line spectrum pair (LSP) and speech data compression. International Conference on Acoustics, Speech, and Signal Processing[C]. USA: IEEE, 1984:37-40.