Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 9, No. 2, 2024

Research on Speech Emotion Recognition Analysis Based on Deep Learning

Chunhao Li*, Ailiang Yin[2]

School of Electrical Engineering, Southwest Minzu University, Chengdu, China
* Corresponding author

Abstract: Speech is a common way of communication in human daily life, which is very natural and easy to understand. In the process of communication, one will also express their emotions to the other party. Through emotional expression, the distance between people is smaller and communication is more convenient. Speech emotion recognition is also the key to human-computer emotional interaction, and effective recognition of speech emotions can improve speech intelligibility. The number of speech datasets is usually large, with several G or even larger. This article proposes a method of preprocessing the speech dataset to extract feature data folders, and then using the feature data folder directory as input to train the model. This method helps to improve the efficiency, repeatability, and manageability of data processing.

Keywords: Human-computer emotional interaction, Preprocessing, Speech emotion recognition.

1. Introduction

Speech emotion recognition has received widespread attention in recent years [1], especially in the field of artificial intelligence. For example, in the field of education, speech emotion recognition can be used to monitor the emotional state of students to improve the teaching experience. Educational software can automatically adjust content and teaching methods based on students' emotional needs; In the automotive field, speech emotion recognition can be used to detect the emotional state of drivers, in order to improve vehicle safety and driver attention; In the field of security, voice emotion recognition can be used to detect potential threats, such as terrorist attacks or criminal activities. Security systems can detect abnormal situations by analyzing emotions in sound signals.

The most important step in speech emotion recognition is the extraction of emotional features. Traditional speech emotion features include spectral correlation features, ultrasound quality features, and speech quality features [2], while we usually use spectral features. Spectral features can capture the energy distribution of sound signals at different frequencies, which is crucial for expressing emotional content in speech, as speech with different emotional states usually has significant differences in spectral characteristics. Spectral features also provide resolution in time and frequency, which is very helpful for capturing instantaneous emotional changes in speech signals and the details of spectral characteristics. The spectral features are more stable compared to the original waveform signal. Moreover, the dimensions of spectral features are usually two-dimensional, making it more convenient to train in deep learning networks.

Deep learning networks [3] have shown outstanding performance in expressing features in recent years, and the latest research has provided a better deep model for describing emotional states in speech for emotion recognition. At present, a large number of neural networks and deep learning methods are applied in speech emotion recognition [4]. By designing and developing corresponding network models, the research level of speech emotion recognition has been improved, and some related networks have been successfully applied to feature learning in speech emotion recognition [5-7]. However, many previous studies on emotion recognition have focused on training individual speech signals into the model, which is time-consuming and places significant pressure on computers due to the large size of the speech dataset.

This article proposes a training method for speech emotion recognition, which first extracts the features of speech and saves them in a .NPY format folder, and then trains them in the model to verify its feasibility The .NPY file is saved in binary format, which makes data reading and writing more efficient. Compared to the .WAV file format, it occupies less storage space. Due to its binary format,. NPY files have a faster read and write speed than text file formats (such as .TXT), especially for large datasets NPY files can also easily save multidimensional arrays, such as images, audio, text, model parameters, etc. So we can first extract the audio features from the speech database, which is the Mel spectrum, and save it as a .NPY format speech database. In order to verify the effectiveness of this idea, this study refers to general practices and uses the publicly available CASIA dataset for validation, and then trains it in a speech recognition model.

2. Research on Speech Emotion Recognition

The research on speech emotion recognition needs to be driven by data, and the quality of the database will affect the final recognition results of the model. Therefore, it is necessary to develop a high-performance emotion recognition system based on a high-quality emotional speech database. At present, there are various types of emotional speech datasets available, as shown in Table 1. From Table 1, we can see that there are multiple languages such as English, German, Chinese, etc. Due to cultural differences between languages, there are biases in the process of creating speech emotion recognition databases, resulting in different performance of recognition models on different speech emotion databases. Therefore, this article chooses the CASIA Chinese Sentiment Database to complete the research of this project.

Table 1. Common Emotional Speech Database

Database	Languages	Types	Scale	Number of participants
EMO-DB	German	perform	800 sentences, 7 emotions	5 males and 5 females
SAVEE	English	perform	480 sentences, 7 emotions	4 males
RAVDESS	English	perform	7356 samples, 8 emotions	12 males and 12 females
EMOVO	Italian	perform	588 sentences, 7 emotions	3 males and 3 females
CASIA	Chinese	perform	9600 sentences, 6 emotions	2 males and 2 females
TESS	English	perform	2800 samples, 7 emotions	2 women
FAU AIBO	German	nature	48401 words, 9.2h, 11 emotions	21 males and 30 females

2.1. CASIA Chinese Speech Emotion Database

CASIA Chinese database was recorded by the Institute of Automation, Chinese Academy of Sciences. It includes six emotions: anger, happiness, fear, sad, surprise, and neutral, and is currently the most commonly used Chinese phonetic emotion dataset. This database is composed of four professional speakers who record emotional speech in a quiet and noise free recording studio. The four speakers read the same text information with different emotions, resulting in 1200 emotional speech data[8]. By extracting the phonetic and prosodic features from the speech data, the emotional state of the speech data can be analyzed. Multiple testers were also selected to manually distinguish the recorded speech, achieving high recognition and demonstrating the feasibility of the database for scientific research.

2.2. Extraction of speech emotion data features

In the process of collecting voice emotion data, some factors may lead to incomplete data or excessive noise in the data, which is not conducive to the recognition of voice emotions. So the first step in obtaining emotional data from speech is to preprocess the speech, which mainly includes: pre-emphasis[9], framing, windowing, endpoint detection, noise reduction, feature selection and dimensionality reduction, normalization, etc.

Mel spectrogram is a commonly used speech feature extraction technique, which is a non-linear frequency scale that simulates the human ear's perception of sound frequency. The human ear is more sensitive to low-frequency sound and less sensitive to high-frequency sound, so using the Mel frequency scale can better reflect the characteristics of the human auditory system. This non-linear transformation can improve the robustness of speech emotion recognition systems. The application of Mel spectrum also helps to overcome the challenge of speech data not having a two-dimensional structure of image data. Mel spectra are usually presented in graphical form, where the horizontal axis represents time, the vertical axis represents frequency, and color or brightness represents energy at the corresponding frequency. This graph can provide intuitive information about the distribution of audio signals at different frequencies, providing a more informative spectral representation for subsequent audio analysis and processing, as shown in Figure 2-1.

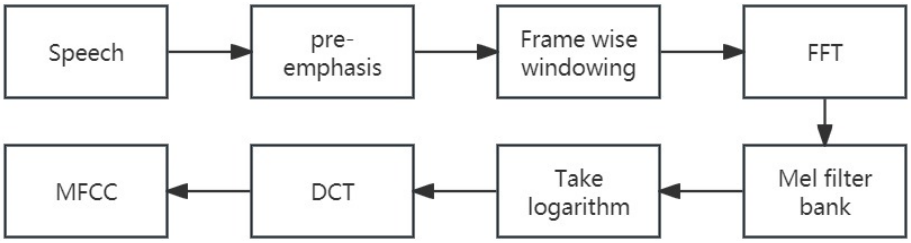


Figure 2-1. MFCC extraction process diagram

2.2.1. Pre-emphasis

For speech signals, in practice, most of the energy of the speech signal is distributed in the low-frequency part. As the power spectrum decreases with increasing frequency, the attenuation of the high-frequency component of the speech leads to a low signal-to-noise ratio. In order to improve the signal-to-noise ratio of the high-frequency component, it is often necessary to strengthen the high-frequency component of the speech. The purpose of pre emphasis is to enhance the high-frequency part, make the signal spectrum flat, and also to eliminate the effects of vocal cords and lips during the vocal process, thereby providing better input for subsequent speech recognition tasks. Pre emphasis processing is actually passing the speech signal through a high pass filter:

$$H(z) = 1 - \mu z^{-1} \tag{2-1}$$

The range of the pre emphasis coefficient is 0.9~1, and we usually take it as 0.97.

2.2.2. Frame wise windowing

In general, we consider data from 20 to 40ms as one frame. If the length of the frame is too short, we do not have enough samples to obtain reliable spectrum estimation; If the length of the frame is too long, the signal will vary more throughout the entire frame. And our commonly used frame length is 20-25ms, with a frame shift of 10ms.

After framing processing, it is generally necessary to perform windowing on the speech signal, which is a classic method in signal processing. Because during the framing process, the data is disconnected, but our signal processing methods require the signal to be continuous. Therefore, we need a window function to weight each frame and multiply it with the input speech signal. The window function tends to zero at the boundary, and the multiplied result also tends to zero at the boundary, thus avoiding discontinuity. The commonly used window functions include rectangular windows, triangular windows, Hanning windows, and Hamming windows [10].

$$S_\omega(n) = \omega(n)*s(n) \quad (2\text{-}2)$$

Among them, is the window function, and s (n) is the speech signal convolved to obtain the windowed speech signal.

Rectangular window

$$\omega(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & \text{other} \end{cases} \quad (2\text{-}3)$$

Hamming

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos[2\pi n(N-1)] & 0 \le n \le N-1 \\ 0 & \text{other} \end{cases} \quad (2\text{-}4)$$

Through analysis, it can be concluded that rectangular windows have good spectral smoothing characteristics, but they may lose high-frequency components, leading to energy leakage. The effect of the Hamming window and the Hanning window is similar. The Hamming window contains more high-frequency speech information, which can overcome the problem of energy leakage. Therefore, this experiment uses a Hamming window to add a window to the speech signal.

2.2.3. Fourier transform

After completing the framing and windowing, the next step is to perform Fourier transform on each frame of data. Fourier transform is a mathematical operation that converts a signal from time domain to frequency domain. It decomposes the signal into sine and cosine components of different frequencies, allowing us to analyze the spectral characteristics of the signal.

The data we use are all discrete numerical signals, so the Fourier transform performed on discrete numerical signals is also known as Discrete Fourier Transform (DFT). Generally, more efficient and faster discrete Fourier transforms are used in computers, which we refer to as Fast Fourier Transform (FFT), thereby converting signals from one-dimensional time domain to two-dimensional frequency domain.

2.2.4. Mel filter bank

Mel Filter Bank is a set of filter banks designed to simulate the characteristics of human hearing. Research has shown that human perception of frequency is not linear and is more sensitive to low-frequency signals than high-frequency signals. For frequencies below 1kHz, there is a linear relationship with frequency, and for frequencies above 1kHz, there is a logarithmic relationship with frequency. The higher the frequency, the poorer the perception ability. For example, people can easily distinguish between 500 and 1000Hz, but it is difficult to distinguish between 7500 and 8000Hz. Therefore, the Mel filter simulates the non-linear perception ability of humans towards sound frequency and further extracts features from the signal. The conversion relationship between Mel (m, unit: Mels) and frequency (f, unit: HZ) is as follows:

$$m = 2595\log_{10}\left(1 + \frac{f}{700}\right) \quad (2\text{-}5)$$

$$f = 700\left(10^{\frac{m}{2595}} - 1\right) \quad (2\text{-}6)$$

We usually use triangular filters to design Mel filter banks, but other filters can also be used. Triangle is the most commonly used filter in Mel design.

2.2.5. MFCC

Then, the previously obtained data is filtered using a Mel filter bank to obtain the Mel Spectra. After obtaining the Mel Spectra, the energy of each Mel frequency band is logarithmically taken to obtain the MFCC, as shown in Figure 2-2.

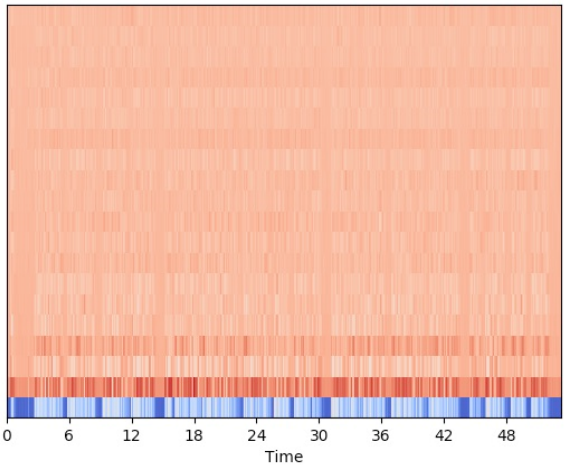


Figure 2-2. MFCC

3. Research on Speech Emotion Recognition

There are generally two types of speech emotion recognition algorithms. One is based on traditional machine learning, and commonly used models include Gaussian mixture models[11], hidden Markov models[12], support vector machines[13], multi-layer perceptrons[14], and recurrent neural networks[15]. Another type based on deep learning[16], with the development of deep learning this year, deep learning models have been widely used in the field of speech emotion recognition with high recognition accuracy advantages. Common examples include Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). This article will use a special network model in Recurrent Neural Network (RNN) called Long Short Term Memory (LSTM). Because RNN network models may encounter certain gradient explosion and vanishing problems during training, and LSTM can precisely avoid these problems.

3.1. Long short-term memory network

LSTM is a special type of recurrent neural network (RNN)[17]. Due to the continuous and temporal nature of speech, RNN is widely used in natural language processing (NLP). However, there are problems with gradient explosion[18] and gradient loss during RNN training. Therefore, in order to solve these problems, RNN has been improved to obtain LSTM.

LSTM has several key advantages over traditional recurrent neural networks:

1. Dealing with long-term dependencies: LSTM is designed to address the issue of long-term dependencies in traditional RNNs. In long sequences, RNNs may face the problem of vanishing or exploding gradients, resulting in the inability to effectively capture long-range dependencies. LSTM introduces a gating mechanism to better control the flow of information, thus being able to handle long-term dependencies.

2. Gating mechanism: LSTM includes forget gates, input gates, and output gates, which allow the model to decide when to forget, store, and output information. This enables LSTM to selectively retain and update information, thereby better adapting to different sequence data.

3. Prevent gradient vanishing: Through gating mechanisms, LSTM can effectively prevent gradient vanishing problems, making it more stable when training deep networks. This is particularly useful when dealing with very deep neural networks.

4. Excellent performance: LSTM performs well in many sequence modeling tasks, including natural language processing, speech recognition, time series prediction, etc. They are usually able to capture more complex relationships in these tasks, thereby improving model performance.

LSTM was proposed by Hochreiter&Schmidhuber (1997) and recently improved and promoted by Alex Graves. LSTM has achieved significant success in many issues and has been widely used in the field of speech emotion recognition[19-20]. LSTM has three gates in total, namely forget gate, input gate (some also called update gate), and output gate. Control which information should be forgotten or discarded, which information should be retained, or kept unchanged through these three gates. Just like a pipeline valve, it controls the flow but not the flow.

(1) Forgetting Gate

The target of the forgetting gate is the cellular state, which controls the selective forgetting of information in the cellular state, determining which needs to be discarded and which needs to be retained. The formula for the forget gate is as follows: represents bias and weight respectively; σ Is the sigmoid activation function;

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3-1}$$

(2) Input Gate

The input gate is mainly used to determine which information is added to the Cell state at the current time, consisting of tanh activation function and sigmoid activation function. In the formula (3-2), it is mainly determined by the output of the hidden state at the previous time and the input at the current time. The value range of the element is between -1 and+1, representing which information is updated to the Cell at the current time; In the formula, (3-3) represents the current moment of information, used to update the cell state. The output formulas of the tanh activation function and the sigmoid activation function are shown in equations (3-2) and (3-3)

$$i_t = \sigma[W_i(h_{t-1}, x_t + b_i)] \tag{3-2}$$

$$C_t^{\sim} = \tanh[W_c(h_{t-1}), x_t + b_c] \tag{3-3}$$

In the formula σ—— Activation function sigmoid.

(3) Output gate

The main purpose of the output gate is to update the cell state and determine the final output value of the gate through the cell state. Firstly, the sigmoid layer determines which part of the cell state to output, and then multiplies this state with the output state of the previous layer through the tanh layer to obtain a value from -1 to+1, ultimately determining the output. The calculation formula is shown in (3-4), (3-5), and (3-6):

$$C_t = f_t * C_{t-1} + i_t * C_t^{\sim} \tag{3-4}$$

$$o_t = (W_o[h_{t-1}, x_t] + b_o) \tag{3-5}$$

$$h_t = o_t * tanh(C_t) \tag{3-6}$$

In the formula——current layer output;

——Hidden layer state.

4. Pre-processing Design Section

Firstly, download 1200 publicly available CASIA voice emotion datasets online. As shown in Figure 4-1, the size of this dataset is 69.8MB. This database is composed of four professional vocalists recording emotional speech in a quiet and noise free recording studio, as shown in Figure 4-2.

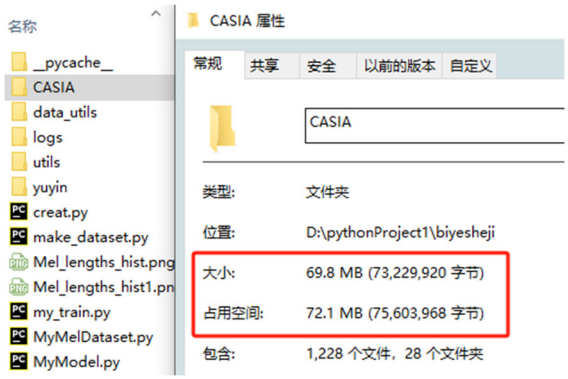


Figure 4-1

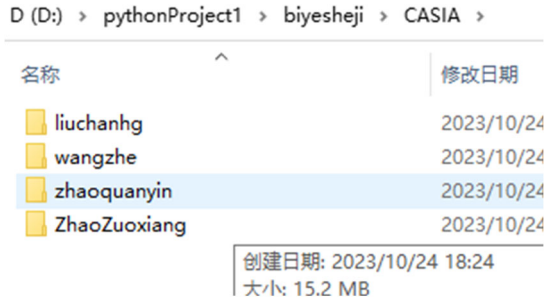


Figure 4-2

Four speakers read the same text information with different emotions and obtained six types of emotions: anxiety, fear, happiness, neutral, sad, and surprise. As shown in Figure 4-3.

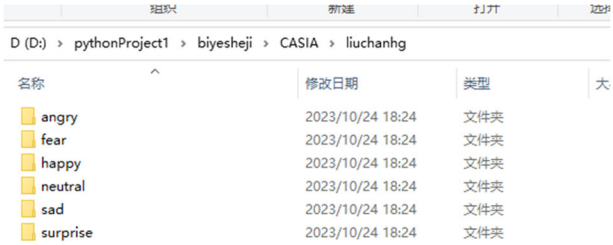


Figure 4-3

Opening the first angle folder as shown in Figure 4-4, we can find that it contains audio files stored in .WAV format.

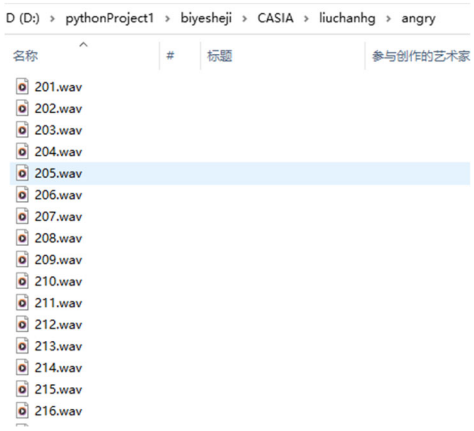


Figure 4-4

Then, we can extract Mel spectral features from each voice in it through the steps shown in 2-1 above and save them in a new folder with the same directory as it. We will name it yuyin, as shown in Figures 4-5 and 4-6.

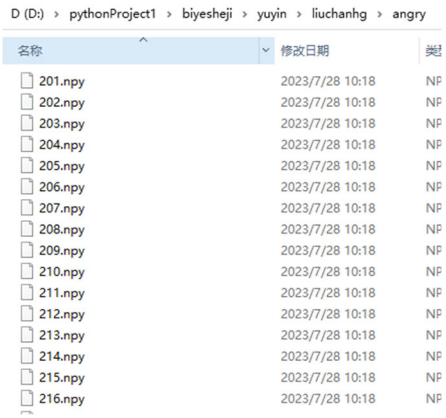


Figure 4-5

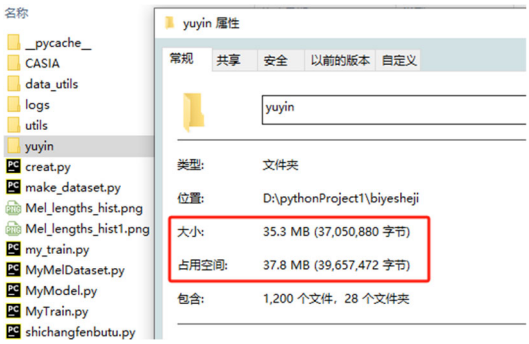


Figure 4-6

From Figures 3-6, we can see that the size of the feature folder for this speech dataset is 35.3MB, which is half the size of the original dataset. Speech datasets are usually quite large, with several gigabytes, or even dozens of gigabytes. This method can shorten memory by half and reduce computational pressure on computers. It also lays the foundation for future training. Although there were errors in the subsequent steps, there is no need to re extract features from the speech dataset.

The .NPY format file in Figures 4-5 can be debugged using PyCharm to see the matrix data saved in .NPY, as shown in Figures 4-7.

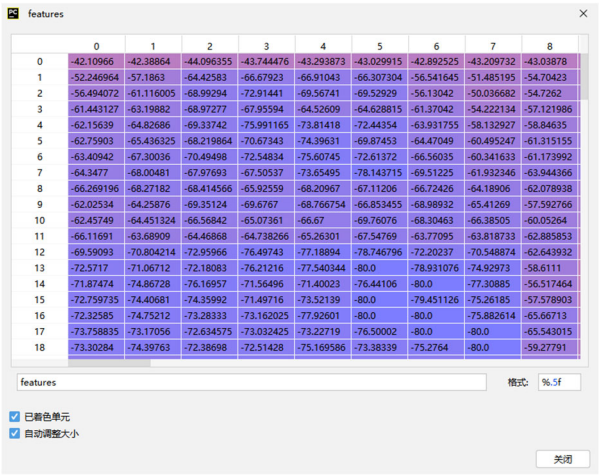


Figure 4-7

The duration of each speech in the speech database varies from 1 second to 4 seconds. However, in AI classification tasks, Pytorch's neural network model usually accepts a tensor of [B, D, T], which makes it inconvenient for batch processing. Therefore, after obtaining the .NPY format dataset, we also need to analyze the duration of these 1200 speech features, as shown in Figures 4-8.

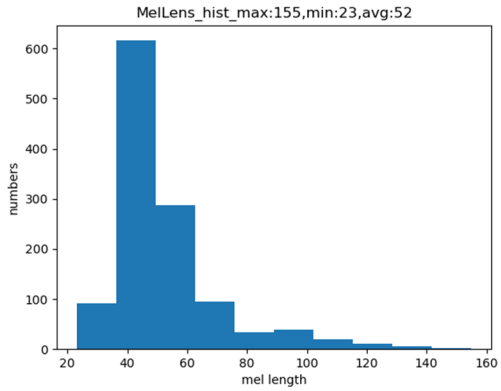


Figure 4-8

In Figures 4-8, the horizontal axis Mel length represents the length of a Mel spectrum, while the vertical axis numbers represent the number of speech at this length. From Figure 4-8, it can be seen that the longest frame length of Mel spectrum is 155 frames, and the shortest frame number is 23 frames, so we need to fill in the short frame numbers. This experiment uses the np. pad function to fill the array. Add additional zeros around the array to adjust its shape or size. Fill the frame rate of each speech into 155 frames, and finally obtain tensor data in the form of (4,128,155).

Then pack and save all the filled datasets, and in this experiment, save them as X_ All. npy, as shown in Figures 4-9.

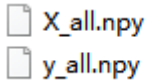


Figure 4-9

Among them, y_ All saves classified labels.

This experiment is based on the Ryzen 7 5800H processor and RTX-3060 graphics card hardware. The saved data features are (128155), and padding zeros are used for frames

that do not reach 155. The LSTM model is used, and two layers of LSTM are set in this experiment. The hidden layer in each layer is 256 cell units, and the output is 6 types of emotions. Each experiment uses three-quarters of the dataset as the training experiment, and selects one quarter as the testing experiment. The network training parameters are: a batch size of 4 for a single batch iteration, and a training batch of 1000 for an Epoch; Adam optimization function, learning rate parameter Lr is 0.0001, Beta_ 1 is 0.9, Beta_ 2 is 0.999, Dropout is set to 0.5.

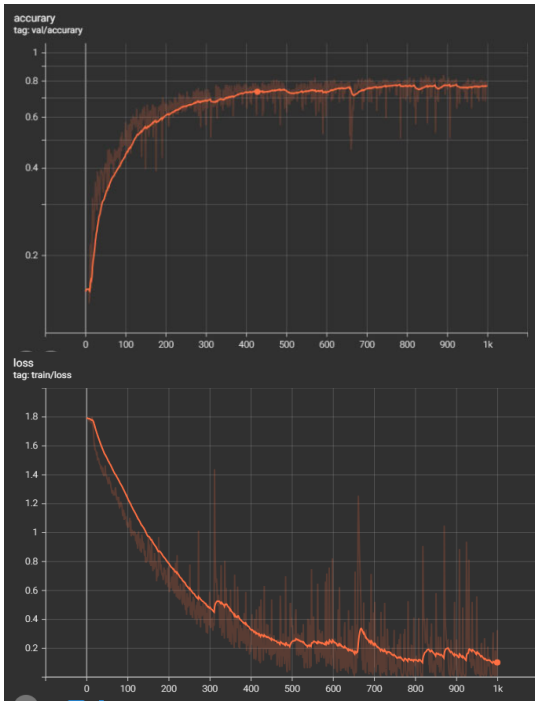


Figure 4-10

The network has undergone 1000 iterations. From the training results in Figures 4-10, it can be seen that the training loss approaches around 0.2 and the accuracy approaches 80%. The effect can meet the design and implementation requirements of the network, indicating that this method is feasible.

5. Summary

This article proposes a training method that first converts audio data in .WAV format from a speech database into a .NPY format data file, and extracts the features of the .WAV audio data to save in .NPY format. Improved storage efficiency means that audio data can be stored more compactly, saving disk space. In addition,.NPY files are also loaded into memory more quickly, which is very useful for large-scale datasets and real-time processing. Then train the file through the LSTM network model. The experiment proved the feasibility of this method.

References

[1] Zhang Shiqing, Zhao Zhijin, Dai Yuliang, etc Research on the Application of Support Vector Machine in Speech Emotion Recognition [J] Acoustic Technology, 2008, 27 (1): 87-90 ZHANG Shiqing, ZHAO Zhijin, DAI Yuliang, et al. A study of support vector machine for speech emotion recognition [J] Technical Acoustics, 2008, 27 (1): 87-90.

[2] Tao Huawei Research on Several Issues of Speech Emotion Recognition Based on Spectral Features [D] Nanjing: Southeast University, 2017.

[3] HINTON G E. Reducing the dimensionality of data with neural networks [J] Science, 2006, 313 (5786): 504-507.

[4] Zhou Yongdong, Zhang Yun, Cao Yanrong, etc Emotional Analysis Based on Feature Fusion Segmented Convolutional Neural Networks [J] Computer Engineering and Design, 2019, 40 (10): 3009-3013, 3029 Zhou Yongdong, ZHANG Yun, CAO Yanrong, et al. Sentiment analysis based on pipeline convolutional neural network combined with features [J] Computer Engineering and Design, 2019, 40 (10): 3009-3013, 3029.

[5] MAO Q R, DONG M, HUANG Z W, et al. Learning salient features for speech emotion recognition using convolutional neural networks [J] IEEE Transactions on Multimedia, 2014, 16 (8): 2203-2213.

[6] ZHANG S Q, ZHANG S L, HUANG T J, et al. Speech emotion recognition using deep convolutional neural network and discriminative temporary pyramid matching [J] IEEE Transactions on Multimedia, 2018, 20 (6): 1576-1590.

[7] CHEN M Y, HE X J, YANG J, et al. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition [J] IEEE Signal Processing Letters, 2018, 25 (10): 1440-1444.

[8] Burkhardt F, Paeschke A, Rolves M, et al. A database of German emotional speech [C] INTERSPEECH, Lisbon, 2005: 1517-1520.

[9] Chen Weiliang, Sun Xiao Speech emotion recognition based on MFCCG-PCA [J]. Journal of Peking University (Natural Science Edition), 2015, 51 (2): 269-274. CHEN Weiliang, SUN Xiao Mandarin speech emotion recognition based on MFCCG-PCA [J] Acta Scientiarum Naturalium University Pekinensis, 2015, 51 (2): 269-274.

[10] Hann (Hanning) window - MATLAB Hann Www.mathworks.com. Retrieved 2020-02-12.

[11] SIDDIQI M H. An improved Gaussian mixture hidden conditional random fields model for audio based emotions classification [J] Egyptian Informatics Journal, 2020,22 (1): 45-51.

[12] GERAVANCIZADEH M, FOROUHANDEH E, BASHIRPOUR M. Feature compensation based on the normalization of vocal tract length for the improvement of emotion - affected speech recognition [J] EURASIP Journal on Audio Speech and Music Processing, 2021 (1): 2-19.

[13] KE X, ZHU Y, WEN L, et al. Speech emotion recognition based on SVM and ANN [J] International Journal of Machine Learning and Computing, 2018,8 (3): 198-202.

[14] Pan Guozhuang. Research on rapid monitoring of controller fatigue status based on real-time land air communication emotion recognition [D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2020.

[15] PAN Guozhuang Research on rapid monitoring of controller fatigue based on real time land air communication emotion recognition [D] Nanjing: Nanjing University of Aeronautics and Astronautics, 2020.

[16] Han Wenjing, Li Haifeng, Ruan Huabin, et al. Review of research progress on speech emotion recognition [J]. Journal of Software, 2014,25 (1): 37-50. HAN Wenjing, LI Haifeng, RUAN Huabin, et al. Review on speech emotion recognition [J] Journal of Software, 2014,25 (1): 37-50.

[17] Rumelhart D E, Hinton G E, Williams R J. Learning presentations by back propagating errors [J] Nature, 1986, 323 (6088): 533-536.

[18] Hochreiter S, Schmidhuber J. Long short term memory [J]. European Calculation, 1997, 9 (8): 1735-1780.

[19] DANPOL R, ALASDOON A, PRASAD P, et al. Speech emotion recognition using convolutional neural network and long short term memory [J] Multimedia Tools and Applications, 2020,79 (43/44): 32917-32934.

[20] Hu Tingting, Feng Yaqin, Shen Lingjie, et al. LSTM speech emotion based feature selection based on attention mechanism [J]. Acoustic Technology, 2019,38 (4): 414-421. HU Tingting, FENG Yaqin, SHEN Lingjie, et al. Main feature selection of LSTM speech emotion based on attention mechanism [J] Technical Acoustics, 2019,38 (4): 414-421.