Audio Emotion Detection Based on CNN & CRNN

ID:489716 Gorden Li

**Abstract:** Emotion is exposed more or less in people’s daily life. The most common format of information people output is audio information. It makes emotion of what they speak the most common source we can get to know what’s going on. Audio emotion detection is important since there is a great bunch of hidden information we can extract from audio emotion. In this research, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) data set is used. It is an audio data set composed of 24 actors simulating 8 emotions defined in psychology which are neutral, calm, happy, sad, angry, fearful, disgust and surprised. CRNN is applied in this study. Different from traditional audio context analysis which is based on extracting feature from time domain and frequency domain, CRNN uses CNN to directly extract feature from a spectrogram and then use LSTM to process the information on a sequence basis. With CRNN, audio information is processed from a computer vision’s point of view. By transferring audio information to a computer vision problem, we can detect the emotion without any manual labelling. It can reduce the cost and increase the accuracy of detection. As a result, CNN model gain a 70+% of accuracy on validation set. CRNN is still needed to be improved with accuracy rate at 35%.

**Keywords:** Audio Emotion Detection, 8 Emotions Defined In Psychology, CNN, CRNN

Presentation Link:

https://wustl.zoom.us/rec/share/7Cs60S-2JUvGqO8GhYJ6Ses7-4Wu-kyBT\_-1SVx24VEtwpqxiB1MOy2xx-gF5WLJ.HRMvw67sllC0i6Nr?startTime=1620447841000

1. **Introduction**
   1. **Motivation of the study**

Audio is the most common type of emotion in daily life. Compared to vision, audio is processed more to get information. Around 90% of information we get is from audio. In business, audio information is also a very essential source of information but sometimes not treated seriously. For example, when an insurance buyer is going to purchase a high-compensation-rate insurance, the company might want to identify if he is hiding some essential information. Audio emotion capture will help during an interview to find out if there is unusual vibration of emotion indicating he or she might be lying. The extraction of these hidden information is valuable to generate a comprehensive business insight.

* 1. **Goal of the study**

In the study, final purpose is to generate a model can be used to identify audio emotion. The audio data set we choose is RAVDESS. CRNN is used to train the model. By directly analyzing from a view point of computer vision, the model can help release traditional audio emotion analysis from costly labeling and data transferring.

1. **Literature Review**

Audio has already been studied by a lot of researchers. Traditional researches use simple model to deal with transferred audio information like GMM, SVM, CRF. In these methods, researchers gather time-domain and frequency-domain information from spectrogram. Short-time energy (STE) and zero-cross ratio (ZCR) are 2 very common features[1]. Nowadays, computer vision is introduced in audio analysis more often. CRNN (CNN + LSTM) is applied in field of music style detection. This method is successful in distinguishing music style like Disco, Blues, Classic, Hiphop etc[2]. Ling Yuhui (2018), use CNN to extract feature of Mel-spectrogram and use RNN to classify the class of different music[3]. These are the studies on classification of music, they all share the same idea that first use CNN to gather information from the spectrogram on a vision basis instead of extracting information from digital feature of a certain audio. Then RNN, LSTM and Bidirectional-LSTM are used to finally classify the style of music. The input of RNN, LSTM and Bidirectional-LSTM is the output of CNN, which is the refined feature of the audio. Later, there are researches started studies covering the emotion detection of music. Chen Changfeng (2019) use CNN-LSTM model to identify style of music as well as the emotion feature of a certain music[4]. So far, there are a bunch of studies regarding to music classification since it is a popular field which can be easily applied in music recommendation system of some music Apps. Not many researches were focus on emotion detection of human audio. On the other hand, human emotion can be potentially applied to business scenario since business’ nature is sell stuff to human. How they react can transfer inner information of his mind, which could be high value in daily business.

1. **Research Design**

**3.1 Problem Description**

The purpose of the study is to identify the emotion based on the audio. In the study, audio context is transferred to spectrogram. Thus, the actual problem this research going to solve is to classify the spectrogram to specific emotion. Emotion here are defined into 8 class from human psychology’s perspective, including neutral, clam, happy, sad, angry, fearful, disgust, surprised.

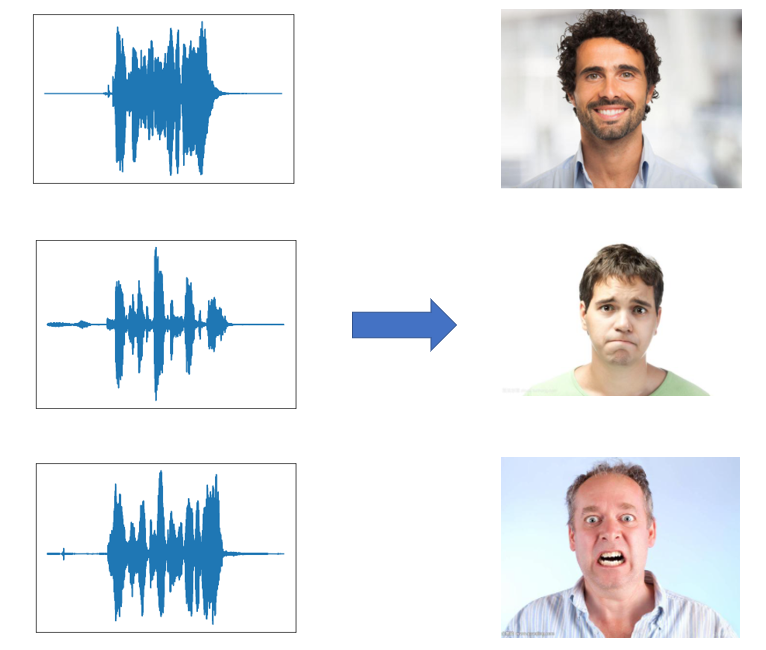


Figure 1 Emotion Detection Sample

In Figure 1, there are 3 examples. On the left-hand side is the spectrogram. After the detection, the emotion of the spectrogram is labeled, which is on the right-hand side.

**3.2 Methodology**

Compared to traditional method used on audio analysis, in this research, the audio emotion detection process is designed on a CRNN basis. CRNN includes CNN and LSTM. Figure 2 describes the process of the analysis. First a wav file will be transferred to a spectrogram. Then by several deep convolution layers, the feature of spectrogram will be extracted as a 1\*Width\*N tensorflow (1 is Height, N is batch size of tensorflow) by using 1 dimension convolution. It is because the figure is spectrogram which is a sequence on a time scale. Convolution on time and frequency have different meaning. Thus, in this research, 1 dimension convolution is applied on frequency. After that, the feature tensorflow will be transferred to sequence and put into LSTM model. Finally LSTM will have a result on classification of emotion.

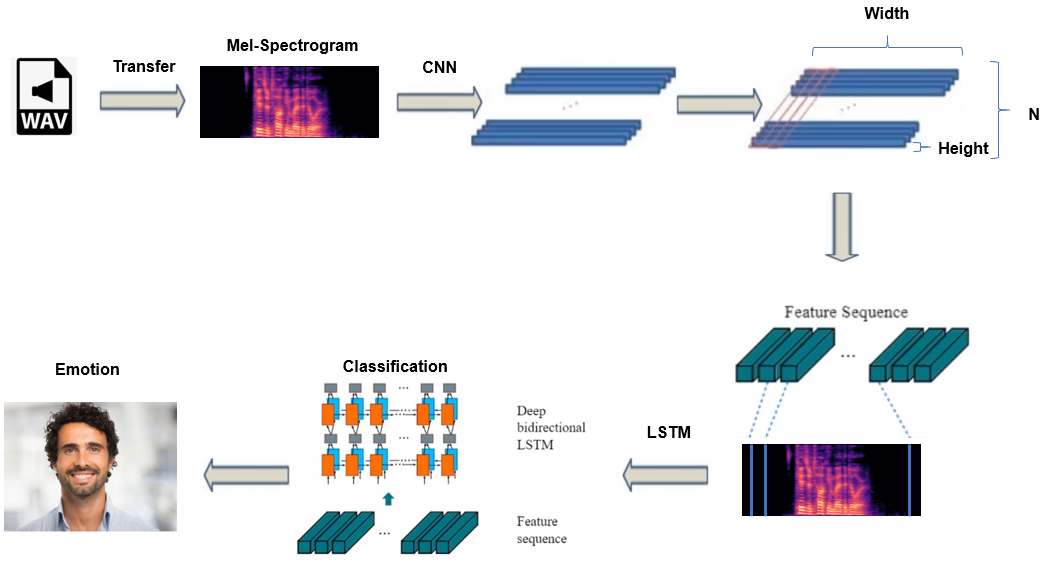


Figure 2 Steps of Analysis

The advantage of this CRNN based method can be divided into 2 parts. First, CNN is used to filter vision-based audio information, it directly analyzes the spectrogram. The merit is one spectrogram indicates one audio. It has uniqueness[2]. In LSTM, the feature map is treated as sequence. It shows the property of audio information very well since information on different time period is not separated with each other. Human emotion is continuous. LSTM can solve the information relation on the time scale well by setting forgetting gate and information gate. Moreover, LSTM can solve vanishing gradient and exploding gradient. Combining CNN and LSTM, it is enabled to deal human audio emotion feature on a time scale and identify the corresponding emotion.

1. **Data, Model and Results**

**4.1 Data Selection and Preprocess**

In this research, data set chosen is Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) from www.kaggle.com. RAVDESS is a data set containing audio with 8 psychological defined emotions-neutral, calm, happy, sad, angry, fearful, disgust, surprised. Another reason it’s chosen is that the audio is simulated by professional actors to eliminate the difference between emotion in real life and lab. Totally, there are 1440 samples. Then, the audio is transferred to computer vison by using spectrogram. In this research, Mel-spectrogram is applied. The reason is the tone human hear is not linear related with frequency of the voice. When the frequency of the voice doubles, the tone human hear is less than the doubled voice. To simulate this situation, Mel-spectrogram is applied. Formular 1 shows the mathematic relation between Mel frequency and frequency. Mel frequency is a type of logarithm calculation of frequency.

Formular 1 Relation Between Mel Frequency and Frequency

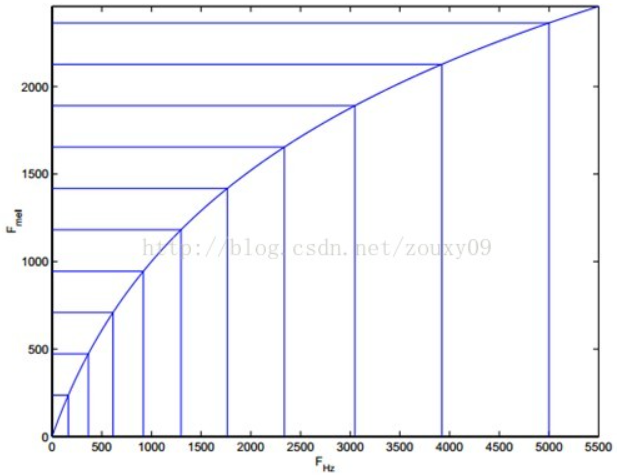


Figure 3 Frequency transferred by Mel-scale filter

In Figure 3, it shows the corresponding process of transferring frequency of voice to Mel frequency according to Formular 1. The x axis is the frequency of normal voice and y axis is the frequency transferred by Mel-scale filter. A non-linear relation is simulated by applying logarithm to normal frequency. After transferring, Mel-spectrogram is generated from frequency wave of original audio. Figure 4 shows the transformation visually. Mel-spectrogram will be the output of data preprocess and used in later CNN as input picture.

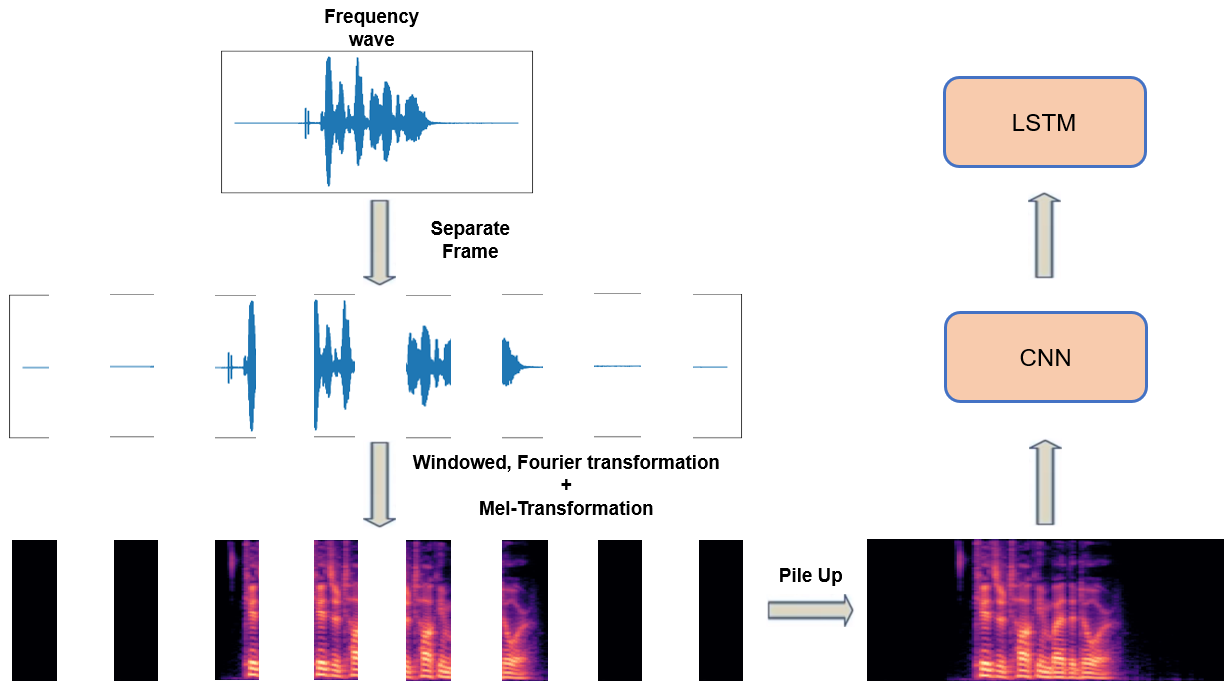


Figure 4 Mel-scale transformation

**4.2 Model Design and Results**

In the model, CRNN structure is used. It contains 2 part-CNN and LSTM. In CNN, 1 dimension convolution is applied on time scale.

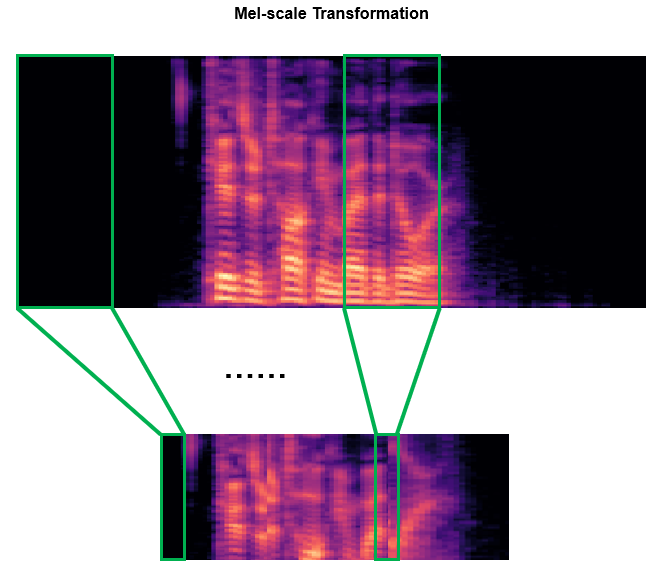


Figure 5 1 dimension convolution on time

In LSTM, Bidirectional-LSTM is used since human emotion is continuous and emotions are binarily connected with each other. For example, when someone is happy, his or her tone will increase. During his happiness status, happiness will not go away in as sudden. It will go away gradually. In this research, structure with only CNN and structure with CNN+Bidirectional-LSTM(CRNN) will be used in training to have a comparison. Table 1 shows the structure of only CNN model. Table 2 shows the structure of CNN+Bidirectional-LSTM model. Regarding to Baoguang Shi, Xiang Bai and Cong Yao[6], before going into Bidirectional-LSTM, we need to generate the feature map of Mel-spectrogram into a 1\*N\*M(Height:1, Width:N, Channels: M) tensorflow during CNN by setting . Then the feature will be transferred into sequence and put into Bidirectional-LSTM. The width of the sequence refers to the length of the audio but is a result of feature extracted. The final out put is a 8-class probability vector, numbers in which respectively refer to probability of the audio to be neutral, calm, happy, sad, angry, fearful, disgust and surprised.

|  |  |
| --- | --- |
| Type | Configurations |
| Input | 288\*720\*1 |
| Convolution | Maps:32, K:3X3, S:1, P:1, relu |
| Convolution | Maps:64, K:3X3, S:1, P:1, relu |
| MaxPooling | Window:3X3, S:3 |
| Dropout | Dropout:0.4 |
| Convolution | Maps:128, K:3X3, S:1, P:1, relu |
| Convolution | Maps:256, K:3X3, S:1, P:1, relu |
| MaxPooling | Window:3X3, S:3 |
| Dropout | Dropout:0.4 |
| Convolution | Maps:512, K:3X3, S:1, P:1, relu |
| Convolution | Maps:512, K:3X3, S:1, P:1, relu |
| MaxPooling | Window:3X3, S:3 |
| Flatten() | --- |
| Dropout | Dropout:0.25 |
| Dense | Neural: 128, relu, L2 regularization |
| Dropout | Dropout:0.25 |
| Dense | Neural: 32, relu |
| Dense | Neural: 8, softmax |
| Training | EarlyStop, categorical\_crossentropy |

Table 1 CNN Structure

To optimize the CNN model, first, Relu function is used as activation in every convolution layer to avoid vanishing gradient. After transferring feature maps into flatten tensorflow, dropout is set at a rete of 0.2 to avoid overfitting. Finally in training, EarlyStop is set to avoid overfitting and categorical crossentropy is used as Loss function in the multiple classification task to increase accuracy. The result of CNN model is shown in Figure 6. The model achieves an accuracy rate at 70% with categorical crossentropy.



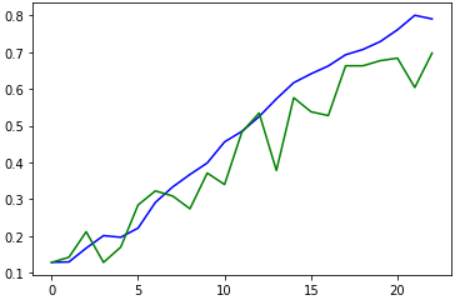


Figure 6 Result

|  |  |
| --- | --- |
| Type | Configurations |
| Input | 288\*720\*1 |
| Convolution | Maps:32, K:3X1, S:1, P:1, relu |
| MaxPooling | Window:3X2 |
| Convolution | Maps:64, K:3X1, S:1, P:1, relu |
| MaxPooling | Window:3X2 |
| Dropout | Dropout:0.4 |
| Convolution | Maps:128, K:3X1, S:1, P:1, relu |
| MaxPooling | Window:3X1 |
| Convolution | Maps:256, K:3X1, S:1, P:1, relu |
| MaxPooling | Window:3X1 |
| Dropout | Dropout:0.4 |
| Convolution | Maps:512, K:3X1, S:1, P:1, relu |
| Convolution | Maps:512, K:3X1, S:1, P:1, relu |
| MaxPooling | Window:3X1 |
| Dropout | Dropout:0.3 |
| Flatten() | --- |
| Map\_to\_Sequence | 1\*180\*512 |
| Bidirectional-LSTM | Hidden\_units:256, Dropout:0.2, Recurrent\_dropout:0.2 |
| Bidirectional-LSTM | Hidden\_units:64, Dropout:0.2, Recurrent\_dropout:0.2 |
| Bidirectional-LSTM | Hidden\_units:16 |
| Dense | Neural: 8, softmax |
| Training | EarlyStop, categorical\_crossentropy |

Table 2 CRNN(CNN+Bidirectional-LSTM) Structure

Also, to optimize CRNN(CNN+ Bidirectional-LSTM), Relu function is used in convolution layers. In Bidirectional-LSTM layers, dropout is set at 0.2 and recurrent\_dropout is set at 0.2 to avoid overfitting. Softmax is chosen as activation function in the dense layer. The result of CRNN model is shown in Figure 7. The model achieves an accuracy rate at % with categorical crossentropy just at 35%.

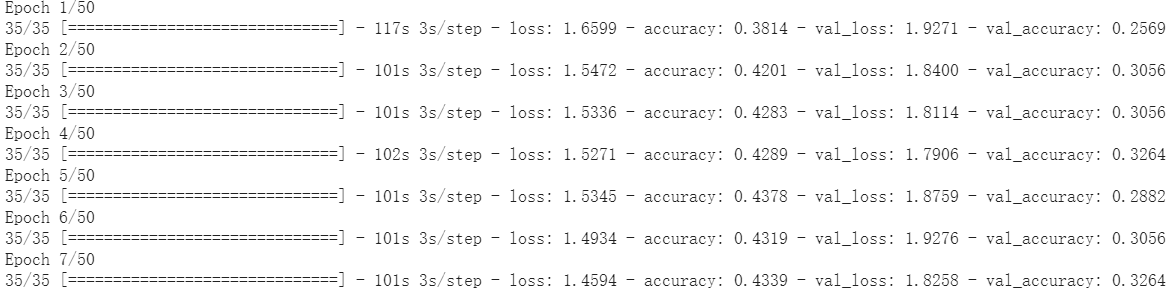


Figure 7 Result of CRNN

**4.3 Model Application**



Figure 8 Result

Figure 8 shows how to save the model. In real business, this model can be used in audio emotion detection in many scenarios. For example, in a scenario of an interview of clients on their opinion on the car they purchased. The researches of the car company can use the audio record to detect the hidden emotion and get a comprehensive understanding of what clients really think about their products by using this model.

1. **Conclusion and Prospect**

In the research, 1440 3-second audio wav samples is used, which cover 8 psychology defined emotions, neutral, calm, happy, sad, angry, fearful, disgust and surprised. By using pure CNN structure & CRNN structure including deep convolution(CNN) feature extraction and bidirectional long-short term memory structure(Bidirectional-LSTM) dealing with time sequence tensorflow. CNN model reaches an accuracy at 70%. CRNN reaches an accuracy at 35%???.

The model will be useful in today’s scenarios of AI, the robots can identify the speaker’s emotion and generate corresponding words to them. If the speakers are sad, robots will tell some joke. If the speakers are fearful, robots will say something to comfort them. Moreover, using this model, we are abled to generate Mel-spectrogram simulating emotions and Mel-spectrogram can be transfer to wav using python. In other word, we can teach AI how to speak. Combining with NLP text generation, what to speak, the model can equip those generated text with emotion, which enable the robot to generate more human-like communication automatically. This kind of technology can be used in ways benefiting the world. For example, in orphanage, AI designed with function of verbal communication with emotion can help relieve small children’s negative emotion like loneliness or fear. Sure, more analysis need to be done before the ultimate model come into its birth. As a trial in my research, it’s purpose is to investigate the possibility of enabling AI-made audio emotions.

**Reference:**

[1]. J. Futrelle and J. S. Downie. Interdisciplinary communities and research issues in music information retrieval. InProc. Int’l Conf. Music Information Retrieval, pages 121–131, 2002.

[2]. Xu Youzheng. Research and Application of Music Classification Based on Convolutional Neural Network[D], Nanjing University of Posts and Telecommunications, pages 3–20, 2018

[3]. Ling Yuhui. Content-Based Music Recommender System Using Deep Learning[D]. Xiamen University, pages 20–30, 2018

[4]. Chen Changfeng. Music Audio Sentiment Classification based on CNN-LSTM[J]. Communications Techonology

[5]. Bo Hongjian. Research and Implementation on Deep-learning-Based Facial Recognition[D]. Harbin Institute of Technology, pages 26-40. 2019

[6]. Baoguang Shi, Xiang Bai and Cong Yao. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition. School of Electronic Information and Communications Huazhong University of Science and Technology[D]. Wuhan, China, pages 5-6, 2015