Novel Speech Emotion Recognition Using Audio and Text

Rui Sun, Gang Yang, Boxun Xu, Wei Hu, Chuxuan Chen
{rayss, gangyang, boxunxu, huwei, cxchen}@umich.edu

Abstract

Speech emotion recognition is a challenging task, and extensive reliance has been placed on conventional models that use audio features in building well-performing classifiers. To achieve higher accuracy of emotion recognition, this paper proposes a novel model utilizing audio signal and text data simultaneously to obtain a better understanding of speech emotion. Text sentimental analysis in virtue of speech text help understand the emotion in the speech. First, this paper builds the neural network models for text and audio respectively, and checks their performance. Then, two models are combined in parallel to build an Audio-Text Mixed (ATM) neural network architecture. Based on our experiment results, the ATM model achieves an average accuracy of 68%, which outperforms previous state-of-the-art models in assigning speech sentiment to one of four emotion categories(i.e., angry, happy, sad and neutral) when applied to the IEMOCAP dataset.

1 Introduction

Speaking is a fast and straightforward method of communication. Besides, Speech signals are more accessible and more economical than other biological signals. Hence, it can be a good resource to provide high-quality interactions between humans and machines. In the past decades, there have been significant breakthroughs in the area of voice synthesis, speech recognition, and text-to-speech [1] [2] [3]. However, the machine cannot recognize the emotion of the speaker with high accuracy. Since the emotion conveys a wealth of information about an individual's mental state, it enhances the communication between not only people but man and machine. Speech emotion recognition (SER) enables machines to detect a speaker's emotions. It aims to predict and classify the emotion content of the speech. SER can be utilized in applications that involve man-machine interaction. It is also useful for companies to gauge customer mood towards their product or brand, e.g., in medicine, entertainment, education, robotics engineering, and call center applications.

Various designs have been proposed for SER. However, they focus on the audio features only, and the accuracy is limited. To improve the performance, this paper proposes a novel approach that employs audio and text contents of the speech simultaneously, and we build the audio-text mixed neural network architecture to classify the emotion. The model performance is not bad even though the size of the dataset is relatively small. Given achievements in the field of automatic speech recognition (ASR) [4], text content of speech can be obtained from audio signals with high quality, which enables the deep learning network to learn the characteristics of emotion from not only the low-level audio signals but high-level text content. In this paper, we build our model based

on IEMOCAP (Interactive Emotional Dyadic Motion Capture) [5] dataset. IEMOCAP is a gender balanced database which contains audio and video recordings by 10 actors. It utilizes statements that guarantee the emotion of speech; therefore, text-based SER can be applied to it. For the details about the dataset, we would introduce in the later section.

2 Related Work

Speech emotion recognition is an attractive and challenging topic in deep learning, and many researchers have engaged in this topic and made extraordinary achievements. There are several approaches to recognize speech emotion. One novel approach is to develop a deep dual recurrent encoder model that simultaneously utilizes audio and text data in recognizing emotions from speech. In Multi-modal Speech Emotion Recognition Using Audio and Text[6], researchers employ this approach to deal with the speech emotion recognition. The general framework of their work is building the recurrent encoder model for the audio and text modalities respectively, and then proposing a multi-modal approach that encodes both audio and textual information simultaneously via a dual recurrent encoder. To be specific, in the recurrent audio encoder, they extract the Mel-frequency cepstral coefficient (MFCC) features from the signal and put it into the recurrent neural network (RNN). In the recurrent text encoder, they first obtained the text transferred from audio signals by ASR technologies and then apply the Natural Language Toolkit and word embeddings to preprocess the text. Then, they put the embedded tokens into the RNN. Finally, they combine the results of the final hidden states from the above two RNN. In our paper, we would also use the deep dual model to recognize the speech emotion as one of the approaches. Comparing to the above paper, we prefer to use the Bi-LSTM instead of the RNN in our neural network.

Besides using the deep dual model, we also want to use a more typical approach to recognize the speech emotion, which is merely based on the audio and without the assistance from the text. The mainstream model of this approach is generally using CNN+LSTM architecture. C. Etienne et al. [7] demonstrate the great performance of this model. First, they augment the dataset using vocal tract length perturbation (VTLP) and minor class oversampling. Then, they try the model with different convolutional and recurrent layers and find out the combination of 4 convolutional layers and 1 Bi-LSTM layer performs the best. In the reference paper, they use the deep CNN architecture. However, in our experiments, using one CNN layer is also able to do the classification task, but the variation of accuracy is large. The final classification accuracy with this simpler network is similar to their result.

3 Methodology

3.1 Text model

3.1.1 Text Data Preprocess

Texts and labels are located in different files. The first thing is to extract and pair texts and labels. Once getting the text-label pairs, we need to clean our texts, like setting the words to the lowercase, expanding the abbreviation by using the contraction map (i.e., it's $\rightarrow$ it is, I'm $\rightarrow$ I am), and removing punctuation. Normally, data cleaning process should remove stop words from the texts. However, since this research is based on the daily dialogue, stop words are kept; otherwise

the length of the texts would decrease dramatically. Meanwhile, we find that 95% text sequences having length less than or equal to 34. Then, we should transfer the text sequence to the size-fixed integer sequence. Here, we mainly use the preprocessing modules "Tokenizer" and "Pad_sequence" from TensorFlow. The "Tokenizer" would create a dictionary based on the tokens in the text, whose keys are the tokens and the values are the frequency ranking of the tokens. Then, the text sequence would transfer to the integer sequence. Also, to keep all the integer sequence the same length, we would use "Pad_sequence" to supplement zero at the end of the integer sequence if it is shorter than our setting value. The setting value is 35, since the majority of the text have length less than or equal to 34. Furthermore, as the labels are string type, we should encode it to the integer type. By now, we have done the text data preprocessing.

3.1.2 BiLSTM Model

We mainly use the Pytorch to construct the TER (text emotion recognition) encoder. Figure 2 shows our TER model. The first layer is the embedding layer, which could change the integer sequence into vectors. Then, we apply the Bi-LSTM layer, which applies two regular LSTM from both sides of the text sequence. Comparing to the regular LSTM, BiLSTM can capture more information from the sequence. By now, the text encoder is set up, which will be used in the dual encoder later. To partly check the performance of text-based classification, we apply the SoftMax layer at the end of the neural network, and get a test accuracy of 61.5%.

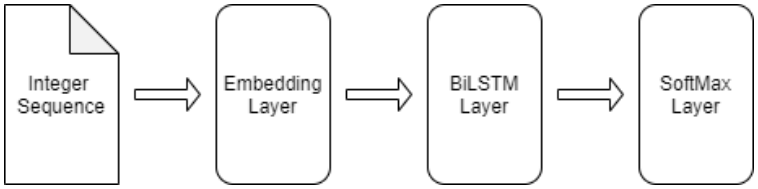

Figure 1: Text Model

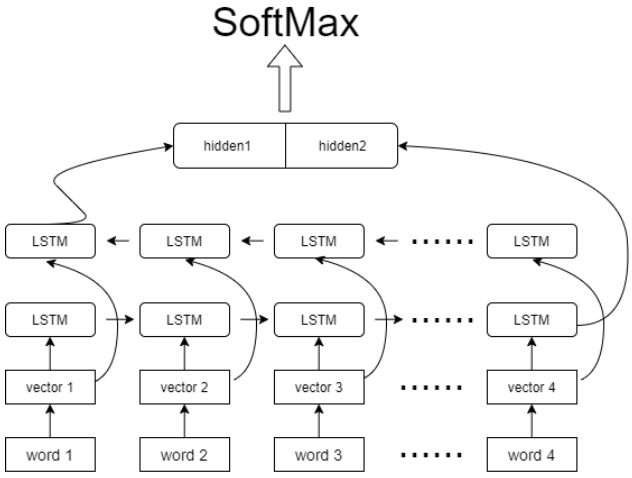

Figure 2: Bidirectional LSTM Model for Text Classification

3.2 Audio model

3.2.1 Audio Data Preprocess

For audio data preprocessing, we referred to other people's work[1]. The first step is loading audio data through Librosa package. Librosa package is a widely used python package in audio tasks. In addition, we can get the sample rate by loading the data. The second step is to do zero padding, which could unify the length of each data. By doing this, every audio data can be divided into the same number of time steps. The last step is to use ¡librosa.feature.mfcc¿ function to extract MFCC features from every time step. The MFCC feature contains a total of 39 features, which include 12 MFCC parameters from the 26 Melfrequency bands and log-energy parameters, 13 delta and 13 acceleration coefficients. In our experiment, we use 39 features and the total number of time steps is 750. Finally, the input data for audio part is a 3-dimension matrix with a shape of [batch_number, feature_number, timesteps_number].

3.2.2 CNN-BiLSTM Model

For AER (Audio Emotion Recognition) encoder, we construct a three layer neural network. The first layer is CNN layer shown in Equation.1 and follows a ReLU function. The second layer is bi-directional LSTM layer shown in Eqution.3 and the last layer is a fully connected layer. In the first layer we perform convolution on time steps to better organize features on every time steps. Figure.3 shows a brief diagram about how we do convolution on time steps. The parameters used in the CNN layer are as follows: 39 input channels, 10 output channels, and 4 kernels. The second layer, LSTM, relying on its three gates structure, is effective in solving the long-term dependence in the neural network and is useful to solve time-dependence problem. For this layer, the input size is 10, and the hidden size is 16. After that, we get 23904 nodes, so the input size of the fully connected layer is 23904 and the output size is 4. In the end, we use a SoftMax layer as shown in Equation.2 to make classification and use cross entropy loss function shown in Equation.4 as total loss to train AER model.

$$O^{oc} = W^{oc} \circledast X + b^{oc} = \sum_{ic=1}^{IC} W^{oc,ic} \circledast X^{ic} + b^{oc} \tag{1}$$

$$Prob(x_a) = softmax(x_a) = exp(x_a) / \sum_{k=0}^{K} exp(x_k) \tag{2}$$

$$h_t = f_\theta(h_{t-1}, x_t) \tag{3}$$

$$L = -\sum_{c=1}^{M} y^{oc} log(Prob^{oc}) \tag{4}$$

[1]https://github.com/david-yoon/multimodal-speech-emotion/blob/master/preprocessing

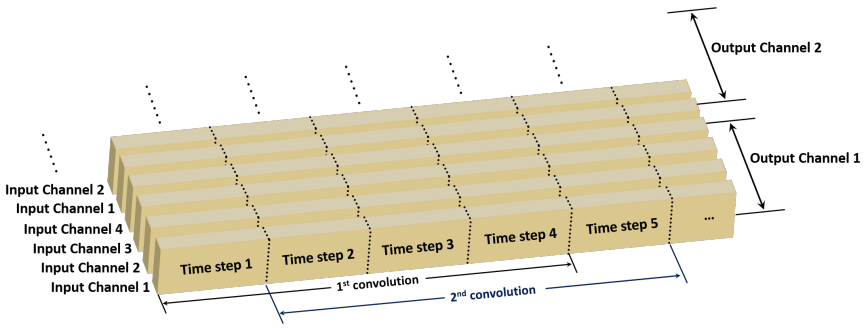

Figure 3: CNN on time steps

3.3 Audio-Text Mixed Network

In this part, we present a novel speech emotion recognition architecture called Audio-Text Mixed(ATM) model to overcome the limitations of existing models. In this model, we consider multiple modalities such as MFCC features from audio samples and scripts contain sequential audio and information textual information respectively. Types of data are the same as those used in the previous AER and TER units. The ATM model employs combine of two networks to work in parallel to encode data from the audio signal and textual inputs independently.

The AER unit encodes MFCC features from the audio signal by 1 convolution layer and 1 bidirectional-LSTM layer to generate a vector, and the vector is passed through a fully connected neural network layer to form the audio encoding vector **A**. On the other hand, the text emotion recognition unit encodes the word sequence of the transcript by 1 embedding layer and 1 bidirectional-LSTM layer. The final hidden states of the text emotion recognition unit are also passed through another fully connected neural network layer to form a textual encoding vector **T**. Finally, the emotion is predicted by applying the softmax function(5) to the concatenation of the vectors **A** and **T**. The overall architecture is shown as figure 4. The training objective is the same as the ARE model, and the predicted probability distribution for the target class is shown in next section.

$$y = softmax(concat(A, T)^T M + b) \tag{5}$$

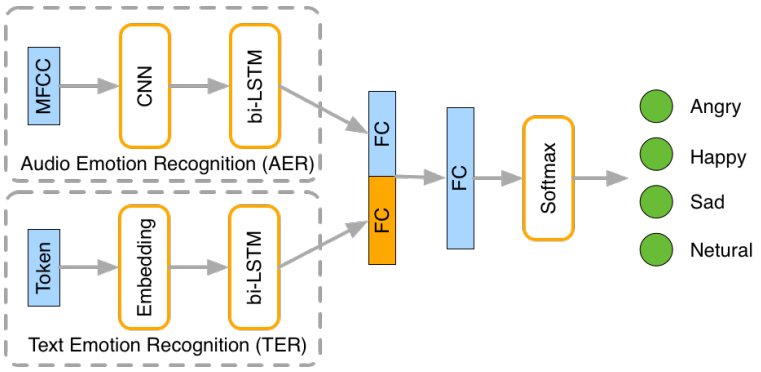

Figure 4: Architecture of Audio-Text Mixed(ATM) neural network for speech emotion sensing

4 Experiments

4.1 Problem Definition

In this paper, our purpose is using both the audio and the speech transcripts generated from the audio to predict the emotion of the audio. Thus, the input of the system comes from two sources – one is the audio itself, the other is the corresponding text; the output of the system is the predicted emotion.

4.2 DataSet

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database, collected by USC, consists of about 12 of hours audio data and scripts performed by 10 actors (5 men and 5 women). The original emotional categories in IEMOCAP are unbalanced. We pick four most represented emotions: neutral, sad, angry and happy (happy is combined with original excited and happy categories), and the distribution of each emotion is shown in Figure.5. 90% of the data is applied for training and validation, and 10% of the data is applied for testing.

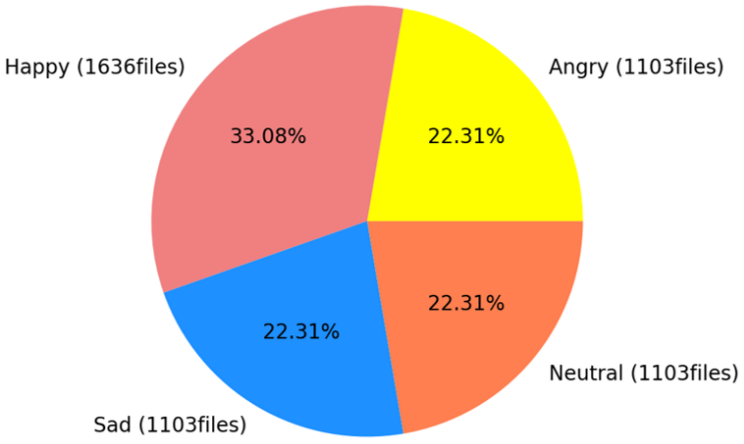

Figure 5: Data distribution of the 4 most represented emotions picked

4.3 Evaluation and Results

We analyse the predictions of AER model and ATM model. Our experiment results are shown in figure 6. For the AER model, the best classification accuracy is 65% for neutral and the worst case is 42% for happy. The overall accuracy is only 53%. The result generated by only AER model is not optimal. However after improved by using ATM model, the predict accuracy is higher than the AER model in all four emotions respectively. The best accuracy achieves 71% for happy and the worst case is 56% for angry. In the end, the overall accuracy for ATM model achieves 68%, which is increased by 15% compared to AER model only. Meanwhile, we compared our results to other models built by other researchers with the same dataset(shown in table 1), our ATM model outperforms those models. Thus, it is evident that applying the text features in the model could help predict the speech emotion.

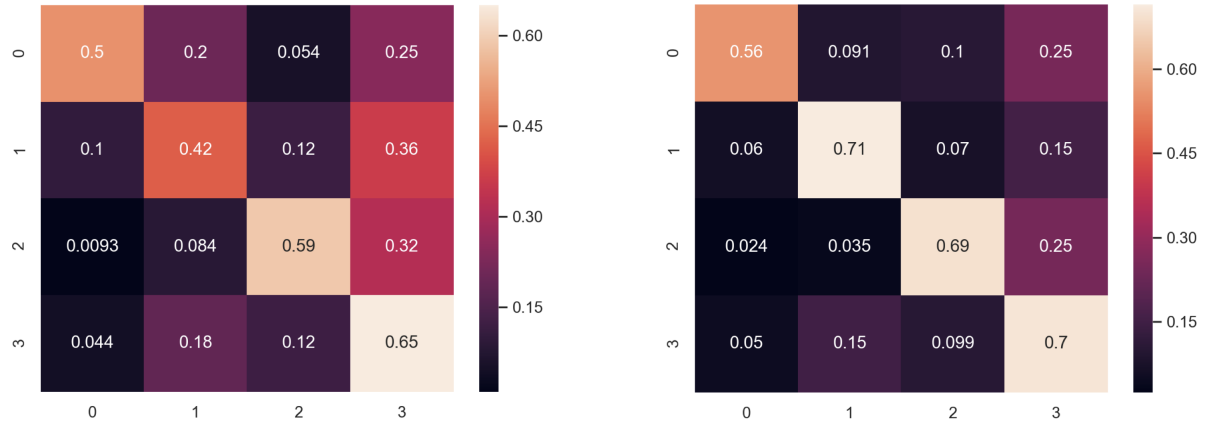

Figure 6: Confusion matrix of each model with Audio-only model on left and Audio-Text Mixed on right where x-axis and y-axis indicate predicted and true class respectively, and 0,1,2,3 indicate angry, happy, sad and neutral respectively

Table 1: The comparision of different methods

Method	Weighted Accuracy(%)
ACNN[8]	56.1
RSM+SVM[9]	62.0
LLD RNN-attn[10]	63.5
RNN(prop.)-ELM[11]	62.8
Our AER model	**53.0**
Our ATM model	**68.0**

5 Conclusion

This project proposes a multimodal method which simultaneously utilizes text data as well as audio information to sense emotion (angry, happy, sad and neutral) which permits the better understanding of speech data, improves the accuracy from 53.0% to 68.0% and outperforms most state-of-the-art methods in classifying the four emotion categories. For the text part, we use the embedding and Bi-LSTM layers; for the audio part, we use the structure of CNN-LSTM. With the high accuracy in speech emotion recognition, we could apply our model in lots of areas, such as intelligent customer service system, psychological counseling, etc..

6 Future Work

Using pre-trained model gradually becomes an important approach in the machine learning. In our project, the size of the dataset is relatively small, it is highly possible that some words in

the test data is not covered in the training data. So, if we use the pretrained word embedding, like BERT, we would not be worried about the gap between the training set and the testing set. Furthermore, we aim to extend the modalities to audio, text and video inputs, and for audio part, prosody features will be extracted to imporove the performance of our architecture.

7 Author Contribution Statement

Rui Sun: Architecture Modeling, Audio data pre-processing, Audio network training
Gang Yang: Text data pre-processing, text network training, Demo testing
Boxun Xu: Architecture Modeling, Audio-text mixed network training, Data visualization
Wei Hu: Literature review, Demo testing, Data visualization
Chuxuan Chen: Literature review, Data Collection, Data Visualization

References

[1] M. Mutsuno and T. Fukada, "Text structure for voice synthesis, voice synthesis method, voice synthesis apparatus, and computer program thereof," Feb. 3 2009, uS Patent 7,487,093.

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[3] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 195–204.

[4] P. Lange and D. Suendermann-Oeft, "Tuning sphinx to outperform google's speech recognition api," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2014*, pp. 32–41, 2014.

[5] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[6] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[7] C. Caroline Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "Cnn+ lstm architecture for speech emotion recognition with data augmentation," *arXiv preprint arXiv:1802.05630*, 2018.

[8] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *Processing of Interspeech*, pp. 1263–1267, 2017.

[9] C. M.Shah and A.Spanias, "A mult-modal approach to emotion recognition using undirected topic models," *Processing of the 2014 IEEE International Symposium on Circuits and Systems(ISCA)*, pp. 754–757, 2014.

[10] E. B. Seyedmahdad Mirsamadi and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, p. 2227–2231, 2017.

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