A Multi-modal Approach to Speech Emotion Recognition

Jiayu Zhou, Paras Sibal, Rohit Bernard, Sanjana Mallikarjuna, Swapnil Mallick
University of Southern California
{zhoujiay,psibal,rpbernar,moodbagi,smallick}@usc.edu

1 Introduction

Speech Emotion Recognition (SER) attempts to identify emotions from speech irrespective of the semantic content. The domain of emotion recognition from natural language is quite extensive. Emotions can be recognised from multiple sources ranging from audio data, facial images to the utterances or text.

Emotion recognition has found great importance in the field of human-computer interaction. This objective motivates the use of automatic emotion recognition systems. The advancements in the field of emotion recognition have made it useful for monitoring psycho-physiological state, improving customer service and forensics.

Having said that, emotion recognition from natural language is still considered to be a challenging task. This is because human emotion can be very subjective. Sometimes, it can be difficult for even human beings to accurately identify the emotion attached to a particular conversation. Thus, considerable research work is still being carried out to devise automatic emotion recognition systems that can efficiently identify the emotion in a conversation.

In this paper, an attempt has been made to come up with a novel multi-modal architecture for emotion recognition from natural language. Different modalities can be used when it comes to emotion recognition. For our work, we are using audio and text as the modalities and IEMOCAP (Busso et al., 2008) as the primary source of data. First, we try to build models that are capable to identify emotions separately from each of the two modalities. Consequently, we try to concatenate these two models to see the effect use of different modalities has on emotion recognition. Adding different modalities makes our emotion recognition model more powerful and robust.

2 Related Work

Attempts to build efficient systems for emotions recognition has been an active area of research over the years. But the main focus of most of the previous attempts have been on either text or audio data separately. Research on multi-modal approach to emotion recognition is still in its latent stages. One of the reasons for insufficient research may be the lack of reliable and complete datasets. However, IEMOCAP can be considered as a benchmark dataset in this regard.

Though IEMOCAP provides as a good source of data with multiple modalities, a significant amount of preliminary research on IEMOCAP dataset has been focussed on speech (or audio) data only. Majority of the architectures are based on neural networks. (Han et al., 2014) is considered to be one of the earliest works in this regard. In their work, the authors have proposed an MLP based architecture to solve the problem. For every speech segment, an emotion state probability distribution is produced. These segment level features are then fed to the proposed model.

Another seminal work (Lee and Tashev, 2015) suggests a Bidirectional Long Short-Term Memory (Bi-LSTM) architecture. For every frame, 32 features are extracted - F0 (pitch), voice probability, zero-crossing rate, 12-dimensional Mel-frequency cepstral coefficients (MFCC) with log energy, and their first-time derivatives. These 32-dimesional vectors are expanded to 800-dimensional vectors and fed into the Bi-LSTM model which contains 2 hidden layers with 128 Bi-LSTM nodes.

(Hazarika et al., 2018) have proposed a multi-modal architecture tested on IEMOCAP dataset. The model is based on Gated Recurrent Unit (GRU). The authors have used different combinations of modalities and compared the accuracy of their model to see how different modalities affect the performance of the model. In our work, we

propose a novel method for multi-modal emotion recognition and achieve either similar or superior performance.

3 IEMOCAP

3.1 Existing dataset

The dataset used to build the multi-modal architecture is USC's Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) dataset. To obtain the data, we requested USC's SAIL laboratory. IEMOCAP is an audio-visual dataset used in emotional learning. It contains 12-hour record of simulated scenarios from ten actors. The scenarios are designed to draw out a set of emotions. There are 10039 utterance files generated from ten actors. Six annotators classify each utterance based on their emotions into ten categories: anger, disgust, excited, fear, frustration, happiness, neutral, sadness, surprise, and other. The files were annotated in such a way that each utterance was annotated by three annotators. For each utterance, there is audio-video file and transcriptions. One utterance can have as many as four emotions. Because it is possible that one annotator labels an utterance with more than one emotion. This shows that nature of human interaction from the hearing can lead to different meaning.

3.2 Dataset creation

From the IEMOCAP dataset, we extract the data required for our modelling. For each utterance, both audio as well as the transcriptions are used. Labels given by the three annotators are converted to percentages. For example, if one utterance was labelled 'anger' by first annotator, 'frustrated' by second annotator and 'anger' by third annotator, then the utterance was assigned values 0.66 anger and 0.33 frustration. The dataset also had one label assigned. The assessment for each utterance is decided by the simple majority vote. For simplicity and for better performance of the multi-modal architecture - 5 emotion categories - Anger, Frustration, Happiness, Neutral and Sadness were considered for model training and testing. Excited was converted to Happiness and utterances with other labels were ignored. On training our model using 8 categories vs 5 categories, 5 classes gave us a more stable model with lesser variance.

Emotion	No. of Utterances
Anger	1103
Frustration	1849
Happiness	1636
Neutral	1708
Sadness	1084

Table 1: No. of Utterances of each category

3.3 Data Split

Our experiment required splitting of data into training, validation and testing set. The IEMOCAP dataset does not have these splits. To train and validate our model, and finally to make predictions on the test data, we split the dataset using sklearn library's train-test-split into 70% training data, 20% validation data and 10% testing data. There is no overlapping occurrence between the sets. Splitting is such that there are approximately same proportion of data of each category in each of the splits.

4 Data Preprocessing

For the development of a speech emotion analyzer, preprocessing is the crucial phase to tune the audio data to create feature vectors used as an input to the model. Nowadays, most researchers utilize deep learning techniques for SER using Mel-frequency cepstrum coefficients and Mel-scale filter bank speech spectrogram as input features to their models. Similarly, we utilized the transfer learning strategies for SER using speech spectrograms to extract the salient and discriminate features.

The audio data obtained from the raw dataset had non-uniform dimensions varying in audio length. The raw data was sampled to have uniform standardized dimensions by adding zero paddings to shorter audio clips to get an equal number of features. Furthermore, we checked for the signal-to-noise ratio to assure that there was less ambient distorted noise in the data. Moreover, the sampling rate for each audio file was set to 22 kHz. This allowed each audio sample to get all features that helped classify the audio file while keeping noise minimum.

When we do speech recognition tasks, Mel-frequency cepstrum coefficients (MFCCs) are the state-of-the-art features of the training model. For this project, we generated mel spectrograms using tuned data clips with the help of python librosa library and converted them to mel scales so that

the perpetual difference becomes smaller to deal with human frequencies. Furthermore, the log-mel spectrogram, given as input to the model, is normalized by the mean and variance of the training set. Moreover, we compressed the mel spectrogram into a short-time spectrum to extract only the essential coefficients, which only correspond to human frequency ranges.

In addition, we performed Data Augmentation by adding noise, pitch, stretch, and shift to the original signal. This immensely improved the accuracy of our model.

5 Design

In this paper, we compare the accuracies achieved by text-only based models, audio-only based models, and our novel multimodal model. In order to compare these accuracies, our multimodal architecture combines the results generated by two separate models, each trained using a different modality.

5.1 Emotion Recognition from text

In order to predict emotions from text, we fine tune BERT (Devlin et al., 2019) on our dataset to perform 5-class text classification. We make use of the pre-trained transformer model 'bert-base-cased'. To fine tune the model, we use CrossEntropy loss and Adam optimizer over 15 epochs. The results obtained by the text-only model were documented.

5.2 Emotion Recognition from audio

In order to predict emotions from audio, we train a Convolutional Neural Network using VGG frame from scratch. The model consists of 5 Convolutional layers with Max Pooling, 2 Linear layers, LeakyReLU activation, Dropout, and a Softmax head. The model is trained to classify 5 classes of emotions using the MFCC generated from the audio files. We use CrossEntropy loss and Adam optimizer over 15 epochs. The results obtained by the text-only model were documented.

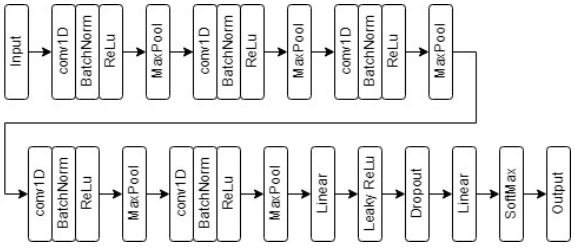

Figure 1: CNN in VGG Frame Architecture

5.3 Multimodal Design

To create our multimodal architecture, we combine our fine-tuned BERT model with our CNN in parallel, and combine the results of these two models. For each datapoint, both models produce 5x1 vectors containing the probabilities for each class. We pass each vector through a Linear layer, after which they are summed up and passed through a Softmax layer to generate the final predictions. This multimodal architecture is trained again using CrossEntropy loss and Adam optimizer over 15 epochs. The results obtained by the text-only model were documented.

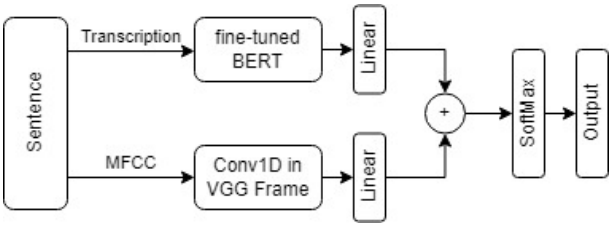

Figure 2: Multimodal model Architecture

6 Experiment

6.1 Experimental Setup

We evaluate our proposed model on IEMOCAP dataset, and consider five broad emotion categories - Anger, Frustration, Happiness, Neutral and Sadness for examining the performance of our architecture.

Baseline Methods: We compare our proposed model with different state-of-the-art multi-modal emotion recognition architectures:

- **ICON:** Interactive COnversational memory Network (ICON) (Hazarika et al., 2018) is a multi-modal emotion recognition model. It uses Gated Recurrent Unit (GRU) to hierarchically model self- and inter-speaker emotional influences into global memories.

- **bc-LSTM:** (Poria et al., 2017) have proposed a bidirectional LSTM with hierarchical fusion. The model uses context features from unimodal LSTM's. The features are concatenated and fed as input to the final LSTM.

Evaluation Protocols: We use 20% of our training set as the validation set for hyper-parameter tuning. We train each of the three models for 15

epochs. We use Adam optimizer to train our models with a learning rate of 1.5e-4 and a weight decay of 1e-3.

Hyper-parameter	Value
Optimizer	Adam
Learning Rate	1.5e-4
Weight Decay	1e-3
Loss Function	CrossEntropy
Activation	ReLU, LeakyReLU
Dropout	0.6
# Epochs	15

Table 2: Hyper-Parameters

For audio-only model, we have used different architectures of CNN before coming up with the final model. We also used Wav2Vec (Schneider et al., 2019) method. For text-only model, we used different transformers like, BERT and RoBERTa (Liu et al., 2019) to see which model performs better.

6.2 Results and Discussion

We compare the performance of our proposed multi-modal architecture with that of the previously mentioned baseline models and find some interesting results. The results obtained from our experiment can be found in the table below.

Modality	Models		
	ICON	bc-LSTM	Our Model
Text	58.3	**73.6**	72.4
Audio	50.7	**57.1**	47.8
Text + Audio	63.8	75.6	**76.9**

Table 3: Comparison of accuracy of proposed model

When we take text as the modality, our model performs significantly better than ICON and comes pretty close to the performance of bc-LSTM. However, our model fails to outperform both the baseline models when audio is chosen as the modality. Finally if we use both text and audio as modalities then we find that our model performs notably better than ICON and achieves a small improvement over bc-LSTM.

If we examine our own model, we find that the multi-modal architecture performs slightly better than the text-only model. This, further, indicates that using multiple modalities helps in improving the performance of automatic emotion recognition systems.

7 Conclusion

In this paper, we have attempted to design a novel multimodal approach to Emotion Recognition, and compare the effectiveness of multimodality against any single modality. We can see that our multimodal model performs better than the baseline models. The results of the experiment show that we can achieve a good accuracy by fine-tuning a pre-trained model to predict emotion solely from text. However, a from-scratch, purely audio-based model does not perform as well as a text-based model. But, combining the predictions of both models before the final layer of the network does help increase the performance, which shows that speech audio does have some useful information that cannot be picked up from text alone, and thus showing that multiple modalities of data help achieve better accuracy than any single modality.

Team responsibilities.

Our project had various aspects that needed to work upon, from the preprocessing to the model implementation. Each individual carried out their respective task and helped other team members accomplish the results. Paras Sibal and Sanjana Mallikarjuna were in charge of Data Preparation, Data Preprocessing, and Data Augmentation. Rohit Bernard was in charge of the speech-to-emotion classification model and Swapnil Mallick for the text-to-emotion classification model. Jiayu Zhou carried out the final task of designing the multimodal architecture to combine the results from the speech-to-emotion and text-to-emotion classification models. In addition, each team member worked on a separate speech-to-emotion classification model based on their understanding to find the perfect fit model for our project.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA,*

June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 223–227. ISCA.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2594–2604. Association for Computational Linguistics.

Jinkyu Lee and Ivan Tashev. 2015. High-level feature representation using recurrent neural network for speech emotion recognition. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 1537–1540. ISCA.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 873–883. Association for Computational Linguistics.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 3465–3469. ISCA.