The 15th International Conference on Auditory-Visual Speech Processing
10-11 August 2019, Melbourne, Australia

Multi-Modal Speech Emotion Recognition Using Speech Embeddings and Audio Features

Krishna D N, Sai Sumith Reddy

YouPlus India

krishna@youplus.com, saisumith@youplus.com

Abstract

In this work, we propose a multi-modal emotion recognition model to improve the speech emotion recognition system performance. We use two parallel Bidirectional LSTM networks called acoustic encoder (ENC1) and speech embedding encoder (ENC2). The acoustic encoder is a Bi-LSTM which takes sequence of speech features as inputs and speech embedding encoder is also a Bi-LSTM which takes sequence of speech embeddings as input and the output hidden representation at the last time step of both the Bi-LSTM are concatenated and passed into a classification which predicts emotion label for that particular utterance. The speech embeddings are learned using the encoder-decoder framework as described in [1] using skipgram [2] training. These embeddings are shown to outperform word embeddings(word2vec) in many word similarity benchmarks. Speech embeddings are shown to capture semantic information present in speech and speech embeddings have the capabilities to handle speech variabilities which is not possible by plain text. We compare our model with the word embedding based model where we feed Word2Vec to ENC2 and speech features to ENC1 and we observe that the speech embedding based model gives better results compared to word embedding based model. We compare our system to previous multi-modal emotion recognition models which use text and speech features and we get absolute 2.59% improvement over the previous systems[8] on IEMOCAP dataset. We also compare our system performance with different speech embedding dimensions of [50,100,200,300] and we observe the speech embedding of 50 dimensions is achieving 68.59% accuracy.

Index Terms: Speech2Vec, Speech emotion recognition, Multi-modal SER

1. Introduction

Speech emotion recognition is about identifying the emotion of a speech utterance. Due to the rapid development of deep learning, today we are able to build high capacity models which can ingest a huge amount of data and train the models to do better prediction. Today, due to the availability of large data and GPUs we can train better models much more quickly. The deep neural network has shown state of the art results for speech related problems like speech recognition [3], speaker identification[4], language indentification [5], etc. Recently a lot of work has been done in using neural networks for speech emotion recognition model and these models are shown to outperform classical machine learning models like HMM, SVM, and decision tree based methods. Hidden Markov models(HMM) and support vector machines(SVM) are used in the work done by Yi-Lin Lin et. al [6] for speech emotion classification. Recent work by K.Han et. al [9] shows how to use deep neural networks and extreme learning machines to first classify segment level and then classify utterance level emotion of speech. Convolution neural networks are used in the paper by Dario Bertero et. al [10] for speech emotion recognition. Work by Abdul Malik Badshah et. al [11] shows how deep convolutional networks can be used for speech emotion recognition and also they show how pretrained Alex-net model can be used as means of transfer learning for the emotion recognition task. Recently many works have been done on improving speech emotion recognition using multi-modal techniques. Researchers have shown that we can improve the accuracy of an emotion recognition model if we have text data of an utterance. If we provide text data as a guiding signal along with speech features to our model, the model performance can be greatly improved. Since face emotion and speech emotions are correlated with each other, work by Samuel Albanie et. al [7] tried using cross model transfer from face domains to speech domain using knowledge distillation. Using face features as one of the guiding signals during speech emotion prediction is shown to increase the performance of speech emotion recognition. Recent work by Jaejin Cho et .al[8] has shown how to combine text and audio signals for classifying speech utterance into emotion labels by encoding transcript using multi-scale CNN and encoding speech using LSTM networks.

In this work, we study emotion recognition using multi-modal setup using both audio and text information. The model consists of 2 encoders called acoustic encoder (ENC1) and speech embedding encoder (ENC2). The acoustic encoder (ENC1) is a Bi-LSTM which takes a sequence of speech features like MFCC,chroma-gram and,zero-crossing rate,short-term energy,short-term entropy of energy,spectral centroid and spread, spectral entropy, spectral flux and spectral roll-off computed for every 100ms and generates hidden representation at the last time step. The speech embedding encoder (ENC2) is also a Bi-directional LSTM which takes a sequence of speech embeddings at word level and generates hidden representation at the last time step. We fuse the hidden representation from ENC1 and ENC2 and finally send it to an MLP for classification. The ENC2 tries to encode the semantic information present in the audio signal whereas the ENC1 tries to encode the emotion-related features from the audio signal. Fusing these two networks for the final task of emotion classification gives better results compared to using only one them. The speech embeddings are trained using the encoder-decoder framework where the task is to predict the context(left+right) word given the center word. The center words are converted into a sequence of MFCC features and are fed into encoder LSTM and decoder network tries to generate MFCC for the context words and the model is trained using mean squared error loss. The setup is similar to the skip-gram model setup in training word embeddings.

The organization of the paper is as follows. In section 2 we explain out proposed idea. In section 3 we explain the dataset structure and in section 4, we explain our experimental analysis.

10.21437/AVSP.2019-4

2. Proposed model

In this work, we propose a multi-modal speech emotion recognition model which tries to capture acoustic features in one encoder and semantic information of the utterance in another encoder. This proposed model is shown in figure Figure 1. The model will have dual encoder architecture which fuses information from both the encoder streams and predicts the emotion labels for that particular utterance. We feed the acoustic features into an Bi-LSTM and generate utterance level presentation at the last time stamp of the Bi-LSTM and also we generate the utterance level semantic information by feeding the speech embeddings of the words into Bi-LSTM. Finally we concatenate these features. This concatenated speech feature carries information from acoustic signal and context information from speech embeddings. We pass these fused features through the dense layer and finally, we apply the softmax layer to predict the emotion class label. We describe our acoustic feature extraction and speech embeddings extraction in section 2.1 and section 2.2 respectively. We describe our full architectures of the proposed model in section 2.3.

2.1. Acoustic feature extraction

We use frequency and energy based features for our model. We compute 34 features which include 13 MFCC, 13 chromagram and 8 spectral features. The spectral features include zero crossing rate, short-term energy, short-term entropy of energy, spectral centroid and spread, spectral entropy, spectral flux, spectral roll-off. We compute these features for every 100ms windows with no overlap. The total number of feature vectors are limited to 200 which covers 20-sec audio data. These 34 dimensional feature sequence goes as input to the Bi-LSTM network and we tap the last timestep hidden representation to fuse it with speech embedding networks(ENC2) hidden representation.

2.2. Speech Embeddings

Speech embeddings are generated using a skip-gram model which is trained to predict the context words given the center word. The skip-gram model for speech embedding generation is described in the work by Yu-An Chung et. al [1]. The Speech2Vec model is shown in Figure 2. The speech2vec model is built using the encoder and decoder framework where the encoder is an RNN and decoder is also an RNN. The encoder in Speech2Vec takes MFCC features sequence of the center word and tries to encode these features into high-level representation using LSTM. The decoder takes the encoder representation and tries to generate the context words using LSTM. The speech embeddings are the hidden representations generated by encoder LSTM at the last LSTM time step. We use the speech and word embeddings given by Yu-An Chung et. al [1].

2.3. Acoustic encoder

The acoustic encoder ENC1 is a Bi-directional LSTM model which takes 34 dimensional acoustic features as input at every time step. This 34 dimensional features includes 13 MFCC, 13 chromogram and 8 spectral features like zero crossing rate, short-term energy, short-term entropy of energy, spectral centroid and spread, spectral entropy, spectral flux, spectral roll-off. The Bi-LSTM processes sequences from both forward and backward direction and has gating mechanism which controls what information is to be stored or discarded from memory. The acoustic encoder at t-th time step takes the acoustic features $\boldsymbol{x_t^A}$ and previous hidden state representation $\boldsymbol{h_{t-1}^A}$ to compute hidden state representation $\boldsymbol{h_t^A}$ at time t as follows.

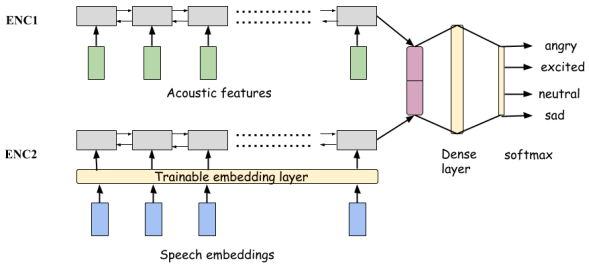

Figure 1: *Proposed multi-modal emotion recognition model*

$$\boldsymbol{h_t^A} = f(\boldsymbol{x_t^A}, \boldsymbol{h_{t-1}^A}, \boldsymbol{\Theta_A}) \tag{1}$$

where $\boldsymbol{f}$ is the Bi-LSTM function with parameters $\boldsymbol{\Theta_A}$,$\boldsymbol{x_t^A} \in R^{34}$ is the input feature from the input feature sequence $\boldsymbol{X^A} = [\boldsymbol{x_1^A}, \boldsymbol{x_2^A}.....\boldsymbol{x_T^A}]$ and t $\in [1, 2.....T]$, where T is the total number of time steps. Since, we are using Bi-directional LSTM, the network has the capacity to process our feature sequence both from forward direction and backward direction and the final hidden representation $\boldsymbol{h_T^A}$ is calculated by concatenating forward and backward hidden representation at the last time step.

2.4. Speech Embeddings encoder

Speech embedding encoder ENC2 is a Bi-LSTM model which takes sequence speech embedding vectors and calculates hidden layer output at every time step. These embeddings are designed to capture the context information in the utterance. The speech embeddings play an important role in capturing the semantic and syntactic information in the utterance. We add an embedding layer before the Bi-LSTM and initialize that embedding layer by the pretrained speech embeddings. These embeddings are generated from speech2vec model as described above where we get a fixed dimensional embedding for every word. The embedding layer is a trainable layer and this layer parameters are updated along with Bi-LSTM model parameters during backpropagation. The model takes a sequence of speech embeddings $\boldsymbol{X^S} = [\boldsymbol{x_1^A}, \boldsymbol{x_2^A}.....\boldsymbol{x_N^A}]$ as inputs to the Bi-LSTM model and generates a fixed dimensional embedding at the last time step.

$$\boldsymbol{h_n^S} = g(\boldsymbol{x_n^S}, \boldsymbol{h_{n-1}^S}, \boldsymbol{\Theta_S}) \tag{2}$$

where $\boldsymbol{g}$ is the Bi-LSTM function with parameters $\boldsymbol{\Theta_S}$ of speech embeddings model ,$\boldsymbol{x_t^S}$ is the input speech embedding from the speech embedding sequence $\boldsymbol{X^S} = [\boldsymbol{x_1^S}, \boldsymbol{x_2^S}.....\boldsymbol{x_N^S}]$ and n $\in [1, 2.....N]$, where N is the total number of words in that sequence. Since we are using Bi-directional LSTM, the network has the capacity to process our feature sequence both from forward direction and backward direction and the final hidden representation $\boldsymbol{h_N^S}$ is calculated by concatenating forward and backward hidden representation.

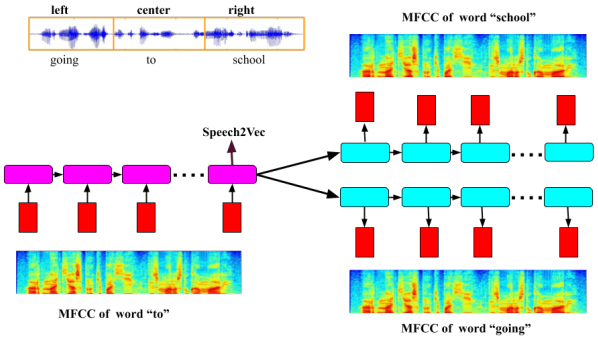

Figure 2: *Speech2Vec model*

2.5. Speech Embeddings and acoustic features for emotion recognition

In this section, we describe our approach of combining acoustic encoder representation and speech embedding encoder representation for predicting emotion label for an utterance. We extract the last time step hidden representation of the acoustic encoder $\boldsymbol{h}_T^A$ and last time step hidden representation of the speech embedding encoder $\boldsymbol{h}_N^S$. We then fuse these two features using early fusion. This fused features are passed to a dense layer and finally to softmax for emotion label prediction. This process is shown in Figure 2. The fused feature is represented as follows.

$$\boldsymbol{F} = concat[\boldsymbol{h}_A^T, \boldsymbol{h}_S^N] \tag{3}$$

This fused feature will have information from both acoustic features and semantic information of the contextual words from speech embeddings. This fused feature is passed through a dense layer and finally, softmax is applied to predict the label class. The loss is computed based on the predicted outputs and ground truth outputs. This is described as follows.

$$\boldsymbol{y}^* = softmax(\boldsymbol{F}^T \boldsymbol{M} + \boldsymbol{b}) \tag{4}$$

$$\boldsymbol{Loss} = CE(\boldsymbol{y}^*, \boldsymbol{y}) \tag{5}$$

3. Dataset

We conduct all our experiments on IEMOCAP dataset which is publicly available for research. The dataset is a 12hrs of multimodal dataset consist of audio, video and text information. The data is collected in order to simulate the natural interaction between actors. The dataset is collected in 5 sessions from 10 different people. Each session contains utterances of one male and one female in a conversation. These sentences were manually segmented and labeled by professionals with multiple validations. The dataset consists of total 5531 utterances from all 5 sessions. We use 4 sessions for training and the last session of testing. We use 4 emotions *angry*(1103), *excited*(1636), *neutral*(1708) and *sad*(1084) and data belonging to *happy* category is merged into *exited* category. For training we take 4290 utterances from first 4 sessions and we take 1241 utterances from last session for testing.

4. Experimental analysis

4.1. Acoustic feature based emotion recognition

Audio-based emotion recognition model is a single-layered Bi-directional LSTM trained using acoustic features only. We extract 34 acoustic features for every 100ms audio with no overlap and these features sequences are padded with zeros to make the sequence length to be equal to 200 and this corresponds to the 20sec worth of audio data. The speech files are sampled at 16KHz sampling frequencies. The model is single layer Bi-LSTM having the cell dimension of 128 and the Bi-LSTM layers are trained with a dropout of p=0.2. The Bi-LSTM layer output is fed to a dense layer of dimension 256 and finally, the softmax layer is applied to predict the emotion label for the utterance. The softmax layer predicts probability scores for 4 different classes and we choose the class with maximum probability unit as the prediction. This model is trained using the cross-entropy objective for up to 20 epochs using Adam optimizer. We use a batch size of 32 along with learning rate 0.001. We use 4290 sentences for training and the remaining 1241 utterances are used for testing. The loss function is optimized using Adam optimizer with a learning rate of 0.001. The trained model works with 55.10% accuracy on the test data as shown in Table 1(Row - 1).

Table 1: *Uni-Modal emotion recognition models*

Model	Unweighted Accuracy
Acoustic feature only model	55.01%
Speech embedding only model	60.03%
Word embedding only model	60.68%

4.2. Speech embedding based emotion recognition

The speech embedding based emotion model is a single layer Bi-LSTM with 2 fully connected layers on top. The input to the model is a sequence of 50-dimensional speech embedding extracted for every word in the utterance. Each utterance will have N words and we generate embeddings for every word using Speech2Vec model as described in section 2.2 and we pad these sequences with zero vectors to make every text sequence length to be equal to 128. We create a trainable embedding layer and it is initialized with pretrained speech embeddings extracted from Speech2Vec model. Finally, we attach Bi-LSTM layer with 2 fully connected layers on top of this embedding layer and train the network to minimize the cross-entropy loss. The Bi-LSTM layer has 128 dimension cell and they are trained with a dropout of 0.2. The final hidden state of Bi-LSTM (of dimension 256) goes to 2 fully connected layers with hidden layer size of 256. Dropout of p=0.3 is applied after every fully connected layer and finally, softmax layer predicts the probability scores for emotion classes. The model is trained to optimize the categorical cross entropy with Adam optimizer having a learning rate of 0.001. This standalone network achieves an accuracy of 60.03% accuracy on the test data as shown in Table 2. We also compare our speech embedding only model with the word embedding based model. We use the same setup but instead of initializing the embedding layers with speech embeddings, we initialize it with word embeddings. We see that word embedding based model performs slightly better than speech embedding based model as shown in row 3 of Table 1.

4.3. Combining Speech embeddings and acoustic features for speech emotion recognition

This model is a multi-modal network takes 2 streams of information, one is acoustic information and the other is speech embeddings information. Speech embeddings are fed as input to an encoder ENC2 which takes a sequence of speech embedding computed for every word in the transcript and generates and representation which will have semantic and context information of the utterance. Acoustic features are computed for every 100ms which are of dimension 34 and are fed to ENC1. The acoustic encoder ENC1 consist Bi-LSTM with cell dimension of 128 and a fully connected layer on top having dimension 256 and speech embedding encoder ENC2 will have single Bi-LSTM with a hidden dimension of 128 and a fully connected layer on top having dimension 256 as shown in Figure 2. The input ENC1 is (200,34) where 200 is the number of timesteps and 34 is acoustic feature dimensions. Similarly, the input to encoder ENC2 is (128,50), where 50 is the dimension of speech embeddings. The output of the dense layers from ENC1 and ENC2 which are of dimension 256 are concatenated to get a fused feature which will have both acoustic information and context information. This fused feature is passed through a fully connected layer of dimension 256 followed by a softmax to predict the class score. This whole network is jointly optimized using categorical cross entropy objective function. The Bi-LSTM layers are regularized using recurrent dropout of 0.2 and dense layers are regularized using dropout of 0.3. The objective function is minimized using Adam optimizer with a learning rate of 0.001. The model achieves an accuracy of 68.49% on the test data for speech embedding of size 50. We compare our results with previously published models in Table 2.

Table 2: *Performance (%) of single systems and their fusion on IEMOCAP dataset*

System	Unweighted Accuracy
Bi-LSTM	55.03%
E-Vector	57.25%
MCNN + LSTM	64.33%
E-vector + MCNN + LSTM	65.90%
Our system	**68.49**%

Table 2 compares the results from a single system and fusion of multiple systems. The acoustic feature based LSTM can give only 55.03% unweighted accuracy as shown in row 1 of Table 2. The E-Vector[12] system which combines both lexical features and acoustic features gives us 57.25% (row 2) accuracy as reported in [8]. A multi-modal emotion recognition system which combines both speech and transcript called MCNN(multiscale CNN)[8] gives unweighted accuracy of 64.3% as in row 3 of Table 3. Our speech embedding based multi-modal emotion recognition model gives the best results among all the above models achieving 68.49% unweighted accuracy as shown in row 5 of Table 2.

We conduct experiments for different dimensionalities of word embeddings to compare the performance with speech embeddings. We use pretrained word embeddings of dimensions 50,100,200 and 300. We run our multi-modal emotion recognition system using word embeddings instead of speech embeddings. We tabulate our results in Table 3. We can clearly see that speech embeddings are better than word embeddings.

Table 3: *Performance (%) of our system for different word embedding and speech embedding size*

Embedding dimension	Speech embedding	Word embedding
50	**68.49**%	67.12%
100	**66.08**%	66.01 %
200	64.87%	**66.80**%
300	**65.75**%	61.32%

5. Conclusion

Speech emotion recognition is a very challenging problem and it is still one of the unsolved problems in the speech community. In this paper, we propose a multi-modal speech emotion recognition system which uses information from both audio and speech embedding to accurately classify the utterances into emotion classes. The proposed approach uses speech embeddings which captures semantic and context information in the utterance and we use it as an additional guiding signal during speech emotion classification. This proposed approach outperforms previously published multi-modal emotion recognition models by 2.59% absolute improvement. We compared our speech embedding based model with word embedding based model and we see that speech embeddings are better than word embeddings for speech emotion recognition problem because speech embeddings have the capability to capture speech variations like speaker and channel variabilities whereas just plain text cannot capture these speech variabilities. In future work, we would like to see how to make use of attention mechanism and other techniques which can capture emotion cues present in both acoustic and embeddings modalities.

6. Ackowledgements

I'd like to thank YouPlus, India for supporting and funding this project. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the YouPlus India. We would like to thank Sumanth and Shweta Bhatt for useful discussions.

7. References

[1] Chung, Y., Glass, J, "Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech," *Proc. Interspeech 2018,*

[2] Tomas Mikolov and Ilya Sutskever and Kai Chen and Gregory S. Corrado and Jeffrey Dean,"Distributed Representations of Words and Phrases and their Compositionality," in *NIPS 2014 ,*

[3] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 49604964. IEEE, 2016

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018,* April 2018, pp. 53295333.

[5] David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell,Daniel Povey, Sanjeev Khudanpur, "Spoken Language Recognition using X-vectors," *Odyssey 2018,*

[6] Y.L. LI, G. Wei., "Speech emotion recognition based on HMM and SVM ," *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics,* Vol.8, 18-21 Aug. 2005, pp.4898 4901

[7] Albanie, S. and Nagrani, A. and Vedaldi, A. and Zisserman, A., "Emotion Recognition in Speech using Cross-Modal Transfer in the Wild," *ACM Multimedia 2018*

[8] Cho, J., Pappagari, R., Kulkarni, P., Villalba, J., Carmiel, Y., Dehak, N, "Deep Neural Networks for Emotion Recognition Combining Audio and Transcripts" *Proc. Interspeech 2018.*

[9] Han, Kun Yu, Dong Tashev, Ivan, "Speech emotion recognition using deep neural network and extreme learning machine," *in INTERSPEECH-2014, 223-227.*

[10] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017),* New Orleans, LA, USA, March 2017, pp. 51155119.

[11] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, Speech emotion recognition from spectrograms with deep convolutional neural network,, in *Proc. of 2017 Intern. Conf. on Platform Technology and Service,* Busan, South Korea: IEEE, Feb 2017, pp. 15.

[12] Q. Jin, C. Li, S. Chen, and H. Wu,"Speech emotion recognition with acoustic and lexical features,", *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* IEEE, 2015, pp. 47494753

[13] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation,* vol. 42,no. 4, pp. 335, 2008

[14] Samarth Tripathi, Homayoon Beigi, " Multi-modal emotion recognition on IEMOCAP with neural networks," *arXiv:1804.05788*

[15] Diederik P. Kingma,Jimmy Lei Ba "Adam: A method for stochastic optimization,"in *ICLR 2015*