Implementation of voice-based emotion recognition using librosa,SVM,MLP

Priyanshu Singh[1], Prateek Singh[1], Anand Prakash Dwivedi[3]

Department of Computer Science and Engineering, [1] Maharana Pratap Engineering College, Kanpur, Uttar Pradesh, India

[1]singhpriyanshu@gmail.com

[2]prateek9839052395@gmail.com

[3]apdwivedi71@gmail.com

ABSTRACT

Speech emotion recognition is a challenging task, and extensive reliance has been placed on models that use audio features in building well-performing classifiers. In this paper, we propose a novel deep dual recurrent encoder model that utilizes text data and audio signals simultaneously to obtain a better understanding of speech data. As emotional dialogue is composed of sound and spoken content, our model encodes the information from audio and text sequences using dual recurrent neural networks (RNNs) and then combines the information from these sources to predict the emotion class. This architecture analyzes speech data from the signal level to the language level, and it thus utilizes the information within the data more comprehensively than models that focus on audio features. Extensive experiments are conducted to investigate the efficacy and properties of the proposed model. Our proposed model outperforms previous state-of-the-art methods in assigning data to one of four emotion categories (i.e., angry, happy, sad and neutral) when the model is applied to the IEMOCAP dataset, as reflected by accuracies ranging from 68.8% to 71.8%.

INTRODUCTION

Recently, deep learning algorithms have successfully addressed problems in various fields, such as image classification, machine translation, speech recognition, text-to-speech generation and other machine learning related areas [1, 2, 3]. Similarly, substantial improvements in performance have been obtained when deep learning algorithms have been applied to statistical speech processing [4]. These fundamental improvements have led researchers to investigate additional topics related to human nature, which have long been objects of study. One such topic involves understanding human emotions and reflecting it through machine intelligence, such as emotional dialogue models [5, 6].

In developing emotionally aware intelligence, the very first step is building robust emotion classifiers that display good performance regardless of the application; this outcome is considered to be one of the fundamental research goals in affective computing [7]. In particular, the speech emotion recognition task is one of the most important problems in the field of paralinguistics. This field has recently broadened its applications, as it is a crucial factor in optimal humancomputer interactions, including dialog systems. The goal of speech emotion recognition is to predict the emotional content of speech and to classify speech according to one of several labels (i.e., happy, sad, neutral, and angry). Various types of deep learning methods have been applied to increase the performance of emotion classifiers; however, this task is still considered to be challenging for several reasons. First, insufficient data for training complex neural network-based models are available, due to the costs associated with human involvement. Second, the characteristics of emotions must be learned from low-level speech signals. Feature-based models displaylimited skills when applied to this problem.

To overcome these limitations, we propose a model that uses high-level text transcription, as well as low-level audio signals, to utilize the information contained within low-resource datasets to a greater degree. Given recent improvements in automatic speech recognition (ASR) technology [8, 3, 9, 10], speech transcription can be carried out using audio signals with considerable skill. The emothe emotion words contained in a sentence [11].

RELATED WORK

Classical machine learning algorithms, such as hidden Markov models (HMMs), support vector machines (SVMs), and decision tree-based methods, have been employed in speech emotion recognition problems [12, 13, 14]. Recently, researchers have proposed various neural network-based architectures to improve the performance of speech emotion recognition. An initial study utilized deep neural networks (DNNs) to extract high-level features from raw audio data and demonstrated its effectiveness in speech emotion recognition [15]. With the advancement of deep learning methods, more complex neuralbased architectures have been proposed. Convolutional neural network (CNN)-based models have been trained on information derived from raw audio signals using spectrograms or audio features such as Mel-frequency cepstral coefficients (MFCCs) and low-level descriptors (LLDs) [16, 17, 18]. These neural network-based models are combined to produce higher-complexity models [19, 20], and these models achieved the best-recorded performance when applied to the IEMOCAP dataset.

Another line of research has focused on adopting variant machine learning techniques combined with neural networkbased models. One researcher utilized the multiobject learning approach and used gender and naturalness as auxiliary tasks so that the neural network-based model learned more features from a given dataset [21]. Another researcher investigated transfer learning methods, leveraging external data from related domains [22].

As emotional dialogue is composed of sound and spoken content, researchers have also investigated the combination of acoustic features and language information, built belief network-based methods of identifying emotional key phrases, and assessed the emotional salience of verbal cues from both phoneme sequences and words [23, 24]. However, none of these studies have utilized information from speech signals and text sequences simultaneously in an end-to-end learning neural network-based model to classify emotions.

MODELS

This section describes the methodologies that are applied to the speech emotion recognition task. We start by introducing the recurrent encoder model for the audio and text modalities individually. We then propose a multimodal approach that encodes both audio and textual information simultaneously via a dual recurrent encoder.

Audio Recurrent Encoder (ARE)

Motivated by the architecture used in [25, 26], we build an audio recurrent encoder (ARE) to predict the class of a given audio signal. Once MFCC features have been extracted from an audio signal, a subset of the sequential features is fed into the

RNN (i.e., gated recurrent units (GRUs)), which leads to the formation of the network's internal hidden state h_t to model the time series patterns. This internal hidden state is updated at each time step with the input data x_t and the hidden state of the previous time step h_{t-1} as follows:

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

where f_θ is the RNN function with weight parameter θ, h_t represents the hidden state at t-th time step, and x_t represents the t-th MFCC features in $x = \{x_{1:t_a}\}$. After encoding the audio signal x with the RNN, the last hidden state of the RNN, h_{t_a}, is considered to be the representative vector that contains all of the sequential audio data. This vector is then concatenated with another prosodic feature vector, p, to generate a more informative vector representation of the signal, $e = \text{concat}\{h_{t_a}, p\}$. The MFCC and the prosodic features are extracted from the audio signal using the openSMILE toolkit [27], $x_t \in \mathbb{R}^{39}$ and $p \in \mathbb{R}^{35}$, respectively. Finally, the emotion class is predicted by applying the softmax function to the vector e. For a given audio sample i, we assume that y_i is the true label vector, which contains all zeros but contains a one at the correct class, and $\hat{y}_i$ is the predicted probability distribution from the softmax layer. The training objective then takes the following form:

$$\hat{y}_i = \text{softmax}(e^{\intercal} M + b),$$

$$\mathcal{L} = -\log \prod_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}),$$

where e is the calculated representative vector of the audio signal with dimensionality $e \in \mathbb{R}^d$. The $M \in \mathbb{R}^{d \times C}$ and the bias b are learned model parameters. C is the total number of classes, and N is the total number of samples used in training. The upper part of Figure 1 shows the architecture of the ARE model.

Text Recurrent Encoder (TRE)

We assume that speech transcripts can be extracted from audio signals with high accuracy, given the advancement of ASR technologies [8]. We attempt to use the processed textual information as another modality in predicting the emotion class of a given signal. To use textual information, a speech transcript is tokenized and indexed into a sequence of tokens using the Natural Language Toolkit (NLTK) [28]. Each token is then passed through a word-embedding layer that converts a word index to a corresponding 300-dimensional vector that contains additional contextual meaning between words. The sequence of embedded tokens is fed into a text recurrent encoder (TRE) in such a way that the audio MFCC features are encoded using the ARE represented by equation 1. In this case, xt is the t-th embedded token from the text input. Finally, the emotion class is predicted from the last hidden state of the text-RNN using the softmax function.

We use the same training objective as the ARE model, and the predicted probability distribution for the target class is as follows:

$$\hat{y}_i = \mathrm{softmax}(h_{last}^{\intercal} M + b),$$

where hlast is last hidden state of the text-RNN, hlast ∈ Rd, and the M ∈Rd×C and bias b are learned model parameters. The lower part of Figure 1 indicates the architecture of the TRE model.

Multimodal Dual Recurrent Encoder (MDRE)

We present a novel architecture called the multimodal dual recurrent encoder (MDRE) to overcome the limitations of existing approaches. In this study, we consider multiple modalities, such as MFCC features, prosodic features and transcripts, which contain sequential audio information, statistical audio information and textual information, respectively. These types of data are the same as those used in the ARE and TRE cases. The MDRE model employs two RNNs to encode data from the audio signal and textual inputs independently. The audio-RNN encodes MFCC features from the audio signal using equation 1. The last hidden state of the audio-RNN is concatenated with the prosodic features to form the final vector representation e, and this vector is then passed through a fully connected neural network layer to form the audio encoding vector A. On the other hand, the text-RNN encodes the word sequence of the transcript using equation 1. The final hidden states of the text-RNN are also passed through another fully connected neural network layer to form a textual encoding vector T.

Finally, the emotion class is predicted by applying the softmax function to the concatenation of the vectors A and T. We use the same training objective as the ARE model, and the predicted probability distribution for the target class is as follows:

$$A = g_\theta(e), \quad T = g^0{}_\theta(h_{last}), \quad \hat{y}_i = \mathrm{softmax}(\mathrm{concat}(A,T)^{\intercal} M + b),$$

where gθ,g0θ is the feed-forward neural network with weight parameter θ, and A, T are final encoding vectors from the

audio-RNN and text-RNN, respectively. M ∈Rd×C and the bias b are learned model parameters.

Inspired by the concept of the attention mechanism used in neural machine translation [29], we propose a novel multimodal attention method to focus on the specific parts of a transcript that contain strong emotional information, conditioning on the audio information. Figure 2 shows the architecture of the MDREA model. First, the audio data and text data are encoded with the audio-RNN and text-RNN using equation 1. We then consider the final audio encoding vector e as a context vector. As seen in equation 5, during each time step t, the dot product between the context vector e and the hidden state of the text-RNN at each t-th sequence ht is evaluated to calculate a similarity score at. Using this score at as a weight parameter, the weighted sum of the sequences of the hidden state of the text-RNN, ht, is calculated to generate an attention-application vector Z. This attention-application vector is concatenated with the final encoding vector of the audioRNN A (equation 4), which will be passed through the softmax function to predict the emotion class. We use the same training objective as the ARE model, and the predicted probability distribution for the target class is as follows:

is as follows:

$$a_t = \frac{\exp(\mathbf{e}^{\intercal}\mathbf{h}_t)}{\sum_t \exp(\mathbf{e}^{\intercal}\mathbf{h}_t)}, \quad \mathbf{Z} = \sum_t a_t \mathbf{h}_t, \quad \hat{y}_{i,j} = \mathrm{softmax}(\mathrm{concat}(\mathbf{Z}, \mathbf{A})^{\intercal} M + b)$$

where $M \in R^{d\times C}$ and the bias b are learned model parameters.

EXPERIMENTAL SETUP AND DATASET

Dataset

In this study, two standard emotion data sets (RAVDESS and TESS) and self-recorded custom data (custom-DB) were used which cover a wide range of acted and spontaneous emotions [3]. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 24 professional actors (12 females, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. The Toronto Emotional Speech Set (TESS) was modelled on Northwestern University Auditory Test No. 6. A set of 200 target words were spoken by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 stimuli in total. Both actresses are native English speakers and have thresholds within the normal range. The custom-DB dataset is an unbalanced noisy dataset created by self-recording samples and converting the raw file to 16 kHz and tagging the emotion to the filename. In order to compare ASER results across data sets a sampling rate of 16 kHz was chosen.

Audio Coding Format (Codec)

MPEG-2 Audio Layer III (MP3) is a lossy audio codec, introduced by Fraunhofer Institute in the year 1993 [4]. The audio compression is acquired by perceptual coding: certain regions of the original sound signal, considered beyond the auditory resolution ability, are discarded [6]. Next, the remaining information is stored in an effective manner using Huffman-coding. The bit rates range from 8-320 kbit/s. In the current study, the bit data was encoded at the following bitrates: 256, 192, 128, 96, 64, 24, 16, 8 kbit/s. MP3 is a popular codec and is freely available, allowing for easy reproduction of results presented.

Speech emotion recogntion algorithms

Feature extraction

Previous studies have used the emobase feature set [3] from the openSMILE toolkit. Feature extraction is one of the most key aspects of ASER systems. It converts the speech waveform to parametric representation at a very low data rate. In this study, we have used the features available in the LibROSA library such as mel-frequency cepstral coefficient (MFCC), mel spectrogram frequency (mel), chromagram [13]. LibROSA is a python open-source library for music and audio file analysis. It helps to extract the necessary audio features for developing ASER systems. In audio file processing, the melfrequency cepstrum (MFC) is defined as the short-term power spectrum of a sound, based in a linear cosine transform of a log power spectrum on a non-linear mel scale of frequency. MFCC are coefficients that collective build an MFC. MFCC is the most widely used feature for emotion recognition from audio speech data sets.

Classification algorithm

In this study, we used support-vector machine (SVM), multilayer perceptron (MLP) and Long shortterm memory (LSTM) as our classification algorithms for building our ASER models. We normalized our data using a mean and standard deviation normalization which was computed on the training set and applied to test set. In machine learning hyper parameter optimization is an act of choosing a set of optimal parameters for a learning algorithm. A hyper parameter is a parameter whose value is used to control the learning process. The traditional way of performing hyper parameter optimization has been grid search. In grid search we do an exhaustive searching through a manually specified subset of hyper parameter space of a learning algorithm. A grid search was performed on each of the specified models using cross validation on the training set as a performance matrix. The best parameters of each of the classification algorithms were used for the analysis. MLP is class of feed-forward artificial neural network. The following parameters were set to build the MLP classifier: activation function - ReLU, learning rate (0.001), learning rate (adaptive), batch size (1024), number of hidden layers – 300, validation fraction (0.1) and maximum epoch (50). We chose SVM with linear kernel and gamma value of 0.001. LSTM is an artificial recurrent neural network architecture for various classification applications. The first part of the model is a feature extractor. The output is normalized and fed to two uni-directional LSTM layers [3], with a batch size of 64. Other parameters of the end-to-end LSTM based ASER model are as follows: dropout coefficient of 0.5 to prevent over-fitting, learning rate 0.001, epoch 1000.

EMPIRICAL RESULTS

Performance evaluation

As the dataset is not explicitly split beforehand into training, development, and testing sets, we perform 5-fold cross validation to determine the overall performance of the model. The data in each fold are split into training, development, and testing datasets (8:0.5:1.5, respectively). After training the model, we measure the weighted average precision (WAP) over the 5-fold dataset. We train and evaluate the model 10 times per fold, and the model performance is assessed in terms of the mean score and standard deviation.

We examine the WAP values, which are shown in Table 1. First, our ARE model shows the baseline performance because we use minimal audio features, such as the MFCC and prosodic features with simple architectures. On the other hand, the TRE model shows higher performance gain compared to the ARE. From this result, we note that textual data are informative in emotion prediction tasks, and the recurrent encoder model is effective in understanding these types of sequential data. Second, the newly proposed model, MDRE, shows a substantial performance gain. It thus achieves the state-of-the-art performance with a WAP value of 0.718. This result shows that multimodal information is a key factor in affective computing. Lastly, the attention model, MDREA, also outperforms the best existing research results (WAP 0.690 to 0.688) [20]. However, the MDREA model does not match the performance of the MDRE model, even though it utilizes a more complex architecture. We believe that this result arises because insufficient data are available to properly determine the complex model parameters in the MDREA model. Moreover, we presume that this model will show better performance when the audio signals are aligned with the textual sequence while applying the attention mechanism. We leave the implementation of this point as a future research direction.

To investigate the practical performance of the proposed models, we conduct further experiments with the ASRprocessed transcript data (see “-ASR” models in Table 1). The label accuracy of the processed transcripts is 5.53% WER. The TRE-ASR, MDRE-ASR and

MDREA-ASR models reflect degraded performance compared to that of the TRE, MDRE and MDREA models. However, the performance of these models is still competitive; in particular, the MDREASR model outperforms the previous best-performing model, 3CNN-LSTM10H (WAP 0.691 to 0.688).

Error analysis

We analyze the predictions of the ARE, TRE, and MDRE models. The confusion matrix of each model. The ARE model incorrectly classifies most instances of happy as neutral (43.51%); thus, it shows reduced accuracy (35.15%) in predicting the the happy class. Overall, most of the emotion classes are frequently confused with the neutral class. This observation is in line with the findings of , who noted that the neutral class is located in the center of the activation-valence space, complicating its discrimination from the other classes prediction gains in predicting the happy class when compared to the ARE model (35.15% to 75.73%). This result seems plausible because the model can benefit from the differences among the distributions of words in happy and neutral expressions, which gives more emotional information to the model than that of the audio signal data. On the

CONCLUSION

In this paper, we propose a novel multimodal dual recurrent encoder model that simultaneously utilizes text data, as well as audio signals, to permit the better understanding of speech data. Our model encodes the information from audio and text sequences using dual RNNs and then combines the information from these sources using a feed-forward neural model to predict the emotion class. Extensive experiments show that our proposed model outperforms other state-of-the-art methods in classifying the four emotion categories, and accuracies ranging from 68.8% to 71.8% are obtained when the model is applied to the IEMOCAP dataset. In particular, it resolves the issue in which predictions frequently incorrectly yield the neutral class, as occurs in previous models that focus on audio features.

opposites of one another. The MDRE model compensates for the weaknesses of the previous two models (ARE and TRE) and benefits from their strengths to a surprising degree. The values arranged along the diagonal axis show that all of the accuracies of the correctly predicted class have increased. Furthermore, the occurrence of the incorrect “sad-to-happy” cases in the TRE model is reduced from 16.20% to 9.15%.

In the future work, we aim to extend the modalities to audio, text and video inputs. Furthermore, we plan to investigate the application of the attention mechanism to data derived from multiple modalities. This approach seems likely to uncover enhanced learning schemes that will increase performance in both speech emotion recognition and other multimodal classification tasks.

REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014.

[3] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang 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.

[4] Alex Graves, Santiago Fernandez, Faustino Gomez, and Jurgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in Proceedings of the 23rd international conference on Machine learning. ACM, 2006,pp. 369–376.

[5] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu, “Emotional chatting machine: Emotional conversation generation with internal and external memory,” 2018.

[6] Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri, “Automatic dialogue generation with expressed emotions,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, vol. 2, pp. 49–54.

[7] Carlos Busso, Murtaza Bulut, and Shrikanth Narayanan, “Toward effective automatic recognition systems of emotion in speech,” Social Emotions in Nature and Artifact, p. 110, 2014.

[8] Dong Yu and Li Deng, AUTOMATIC SPEECH RECOGNITION., Springer, 2016.

[9]Google, “Cloud speech-to-text,” http://cloud.google.com/speech-to-text/, 2018.

[10] Microsoft, “Microsoft speech api,” http://docs.microsoft.com/enus/azure/cognitiveservices/speech/home, 2018.

[11] Linhong Xu, Hongfei Lin, Yu Pan, Hui Ren, and Jianmei Chen, “Constructing the affective lexicon ontology,” Journal of the China Society for Scientific and Technical Information, vol. 27, no. 2, pp. 180–185, 2008.

[12] Thapanee Seehapoch and Sartra Wongthanavasu, “Speech emotion recognition using support vector machines,” in Knowledge and Smart Technology (KST), 2013 5th International Conference on. IEEE, 2013, pp. 86–91.

[13] Bjorn Schuller, Gerhard Rigoll, and Manfred Lang, ¨ “Hidden markov model-based speech emotion recognition,” in Multimedia and Expo, 2003. ICME’03. Proceedings. 2003 International Conference on. IEEE, 2003, vol. 1, pp. I–401.

[14] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” Speech Communication, vol. 53, no. 9-10, pp. 1162– 1171, 2011.

[15] Kun Han, Dong Yu, and Ivan Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[16] Dario Bertero and Pascale Fung, “A first look into a convolutional neural network for speech emotion detection,” in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on.

IEEE, 2017, pp. 5115–5119.

[17] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik, “Speech emotion recognition from spectrograms with deep convolutional neural network,” in Platform Technology and Service (PlatCon), 2017 International Conference on. IEEE, 2017, pp. 1–5.

[18] Zakaria Aldeneh and Emily Mower Provost, “Using regional saliency for speech emotion recognition,” in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 2741–2745.

[19] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzade , 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.

[20] Aharon Satt, Shai Rozenberg, and Ron Hoory, “Efficient emotion recognition from speech using deep learning on spectrograms,” Proc. Interspeech 2017, pp. 1089–1093, 2017.

[21] Jaebok Kim, Gwenn Englebienne, Khiet P Truong, and Vanessa Evers, “Towards speech emotion recognition” in the wild” using aggregated corpora and deep multi-task learning,” in 18th Annual Conference of the International Speech Communication Association, INTERSPEECH 2017: Situated interaction. International Speech Communication Association (ISCA), 2017.

[22] John Gideon, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost, “Progressive neural networks for transfer learning in emotion recognition,” Proc. Interspeech 2017, pp. 1098–1102, 2017.

[23] Bjorn Schuller, Gerhard Rigoll, and Manfred Lang, ¨

“Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture,” in Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP’04). IEEE International Conference on. IEEE, 2004, vol. 1, pp. I–577.

[24] Kalani Wataraka Gamage, Vidhyasaharan Sethu, and

Eliathamby Ambikairajah, “Salience based lexical features for emotion recognition,” in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 5830–5834.

[25] Yun Wang, Leonardo Neves, and Florian Metze, “Audio-based multimedia event detection using deep recurrent neural networks,” in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016, pp. 2742–2746.

[26] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 2227–2231.

[27] Florian Eyben, Felix Weninger, Florian Gross, and Bjorn Schuller, “Recent developments in opensmile, the ¨ munich open-source multimedia feature extractor,” in Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013, pp. 835–838.

[28] Steven Bird and Edward Loper, “Nltk: the natural language toolkit,” in Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004, p. 31.

[29] Thang Luong, Hieu Pham, and Christopher D Manning, “Effective approaches to attention-based neural machine translation,” in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1412–1421.

[30] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” arXiv preprint arXiv:1412.3555, 2014.

[31] Michael Neumann and Ngoc Thang Vu, “Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,” Proc. Interspeech 2017, pp. 1263–1267, 2017.

[32] Andrew M Saxe, James L McClelland, and Surya Ganguli, “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks,” arXiv preprint arXiv:1312.6120, 2013.

[33] Jeffrey Pennington, Richard Socher, and Christopher Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[34] Jinkyu Lee and Ivan Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” in Sixteenth Annual Conference of the International Speech Communication Association, 2015.