SPEECH EMOTION RECOGNITION USING MULTI-HOP ATTENTION MECHANISM

Seunghyun Yoon[1], Seokhyun Byun[1], Subhadeep Dey[2] and Kyomin Jung[1]

[1]Department of Electrical and Computer Engineering,
Seoul National University, Seoul, Korea
[2]Idiap Research Institute, Martigny, Switzerland
mysmilesh@snu.ac.kr, byuns9334@snu.ac.kr, subhadeep.dey@idiap.ch, kjung@snu.ac.kr

ABSTRACT

In this paper, we are interested in exploiting textual and acoustic data of an utterance for the speech emotion classification task. The baseline approach models the information from audio and text independently using two deep neural networks (DNNs). The outputs from both the DNNs are then fused for classification. As opposed to using knowledge from both the modalities separately, we propose a framework to exploit acoustic information in tandem with lexical data. The proposed framework uses two bi-directional long short-term memory (BLSTM) for obtaining hidden representations of the utterance. Furthermore, we propose an attention mechanism, referred to as the multi-hop, which is trained to automatically infer the correlation between the modalities. The multi-hop attention first computes the relevant segments of the textual data corresponding to the audio signal. The relevant textual data is then applied to attend parts of the audio signal. To evaluate the performance of the proposed system, experiments are performed in the IEMOCAP dataset. Experimental results show that the proposed technique outperforms the state-of-the-art system by 6.5% relative improvement in terms of weighted accuracy.

Index Terms— speech emotion recognition, computational paralinguistics, deep learning, natural language processing

1. INTRODUCTION

In this era of high-performance computing, human-computer interaction (HCI) has become pervasive. To enrich the user experience, the system is often required to detect human emotion and produce a response with proper emotional context [1, 2]. The first step in such an HCI involves building a system that recognizes emotion from the speech utterance. A speech emotional system aims to identify audio recording as belonging to one of the categories, like happy, sad, angry or neutral. Beside HCI, the output of emotion recognition engine is beneficial in the paralinguistic area as well [3]. In this paper, we build a speech emotion recognition system that uses acoustic and textual information in tandem.

Various approaches to address emotion recognition have been investigated in the literature. Most of the techniques involve extracting low-level or high-level acoustic features for this task [4]. In emotion recognition, the lexical content of the audio recording is an important source of information that is usually ignored. For example, the presence of words such as "gorgeous" and "stunning" in the utterance would indicate that the person is happy. Recently researchers have also explored the application of textual content of the speech signal for this task. In [5], frame and supra-segmental level features (such as pitch and spectral contours) are derived from the speech signal. Textual information is used by spotting keywords that emphases

the emotional states of the speaker. The work in [6] also presents an approach to exploit the acoustic and lexical content. In particular, they explored conventional acoustic features from the speech signal while the textual information is derived from the bag of word representation.

Recently, deep neural network (DNN) has shown to provide good results for modeling acoustic and textual information for emotion identification. In [5], textual and acoustic information of the utterance are used by a DNN to obtain hidden feature representations for both the modality. These features are then concatenated to represent the utterance and subsequently used to classify the emotion of the speaker. Experimental evidence shows the potential of the approach. In our previous work [7], we applied a dual RNN in order to obtain a richer representation by blending the content and acoustic knowledge.

In this paper, we improve upon our earlier work by incorporating an attention mechanism in the emotion recognition framework. The proposed attention mechanism is trained to exploit both textual and acoustic information in tandem. We refer to this attention method as the multi-hop. The multi-hop attention is designed to select relevant parts of the textual data, which is subsequently applied to attend to the segments of the audio signal for classification. We hypothesize that this approach would automatically detect the segments that contain information relevant for the task. The emotion recognition experiments are performed on the standard IEMOCAP dataset [8]. Experimental results indicate that the proposed approach outperforms the state-of-the-art system published in the literature on this database by 6.5% relative improvement in terms of weighted accuracy.

This paper is organized as follows. Section 2 provides a brief literature review on speech emotion recognition. In Section 3, we start by describing the baseline bidirectional recurrent encoder model considered in this paper, then introducing the proposed technique in detail. Experimental setup for evaluating the system and discussion of the achieved results by various systems are presented in Sections 4. Finally, the paper is concluded in Section 5.

2. RELATED WORK

Along with classical algorithms based models such as support vector machine (SVM), hidden markov model (HMM) and decision tree [9, 10, 11], various neural network architectures have been recently introduced for the speech emotion recognition task. For example, convolutional neural network (CNN)-based models were trained on spectrograms or audio features such as mel-frequency cepstral coefficients (MFCCs) and low-level descriptors (LLDs) [12, 13, 14]. More complex models such as [15] were designed to better learn nonlinear decision boundaries of emotional speech and achieved the

best-recorded performance in audio modality models on IEMOCAP dataset [8]. Several neural network models with attention mechanism have been proposed to efficiently focus on a prominent part of speech and learn temporal dependency within whole utterance [16, 17].

Multi-modal approaches using acoustic features and textual information have been investigated. [5] identified emotional key phrases and salience of verbal cues from both phoneme sequences and words. Recently, [7, 18] combined acoustic information and conversation transcripts using a neural network-based model to improve emotion classification accuracy. However, none of these studies utilized attention method over audio and text modality in tandem for contextual understanding of the emotion in audio recording.

3. MODEL

This section describes the methodologies that are applied to the speech emotion recognition task. We start by introducing a baseline model, the bidirectional recurrent encoder, for encoding the audio and text modalities individually. We then propose an approach to exploit both audio and text data in tandem. In this technique, multi-hop attention is proposed to obtain relevant parts of audio and text data automatically.

3.1. Bidirectional Recurrent Encoder

Motivated by the architecture used in [7, 17, 19], we train a recurrent encoder to predict the categorical class of a given audio signal. To model the sequential nature of the speech signal, we use a bidirectional recurrent encoder (**BRE**) as shown in the Figure 1(a). We also added a residual connection to the model for promoting convergence during training [20]. A sequence of feature vectors is fed as input to the **BRE**, which leads to the formation of hidden states of the model as given by the following equation:

$$
\begin{aligned}
\overrightarrow{\mathbf{h}}_t &= f_\theta(\overrightarrow{\mathbf{h}}_{t-1}, \overrightarrow{\mathbf{x}}_t) + \overrightarrow{\mathbf{x}_t}, \\
\overleftarrow{\mathbf{h}}_t &= f'_\theta(\overleftarrow{\mathbf{h}}_{t+1}, \overleftarrow{\mathbf{x}}_t) + \overleftarrow{\mathbf{x}}_t, \\
\mathbf{o}_t &= [\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t], \\
\mathbf{o}_t^A &= [\mathbf{o}_t; \mathbf{p}],
\end{aligned}
\tag{1}
$$

where f_θ, f'_θ are the forward and backward long short-term memory (LSTM) with weight parameter θ, $\mathbf{h}_t$ represents the hidden state at t-th time step, and $\mathbf{x}_t$ represents the t-th MFCC features in audio signal. The hidden representations ($\overrightarrow{\mathbf{h}}_t$, $\overleftarrow{\mathbf{h}}_t$) from forward/backward LSTMs are concatenated for produce the feature, $\mathbf{o}_t$. To follow previous research [7], we also add another prosodic feature vector, $\mathbf{p}$, with each $\mathbf{o}_t$ to generate a more informative vector representation of the signal, $\mathbf{o}_t^A$. Finally, an emotion class is predicted from the acoustic signal by applying a softmax function to the final hidden representation at the last time step, $\mathbf{o}_{\text{last}}^A$. We refer this model as **audio-BRE** with the objective function as follows:

$$
\begin{aligned}
\hat{y}_c &= \text{softmax}(\, (\mathbf{o}_{\text{last}}^A)^\mathsf{T}\, \mathbf{W} + \mathbf{b} \,), \\
\mathcal{L} &= -\log \prod_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}),
\end{aligned}
\tag{2}
$$

where $y_{i,c}$ is the true label vector, and $\hat{y}_{i,c}$ is the predicted probability distribution from the softmax layer. The $\mathbf{W}$ and the bias $\mathbf{b}$ are learned model parameters. C is the total number of classes, and N is the total number of samples used in training.

Next, we attempt to use the processed textual information as another modality in predicting the emotion class of a given signal. To obtain textual hidden representation, $\mathbf{o}_t^T$, we tokenize the transcript and feed it into the **BRE** in such a way that the acoustic signals are encoded by equation (1). We refer this model as **text-BRE**. The training objective for the **text-BRE** is same as the **audio-BRE** in equation (2).

3.2. Proposed Multi-Hop Attention

We propose a novel multi-hop attention method to predict the importance of audio and text, referred to multi-hop attention (**MHA**). Figure 1 shows the architecture of the proposed **MHA** model. Previous research used multi-modal information independently using neural network model by concatenating features from each modality [7, 21]. As opposed to this approach, we propose a neural network architecture that exploits information in each modality by extracting relevant segments of the speech data using information from the lexical content (and vice-versa).

First, the acoustic and textual data are encoded with the **audio-BRE** and **text-BRE**, respectively, using equation (1). We then consider the final hidden representation of **audio-BRE**, $\mathbf{o}_{\text{last}}^A$, as a context vector and apply attention method to the textual sequence, $\mathbf{o}_t^T$. As this model is developed with a single attention method, we refer to the model as **MHA-1**. The final hidden representation of the **MHA-1** model, $\mathbf{H}$, is calculated as follows:

$$
\begin{aligned}
a_i &= \frac{\exp(\, (\mathbf{o}_{\text{last}}^A)^\mathsf{T}\, \mathbf{o}_i^T \,)}{\sum_i \exp(\, (\mathbf{o}_{\text{last}}^A)^\mathsf{T}\, \mathbf{o}_i^T \,)}, \ (i = 1, ..., t) \\
\mathbf{H}^1 &= \sum_i a_i\, \mathbf{o}_i^T, \ \ \mathbf{H} = [\mathbf{H}^1; \mathbf{o}_{\text{last}}^A].
\end{aligned}
\tag{3}
$$

The $\mathbf{H}^1$ (equation 3) is a new hidden representation for textual information with consideration of audio modality. With this information, we apply 2nd-hop attention, referred to **MHA-2**, to the audio sequence. The final hidden representation of the **MHA-2** model, $\mathbf{H}$, is calculated as follows:

$$
\begin{aligned}
a_i &= \frac{\exp(\, (\mathbf{H}_1)^\mathsf{T}\, \mathbf{o}_i^A \,)}{\sum_i \exp(\, (\mathbf{H}_1)^\mathsf{T}\, \mathbf{o}_i^A \,)}, \ (i = 1, ..., t) \\
\mathbf{H}^2 &= \sum_i a_i\, \mathbf{o}_i^A, \ \ \mathbf{H} = [\mathbf{H}^1; \mathbf{H}^2],
\end{aligned}
\tag{4}
$$

where $\mathbf{H}^2$ is a new hidden representation for audio information with the consideration of textual modality obtained from the **MHA-1**.

Similarly to **MHA-1**, we further apply 3rd-hop attention to textual sequence, referred to **MHA-3**, with the new audio hidden representation $\mathbf{H}^2$ (equation 4). The final hidden representation of the **MHA-3** model, $\mathbf{H}$, is calculated as follows:

$$
\begin{aligned}
a_i &= \frac{\exp(\, (\mathbf{H}_2)^\mathsf{T}\, \mathbf{o}_i^T \,)}{\sum_i \exp(\, (\mathbf{H}_2)^\mathsf{T}\, \mathbf{o}_i^T \,)}, \ (i = 1, ..., t) \\
\mathbf{H}^3 &= \sum_i a_i\, \mathbf{o}_i^T, \ \ \mathbf{H} = [\mathbf{H}^3; \mathbf{H}^2],
\end{aligned}
\tag{5}
$$

where $\mathbf{H}^3$ is updated representative vector of the textual information with the consideration of audio modality one more time.

In each case, the final hidden representation, $\mathbf{H}$, is passed through the softmax function to predict the four-categories emotion class. We use the same training objective as the **BRE** model with equation (2), and the predicted probability distribution for the target class, $\hat{y}_c$ is as follows:

$$
\hat{y}_c = \text{softmax}((\mathbf{H})^\mathsf{T}\, \mathbf{W} + \mathbf{b} \,),
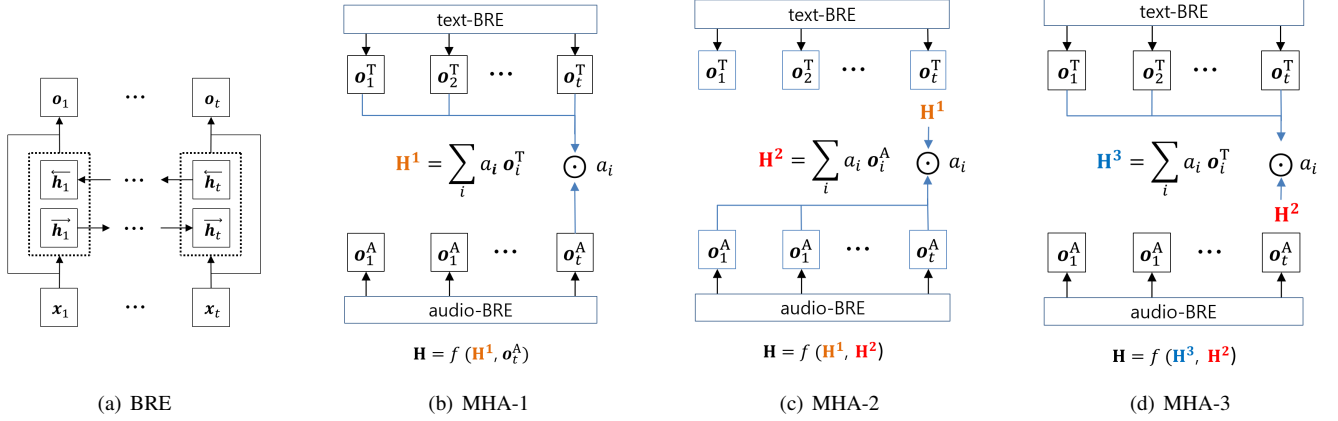\tag{6}
$$

Fig. 1. Architecture of the bidirectional recurrent encoder (BRE), and multi-hop attention model (MHA).

where projection matrix $\mathbf{W}$ and bias $\mathbf{b}$ are leaned model parameters.

4. EXPERIMENTS

4.1. Dataset and Experimental Setup

To train and evaluate our model, we use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [8] dataset, which includes five sessions of utterances between two speakers (one male and one female). Total 10 unique speakers participated in this work. For consistent comparison with previous works [7, 18], all utterances labeled "excitement" are merged with those labeled "happiness". We assign single categorical emotion to the utterance with majority of annotators agreed on the emotion labels. The final dataset contains 5,531 utterances in total (1,636 *happy*, 1,084 *sad*, 1,103 *angry* and 1,708 *neutral*). In the training process, we perform 10-fold cross-validation where each 8, 1, 1 folds are used for the train set, development set, and test set, respectively.

4.2. Feature extraction and Implementation details

As this research is extended work from previous research [7], we use the same feature extraction method as done in our previous work. After extracting 40-dimensional Mel-frequency cepstral coefficients (MFCC) feature (frame size is set to 25 ms at a rate of 10 ms with the Hamming window) using Kaldi [22], we concatenate it with its first, second order derivates, making the feature dimension to 120. We also extract prosodic features by using OpenSMILE toolkit [23] and appending it to the audio feature vector.

In preparing the textual dataset, we first use the ground-truth transcripts of the IEMOCAP dataset. In a practical scenario where we may not access to transcripts of the audio, we obtain all of the transcripts from the speech signal using a commercial ASR system [24] (The performance of the ASR system is word error rate (WER) of 5.53%). We apply word-tokenizer to the transcripts and obtain sequential data for textual input.

The maximum length of an audio segment is set to 750 based on the implementation choices presented in [25] and 128 for the textual input which covers the maximum length of the tokenized transcripts. We minimize the cross-entropy loss function using (equation (2)) the Adam optimizer [26] with a learning rate of 1e-3 and gradients clipped with a norm value of 1. For the purposes of regularization, we apply the dropout method, 30%. The number of hidden units and the number of layers in the RNN for each model (**BRE** and **MHA**) are optimized on the development set.[1]

4.3. Performance evaluation

To measure the performance of systems, we report the weighted accuracy (WA) and unweighted accuracy (UA) averaging over the 10-fold cross-validation experiments. We use the same dataset and features as other researchers [7, 18].

Table 1 presents performances of proposed approaches for recognizing speech emotion in comparison with various models. To compare our results from previous approaches, we first use ground-truth transcripts included in the dataset in training textual modality. From the previous model, **E_vec-MCNN-LSTM** encodes acoustic signal and textual information using a neural network (RNN and CNN, respectively) and then fuse each result by concatenating and feeding them into following (SVM) to predict emotion labels. On the other hand, **MDRE** model use dual-RNNs to encode both the modalities and merge the results using another fully-connect neural network layer. This **MDRE** approach applies end-to-end learning and outperforms **E_vec-MCNN-LSTM** by 10.6% relative (0.649 to 0.718 absolute) in terms of WA.

Among our proposed system, the **audio-BRE** model that uses an acoustic signal with bidirectional-RNN architecture achieves WA 0.646. Interestingly, the **text-BRE** model that use textual information shows higher performance than that of **audio-BRE** by 8% relative (0.646 to 0.698) in WA. The multi-hop attention model, **MHA-N**, ($N = 1, 2, 3$), shows a substantial performance gain. In particular, the **MHA-2** model (best performing system among MHA-N) outperformed the best baseline model, **MDRE**, by 6.5% relative (0.718 to 0.765) in WA. Although we observe performance degradation in the **MHA-3** model, we believe that this could be due to the limited data for training.

In a practical scenario, we may not access the audio transcripts. We describe the effect of using ASR-processed transcripts on the proposed system. From table 1, we observe performance degradation in **text-BRE-ASR** and **MHA-2-ASR** (our best system), compared to that of **text-BRE** and **MHA-2** by 6.6% (0.698 to 0.652) and 4.6% (0.765 to 0.730) relative in WA, receptively. Even with the erroneous transcripts (WER = 5.53%), however, the proposed approach

---

[1]The code will be made available via a public web repository.

Model	Modality	WA	UA
Ground-truth transcript			
E_vec-MCNN-LSTM [18]	A+T	0.649	0.659
MDRE [7]	A+T	0.718	-
audio-BRE (ours)	A	0.646	0.652
text-BRE (ours)	T	0.698	0.703
MHA-1 (ours)	A+T	**0.756**	**0.765**
MHA-2 (ours)	A+T	**0.765**	**0.776**
MHA-3 (ours)	A+T	0.740	0.753
ASR-processed transcript			
text-BRE-ASR (ours)	T	0.652	0.658
MHA-2-ASR (ours)	A+T	0.730	0.739

Table 1. Model performance comparisons. The top 2 best-performing models (according to the WA) are marked in bold. The "-ASR" models use ASR-processed transcripts from the Google Cloud Speech API. The "A" and "T" in modality indicate "Audio" and "Text", receptively.

(**MHA-2-ASR**) outperforms the best baseline system (**MDRE**) by 1.6% relative (0.718 to 0.730) in terms of WA.

4.4. Error analysis

Figure 2 shows the confusion matrices of the proposed systems. In **audio-BRE** (Fig. 2(a)), most of the emotion labels are frequently misclassified as *neutral* class, supporting the claims of [7, 25]. The **text-BRE** shows improvement in classifying most of the labels in Fig. 2(b). In particular, *angry* and *happy* classes are correctly classified by 32% (57.14 to 75.41) and 63% (40.21 to 65.56) relative in accuracy with respect to **audio-BRE**, receptively. However, it incorrectly predicted instances of the *happy* class as *sad* class in 10% of the time, even though these emotional states are opposites of one another.

The **MHA-2** (our best system, Fig. 2(c)) compensates for the weaknesses of the single modality models and benefits from their strengths. It shows significant performance gain for *angry*, *happy*, *sad* and *neutral* classes by 6%, 20%, 15% and 13% relative in accuracy with respect to **text-BRE**. It also correctly classify *neutral* class similar to that of **audio-BRE** (81.63 and 78.00 for **audio-BRE** and **MHA-2**, receptively). Interestingly, although **MHA-2** shows supe-rior discriminating ability among emotion classes, it still shows the tendency such that most of the incorrect cases are misclassified into *neutral* class. We consider this observation as a future research direction.

5. CONCLUSIONS

In this paper, we propose a multi-hop attention model to combine acoustic and textual data for speech emotion recognition task. The proposed attention method is designed to select relevant parts of the textual data, which is subsequently applied to attend to the segments of the audio signal for classification. Extensive experiments show that the proposed **MHA-2** outperforms the best baseline system in classifying the four emotion categories by 6.5% (0.718 to 0.765 absolute) in terms of WA when the model is applied to the IEMOCAP dataset. We further test our model with ASR-processed transcripts and achieve WA 0.73 that shows the reliability of the proposed system (**MHA-2-ASR**) in the practical scenario where the ground-truth transcripts are not available.

Acknowledgments

6. REFERENCES

[1] Rosalind W Picard, "Affective computing: challenges," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003.

[2] Carlos Busso, Murtaza Bulut, Shrikanth Narayanan, J Gratch, and S Marsella, "Toward effective automatic recognition systems of emotion in speech," *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds*, pp. 110–127, 2013.

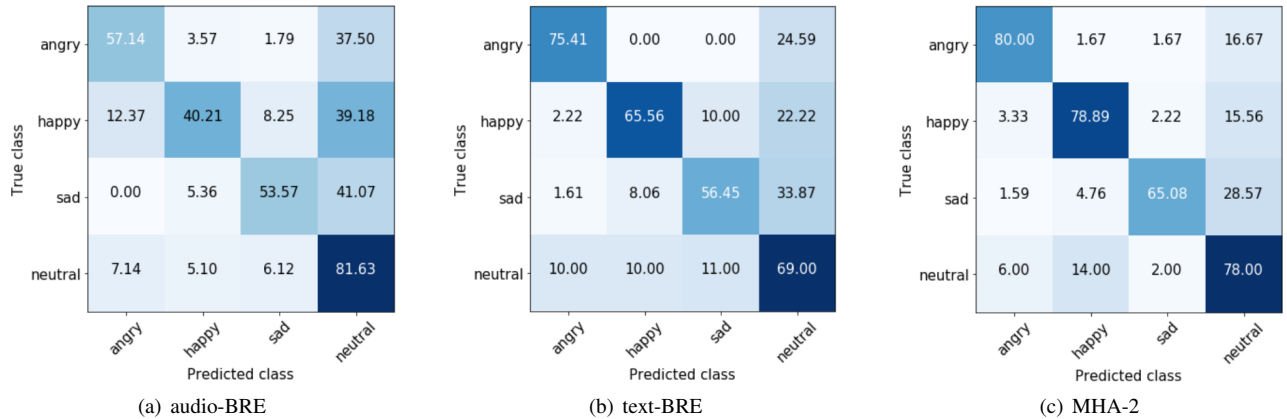(a) audio-BRE   (b) text-BRE   (c) MHA-2

Fig. 2. Confusion matrix of each model.

[3] Agata Kołakowska, Agnieszka Landowska, Mariusz Szwoch, Wioleta Szwoch, and Michal R Wrobel, "Emotion recognition and its applications," in *Human-Computer Systems Interaction: Backgrounds and Applications 3*, pp. 51–62. Springer, 2014.

[4] 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.

[5] Björn 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 (ICASSP), IEEE International Conference on*. IEEE, 2004, pp. I–577.

[6] Qin Jin, Chengxin Li, Shizhe Chen, and Huimin Wu, "Speech emotion recognition with acoustic and lexical features," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4749–4753.

[7] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, "Multi-modal speech emotion recognition using audio and text," *arXiv preprint arXiv:1810.04635*, 2018.

[8] 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.

[9] 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.

[10] Björn 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] Pengcheng Li, Yan Song, Ian McLoughlin, Wu Guo, and Lirong Dai, "An attention pooling based representation learning method for speech emotion recognition," *Proc. Interspeech 2018*, pp. 3087–3091, 2018.

[17] 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.

[18] Jaejin Cho, Raghavendra Pappagari, Purva Kulkarni, Jesús Villalba, Yishay Carmiel, and Najim Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," *Proc. Interspeech 2018*, pp. 247–251, 2018.

[19] 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.

[20] Yiren Wang and Fei Tian, "Recurrent residual learning for sequence classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 938–943.

[21] Samarth Tripathi and Homayoon Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," *arXiv preprint arXiv:1804.05788*, 2018.

[22] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," Tech. Rep., IEEE Signal Processing Society, 2011.

[23] Florian Eyben, Felix Weninger, Florian Gross, and Björn 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.

[24] Google, "Cloud speech-to-text," http://cloud.google.com/speech-to-text/, 2018.

[25] 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.

[26] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.