INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Speech Emotion Recognition in Dyadic Dialogues with Attentive Interaction Modeling

*Jinming Zhao, Shizhe Chen, Jingjun Liang, Qin Jin**

School of Information, Renmin University of China
{zhaojinming, cszhe1, liangjingjun, qjin}@ruc.edu.cn

Abstract

In dyadic human-human interactions, a more complex interaction scenario, a person's emotional state can be influenced by both self emotional evolution and the interlocutor's behaviors. However, previous speech emotion recognition studies infer the speaker's emotional state mainly based on the targeted speech segment without considering the above two contextual factors. In this paper, we propose an Attentive Interaction Model (AIM) to capture both self- and interlocutor-context to enhance the speech emotion recognition in the dyadic dialog. The model learns to dynamically focus on long-term relevant contexts of the speaker and the interlocutor via the self-attention mechanism and fuse the adaptive context with the present behavior to predict the current emotional state. We carry out extensive experiments on the IEMOCAP corpus for dimensional emotion recognition in arousal and valence. Our model achieves on par performance with baselines for arousal recognition and significantly outperforms baselines for valence recognition, which demonstrates the effectiveness of the model to select useful contexts for emotion recognition in dyadic interactions.

Index Terms: Speech Emotion Recognition, Dyadic Interactions, Self-attention

1. Introduction

Automatic speech emotion recognition has been an active research area in recent years, which has a wide range of applications in audio/video chatting [1], call-center dialogue systems [2], conversational agents [3] and so on.

Most of previous studies perform speech emotion recognition on single speech segment. Among them, the CNN-LSTM network has achieved the state-of-the-art performance to predict the emotion of a single utterance [4, 5, 6]. However, emotion is not an instantaneous state. It is an evolutionary state that is affected by the context in the real dynamic interaction scenario. For example, as shown in Figure 1, the speaker's emotional state of the current turn will affect the emotional state in his/her following utterance. The interlocutor's behaviours in the dialog also have influence on the speaker's emotional state. Additionally, in offline scenarios, speaker's and interlocutor's behaviors in the future can also offer hints of the speaker's previous emotional states. Therefore, in this paper, we investigate how to utilize the interaction context information from the speaker and the interlocutor to improve the emotion recognition performance in the dialog interaction scenario.

There have been few endeavors to explore interaction contexts for speech emotion recognition. Dynamic Bayesian Network (DBN) and Hierarchical Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) have been applied to model the dependency between two interacting partners' emotional states in a dialog [7, 8]. However, these models only capture short-term dependency of contexts, which may not be sufficient since emotional states are developing and changing in the dialogue. Bidirectional LSTM model has been used to learn the long-term relevant emotional context from past and future observations during the conversation, and has shown better performance than GMM-HMM models [8, 4]. Zhang *et al.* [9] propose an interaction-and-transition model, which applies temporal pooling to encode long-term speaker's past contexts for emotion recognition. Different contexts are treated equally in the temporal pooling. However, different contexts are of different importance to the targeted utterance. Equally combining all contexts can bring negative effects for the emotion recognition due to the use of irrelevant or noisy contexts.

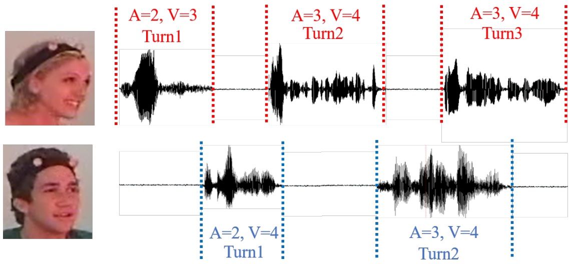

Figure 1: *The interaction of speaker and interlocutor in the dialog. "A=·" denotes the emotion label in the arousal (a measure of affective activation) dimension and "V=·" denotes the emotion label in the valence (a measure of pleasure) dimension. The speaker's emotional state can be affected by both self and interlocutor's behaviors in the dialog.*

In this paper, we propose a novel attentive interaction model (AIM) to dynamically select relevant contextual information in the dialogue to enhance the emotion recognition of the targeted speech utterance. Our model can take advantage of the bi-directional contexts of the speaker and interlocutor in the past and future, which captures long-term context dependencies. In order to focus on relevant contexts, we explore the self-attention mechanism [10] to dynamically adjust the importance weights of different contexts and then fuse the contexts adaptively with the targeted utterance speech representation to infer the targeted emotional state. We carry out extensive experiments on the widely used IEMOCAP dataset for arousal and valence dimensional emotion recognition. The proposed AIM model significantly outperforms the baselines without considering any contextual information or equally weighting contexts on the valence dimension and achieves on par performance with baselines on arousal dimension.

The contributions of this paper are three folds:

- We propose to harness bi-directional contexts of speaker and his/her interlocutor to infer emotional state of the targeted speech utterance in the interactive dialogue.

*Corresponding author.

http://dx.doi.org/10.21437/Interspeech.2019-2103

- In order to effectively utilize the context information, we propose an Attentive Interaction Model (AIM) to dynamically fuse relevant contexts via self-attention with the targeted speech features.
- The experimental results on arousal and valence emotion recognition demonstrate the effectiveness of the proposed AIM model, which greatly benefits the valence prediction with relevant contexts.

2. Related Work

The emotion recognition is mainly divided into two types, namely discrete and dimensional emotion recognition. The discrete emotion recognition classifies emotion as eight basic classes such as happy, angry, sad etc [11]. The dimensional emotion instead view emotion as point in the dimensional space with axes such as arousal and valence, where the arousal measure the strength of the emotion and valence measures the pleasure [12].

Recently, Convolutional Neural Networks (CNNs) [13] and Long Short-Term Memory (LSTM) [14] have shown improvements across different speech emotion recognition tasks. Neumann et al. [15] train a CNN to extract speech features from cepstral representation and achieve significantly improvement on speech emotion task. Trigeorgis *et al.* and Panagiotis *et al.* [6, 16] propose an end-to-end CNN-LSTM model to capture temporal dynamics in single utterance for emotion prediction. Several recent studies [17, 5, 15, 18, 19] explored the attention mechanism to focus on the emotion-salient frames in an utterance. However, these methods perform speech emotion recognition on single speech segment without considering the context information in the dialogue.

There are only a few works to tackle emotion recognition in dialogues with interaction contexts. Dynamic Bayesian Network (DBN) is used to explicitly model the conditional dependency between two interacting partners' emotional states in a dialog [7]. Metallinou *et al.* [8] have explored the Hierarchical GMM-HMM model and the BLSTM model for modeling emotion evolution within an utterance and between utterances over the course of a dialog. The BLSTM model outperforms the Hierarchical GMM-HMM models, which show the ability of BLSTM model to effectively learn an adequate amount of relevant emotional context from past and future observations. Wöllmer *et al.* [4] have investigated the impact of the number of its past and future utterance-level observations on recognizing the emotional state of the given utterance. Zhang et al. [9] have proposed an interaction-and-transition model which utilizes the time pooling layer to encode the previous speaker and interlocutor emotions to enhance the emotion recognition of the current utterance, which demonstrates the effectiveness of the interactive information. In this paper, we consider four types of contextual information in the dialogue including both speaker and interlocutor's contexts in the past and future. We also employ the self-attention mechanism [10] to effectively encode the contextual information, which differentiates relevant and noisy contexts unlike previous work.

3. The Proposed Approach

3.1. Overview

In order to effectively use different contextual information for emotion recognition in interactive dialogues, we propose the Attentive Interaction Model (AIM), which is illustrated in Figure 2. The AIM model consists of three modules, namely feature embedding layer, context attention layer and context-aware emotion recognition layer. We will describe the details of each layer in Section 3.2.

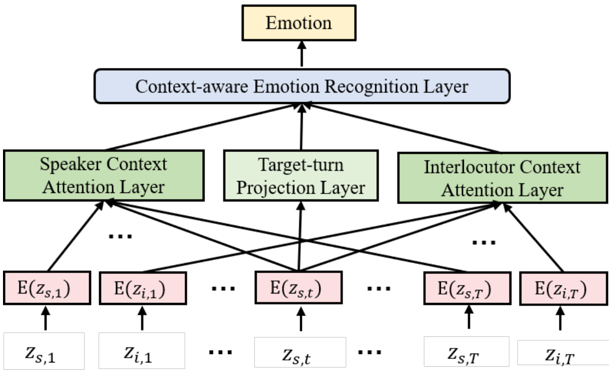
Emotion
Context-aware Emotion Recognition Layer
Speaker Context Attention Layer
Target-turn Projection Layer
Interlocutor Context Attention Layer
...
$E(z_{s,1})$
$E(z_{i,1})$
$E(z_{s,t})$
$E(z_{s,T})$
$E(z_{i,T})$
$z_{s,1}$
$z_{i,1}$
$z_{s,t}$
$z_{s,T}$
$z_{i,T}$

Figure 2: *System framework overview. The context attention layers are used to model different interaction contexts in the framework.*

Assume $\{z_{s,1}, z_{i,1}, \cdots, z_{s,t}, \cdots, z_{s,T}, z_{i,T}\}$ a sequence of speech input features for each turn[1] in the dialogue, where the $z_{s,t}$ represents the speaker's feature at the $(2t-1)$ turn, $z_{i,t}$ represents the interlocutor's feature at the $(2t)$ turn and $2T$ is the total length of turns in the dialogue. The goal is to predict the arousal and valence dimensional emotions for each turn of the speaker. The proposed AIM model firstly projects each turn feature $z_{\cdot,t}$ to an emotional embedding space via the feature embedding layer. Then the context attention layer dynamically attends to relevant contexts from the speaker and interlocutor respectively for the target turn. Finally, the context-aware emotion recognition layer incorporates the attended contextual features and the features of target turn to predict the emotion state.

3.2. Model Architecture

Feature Embedding Layer: The feature embedding layer $E(\cdot)$ is used to extract emotion salient features from the input speech representation z as follows:

$$E(z) = \sigma(W_e z + b_e), \tag{1}$$

where W_e and b_e are weight and bias parameters, and σ is the activation function such as ReLU. The feature embedding layer is shared for all turn features of the speaker and interlocutor.

Context Attention Layer: Since the different contextual information contributes differently to the emotion recognition of the target speech turn, it is necessary to select relevant contexts while alleviate noises from irrelevant contexts. Therefore, we propose the context attention layer which employs self-attention mechanism [10] to dynamically select and weight different contextual features. The detailed structure of the context attention layer is presented in Figure 3.

For the target turn $z_{s,t}$, we consider four types of different contextual turns from two sources (speaker and interlocutor) and two ranges (past and future). Take the speaker's past contextual turns as an example, suppose there are C turns of the speaker before the target turn $z_{s,t}$ as $\{z_{s,t-C}, \cdots, z_{s,t-1}\}$. The context attention layer weights different contextual turn features according to their correlations with the target turn. It firstly maps the target turn $z_{s,t}$ and its contextual turns $z_{s,j}$ into a joint

[1] The turn is defined as the portion of speech that belongs to a single speaker before he/she finishes speaking, and may consists of multiple original segmented utterances.

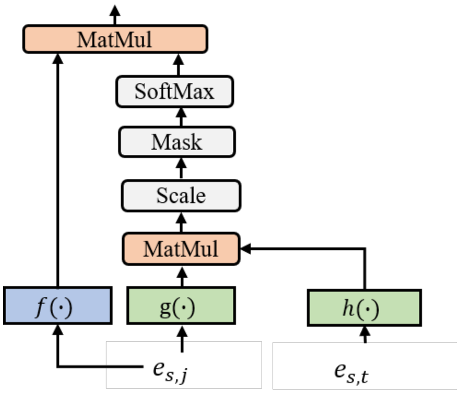

Figure 3: *Illustration of the context attention layer. The blue and green blocks are represent the context feature space, the joint attention space respectively.*

Table 1: *The number of turns in each session of IEMOCAP.*

	Sess1	Sess2	Sess3	Sess4	Sess5	total
#turns	1,510	1,507	1,679	1,801	1,925	8,422

attention space via linear projections[2] $h(\cdot)$ and $g(\cdot)$ respectively. Then the attended context is computed as follows:

$$\hat{z} = \sum_{j=1}^{C} \frac{\exp(\alpha_j)}{\sum_{q=1}^{C} \exp(\alpha_q)} f(e_{s,j}) \tag{2}$$

$$\alpha_j = \frac{g(e_{s,j})h(e_{s,t})}{\sqrt{d}}, \tag{3}$$

where α_j represents attention weight, $e_{s,j}$ represents the encoded features by the feature embedding layer, $f(\cdot)$ is another linear projection for the embedded features, and d is the dimensionality of the projected features.

Since the speaker's context and the interlocutor's context have different influences on the speaker's target emotion state, we utilize different function $h(\cdot), g(\cdot), f(\cdot)$ for different contextual sources such as speaker and interlocutor. But we share function parameters for different ranges such as past and future. For each contextual type, we compute the attended contextual vector $\hat{z}$. Therefore, we generate $\hat{z}_{s,p}$, $\hat{z}_{s,f}$, $\hat{z}_{i,p}$ and $\hat{z}_{i,f}$ as the speaker's past and future context attention features, and interlocutor's past and future context attention features respectively.

Context-aware Emotion Recognition Layer: After the context attention layer, we fuse different types of attended contextual features with the features of the target turn for emotion recognition:

$$\hat{y} = \text{softmax}(W_c c + b_c) \tag{4}$$

$$c = [\hat{z}_{s,p}; \hat{z}_{i,p}; f(z_{s,t}); \hat{z}_{s,f}; \hat{z}_{i,f}] \tag{5}$$

where W_c, b_c are parameters to be learned and $[\cdot]$ represents vector concatenation.

4. Experiments

4.1. Dataset

We use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [20] for evaluation. The IEMOCAP dataset contains recorded audios in 5 dyadic sessions. In each session, there are multiple scripted plays and spontaneous dialogues of a male and a female actor. Each dialog is separated to speech utterances, and each utterance is annotated with arousal and valence dimensional emotion labels from 1 to 5 scales. We merge consecutive utterances of the same person into one turn, so that the dialogue is represented as a sequence of speaker and interlocutor interactions. The turn-level emotional label is the average of labels of its constituent utterances. The number of turns in each session is presented in Table 1.

Following the emotional label processing in previous works [21, 22], we convert the 5 scales of each emotional dimension to 3 classes as three-way classification, where the scales 1-2 as the low class, scale 3 as the neutral class, and scales 4-5 as the high class. Since each turn is labeled by multiple annotators, we follow [22] to generate a fuzzy representation from multiple labels for each turn during training. For instance, if three annotators labeled an turn as [0, 0, 1], [0, 0, 1], and [0, 1, 0] respectively, the fuzzy label representation would be [0, 0.3, 0.7] and the correct class label would be 2. We treat the problem as a three-way classification problem, where the goal is to assign a label from $\{0, 1, 2\}$ to a given turn. However, in the testing phase, only the class of maximum value is used as the ground truth (if the maximum value occurs in different classes, such classes are all considered as the ground truth).

4.2. Experimental Setup

We utilize the OpenSMILE toolkit [23] to extract acoustic features with the IS10 configuration [24, 25]. We apply z-normalization on all the feature vectors from each individual speaker's utterances as in previous works [7, 22].

We compare the proposed AIM model with two types of baselines: 1) **DNN model**: it only utilizes the targeted turn to infer the emotion without considering any contextual information; and 2) **LSTM-based models**: it utilizes the LSTM network to encode contexts in the dialogue which ignores the different contributions of different contexts. For the DNN model, we set the layer as 2 and hidden units as 256 and 128 respectively. For the LSTM-based models, we set the layer as 1 and hidden units size as 128 based on validation performance. We use truncated Back Propagation Through Time (BPTT) with max step of 15 turns to train the LSTM-based models. **For the proposed AIM model**, the units of feature embedding layer and context attention layer are 256, and the size of context-aware emotion recognition layer is 128. We train at most 50 epochs for each model with the Adam optimizer. The learning rate is initialized from 0.0008 and is halved if the accuracy on the validation set is decreased. Dropout is adopted to avoid over-fitting with dropout rate of 0.2.

We use 5-fold leave-one-session-out cross validation to evaluate the models in a speaker independent manner. To demonstrate the robustness of models, we run each model for three times to alleviate influences of random initialization of parameters and apply significance test for model comparison.

4.3. Experimental Results

Table 2 presents emotion recognition performances of different models. The block (A) in Table 2 is the DNN baseline without using any contexts. The block (B) shows the results of LSTM-based models to encode contexts, and block (C) is the proposed AIM model to encode contexts. For valence recognition, fusing with contextual information significantly improves the performance compared with the baseline DNN model with no context. The best LSTM-based model achieves 2.29% absolute

[2] The linear projection functions are same as the function mentioned in Equation (1)

Table 2: *Comparison of different models for arousal and valence emotion recognition on the testing set.* $C = (m, n)$ *denotes the context sizes which are set to* m *and* n *for valence and arousal respectively.*

	model	context source	context range	valence(%)	arousal(%)
(A)	DNN	-	-	62.23 ± 0.42	70.41 ± 0.58
(B)	LSTM	speaker	past	60.94 ± 0.60	69.57 ± 0.24
	BiLSTM	speaker	past + future	63.19 ± 0.16†	69.83 ± 0.69
	LSTM	speaker + interlocutor	past	62.84 ± 0.55†	**70.52** ± 0.45
	BiLSTM	speaker + interlocutor	past + future	64.52 ± 0.29‡	69.92 ± 0.39
(C)	AIM	speaker	past, $C = (6, 2)$	64.97 ± 0.19	70.14 ± 0.51
		speaker	past + future, $C = (6, 2)$	66.41 ± 0.24†	69.53 ± 0.05
		speaker + interlocutor	past, $C = (6, 1)$	67.06 ± 0.23†	69.58 ± 0.49
		speaker + interlocutor	past + future, $C = (6, 4)$	**68.56** ± 0.41‡	70.39 ±0.55

† : significantly ($p < 0.01$) better than just use past speaker information.
‡ : significantly ($p < 0.01$) better than other models in the same block.

gains over the DNN model, while our best AIM model achieves 6.44% absolute gains over the DNN model, which demonstrates that the proposed AIM model can more effectively employ different contexts for valance recognition. However, there only exists marginal difference of models using contexts or not for arousal recognition. Since the arousal dimension measures the strength of emotion expressed in the utterance, it might mainly relies on the target turn to infer the emotion strength and less sensitive to the contexts. But the valence tells whether the emotion is positive or negative, which reflect relatively long-term variations so that both self and interlocutor's contexts are more beneficial for the valence recognition.

In order to ablate contributions of different context sources and ranges, we also report the emotion recognition performance with different contexts in Table 2 for both LSTM-based model and the proposed AIM model. Since the LSTM-based model is less effective to employ contextual information, not all sources and ranges of context are beneficial. However, our proposed AIM model can benefit from all the contexts for valence recognition as shown in block (C), which yield significantly improvements over the LSTM-based model and no-context DNN model. The main reason is that our AIM model learns to pay attention to different contexts dynamically to capture relevant contexts for the targeted speech turn. We can see that both the context of speaker and interlocutor are useful for valence recognition. Only using the past range of the two sources of contexts, we can achieve 4.8% absolute improvements than the baseline DNN without contexts in the real time emotion recognition scenario. In the offline emotion recognition scenario where the contexts from future range can be employed, the AIM model can achieve additional 1.5% absolute gains than the real time scenario.

Furthermore, we explore the influence of context size [3] on valence prediction with the AIM models in Figure 4. We can see that our model can benefit from more contextual turns thanks to our self-attention attention mechanism to select relevant contexts and filter out noisy contexts. When the contextual turns are larger than 4, the valence prediction performance are saturated, which indicates it is sufficient to employ a fixed number of contextual turns such as 6 as shown in Figure 4 to balance the efficiency and accuracy.

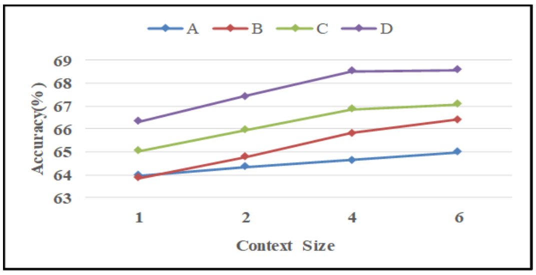

Figure 4: *The influence of context size on valence prediction of our AIM models. The "A" model utilizes self-context in the past, "B": "A" + self-context in the future, "C": "A" + interlocutor's context in the past, "D": both self and interlocutor's context in the past and future.*

5. Conclusion

In dyadic interaction scenario, a person's emotional state can be influenced by both self emotional evolution and the interlocutor's behaviors. In this paper, we propose an Attentive Interaction Model (AIM) to capture context information in the dialog to improve emotion recognition performance. Our proposed model employs the self-attention to capture long-term contexts such as bidirectional contexts from both speaker and interlocutor. The relevant information from different contexts are dynamically selected and fused with the targeted speech turn to infer the emotional state. Our experimental results show that the proposed AIM models achieve on par performance with baselines for arousal recognition and significantly outperform baselines for valence prediction. It demonstrate the usefulness of different contexts for the emotion recognition in the dialogue and effectiveness of the proposal AIM model to employ different contexts. In the future, we will validate the robustness of our proposed method on different datasets and tasks and explore utilizing multimodal cues to infer emotional state in the dialogue.

6. Acknowledgments

This work was supported by National Key Research and Development Plan under Grant No. 2016YFB1001202, Research Foundation of Beijing Municipal Science & Technology Commission under Grant No. Z181100008918002 and National Natural Science Foundation of China under Grant No. 61772535.

[3] Under "B" condition of Figure 4, we set context size as 2, which means 2 turns in the past and 2 turns in the future.

7. References

[1] F. Ringeval, B. W. Schuller, M. F. Valstar, J. Gratch, R. Cowie, and S. Scherer, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 3–9.

[2] M. Danieli, G. Riccardi, and F. Alam, "Emotion unfolding and affective scenes:a case study in spoken conversations," in *ICMI Workshop*, 2015, pp. 5–11.

[3] N. F. Fragopanagos and J. G. Taylor, "Emotion recognition in human-computer interaction," *Neural networks : the official journal of the International Neural Network Society*, vol. 18 4, pp. 389–405, 2005.

[4] M. Wöllmer, A. Metallinou, N. Katsamanis, B. W. Schuller, and S. Narayanan, "Analyzing the memory of blstm neural networks for enhanced emotion classification in dyadic spoken interactions," in *ICASSP*, 2012, pp. 2362–2365.

[5] P. Hsiao and C. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *ICASSP*, 2018.

[6] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*, 2016.

[7] C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *INTERSPEECH*, 2009, pp. 1983–1986.

[8] M. Wöllmer, A. Metallinou, F. Eyben, B. W. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *INTERSPEECH*, 2010, pp. 2362–2365.

[9] R. Zhang, A. Atsushi, S. Kobashikawa, and Y. Aono, "Interaction and transition model for speech emotion recognition in dialogue," in *INTERSPEECH*, 2017, pp. 1094–1097.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *CVPR*, 2017.

[11] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," pp. 205–211, 2004.

[12] S. Marsella and J. Gratch, "Computationally modeling human emotion," *Communications of The ACM*, vol. 57, no. 12, pp. 56–67, 2014.

[13] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *ICASSP*, 2013, pp. 8614–8618.

[14] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *INTERSPEECH*, 2014.

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

[16] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*, 2018.

[17] S. Mirsamadi, E. Barsoum, and C. Zhang, "Seyedmahdad mirsamadi and emad barsoum and cha zhang," in *ICASSP*, 2017.

[18] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *INTERSPEECH*, 2018, pp. 3087–3091.

[19] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio-temporal representations by integrating attention-based bidirectional-lstm-rnns and fcns for speech emotion recognition," in *INTERSPEECH*, 2018, pp. 272–276.

[20] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources Evaluation*, vol. 42, pp. 335–359, 2008.

[21] Z. Aldeneh, S. Khorram, D. Dimitriadis, and E. M. Provost, "Pooling acoustic and lexical features for the prediction of valence," pp. 68–72, 2017.

[22] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," *international conference on acoustics, speech, and signal processing*, pp. 2746–2750, 2017.

[23] F. Eyben, M. Wöllmer, and B. W. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th International Conference on Multimedia*, 2010, pp. 1459–1462.

[24] B. W. Schuller, A. Batliner, S. Steidl, B. Anton, B. Felix, D. Laurence, M. Christian, and N. Shrikanth, "The interspeech 2010 paralinguistic challenge," pp. 2794–2797, 2010.

[25] B. W. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9-10, pp. 1062–1087, 2011.