INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Removing Bias with Residual Mixture of Multi-View Attention for Speech Emotion Recognition

Md Asif Jalal, Rosanna Milner, Thomas Hain, Roger K Moore

Speech and Hearing Group (SPandH), The University of Sheffield
m.a.jalal, rosanna.milner, t.hain, r.k.moore @sheffield.ac.uk

Abstract

Speech emotion recognition is essential for obtaining emotional intelligence which affects the understanding of context and meaning of speech. The fundamental challenges of speech emotion recognition from a machine learning standpoint is to extract patterns which carry maximum correlation with the emotion information encoded in this signal, and to be as insensitive as possible to other types of information carried by speech. In this paper, a novel recurrent residual temporal context modelling framework is proposed. The framework includes mixture of multi-view attention smoothing and high dimensional feature projection for context expansion and learning feature representations. The framework is designed to be robust to changes in speaker and other distortions, and it provides state-of-the-art results for speech emotion recognition. Performance of the proposed approach is compared with a wide range of current architectures in a standard 4-class classification task on the widely used IEMOCAP corpus. A significant improvement of 4% unweighted accuracy over state-of-the-art systems is observed. Additionally, the attention vectors have been aligned with the input segments and plotted at two different attention levels to demonstrate the effectiveness.

Index Terms: speech emotion recognition, attention networks, computational paralinguistics

1. Introduction

Emotion in speech is a fundamental trait in human communication that reflects the meaning and intent. Emotion classification raises the question about 'what is said' and 'how it is said'. There are mainly two different approaches for representing emotions, i.e. categorical and dimensional. In categorical representation, the emotions exist as discrete labels such as happy, angry, sad etc. whereas the dimensional approach emphasises on understanding emotions in terms of valence and arousal. In this work, it has been assumed that emotion is a categorical perception representing discrete sensory events.

Speech emotion recognition (SER) tasks require a front-end for extracting features that hold emotion attributes while being robust to changes in time, frequency, speaker, medium and other external distortions. In practice, the most popular features are Opensmile [1], eGeMaps [2], MFCCs [3] and filterbanks [4]. These features are used with different classifiers such as hidden Markov models (HMMs) [5], support vector machines (SVMs) [6], deep belief networks (DBNs) [7] and deep neural networks (DNNs), and treated as a standard categorical classification task. DNNs learn task-specific abstract feature representations by filtering out unnecessary information and improving generalisation [8, 9, 10]. Research has suggested representation learning by modelling mid to long-term sequence dependencies [11, 12, 13].

The distinction about 'what is said' and 'how it is said' is not overly clear for SER tasks as it has not been well defined. Typically, emotion is represented in either a categorical or a dimensional annotation scheme. Although the duration or the position of emotion are not well defined in a sentence, it is clear that emotion is built upon on either short-term or long-term context [12, 13, 14].

Here, a novel model for speech emotion classification is proposed, which performs a deep level feature transformation. It learns different task-specific feature representations from utterances and performs feature transformation in a high dimensional space. This is followed by projection to the original feature vector. The feature projection aims to remove task-specific bias in the feature space. The experimental results show the effectiveness of the proposed computational model, leading to state-of-the-art results on the IEMOCAP [15] corpus in a 4-class setting.

The rest of this paper is organised as follows. The previous work related to this paper is discussed in Section 2. In Section 3, the components of our framework, i.e. long short term neural networks (LSTMs) and the proposed multi-projection self-attention network, mixture of multi-view attention (MOMA), are described. Section 3 also presents and explains the proposed architecture in terms of representation learning and the motivations behind it. In Section 4, the experimental setup is explained, and the results are presented along with a discussion in Section 5. Finally, Section 6 concludes the paper.

2. Related Work

Different context modelling techniques have been proposed for SER tasks, as mentioned in Section 1. In this paper, acoustic context expansion has been carried out with high dimensional multi-instance feature projection. Philosophically, it has similarity with the context expansion technique in feature-space minimum phone error (fMPE) [16, 17]. Sequential and hybrid-hierarchical models were proposed to learn deep feature representations [12, 14], and task-specific feature clusters [13]. Variants of attention-based mechanisms have been proposed which performed significantly better than the previous models [18, 19, 16]. One of the possible reasons why attention models outperform others is that the models learn the biases for a specific task, or group of tasks, leading to improved generalisation. Recently, a sequence and attention-based domain adversarial system was presented in [20] which investigated whether the information in acted datasets can be learnt to benefit emotion prediction for natural datasets and achieved state-of-the-art results.

3. Representation Learning

The features over time are extracted using a bi-directional long short-term memory network (BLSTM). Then, multiple in-

http://dx.doi.org/10.21437/Interspeech.2020-3005

stances of attention vectors are computed which are projected on to a representation space derived from the same features. The final 'smoothed' projection is applied to attain bias in the original feature space expansion.

3.1. BLSTM Encoder

Long short-term memory (LSTM) networks use the left to right temporal order of the sequence, whereas studies show that future or forward contexts are useful for context-sensitive sequence modelling [21, 22]. BLSTMs model the input sequence forward and backwards in two separate recurrent neural networks (RNNs) as a way to exploit the contextual information from the past and the future [21]. Applying these networks, a temporal feature distribution over the sequence can be obtained in the encoder layer which is stacked. This can be expressed by

$$y^{fwd}[t,h] = [LSTM\left(y^{t}[h], y^{fwd}[t,h-1]\right)] \quad (1)$$

$$y^{bck}[t,h] = [LSTM\left(y^{t}[h], y^{bck}[t,h+1]\right)] \quad (2)$$

$$y[t,h] = [y^{fwd}[t,h], y^{bck}[t,h]] \quad (3)$$

where t is the timesteps, h is hidden dimensions, The output y is stacked over time to form a matrix $\mathbb{Y} \in \mathbb{R}^{(T \times h)}$.

3.2. Mixture of Multi-View Attention (MOMA)

Self-attention networks can flexibly learn representations for long-term inter-sequence dependencies [23]. In this work, the basic attention block is similar to [14, 24]. First, a global contextualised attention mean M is calculated by computing the global mean across time. The mean is then repeated as the same temporal domain length as $\mathbf{Y}$ to form a matrix which has same size as $\mathbf{Y}$. Both $\mathbf{Y}$ and $\mathbf{M}$ are projected on to fully-connected layers, namely $\mathbf{W}_h$ and $\mathbf{W}_m$. These fully-connected layers are multiplied to find non-local positional dependencies and the result is projected to another fully-connected layer, $\mathbf{W}_e$, to produce the attention vector over time frames.

$$E = tanh\left(\boldsymbol{W}_h \boldsymbol{Y}\right) * tanh\left(\boldsymbol{W}_m \boldsymbol{M}\right) \quad (4)$$

$$a_{att1} = Softmax\left(\boldsymbol{W}_e * \boldsymbol{E}\right) \quad (5)$$

where E is positional dependency or self-attention between $\mathbf{W}_h$ and $\mathbf{W}_m$, and a_{att1} is the attention. This attention is projected onto $\mathbf{Y}$ as $\boldsymbol{Y'}$ and added as a skip connection with $\mathbf{Y}$. The skip connection reduces the degradation problem and helps the network attain iterative non-local feature learning [25, 26]. The schematic diagram is shown in Figure 1.

$$\boldsymbol{Y} = \boldsymbol{Y'} + \boldsymbol{Y} \quad (6)$$

Next, multiple attention blocks can be applied, and each of these blocks has different initialization. These are projected to a common space through a control parameter. This acts like an attention mixture model and is referred to as *MOMA*. All the spaces are derived from the same source $\boldsymbol{Y}$. However, they learn different representations.

$$E_n = tanh\left(\boldsymbol{W}_{h_n} \boldsymbol{Y}\right) * tanh\left(\boldsymbol{W}_{m_n} \boldsymbol{M}\right) \forall n = 1,2,3 \quad (7)$$

$$a_n = Softmax\left(\boldsymbol{W}_{e_n} * \boldsymbol{E_n}\right) \forall n = 1,2,3. \quad (8)$$

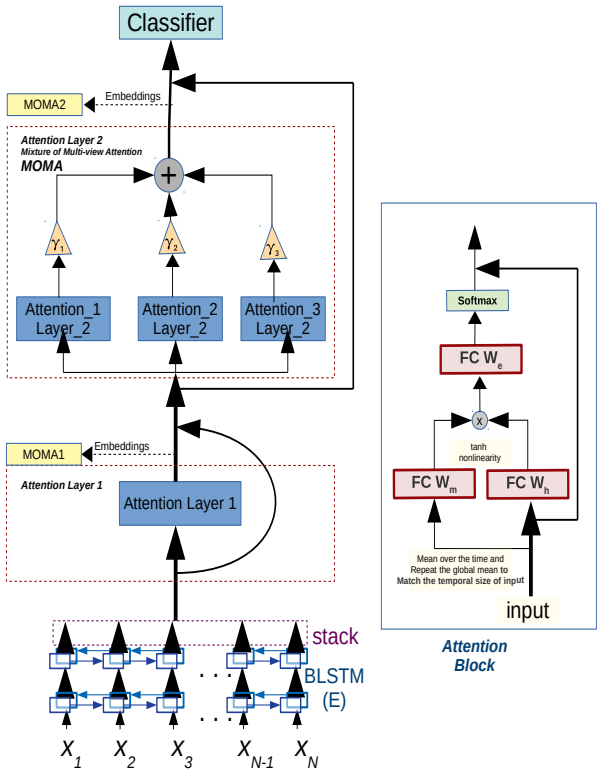

Figure 1: *Schematic diagram of the MOMA architecture.*

where n is the number of attention units and a_n is attention at the nth attention block. Each of $\boldsymbol{W}_{h_n}$ and $\boldsymbol{W}_{m_n}$ are initialized differently but they share a common representation space, $\boldsymbol{Y}$. This means different instances of E_1, E_2, E_3 are obtained from a common representation space. The projection is controlled using γ_1, γ_2, γ_3 as seen in Equation 9. Here $\boldsymbol{W}_{h_n}$, $\boldsymbol{W}_{m_n}$ and $\boldsymbol{W}_{e_n}$ are fully-connected layers and the network weights are trained through back-propagation.

$$a_{att2} = \sum_{n=1}^{3} \left(\gamma_n \cdot a_n\right) \quad (9)$$

where a_{att2} is the attention output from the *MOMA* attention blocks and n is the number of attention blocks in *MOMA* layer. Each of these attention vectors are time aligned with the input segment in the network.

Here it has been hypothesized that by projecting the mixture of attention scores in to the common feature space, the model is learning loosely correlated task-specific attention representations and by adding them the model performs smoothing to improve robustness. To investigate this hypothesis, attention vectors are extracted and analysed with the input segments to investigate attention in the intermediate hierarchies (Figures 2-5). In this work, the γ_1, γ_2, γ_3 are initialised randomly. This layer obtains the non-local dependencies.

3.3. Proposed Architecture

The overall architecture of the proposed framework is shown in Figure 1. The model is a hierarchical attention structure with

LSTMs. The LSTM processes long term temporal sequential dependencies and produce an abstract sequential feature representations. The attention layers attain positional dependencies to capture dynamic acoustic cues.

The *BLSTM Encoder* contains two hidden layers of 512 nodes each. It outputs a stacked matrix of size *[number of frames]* $\times$ 1024. This output of size 1024 is fed into the first attention layer *Attention Layer 1*. The attention mechanism is computing a context vector of size 128. The attention projection is of size *[number of frames]* $\times$ 1024).

The output from the encoder and the attention projection are added as residual skip connections and passed to the *MOMA* layer with three attention blocks i.e. *Attention_1 Layer2, Attention_2 Layer2, Attention_3 Layer2*. Each block in *Attention Layer 2* process it individually and projects it with a control parameter γ. Finally, these attention heads are added, and the result is projected to 1024 nodes. The *Attention Layer 2* obtains task-specific high dimensional features from *Attention Layer 1*'s output feature space and performs smoothing on task-specific multi-view attention. The components are explained in Section 3.2. The $\boldsymbol{W}_y$'s, $\boldsymbol{W}_m$'s and $\boldsymbol{W}_e$'s are fully-connected neural layers and along with the γ's they are trained through back-propagation. This is then passed to the emotion classifier which linearly projects to the number of classes. The cross-entropy loss function is applied, which is preceded by a $softmax$ layer.

4. Experimental Setup

4.1. Dataset

The IEMOCAP corpus [15] is used for validating the proposed framework. The corpus contains utterances from ten speakers (five male and five female) over 12 hours. The sessions are dyadic (between two speakers) and either scripted or improvised for eliciting emotions. Four sessions, containing a total of eight speakers, are used for training. The remaining session, which contains two speakers, is used for testing. In the literature it is common for IEMOCAP to be evaluated with four classes: *happy*, *sad*, *anger* and *neutral* (where *happy* is combined with *excitement*) [27]. The utterances are split into a training set of 4290 samples (Sessions 1-4) and a test set of 1241 samples(Session 5). This is referred to as IEM4 in this paper and in [20].

4.2. Features

Experiments in [13, 20] showed that the sequence model based systems performed best with 23-dimensional log-Mel filterbank features which hence applied to the *MOMA* system as well.

4.3. Implementation

The Adam optimiser [28] is applied to the proposed model with the initial learning rate of 0.0001. As Adam adaptively optimises the learning rate, the PyTorch approach of ReduceLROnPlateau has been investigated. The optimum patience setting was found to be 4 epochs with a multiplicative factor of 0.1. Transfer learning mechanisms are not used.

4.4. Evaluation

Unweighted accuracy (UA) and the weighted accuracy (WA) are used to evaluate the results. The UA calculates accuracy in terms of the total correct predictions divided by total samples, which gives equal weight to each class. As IEM4 is imbalanced across the emotion classes, the WA is calculated as well, which weighs each class according to the number of samples in that class:

$$UA = \frac{TP + TN}{P + N}, \qquad WA = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N}) \tag{10}$$

where P is the number of correct positive instances (equivalent to $TP+FN$) and N is the number of correct negative instances (equivalent to $TN + FP$).

Method	UA%	WA%
Factor analysis [31]	-	56.1
CNN_ LSTM [29]	59.4	-
CNN_ RecCap [12]	58.2	-
CNN_ GRU-SeqCap [12]	59.7	-
Attention Pool [30]	71.8	-
Convolutional self-attention [32]	76.3	68.8
MULTIMODAL: Attention [18]	*78.0*	-
MOMA	**80.5**	**74.8**

Table 1: *Performance of the MOMA model compared to baselines evaluted on the IEM4 dataset in terms of UA and WA.*

4.5. Baseline

The results are compared directly with speech emotion recognition systems that use the IEM4 dataset. Four of these baselines process audio data only. For comparing UA, the results from a CNN-LSTM [29] model, a deep capsule network with GRU [12], and a deep attention pooling [30] based model are presented. The result from [31] has been cited to show WA baseline. A multimodal system [18] carrying out SER on textual as well as audio data is also included to show how much the MOMA model could reach.

5. Result and Discussion

The baseline systems and performance of the proposed model are shown in Table 1. It is clear that the proposed *MOMA* model outperforms the baseline systems, including the multimodal approach which uses lexical and audio data, as opposed to the *MOMA* only using audio data. The proposed system has achieved 80.5% UA and 74.8% WA on segment-level training.

It can be said that the model learns speaker-independent emotional context information. In Equations 7, 8 and 9, the different projections on the same derived feature space $\mathbf{Y}$ learn different variations of the same feature space and the γ's make it more flexible. As a result, the network becomes more robust to speaker and distortion variations (see Section 5.1).

5.1. Hierarchical Attention Weights

To further show how the learned attention representations improve the performance, Figures 2-5 compares the attention at different hierarchies over the same utterance. The audio segments are mapped to the attention to show the relative positions of the attention weights compared to the phones and words.

The projections over two utterances are shown for comparison. The embeddings are extracted from two stages of the network i.e *MOMA1* and *MOMA2*. *MOMA1* (Equation 5) is the extracted attention vector embedding at *Attention Layer 1* and *MOMA2* (Equation 9) is the attention embedding at *Attention Layer 2* from Figure 1. *Attention Layer 2* is the mixture of multi-view attention network.

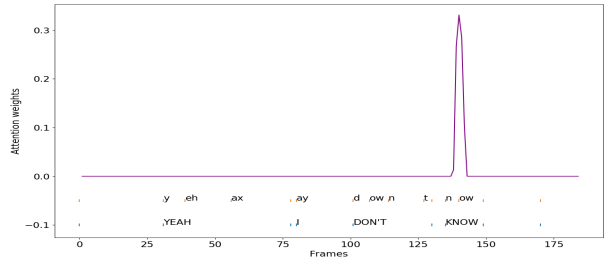

Figure 2: *MOMA1:"Yeah, I don't know"*

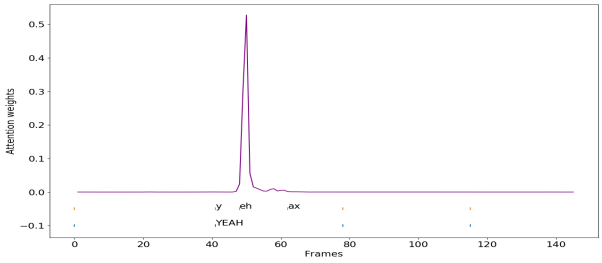

Figure 3: *MOMA1:"Yeah"*

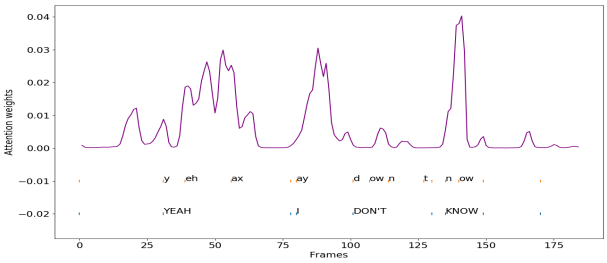

Figure 4: *MOMA2:"Yeah, I don't know"*

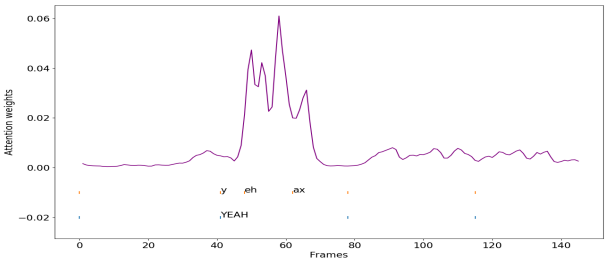

Figure 5: *MOMA2:"Yeah"*

According to Section 3.2, the mixture of attention network is learning loosely correlated task-specific attention representations because the attention blocks are projected onto a common feature space which is added in the end. Thus, it performs smoothing and improves overall robustness which is evident in Figures 2-5. From Figures 2 and 4, it is observed that the attention weights from *Attention Layer 1* embeddings, i.e. *MOMA1* are sensitive to particular regions and phones. However, Figures 3 and 5 show that the attention weights from *Attention Layer 2* embeddings, i.e. *MOMA2* are well distributed over the phone boundaries. Thus, it is evident that there are different representations over the different stages of hierarchy in the network. Also, it can be observed that the attention weights of *MOMA2* are well distributed over the phone boundaries compared to *MOMA1*. Whereas the attention in *MOMA1* is sensitive to some regions, but *MOMA2* is smoothed over the overall boundary. This strongly indicates that *MOMA2* is more robust than *MOMA1*.

5.2. Number of Attention Blocks

Although the mixture of multi-view attention shows a significant improvement of the attention weighting over the segments, the optimal number of such attention blocks is unclear. In this work, three blocks have been applied with three control parameters. A higher number of attention blocks may increase the performance of the model, but it can also overfit the model due to the higher number of parameters. Furthermore, it can cause the *degradation* problem in the model. Therefore, investigating the depth vs. width in this network for SER tasks is an important future research direction.

6. Conclusion

In this paper, a residual mixture of multi-view attention emotional context modelling technique, MOMA, using acoustic feature space expansion has been proposed. The model attains task-specific bias in the feature representation resulting in an improved classifier and *state-of-the-art* performance for this SER task. The model also features hierarchical attention. The interpretability of intermediate states of this particular type of attention mechanism has been explored in order to investigate the hypothesis that by projecting the mixture of attention scores into the common feature space, the model is learning loosely correlated task-specific attention spaces and by adding them, the model performs smoothing to achieve more robustness. This has inspired an empirical way to interpret speech-based emotion perception in computational models by plotting the attention weights with respect to the words and the phones. Exploring this network to adapt to different speech-related tasks would be interesting further work.

7. Acknowledgements

We sincerely thank Voicebase Inc, for their research funding and support.

8. References

[1] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 1459–1462.

[2] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, 2016.

[3] T. L. Nwe, S. W. Foo, and L. C. D. Silva], "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[4] Tin Lay Nwe, Foo Say Wei, and L. C. De Silva, "Speech based emotion classification," in *Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology.*

TENCON 2001 (Cat. No.01CH37239), vol. 1, 2001, pp. 297–301 vol.1.

[5] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 2, pp. 401–404, 2003.

[6] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech and Language*, 2015.

[7] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2013, pp. 216–221.

[8] S. Zhang, T. Huang, and W. Gao, "Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval - ICMR '16*, 2016, pp. 281–284.

[9] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[10] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014.

[11] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Deep temporal models using identity skip-connections for speech emotion recognition," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. New York, NY, USA: ACM, 2017.

[12] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu, and H. Meng, "Speech emotion recognition using capsule networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6695–6699.

[13] M. A. Jalal, E. Loweimi, R. K. Moore, and T. Hain, "Learning temporal clusters using capsule routing for speech emotion recognition," *Proc. Interspeech 2019*, pp. 1701–1705, 2019.

[14] R. Beard, R. Das, R. W. M. Ng, P. G. K. Gopalakrishnan, L. Eerens, P. Swietojanski, and O. Miksik, "Multi-modal sequence fusion via recursive attention for emotion recognition," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 251–259.

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

[16] M. A. Jalal, R. K. Moore, and T. Hain, "Spatio-temporal context modelling for speech emotion classification," *2019 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019.

[17] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fmpe: Discriminatively trained features for speech recognition," in *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005*, 2005, pp. 961–964.

[18] Z. Lian, J. Tao, B. Liu, and J. Huang, "Conversational emotion analysis via attention mechanisms," in *INTERSPEECH 2019*, 2019.

[19] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *INTERSPEECH 2019*, 2019.

[20] R. Milner, M. A. Jalal, R. W. M. Ng, and T. Hain, "A cross-corpus study on speech emotion recognition," in *Proc. ASRU*, 2019.

[21] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM networks," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 4, 2005, pp. 2047–2052.

[22] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Processing*, 1997.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014.

[25] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," *CoRR*, vol. abs/1704.06904, 2017.

[26] S. L. J, D. Arpit, N. Ballas, V. Verma, T. Che, and Y. Bengio, "Residual connections encourage iterative inference," *CoRR*, vol. abs/1710.04773, 2017.

[27] P. L. et al., "An attention pooling based representation learning method for speech emotion recognition," in *Proc. Interspeech*, 2018.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[29] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *INTERSPEECH*, 2017.

[30] P. Li, Y. Song, I. V. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Interspeech*, 2018.

[31] B. Desplanques and K. Demuynck, "Cross-lingual speech emotion recognition through factor analysis," in *Proc. Interspeech*, 2018.

[32] M. A. Jalal, R. K. Moore, and T. Hain, "Spatio-temporal context modelling for speech emotion classification," in *Proc. ASRU*, 2019.