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

WISE: Word-Level Interaction-Based Multimodal Fusion for Speech Emotion Recognition

Guang Shen[1], *Riwei Lai*[1], *Rui Chen*[1*], *Yu Zhang*[2], *Kejia Zhang*[1], *Qilong Han*[1], *Hongtao Song*[1]

[1]College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]Department of Computer Science and Engineering,
Southern University of Science and Technology, Shenzhen, China

{shenguang, lai, ruichen}@hrbeu.edu.cn, zhangy7@sustech.edu.cn,
{kejiazhang, hanqilong, songhongtao}@hrbeu.edu.cn

Abstract

While having numerous real-world applications, speech emotion recognition is still a technically challenging problem. How to effectively leverage the inherent multiple modalities in speech data (e.g., audio and text) is key to accurate classification. Existing studies normally choose to fuse multimodal features at the utterance level and largely neglect the dynamic interplay of features from different modalities at a fine-granular level over time. In this paper, we explicitly model dynamic interactions between audio and text at the word level via interaction units between two long short-term memory networks representing audio and text. We also devise a hierarchical representation of audio information from the frame, phoneme and word levels, which largely improves the expressiveness of resulting audio features. We finally propose WISE, a novel word-level interaction-based multimodal fusion framework for speech emotion recognition, to accommodate the aforementioned components. We evaluate WISE on the public benchmark IEMOCAP corpus and demonstrate that it outperforms state-of-the-art methods.

Index Terms: Speech emotion recognition, dynamic interaction mechanism, hierarchical representation, deep multimodal fusion

1. Introduction

In recent years, speech emotion recognition (SER) has found its way in a wide range of human-computer interaction applications. For example, chatbots have become increasingly popular in various customer services over the phone or the Web. Accurately detecting users' emotions via their utterances is key to better user experience in such scenarios. As another example, SER has been used to help children with autism who may experience significant difficulties to recognize and express emotions to improve their socio-emotional communication skills [1]. Due to its practical importance, SER has received substantial attention from both academia and industry. However, as of now, it still remains a challenging technical problem due to the inherent subtlety of human emotions.

Speech by its nature is multimodal. While there are a large number of studies that consider only a single modality for SER, the latest research results [2, 3, 4, 5, 6, 7, 8, 9, 10] have confirmed the necessity and benefits of leveraging multimodal features. Audio and text are arguably the two modalities most commonly used together for SER. The existing studies normally fuse audio and text features at *the utterance level* and largely neglect the dynamic interplay of features from different modalities along the timeline. We argue that *such interplay at a fine-granular level and its evolution are critical to discriminate human emotions*.

In this paper, we propose to capture the interplay between audio and text at *the word level* as a word is the most natural textual unit in a sentence with meaningful semantics. To enable the interaction between audio and text, we align a word with its corresponding audio clip along the timeline. We explicitly model the interaction between the aligned word and audio clip and leverage their evolution over time, which intuitively unveils human emotions better. While word embeddings are effective to generate word-level textual features, word-level audio features need a more careful design. Existing research typically generates audio features at the frame level. However, our insight is that frame-level audio features tend to introduce emotionally irrelevant noise, leading to worse performance. To this end, we devise a hierarchical representation structure to extract a word-level audio clip's features from its contained phonemes, whose features are further extracted from their contained frames. Finally, we present WISE, a novel word-level interaction-based multimodal fusion framework for speech emotion recognition, which makes use of an attention mechanism to generate a temporally weighted aggregation of word-level fused multimodal features. We summarize our technical contributions as follows.

- We propose a novel interaction mechanism to capture the dynamic interactions and evolution between multimodal features at the word level. To the best of our knowledge, this is the *first* work that considers word-level interaction-based multimodal fusion for SER.
- We design a hierarchical representation of audio at the word, phoneme and frame levels, which forms more emotionally relevant word-level acoustic features.
- We put forward an original deep multimodal fusion framework to accommodate the above components, leading to more accurate SER. We demonstrate that WISE substantially outperforms the best state-of-the-art methods on the benchmark IEMOCAP dataset[1].

2. Related Work

The recent rise of deep learning techniques has been the fuel for SER. There has been a good amount of research [11, 12, 13, 14, 15, 16, 17, 18, 19, 20] using unimodal features to classify speech emotions. Due to the space limit, we focus on reviewing the latest research that considers multimodal features for SER, which has shown improved performance. Yoon *et al.* [2] propose a deep dual recurrent encoder model that represents both

* corresponding author

[1]Our code is available at https://github.com/gshen-heu/WISE.

http://dx.doi.org/10.21437/Interspeech.2020-3131

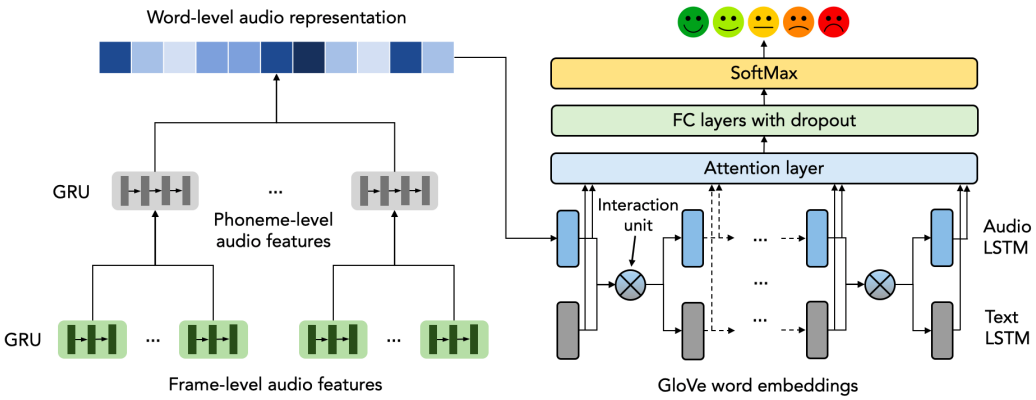

Figure 1: *The architecture of the WISE framework. The left part illustrates the hierarchical representation of word-level audio features; the right part illustrates the word-level multimodal interaction mechanism and the deep fusion network.*

audio and text data as sequences and combines them to recognize emotions from speech. Cho *et al.* [3] propose to use a long short-term memory (LSTM) network to detect emotions from acoustic features and a multi-resolution convolutional neural network (CNN) to detect emotions from word sequences. Liu *et al.* [4] put forward a low-rank multimodal fusion method that performs multimodal fusion using low-rank tensors to improve efficiency. Ghosal *et al.* [5] hypothesize that neighboring utterances of an utterance can contribute to emotion recognition and devise a recurrent neural network based multimodal attention framework. Kim and Shin [6] use deep neural networks to extract bottleneck acoustic features and extract two types of word-level lexical features in the forms of the distributed representation and affective lexicon-based dimensions. Sebastian and Pierucci [7] experiment with various existing fusion strategies over DNN-based unimodal emotion detection systems for audio and text. With a different objective, Lian *et al.* [8] study the problem of conversational emotion analysis. The key idea is to exploit dependencies among different utterances in a conversational dialog. Similarly, Gu *et al.* [9] consider emotion recognition in dyadic communication rather than a single utterance. They design a dyadic fusion network that only relies on an attention mechanism to fuse modality-specific features.

All previous studies consider the fusion between audio and text features at the utterance level, which is insufficient to capture the emotion development process for better classification performance. Xu *et al.* [10] propose to learn the soft alignment between words and acoustic frames. An attention network is used to generate an aligned weighted audio feature vector from all frames for each word. While this method promotes interactions between text and audio at a fine-granular level, aligning words with all audio frames in an utterance could be less effective as previous research [21] has indicated that emotional information is best expressed by *consecutive* frames rather than scattered frames.

3. The WISE Framework

Given an utterance, the proposed WISE framework aligns text and audio data at the word level, creates word-level audio and text features, fuses them dynamically along the timeline, and finally uses an attention mechanism to learn temporally weighted features for emotion classification. The text information in an utterance can be obtained via an automatic speech recognition (ASR) system [22] in real time. The overall architecture of the WISE framework is illustrated in Figure 1.

3.1. Word-Level Multimodal Alignment and Interaction

To achieve multimodal emotion recognition using audio and text, previous studies normally train LSTMs to model audio and text data independently and only concatenate them before feeding into fully connected layers. Such methods cannot learn the temporal correlation among multiple modalities at a fine-granular level, which is vital for more accurate SER. To address this limitation, we propose a novel interaction mechanism that allows to dynamically fuse audio and text features at the word level and that better models the evolution of the emotion embedded in an utterance.

The first step is to create word-level features from audio and text. Extracting words from a sentence is straightforward. Once we have the words in an utterance, we can generate the corresponding audio clip for each word by excerpting the audio chunk between the starting and ending timestamps of the word. We convert each word into a 300-dimensional embedding vector t_i by using pre-trained GloVe vectors [23]. Then the textual word sequence in an utterance (referred to as *text sequence*) can be represented as $[t_1, t_2, \cdots, t_N]$, where N is the number of words in the sentence. Similarly, we can represent the sequence of word-level audio clips (referred to as *audio sequence*) as $[a_1, a_2, \cdots, a_N]$, where a_i is the feature vector of the i-th audio clip. We discuss how to generate a_i by using a hierarchical representation in Section 3.2.

Before introducing the interaction mechanism, we can use LSTMs to model the text and audio sequences at the word level. The hidden state h_{t_i} of the i-th word in the text LSTM is:

$$h_{t_i} = f_{\theta_t}(t_i, h_{t_{i-1}}), \tag{1}$$

where f_{θ_t} is the LSTM function with parameter θ_t, $h_{t_{i-1}}$ is the hidden state of the previous word, and t_i is the GloVe word embedding of the i-th word. Similarly, the hidden state h_{a_i} of the i-th audio clip in the audio LSTM is:

$$h_{a_i} = f_{\theta_a}(a_i, h_{a_{i-1}}), \tag{2}$$

where f_{θ_a} is the LSTM function with parameter θ_a, $h_{a_{i-1}}$ is the hidden state of the previous audio clip, and a_i is the feature vector of the i-th audio clip.

In order to capture the word-level interaction between the $(i-1)$-th text and audio pair, we introduce an *interaction matrix* to fuse $h_{a_{i-1}}$ and $h_{t_{i-1}}$, and the fused results are passed into the i-th text and audio LSTM cells as input. In this way, we enable the interactions between text and audio at the word level

and are able to model the evolution of emotions over time. More specifically, given the hidden states $h_{a_{i-1}}$ and $h_{t_{i-1}}$ at time step $i-1$, the hidden states at time step i are calculated as follows:

$$\begin{bmatrix} \tilde{h}_{a_{i-1}} \\ \tilde{h}_{t_{i-1}} \end{bmatrix} = \begin{bmatrix} W_{aa} & W_{ta} \\ W_{at} & W_{tt} \end{bmatrix} \begin{bmatrix} h_{a_{i-1}} \\ h_{t_{i-1}} \end{bmatrix} \tag{3}$$

$$h_{a_i} = f_{\theta_a}(t_i, \tilde{h}_{a_{i-1}}) \tag{4}$$

$$h_{t_i} = f_{\theta_t}(a_i, \tilde{h}_{t_{i-1}}) \tag{5}$$

Here W_{aa}, W_{ta}, W_{at} and W_{tt} are trainable parameters, which together define the interaction matrix. They are initialized from a normal distribution with zero mean and a standard deviation of 0.05 for stable learning. The learned interaction matrix maps $h_{a_{i-1}}$ and $h_{t_{i-1}}$ into $\tilde{h}_{a_{i-1}}$ and $\tilde{h}_{t_{i-1}}$, respectively, and $\tilde{h}_{a_{i-1}}$ and $\tilde{h}_{t_{i-1}}$ are fed into the next LSTM cells as input. Intuitively, the interaction matrix can map the audio features into the text feature space by the weight matrix W_{at} and map the text features into the audio feature space by the weight matrix W_{ta}. The weight matrices W_{aa} and W_{tt} represent the weights that the LSTM networks should retain and pass to the next time step. So the input of an LSTM cell at the i-th time step consists of two parts: the information from the same modality that is retained from the previous time step and the fused information learned from the multimodal interaction. The interaction matrix is shared by all LSTM cells.

3.2. Hierarchical Representation of Audio Data

Modern linguistics indicates that speech sound can be represented at multiple levels: the sound of a word consists of several phonemes, and a phoneme can be further divided into frames, where a phoneme is the smallest unit of sound that may cause a change of meaning within a human language [24]. This implies a natural hierarchical representation of audio data. Prior studies only focus on frame-level representations of audio data. However, frame-level information is normally noisy for emotion recognition. For example, there are many frames that are simply silent or contain only background noise. Therefore, we propose a hierarchical representation structure of audio data, as illustrated in the left part of Figure 1, which can effectively eliminate noise from low-level acoustic information by leveraging high-level semantic meaning.

We first describe how to generate frame-level audio features. We divide audio signals into frames of 25ms window with 40% overlap. For a given frame, we use the openSMILE toolkit [25] to extract its Mel-Frequency Cepstral Coefficients (MFCCs), including 13 cepstral coefficients, 13 delta coefficients (i.e., first derivatives) and 13 acceleration coefficients (i.e., second derivatives), which form a 39-dimensional frame-level feature vector.

To construct the phoneme-level feature of the i-th phoneme in a word, we represent the frame sequence in the phoneme as $[f_{i1}, f_{i2}, \cdots, f_{iN_i}]$, where f_{ij} is the feature of the j-th frame in the i-th phoneme and N_i denotes the number of frames in the phoneme. A gated recurrent unit (GRU) network [26] is used to capture the contextual information between frames. It can be represented as:

$$h_{ij} = \mathrm{GRU}(f_{ij}), j \in \{1, 2, \cdots, N_i\}, \tag{6}$$

where h_{ij} is the contextual hidden state of the j-th frame. The last hidden state of the GRU, h_{iN_i}, is considered as the representative vector that contains all sequential audio information in the phoneme, and is used as the feature vector of the i-th phoneme.

Similarly, the i-th word-level audio clip can be modeled as a sequence of its contained phoneme feature vectors, $[h_{i1}, h_{i2}, \cdots, h_{iM}]$, where h_{ij} is the feature vector of the j-th phoneme in the audio clip, and M is the number of phonemes in the audio clip. We also apply a GRU to $[h_{i1}, h_{i2}, \cdots, h_{iM}]$ and use the last hidden state as the word-level audio feature vector, which is a_i introduced in Section 3.1. In addition, GRUs used for each level share the same parameters.

It is worth mentioning that we have experimented with various variants of the recurrent neural network (RNN) to generate the hierarchical representation, such as LSTM, Bi-LSTM [27] and Bi-GRU [28]. All of them achieve similar performance, but GRU requires the least number of parameters and trains much faster, which is the main reason to use it.

3.3. The Deep Multimodal Fusion Network

With the aforementioned word-level interaction mechanism and hierarchical representation, we are ready to present the end-to-end deep multimodal fusion framework, WISE. The core of our framework is an attention layer that simultaneously learns the weights of audio and text hidden states and the weights of different word-level audio and text pairs over time.

Given the sequences of audio and text hidden states $[h_{a_1}, h_{a_2}, \cdots, h_{a_N}]$ and $[h_{t_1}, h_{t_2}, \cdots, h_{t_N}]$ learned from Section 3.1, the normalized attention weight α_i of the i-th pair can be calculated as:

$$e_i = \tanh(\omega_a h_{a_i} + \omega_t h_{t_i}) \tag{7}$$

$$\alpha_i = \frac{e^{e_i \omega_c}}{\sum_k e^{e_k \omega_c}}, \tag{8}$$

where e_i is the energy score computed from h_{a_i} and h_{t_i}, and ω_a, ω_t, ω_c are all trainable parameters. ω_c is the context vector, and ω_a and ω_t are the weights used to indicate the relative strengths of h_{a_i} and h_{t_i} for emotion detection. Note that ω_a, ω_t and ω_c are shared by all hidden states and are used to form the joint representations from aligned word-level audio and text features. The final emotion representation p_c fed into the fully connected layers is a weighted sum of all joint representations at different time steps, which allows to selectively focus on the most emotionally relevant audio and text pairs along the timeline:

$$p_c = \sum_i \alpha_i (\omega_a h_{a_i} + \omega_t h_{t_i}) \tag{9}$$

The fully connected layers contain three linear layers, the first two of which are followed by a rectified linear unit (ReLU) layer and a dropout layer. That is,

$$p_1 = \phi(\omega_1^\top r_1 p_c) \tag{10}$$

$$p_2 = \phi(\omega_2^\top r_2 p_1) \tag{11}$$

$$\hat{y}_k = \mathrm{softmax}(\omega_3^\top p_2), \tag{12}$$

where ω_1, ω_2, ω_3 are trainable parameters with ω_i ($i \in \{1, 2, 3\}$) being the weight of each linear layer, r_1 and r_2 are the dropout vectors, $\phi(\cdot)$ is the ReLU function, and $\hat{y}_k$ is the output of the softmax function, which represents the final prediction result. The cross-entropy loss for K-class classification is used as the loss function:

$$\mathcal{L} = \sum_{k=1}^{K} y_k \log(\hat{y}_k). \tag{13}$$

4. Experimental Evaluation

In this section, we experimentally evaluate the performance of our solution by comparing with state-of-the-art competitors.

4.1. Experimental Settings

We use the interactive emotional dyadic motion capture (IEMOCAP) dataset [29] for experiments, which is a standard benchmark dataset widely used for SER. It has five sessions, involving conversations from 10 female and male actors, along with the corresponding labeled speech text (at both phoneme and word levels). We use the ground-truth transcripts to generate word-level textual features. For a fair comparison with previous studies [10, 24, 2], we also consider 4 out of the 9 emotions (*angry*, *happy*, *neutral* and *sad*) for classification[2]. Each utterance in the IEMOCAP dataset is labeled by three annotators, and we assign a single category to each utterance by majority vote. The final dataset contains 5,531 utterances in total, including 1,103 *angry*, 1,636 *happy*, 1,708 *neutral*, and 1,084 *sad*. Similar to the setting in [2, 24], we perform 5-fold cross-validation to do the evaluation. We compare our solution `WISE` with three state-of-the-art multimodal methods that achieve the best performance on IEMOCAP: `LSTM+Attn` [10], `CNN+Phoneme` [24] and `dual RNNs` [2].

4.2. Implementation Details

We implement our model in PyTorch. We set the maximum number of frames contained in a phoneme to 64 because more than 90% of phonemes contain less than 64 frames, the maximum number of phonemes in a word to 14, and the maximum number of words in an utterance to 128. We use 50 hidden units (i.e., the dimension of a hidden state) in a GRU for phoneme-level representations, 100 hidden units in a GRU for word-level representations, 100 hidden units in the audio LSTM and 150 hidden units in the text LSTM. We set the output dimension of the attention layer to 200 and the output dimensions of the three linear layers to 200, 100 and 4, respectively. The WISE model is optimized with the Adam optimizer [30]. We set the learning rate to 0.001 and the decay rate of the learning rate to 0.001. To address overfitting, we use L_2 regularization with the regularization coefficient of 0.0001 and 30% dropout rate. The `PackedSequence` library in PyTorch is used to deal with sequences with variable lengths.

4.3. Results and Discussion

We adopt two widely used evaluation metrics: *weighted accuracy* (WA) that is the overall classification accuracy and *unweighted* accuracy (UA) that is the average recall over the emotion categories. We report the main results in Table 1. First, we confirm that using multimodal features achieves better performance than using unimodal features. Second, `WISE` achieves the best WA and UA scores. It outperforms the best state-of-the-art methods by absolute 2% WA increase and absolute 1.1% UA increase. It is worth mentioning that `CNN+Phoneme` combines the features from audio and spectrogram. The results suggest that leveraging word-level interactions between audio and text features is indeed beneficial for emotion recognition, validating our hypothesis that audio and text together carry strong signals for emotion classification.

To prove the benefits of different components we propose, we perform an ablation study. Below we refer to the word-level multimodal interaction mechanism as `IM`, the hierarchical representation of audio data as `HR` and the deep multimodal fusion network as `DMF`. In Table 2, we show the performance of different variants of `WISE`. The `IM+DMF` method takes out `HR`. By comparing its performance with `WISE`, it can be seen that `HR` brings 3% WA improvement and 2.2% UA improvement, showing that the hierarchical representation is important to distill emotion signals from audio data. The performance difference between `HR+DMF` and `WISE` suggests that `IM` indeed helps more effectively capture emotion evolutions by promoting word-level multimodal interactions, leading to improved performance. `IM+HR` is a simplified version of `WISE` without the attention layer. While not as prominent as `IM` or `HR`, temporal attention leads to additional benefits. Finally, we consider a variant of `HR` that removes the phoneme level from the hierarchy, denoted by `HR`$^-$. We can observe that having phonemes as the bridge between frames and words allows to eliminate emotionally irrelevant information from low-level audio information.

[2] All utterances labeled *excited* are merged into the *happy* category.

Table 1: *Performance comparison between our solution and state-of-the-art multimodal models on the IEMOCAP dataset*

Methods	WA	UA
`Audio only`	0.665	0.657
`Text only`	0.692	0.701
`LSTM+Attn` [10]	0.725	0.709
`CNN+Phoneme` [24]	0.739	0.685
`Dual RNNs` [2]	0.737	0.753
`WISE`	**0.759**	**0.764**

Table 2: *Performance of different variants of the WISE model*

Methods	WA	UA
`IM+DMF`	0.729	0.742
`HR+DMF`	0.734	0.741
`IM+HR`	0.752	0.736
`IM+DMF+HR`$^-$	0.743	0.754
`WISE`	0.759	0.764

5. Conclusion

In this paper, we studied the problem of SER by leveraging multimodal features from audio and text. In view of the limitation of existing methods that they normally fuse multimodal features only at the utterance level, we proposed an interaction mechanism that captures the dynamic interplay between audio and text features at the word level and that models the evolution of emotions as an utterance develops. To support the word-level representation of audio data, we devised a hierarchical representation to derive more emotionally relevant audio features. Based on these modules, we presented our WISE framework which further learns the temporally aggregated features via an attention mechanism. The experimental results show that our solution outperforms the best state-of-the-art methods. Since our solution relies on accurate ASR results, in future work we will investigate the impact of ASR performance on WISE in practical settings.

6. Acknowledgements

This work is supported by Fundamental Research Funds for the Central Universities (Grant No. 3072020CFT2402 and Grant No. 3072020CF0602) and NSFC Grant 61673202.

7. References

[1] E. Marchi, B. Schuller, A. Batliner, S. Fridenzon, S. Tal, and O. Golan, "Emotion in the speech of children with autism spectrum conditions: Prosody and everything else," in *Proceedings of the 3rd Workshop on Child, Computer and Interaction (WOCCI)*, 2012, pp. 17–24.

[2] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 112–118.

[3] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 247–251.

[4] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, pp. 2247–2256.

[5] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual inter-modal attention for multimodal sentiment analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 3454–3466.

[6] E. Kim and J. W. Shin, "DNN-based emotion recognition based on bottleneck acoustic features and lexical features," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6720–6724.

[7] J. Sebastian and P. Pierucci, "Fusion techniques for utterance-level emotion recognition combining speech and transcripts," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 51–55.

[8] Z. Lian, J. Tao, B. Liu, and J. Huang, "Conversational emotion analysis via attention mechanisms," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1936–1940.

[9] Y. Gu, X. Lyu, W. Sun, W. Li, S. Chen, X. Li, and I. Marsic, "Mutual correlation attentive factors in dyadic fusion networks for speech emotion recognition," in *Proceedings of the 27th ACM International Conference on Multimedia (MM)*, 2019, pp. 157–166.

[10] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 3569–3573.

[11] P. Tzirakis1, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5089–5093.

[12] J. Huang, Y. Li, J. Tao, and Z. Lian, "Speech emotion recognition from variable-length inputs with triplet loss function," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 3673–3677.

[13] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[14] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia (TMM)*, vol. 20, no. 6, pp. 1576–1590, 2018.

[15] 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 *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 272–276.

[16] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, and B. Schuller, "Towards temporal modelling of categorical speech emotion recognition," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 932–936.

[17] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 152–156.

[18] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6695–6699.

[19] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7390–7394.

[20] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 2578–2582.

[21] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227–2231.

[22] X. Huang, J. Baker, and R. Reddy, "A historical perspective of speech recognition," *Communications of the ACM*, vol. 57, no. 1, pp. 94–103, 2014.

[23] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[24] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 3688–3692.

[25] 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 (MM)*, 2010, pp. 1459–1462.

[26] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[29] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd Conference on International Conference on Learning Representations (ICLR)*, 2015.