INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Learning Fine-Grained Cross Modality Excitement for Speech Emotion Recognition

*Hang Li, Wenbiao Ding, Zhongqin Wu, Zitao Liu**

TAL Education Group, Beijing, China
{lihang4,dingwenbiao,wuzhongqing,liuzitao}@tal.com

Abstract

Speech emotion recognition is a challenging task because the emotion expression is complex, multimodal and fine-grained. In this paper, we propose a novel multimodal deep learning approach to perform fine-grained emotion recognition from real-life speeches. We design a temporal alignment mean-max pooling mechanism to capture the subtle and fine-grained emotions implied in every utterance. In addition, we propose a cross modality excitement module to conduct sample-specific adjustment on cross modality embeddings and adaptively recalibrate the corresponding values by its aligned latent features from the other modality. Our proposed model is evaluated on two well-known real-world speech emotion recognition datasets. The results demonstrate that our approach is superior on the prediction tasks for multimodal speech utterances, and it outperforms a wide range of baselines in terms of prediction accuracy. Further more, we conduct detailed ablation studies to show that our temporal alignment mean-max pooling mechanism and cross modality excitement significantly contribute to the promising results. In order to encourage the research reproducibility, we make the code publicly available at https://github.com/tal-ai/FG_CME.git.

Index Terms: speech emotion recognition, human-computer interaction, multimodal learning

1. Introduction

Emotion plays an important role in human communication. It has significant effects on information transmission and human interactions. The expression of emotion is usually multimodal, including voice, speech content, facial expressions, etc. With the development of machine learning and deep learning, a large number of speech interaction systems have emerged, such as voice assistants [1, 2, 3], intelligent tutoring systems [4, 5], etc. Speech emotion recognition is particularly useful to this kind of systems to better understand the content of the speech and generate a natural response based on the recognized emotion [6, 7, 8].

Various types of machine learning methods have been applied to improve the performance of speech emotion recognition. However, the majority of them are not sufficient for accurately detecting speech emotions from human due to the following two challenges. First, *emotion is multimodal*. Both the vocal sound and the linguistic content express emotions. It is insufficient to only consider information from single modality when conducting emotion recognition. For example, sentences expressed in different tones will show different emotions even with the same content. Second, *emotion expression is fine-grained*. The emotion of a sentence is often expressed by specific words or voice fragments. The utterance-level emotion recognition approaches treat every words equally and may fail to capture fine-grained and subtle emotions.

Recently, a large number of approaches have been developed to address above challenges [9, 10, 11, 12]. Researchers have jointly combined multiple modalities for speech emotion recognition and achieved better results compared to methods that only consider information from single modality [9, 11]. For example, Schuller et al. combined acoustic features and linguistic information in a hybrid architecture to identify emotional key phrases, and assessed the emotional salience of verbal cues from both phoneme sequences and words [9]. Yoon et al. built deep neural networks to learn vocal representations and text representations and concatenated them for emotion classification [11]. In spite of the success of above methods, there is still large room for improvement. Most existing approaches are based on utterance-level fusion, such as directly concatenating acoustic and semantic features. These methods pay little attention to fine-grained multimodal emotional information and are not able to capture subtle emotions in real-life speech conversations. Therefore, it is very necessary to conduct fine-grained learning of spoken speeches and make full use of relationships between acoustic and semantic modalities.

To overcome above challenges, we propose a novel multimodal neural framework to perform fine-grained emotion recognition. Instead of simply using utterance-level features, we propose a temporal alignment mean-max pooling operation and a cross modality excitation mechanism to capture the interrelations between different modalities. We have conducted extensive experiments on two public available emotion recognition datasets, i.e., Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [1] and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [2]. The results show that our method outperforms all baselines.

Overall this paper makes the following contributions:

- We conduct fine-grained learning with a temporal alignment mean-max pooling operation and cross-modality excitement mechanism to obtain plentiful cross modality information from both voice fragments and sentences.
- We design and evaluate our approach quantitatively by using two well-known real-world emotion recognition datasets. Besides, we present detailed ablation studies to demonstrate the effectiveness of each component in our proposed model.

2. Related Work

Speech emotion recognition has been studied for decades in both the machine learning and speech communities [13, 14, 15, 16, 17]. Following the mainstream direction, researchers have

* corresponding author

[1] https://sail.usc.edu/iemocap/
[2] https://smartlaboratory.org/ravdess/

http://dx.doi.org/10.21437/Interspeech.2021-158

extracted hand-craft features from audio data and applied them to classic supervised learning methods such as support vector machines [15], hidden Markov models [13], Gaussian mixture models [17], Tree-Based models [16], etc.

With recent advances of deep learning, various neural networks have been used in the task of speech emotion recognition [14, 18, 19, 20]. Han et al. extracted speech signal features from audio and used deep neural networks to classify speech emotion [18]. Badshah et al. extracted raw spectrograms features and used convolutional neural network (CNN) to extract high-level features to build the model [19]. Due to the sequential structure of audio, Trigeorgis et al. proposed an end-to-end model that employs long short-term memory (LSTM) network to capture the contextual information [14]. Mirsamadi et al. used recurrent neural networks with local attention mechanism to automatically discover emotionally relevant features from speech and provided more accurate predictions [20].

In addition to tackle the speech emotion recognition problem by directly applying deep neural networks on audio datasets, multimodal learning frameworks that jointly consider emotions implied in different modalities [10, 21, 22]. For instance, Tzirakis et al. provided a multimodal system to perform an end-to-end spontaneous emotion prediction task from speech and visual data [21]. Lee et al. proposed an attention model for multimodal emotion recognition from speech and text data, and provided an effective method to learn the correlation between two output feature vectors from separate yet jointly trained CNNs [22].

The closest work to our research is Xu's [10], where they proposed a fine-grained method to learn alignment between the original speech and the recognized text with attention mechanism. Our approach is different from that: (1) we propose a temporal mean-max alignment pooling method to combine each word and its corresponding voice instead of attention mechanism; and (2) we design a cross modality excitement module to enhance the interactions between the two modalities.

3. Our Approach

The architecture of our purposed model is shown in Figure.1, which is composed of four key components: (1) unimodal embedding module, which captures the acoustic and semantic information respectively; (2) temporal alignment mean-max pooling, which aggregates the acoustic embedding for each word based on its corresponding speech frames, (3) cross modality excitement module, which converts unimodal information into a fine-grained multimodal representation through the cross modality excitement mechanism; and (4) multimodal prediction, which utilizes a bi-directional LSTM network to model the sequential information within the entire speech sentence.

3.1. Unimodal Embedding

3.1.1. Acoustic Embedding

For each utterance, we first transform it into n frames $\{f_i\}_{i=1}^{n}$ of width 25ms and step 10ms. We extract low-level features, i.e., $\mathbf{x}_i^a \in \mathbb{R}^p$, from each frame f_i and the utterance-level feature is $\mathbf{X}^a \in \mathbb{R}^{p\times n}$, $\mathbf{X}^a = [\mathbf{x}_1^a, \cdots, \mathbf{x}_n^a]$. Then we extract the context-aware acoustic embedding $\hat{\mathbf{X}}_1^a \in \mathbb{R}^{q\times n}$ by feeding $\mathbf{X}^a$ into a multi-layer 1-d CNN, i.e., $\hat{\mathbf{X}}_1^a = \mathbf{CNN}(\mathbf{X}^a)$. p and q represent the dimensions of low-level features and the extracted acoustic representations respectively.

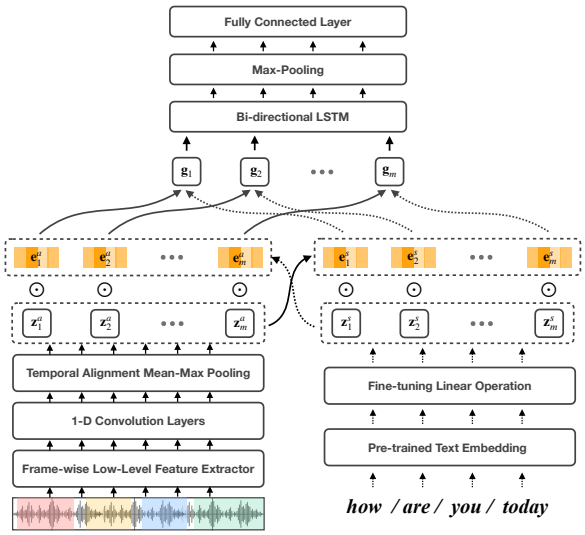

Figure 1: *The proposed neural framework.*

3.1.2. Semantic Embedding

For each utterance, we obtain its transcription by an automatic speech recognition (ASR) service. Then, we extract the linguistic representation $\mathbf{X}^s \in \mathbb{R}^{l\times m}$ via pre-trained language models [23]. m denotes the number of words in the ASR transcription. After that, we apply a linear transformation operator $\mathbf{W}^s \in \mathbb{R}^{t\times l}$ on $\mathbf{X}^s$ to obtain the fine-tuned semantic embeddings $\mathbf{Z}^s \in \mathbb{R}^{t\times m}$, i.e., $\mathbf{Z}^s = \mathbf{W}^s\mathbf{X}^s$, which are used in the downstream emotion recognition task.

3.2. Temporal Alignment Mean-Max Pooling

To capture the subtle emotion expression implied in the utterance, we conduct the fine-grained recognition by extracting the underlying word-level acoustic representation. Hence, we propose a temporal alignment mean-max pooling to aggregate the frame-level acoustic representation $\hat{\mathbf{X}}^a$. Instead of the simply averaging all the frame-level acoustic embeddings like mean pooling operator, our temporal alignment pooling operator uses a word-level binary alignment matrix $\mathbf{A}$ to only select relevant frames' acoustic features for each corresponding word. The binary alignment matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ is a block-wise diagonal matrix, which is calculated by taking an orthographic transcription of an audio file and generating a time-aligned version using a pronunciation dictionary to look up phones for words. The temporal aligned word-level acoustic representation $\mathbf{Z}^a = [\mathbf{z}_1^a, \cdots, \mathbf{z}_m^a]$ can be obtained as follows:

$$\hat{\mathbf{z}}_j^a = \text{meanpool}(\{\mathbf{a}_{i,j} \cdot \hat{\mathbf{x}}_i^a \mid i = 1 \cdots N\})$$

$$\tilde{\mathbf{z}}_j^a = \text{maxpool}(\{\mathbf{a}_{i,j} \cdot \hat{\mathbf{x}}_i^a \mid i = 1 \cdots N\})$$

$$\mathbf{z}_j^a = [\hat{\mathbf{z}}_j^a, \tilde{\mathbf{z}}_j^a]$$

where $a_{i,j}$ represents the (i,j)th entry of alignment matrix A, $\mathbf{z}_j^a$ represents the acoustic feature for jth word, $[\cdot,\cdot]$ denotes the concatenation operation.

3.3. Cross Modality Excitation

To learn the nonlinear interactions between acoustic and semantic modalities, similar to [24], we develop a cross modality ex-

citation (CME) module. The CME module serves as sample-specific activations and is learned for each embedding dimension by a self-gating mechanism based on dimensional dependence. Specifically, we employ a simple gating mechanism with a sigmoid activation, i.e., $\mathbf{E}^a = \delta(\mathbf{W}^a\mathbf{Z}^s), \mathbf{E}^a \in \mathbb{R}^{q\times m}$ and $\mathbf{E}^s = \delta(\mathbf{W}^s\mathbf{Z}^a), \mathbf{E}^s \in \mathbb{R}^{t\times m}$, where $\delta(\cdot)$ denotes the sigmoid function and $\mathbf{W}^a \in \mathbb{R}^{q\times t}, \mathbf{W}^s \in \mathbb{R}^{t\times q}$ represent linear projection operator for the aligned acoustic and semantic representations respectively. $\mathbf{E}^a$ and $\mathbf{E}^s$ are the excitation matrices, which act as the weights adapted to the corresponding features. Then the temporal aligned word-level acoustic and semantic representations are adaptively calibrated by the corresponding cross modality excitation matrices and the cross modality adapted representations are computed as $\bar{\mathbf{Z}}^a = \mathbf{E}^a \odot \mathbf{Z}^a$ and $\bar{\mathbf{Z}}^s = \mathbf{E}^s \odot \mathbf{Z}^s$, where $\odot$ represents the element-wise product.

3.4. Multimodal Prediction

Let $\bar{\mathbf{z}}_i^a$ and $\bar{\mathbf{z}}_i^s$ be the acoustic and semantic representations of ith word in the original utterance, i.e., $\bar{\mathbf{Z}}^a = \{\bar{\mathbf{z}}_i^a\}_{i=1}^m$ and $\bar{\mathbf{Z}}^s = \{\bar{\mathbf{z}}_i^s\}_{i=1}^m$. Each word's multimodal representation $\mathbf{g}_i$ is computed by concatenating $\bar{\mathbf{z}}_i^a$ and $\bar{\mathbf{z}}_i^s$, i.e., $\mathbf{g}_i = [\bar{\mathbf{z}}_i^a, \bar{\mathbf{z}}_i^s]$. In the multimodal prediction layer, we utilize a bi-directional LSTM to capture the word-level sequential patterns on top of the multimodal representation of each word $\mathbf{g}_i$. The resulting hidden state of $\mathbf{h}_i$ is the concatenation of hidden representations from both directions, i.e., $\mathbf{h}_i = [\overrightarrow{\mathbf{h}_i}, \overleftarrow{\mathbf{h}_i}]$. After that, a max-pooling layer is applied to aggregate the sequential information for all the hidden states in the utterance sequence, i.e., $\mathbf{h}_i^* = \mathrm{maxpool}(\{\mathbf{h}_i\}_{i=1}^m)$. Finally, we use a two-layer fully-connected feed forward network (FCN) to conduct the final predictions, i.e.,

$$\mathbf{p}_i = \mathrm{softmax}(\mathrm{FCN}(\mathbf{h}_i^*)) \tag{1}$$

where $\mathbf{p}_i$ is the probabilistic vector that indicates the final probabilities of class memberships of utterance i. In this work, we use the multi-class cross-entropy loss to optimize the prediction accuracy, which is defined as follows:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log p_{i,k} \tag{2}$$

where $p_{i,k}$ is the kth element of $\mathbf{p}_i$, and $y_{i,k} = 1$ if the ith sample belongs to the kth class.

4. Experiments

4.1. Data

To assess our proposed framework, we conduct several experiments with two open source English emotion datasets: Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [25] [3] and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [26] [4].

4.1.1. IEMOCAP

IEMOCAP consists of five sessions of dyadic conversations between different pairs of actors. The total number of utterances in IEMOCAP is 10,039. For each utterance, the audio, transcriptions, video and actor's motion-capturing recordings are collected and emotion labels are annotated by the majority voting results from three experienced evaluators. Following the similar experiment settings with prior studies [10, 11], we use four emotions (*angry*, *happy*, *neutral* and *sad*) for classification evaluation. Besides that, we perform the five-fold cross validation and the average results are reported.

4.1.2. RAVDESS

RAVDESS is a validated multimodal database of emotional speeches and songs, which includes 9 hours of speeches and 1,440 utterance. Unlike the IEMOCAP, the speech content of each utterance in RAVDESS is the same and actors express the different emotions through their different tones. The label of each utterance is validated by multiple raters and the intensities of emotions are conducted based on raters' responses on emotion strength. In our experiment, we use the same four emotions (*angry*, *happy*, *neutral* and *sad*) for classification. The five-fold cross validation is also incorporated for the robustness of the final conclusions.

4.2. Baselines

To evaluate the effectiveness of our proposed framework, we carefully choose the following state-of-the-art emotion classification approaches as our baselines: (1) **Acoustic-CNN+LSTM**($\mathrm{CNN} + \mathrm{LSTM}^a$): We train the similar model proposed by Satt et al., which uses CNN for extracting acoustic embeddings and applies the LSTM layer at the top for aggregating utterance embedding [27]. (2) **Acoustic-LSTM+Attention**($\mathrm{LSTM} + \mathrm{Attn}^a$): A bi-directional LSTM is used for generating the acoustic embeddings from the frame-wise low-level features. An attention layer is used for the final aggregation [20]. (4) **Semantic-LSTM+Attention**($\mathrm{LSTM} + \mathrm{Attn}^s$): Similar to $\mathrm{LSTM} + \mathrm{Att}^a$) but instead only the text features are used. (5) **Multimodal-Utterance+Concat**($\mathrm{UttConcat}^m$): Two individual bi-directional LSTMs are used as unimodal encoders for acoustic and semantic features and the unimodal embeddings are concatenated at utterance level for the final prediction [11]. (6) **Multimodal-AttentionFusion**($\mathrm{AttnFusion}^m$): Similar to $\mathrm{Utt} + \mathrm{Concat}^m$ but instead applying the multi-head attention for fusing the multimodal features at word level [10]. (7) **Multimodal-MultihopAttention**(MHA^m): Similar to $\mathrm{UttConcat}^m$ and $\mathrm{AttnFusion}^m$, except that a multi-hop attention mechanism is used for the inference of correlations between two modalities [28].

4.3. Implementation Details

For the acoustic embeddings, we first transform sample's audio signal into frames with 25ms and step size 10ms. Then, we use a Python Library [29] to extract the 40-dimensional filterbank features from each frame. The 1-d CNN network we used for embedding the acoustic features has 3 layers with kernel sizes (5, 2, 2) and the stride size of each kernel is set to 1 for keeping the length of the frame sequences. The number of the filters are set to (64,128,128) and the dimension of the output acoustic embedding $\mathbf{Z}^a$ is 256, which is generated by concatenating the corresponding word level mean-pooling and max-pooling embeddings.

For the semantic embeddings, we use a 300-dimensional pre-trained GloVe embedding [23] for mapping the words into the fixed-length vectors for each utterance transcription. The dimension of linear operator $\mathbf{W}^s$ we used for fine-tuning the pre-trained word embeddings is 256. To implement the model, we use 256 hidden units in bi-directional LSTM, i.e. d_{h_i}, in the

[3] https://sail.usc.edu/iemocap/

[4] https://smartlaboratory.org/ravdess/

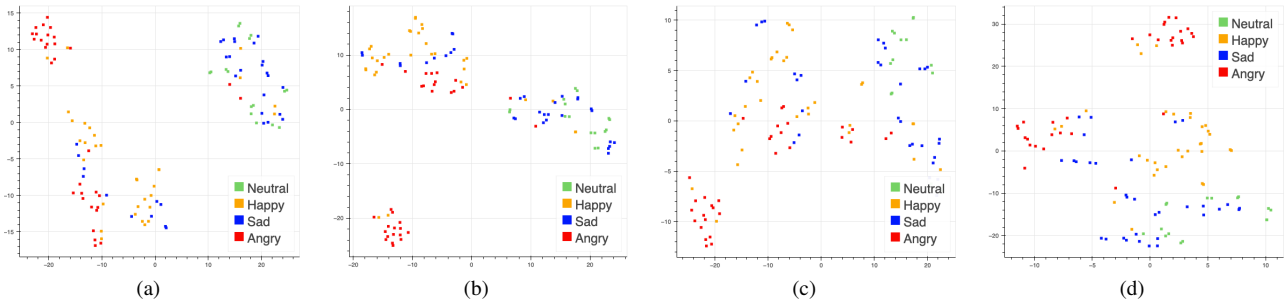

Figure 2: *Visualization of embeddings for validation samples from RAVDESS with different ablated components. (a) w/o both excitement (b) w/o semantic excitement (c) w/o acoustic excitement (d) with both excitement*

multimodal sequence feature extracting layer and the neurons of the two-layer full-connected layer $\mathrm{FCN}(\cdot)$ is (128,4). To train our model, we use Adam optimizer [30] with learning rate of 0.0005.

4.4. Results & Analysis

We evaluate and compare the performance of different method based on two widely used metrics for emotion classification: weighted accuracy (WA) that is the overall classification accuracy and unweighted accuracy (UA) that is the average recall over the emotion categories.

Table 1: *Experimental results on* IEMOCAP *and* RAVDESS *datasets.* UA *and* WA *indicate the unweighted accuracy and weighted accuracy*

	IEMOCAP		**RAVDESS**	
	WA	UA	WA	UA
Acoustic only				
$\mathrm{CNN} + \mathrm{LSTM}^a$	0.635	0.588	0.640	0.611
$\mathrm{LSTM} + \mathrm{Attn}^a$	0.573	0.574	0.665	0.627
Semantic only				
$\mathrm{LSTM} + \mathrm{Attn}^s$	0.640	0.647	–	–
Semantic + Acoustic				
$\mathrm{UttConcat}^m$	0.700	0.677	0.629	0.615
$\mathrm{AttnFusion}^m$	0.712	0.725	0.634	0.593
MHA^m	0.678	0.688	0.615	0.572
Ablation Studies				
w/o Excite^a	0.713	0.722	0.703	0.683
w/o Excite^s	0.721	0.732	**0.720**	0.696
w/o Excite^m	0.712	0.720	0.718	0.700
w/ CTC Align	0.657	0.688	0.582	0.566
Ours	**0.727**	**0.735**	**0.720**	**0.708**

The results of our experiments show that our approach outperforms all other methods on both IEMOCAP and RAVDESS. Specifically, from Table.1, we find the following results: (1) When comparing the unimodal based models ($\mathrm{CNN} + \mathrm{LSTM}^a$, $\mathrm{LSTM} + \mathrm{Attn}^a$,$\mathrm{LSTM} + \mathrm{Attn}^s$) with the multimodal ones ($\mathrm{UttConcat}^m$, $\mathrm{AttnFusion}^m$, MHA^m), we find that the multimodal features provide a huge boost for the models' performance on both WA and UA. (2) Comparing $\mathrm{UttConcat}^m$ and MHA^m with $\mathrm{AttnFusion}^m$, the fusion of multimodal features in a fine-grained manner provides the model with a strong capability to capture the subtle emotions expressed within the utterance level.

4.5. Ablation Studies

Apart from applying the comparisons above, we also present the ablation results of different key components in our proposed algorithm in Table.1. Overall, each component of our proposed model plays an important role in improving the model's performance on both datasets. For model without acoustic excitement (w/o Excite^a), we find its performance drops dramatically in RAVDESS, which is consistent with our expectations since samples in the RAVDESS share the similar text inputs. Comparing model without any excitement (w/o Excite^m) with model without either acoustic excitement or semantic excitement (w/o Excite^s), we observe that the model's performance is boosted by both excitements and the best performance is achieved by using both excitement modules simultaneously. At last, we replace our pre-calculated word-level binary alignment matrix A with the CTC predicting alignment matrix (w/ CTC Align) [31], we find the model's performance suffers a great loss, which indicates that the pre-calculated alignment matrix A is more compatible with our proposed algorithm.

To provide a better view about the effectiveness of each components, we use t-SNE to visualize the embeddings of validation samples from RAVDESS generated by different models in Fig.2. From the figure, we observe that embeddings of different emotion category's samples from the model without both excitement modules is hard to be separated. After employing the CME block, the intra-distance between different categories is enlarged, which makes it easy for the final classifier to generate a robustness split boundaries for each class.

5. Conclusion

In this paper, we propose a novel multimodal deep learning framework to perform fine-grained learning of voice and text in speeches. Through applying the temporal alignment operator combining with the cross modality excitation module, we successfully generate a powerful multimodal representation for each utterance in a fine-grained manner. The experiments on two open source English emotion datasets: IEMOCAP and RAVDESS demonstrate the effectiveness of our purposed algorithm. Besides, the influence of each component of our model is presented via the detailed ablation studies and analysis.

6. Acknowledgements

This work was supported in part by National Key R&D Program of China, under Grant No. 2020AAA0104500 and in part by Beijing Nova Program (Z201100006820068) from Beijing Municipal Science & Technology Commission.

7. References

[1] M. B. Hoy, "Alexa, siri, cortana, and more: an introduction to voice assistants," *Medical reference services quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[2] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *2016 IEEE ICASSP*. IEEE, 2016, pp. 5805–5809.

[3] L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 2666–2670.

[4] C. Tekin, J. Braun, and M. van der Schaar, "etutor: Online learning for personalized education," in *ICASSP*. IEEE, 2015, pp. 5545–5549.

[5] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *ICASSP*. IEEE, 2015, pp. 5058–5062.

[6] M. Ostendorf, E. Shriberg, and A. Stolcke, "Human language technology: Opportunities and challenges," in *ICASSP*, vol. 5. IEEE, 2005, pp. v–949.

[7] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *ICASSP*. IEEE, 2015, pp. 4749–4753.

[8] H. Arsikere, E. Shriberg, and U. Ozertem, "Computationally-efficient endpointing features for natural spoken interaction with personal-assistant systems," in *ICASSP*. IEEE, 2014, pp. 3241–3245.

[9] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *ICASSP*, vol. 1. IEEE, 2004, pp. I–577.

[10] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," *arXiv preprint arXiv:1909.05645*, 2019.

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

[12] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in *ICASSP*. IEEE, 2019, pp. 5866–5870.

[13] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *ICASSP*, vol. 2. IEEE, 2003, pp. II–1.

[14] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*. IEEE, 2016, pp. 5200–5204.

[15] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using support vector machines," in *2013 5th international conference on Knowledge and smart technology (KST)*. IEEE, 2013, pp. 86–91.

[16] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[17] M. M. El Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using gaussian mixture vector autoregressive models," in *ICASSP*, vol. 4. IEEE, 2007, pp. IV–957.

[18] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[19] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*. IEEE, 2017, pp. 1–5.

[20] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *ICASSP*. IEEE, 2017, pp. 2227–2231.

[21] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[22] C. W. Lee, K. Y. Song, J. Jeong, and W. Y. Choi, "Convolutional attention networks for multimodal emotion recognition from speech and text data," *ACL 2018*, p. 28, 2018.

[23] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of EMNLP*, 2014, pp. 1532–1543.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

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

[26] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess)," p. e0196391, Apr 2018.

[27] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *Interspeech*, 2017, pp. 1089–1093.

[28] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2822–2826.

[29] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PLOS ONE*, vol. 10, no. 12, pp. 1–17, 12 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0144610

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

[31] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.