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

Improving Speech Emotion Recognition Using Graph Attentive Bi-directional Gated Recurrent Unit Network

Bo-Hao Su[1,2], *Chun-Min Chang*[1,2], *Yun-Shao Lin*[1,2], *Chi-Chun Lee*[1,2]

[1]Department of Electrical Engineering, National Tsing Hua University
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare

borrissu@gapp.nthu.edu.tw, cmchang@gapp.nthu.edu.tw, astanley18074@gmail.com, cclee@ee.nthu.edu.tw

Abstract

The manner that human encodes emotion information within an utterance is often complex and could result in a diverse salient acoustic profile that is conditioned on emotion types. In this work, we propose a framework in imposing a graph attention mechanism on gated recurrent unit network (GA-GRU) to improve utterance-based speech emotion recognition (SER). Our proposed GA-GRU combines both long-range time-series based modeling of speech and further integrates complex saliency using a graph structure. We evaluate our proposed GA-GRU on the IEMOCAP and the MSP-IMPROV database and achieve a 63.8% UAR and 57.47% UAR in a four class emotion recognition task. The GA-GRU obtains consistently better performances as compared to recent state-of-art in per-utterance emotion classification model, and we further observe that different emotion categories would require distinct flexible structures in modeling emotion information in the acoustic data that is beyond conventional *left-to-right* or vice versa.

Index Terms : speech emotion recognition, graph, attention mechanism, recurrent neural network

1. Introduction

Using speech as the main communication medium has become prevalent in a variety of commercialized applications, e.g., companion robots [1], voice assistants [2], and autonomous agents [3]. The ability to further extract emotional content from speech beyond linguistic message has sparked a large body of works in speech emotion recognition (e.g., [4, 5]). With the surge of deep learning techniques, many speech emotion recognition (SER) models used a variety of network architectures inspired from different research domains, e.g., computer vision and natural language processing, resulting in an improved recognition accuracy beyond conventional machine learning methods. For example, Hazarika et al. developed a deep neural network with memory cell [6], and Suping et al. proposed a semi-supervised multi-path generative neural network [7] for SER tasks, and also Siddique et al. proposed a variational autoencoder based LSTM model [8] as an improved SER model.

The natural course of speech signal results in many variants of SER algorithms to use (B)LSTM network as the main component in capturing emotionally-relevant information of an utterance either in a *left-to-right* and/or *right-to-left* manner, i.e., similar to automatic speech recognition (ASR). Such a mechanism follows closely the development of RNN model, i.e., from simple RNN to complexly considering the bidirectional gated LSTM model. Several recent SER research in this regard includes: Yeh et al. [9] utilized a bidirectional RNN structure to model the temporal relationship between the interlocutors over the interaction; Zadeh et al. proposed to use multi-attention LSTM-based (MARN) model [10] to aggregate multimodal information within an utterance. Han et al. [11] utilized a framework based on LSTM claiming to effectively model emotion over time in an utterance. These works also point to the robustness of SER performances when considering temporal information in the speech data.

While speech, as a time-series signal, could intuitively be modeled as a temporal sequence of local acoustic descriptors to obtains competitive SER performances, both theoretical and empirical evidences have stated that different types of emotions would result in a diverse profile of acoustic manifestation that is beyond *left-to-right* and/or vice versa. In fact, psychologists have stated that emotion is induced through successive self-realizations, which indicates the moment of emotional arousal is critical and its expressive process can either be discrete or continuous [12]. The variable intonation profiles exist between different emotion types has particularly been well documented [13]; Busso et al. also demonstrated that there exists emotionally salient aspect within an utterance by examining different acoustic features [14]. While (B)LSTM based model provides an effective mean of capturing long range time-series information in speech signal, its structure could limit the emotion modeling capacity and likely ignores potential non-sequential yet emotionally-salient information.

In this work, we propose a novel architecture of graph attentive bi-directional gated recurrent neural net (GA-GRU). Instead of only considering the *left-to-right* or *right-to-right* information in the BiGRU model, we further emphasize the structural relationship within a sentence, i.e., imposing a within-utterance frame-wise graph structure that modulates the attention weights on a bi-directional GRU's hidden sequences, to further enable complex modeling of emotion modulation on acoustic profile. GA-GRU combines both the power of attentional long range time series modeling with the salient frame-wise graph structure within an emotional utterance. We evaluate GA-GRU on two widely used benchmark large emotion corpora, the IEMOCAP [15] and the MSP-IMPROV [16]. GA-GRU obtains not only an improved UAR over attention BiGRUs in the 4 emotions classification task (Happiness, Neutrality, Sadness and Anger), but also achieves a state-of-the-art per-utterance emotion recognition rates compared to the well-known recent works. Interestingly, we also observe that emotion class, such as sadness, require a non-sequential graph structure which is critical in improving the recognition performance.

2. Research Methodology

2.1. Dataset and Acoustic Features

In the following sections, we will briefly introduce the two datasets and the acoustic features used in this work.

http://dx.doi.org/10.21437/Interspeech.2020-1733

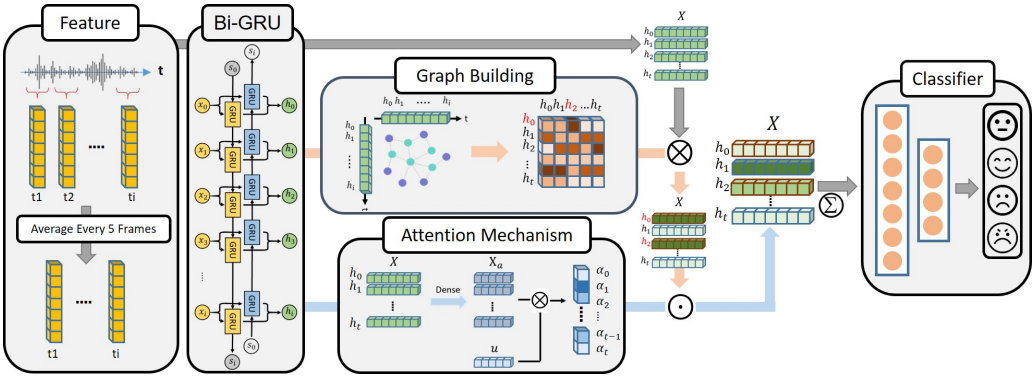


Figure 1: *Architecture of GA-GRU, the input is 78 dimension Emobase LLDs and graph is built by the hidden output of Bi-GRU; After graph attention mechanism, the representation will be classified into 4 categories.*

2.1.1. The IEMOCAP and The MSP-IMPROV

The IEMOCAP and the MSP-IMPROV are affective dyadic interaction English databases. Each session including two actors (1 male and 1 female), and totally 12 hours (five sessions) and 9 hours (six sessions) in the IEMOCAP and the MSP-IMPROV respectively. All utterances were annotated with categorical emotion labels, in this paper, we choose four major emotion categories: neutrality, happiness(including excited), sadness and anger, which results in 5531 utterances in the IEMOCAP and 7798 utterances in the MSP-IMPROV.

2.1.2. Acoustic Low-level Descriptors

We extract 78 dimensional frame level acoustic descriptors using the openSMILE toolkit[17] with the Emobase LLDs config file; it includes low-level descriptors of PCM loudness, Mel-frequency cepstral coefficients (MFCCs), LSP Frequency, F0 Envelope, jitter etc [18]. Per-speaker z-normalization is conducted on all the features, and in order to decrease computation time, we further average the features of every 5 frames to downsample the frame number.

2.2. Graph Attentive Bidirectional GRU (GA-GRU)

In this paper, we propose a graph attentive bidirectional GRU (GA-GRU), which uses attention mechanism to reweight the important time segments and incorporate the use of canonical graph to integrate the cross-frame relationship. The temporal information is captured by a BiGRU network, attention mechanism is applied to enhance the salient segment, and graph enables such saliency to be connected across time steps. We will describe each component in the following.

2.2.1. Bidirectional Gated Recurrent Unit Network

Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are the two most common used recurrent neural networks for time series; however, the computation complexity is much lower for GRU than LSTM, and the performance is often comparable. Bidirectional GRU is used as the basic building block of our GA-GRU network. The input feature of i-th utterance and time step t is encoded into two hidden vectors (one forward and one backward) as the following,

$$\overleftrightarrow{h_{i,t}} = \overleftrightarrow{GRU}(X_{i,t}), t \in [1, T] \tag{1}$$

$$h_i = \overrightarrow{h_{i,t}} + \overleftarrow{h_{i,t}} \tag{2}$$

where X is the feature vector of an utterance, T is the terminated time step of each utterance. In order to reduce the dimension of our features, we sum the two vectors derived from forward and backward GRUs.

2.2.2. Attention Mechanism

Attention mechanism is widely used in time sequence models in order to emphasize those time steps that are discriminative for the given task with learnable weights. In specifics, we define a score function $s(\cdot)$ to transform the hidden dimensions of BiGRU to attention hidden dimensions,

$$s(h_i) = tanh(Linear(h_i)), s(h_i) \in \mathbb{R}^{B \times T \times AT} \tag{3}$$

$$\alpha_t = \frac{exp(u^T \cdot s(h_i))}{\sum_{i=1}^{T} exp(u^T \cdot s(h_i))} \tag{4}$$

where B is batch size, T is time step, AT is attention hidden dimensions, $u \in \mathbb{R}^{1 \times AT}$ is the trainable parameter (which is also a normal distribution with mean equals to zero), and h_i is the feature representation of the Si^{th} utterance.

2.2.3. Graph Modulated Attention

Our proposed use of graph modulated attention is inspired by the recent success of graph convolutional neural network (GCN) [21], which learns a graph-constrained image representation for improved classification task in computer vision [22]. Inspired by GCN, we construct a graph from the time sequence hidden dimensions of BiGRU to integrate emotion saliency not only sequentially but with cross time connection. Specifically, we first build a graph that encodes the distances (dot products) between hidden dimensions of every time frame within an utterance as the edge strength:

$$G^l_{i,j} = \frac{||ReLU(h^l_i {h^l_j}^T)||^2}{\sum_{i'} ||ReLU(h^l_{i'} {h^l_j}^T)||^2 + \epsilon} \tag{5}$$

where h^l_i is the i^{th} time step of l^{th} utterance and ϵ is a small numerical value to avoid division by zero. In the denominator, we sum it over the entire row to make sure i^{th} time step relates to all other time steps has sum equals to 1. We construct this graph at each batch and multiply it through the hidden dimensions with attention before finally passing this graph-enhanced

	IEMOCAP								
	SVM	DNN	CMN	MDNN	CNN-LSTM-DNN[19]	BiGRU+Att	ASRNN[20]	CVAE-LSTM[8]	GA-GRU
Happiness	44.32	52.2	27.78	-	-	59.66	-	-	53.73
Neutrality	52.87	58.61	68.16	-	-	56.56	-	-	58.9
Sadness	73.8	65.13	57.81	-	-	67.34	-	-	71.96
Anger	66.55	61.02	92.94	-	-	66.18	-	-	70.63
WA	56.88	58.47	65.3	61.8	-	61.51	-	-	62.27
UA	59.38	59.24	61.7	62.7	60.23	62.44	62.6	62.8	**63.8**

Table 1: *The table shows the unweighted average recall (UA), weighted accuracy (WA) and each emotion class's recall rate obtained from the baseline model and our proposed model in the IEMOCAP.*

attentive embeddings through the recognition layer,

$$H^l = \sum_t G^l h_t^l \alpha_t^l \tag{6}$$

where H^l is the final embedding of l^{th} sentence after graph-based attention mechanism, and t is the sequence length of each sentence. After obtaining the final embedding in the graph attentive BiGRU, this embedding is fed into common dense layer followed by a softmax layer,

$$Y^l = Softmax(Linear(H^l)) \tag{7}$$

where Y^l is the predicted emotion of the l^{th} utterance.

3. Experimental Setup and Results

3.1. Experimental Setup

The detail settings of our model are as below: there are two layers of BiGRU and each is with 256 hidden dimension, 0.1 dropout rate, and the final emotion recognition layer is one linear layer. The hidden dimension of attention layer is set to 16, epsilon of graph is $1e-15$. Learning rate of both databases is $2e-4$, batch size of these two corpora are both 64 as well. Adam optimizer, crossentropy loss criterion and early stopping are applied on both databases. Due to a more extreme label imbalance condition for the MSP-IMPROV, we further upsample the sadness and anger class by random duplication to balance the overall four class distribution. Finally, to be consistent in comparing with past works, we perform leave-one-person-out cross validation and report unweighted average recall (UAR) and accuracy (WA) as our metric.

3.2. Comparison Models

The following recent per-utterance speech modeling methods for emotion recognition that are used to compare the performance with our proposed model. Only the CMN network is re-implemented by the release source code and other models' results are referring to reference paper.

- **Support Vector Machine (SVM)**: We extract the 88 dimensional eGemaps functional features per utterance and use support vector machine (SVM) with linear kernel (C equals to 1 and class weight parameters is set as balanced).
- **Deep Neural Network (DNN)**: We build 3 dense layers with ReLU activation function (input features are also eGemaps). Early stopping scheme is also utilized.
- **CMN**: This architecture was proposed by Hazarika in 2018, the memory cell was used to retain the conversational context information to help improve emotion recognition [6].
- **MDNN**: This architecture was proposed by Zhou in AAAI 2018, which was a deep neural network that comprise of multiple local classifiers and aggregation of all the local information to a global emotion classifier [7].
- **CNN-LSTM-DNN**: This structure applies convolutional neural network (CNN) to extract features from the speech signal, followed by a LSTM layer which captures the temporal information and aggregates them using a fully connected layer in the final recognition. This method considers the local saliency through CNN and also the global information from LSTM. The model was proposed and applied successfully on categorical emotion recognition in 2019 [19].
- **ASRNN**: This model was proposed in 2020 that combines the 3D convolution in the front-end structure which integrates the information from temporal and spectral to improve the recognition accuracy. A bidirectional LSTM and attention mechanism is applied that helps in strengthening the time step importance through the use of attention weight and then finally predict with a dense layer [20].
- **CVAE-LSTM**: Siddique proposed a conditional VAE in 2017 [8] by using the emotion category as a condition placed on each frame, and this condition enforces the VAE to generate a more emotionally-consistent representation. The performance surpasses conventional VAE and attentive-CNN model.
- **BiGRU+ATT**: This is a BiGRU with attention but without our proposed graph structure. The input feature is 78 dimensions of Emobase, and all other settings are identical to our use of BiGRU with attention.

3.3. Results and Analysis

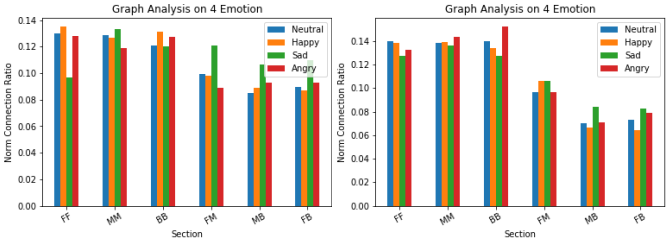
3.3.1. Performance Comparison

Table 1 and Table 2 summarize all of the recognition results in two databases. We see that overall our proposed GA-GRU model performs the best, i.e., 63.8 and 57.47, in the IEMOCAP and the MSP-IMPROV database, respectively. We observe that the performance of our GA-GRU is 1.1 better than the model of BiGRU with attention, MDNN, ASRNN and even CAVE-LSTM, and 4.42 higher than others in the IEMOCAP database. The improvement is much more substantial in the MSP-IMPROV database; the UAR of our GA-GRU is 5.03 higher than CNN-LSTM-DNN and BiGRU with attention, 1.77 higher than ASRNN, and it surpasses the original baseline paper of the MSP-IMPROV by 16.07 [16].

We additionally observe that by including graph structure to enhance the original self attention mechanism, the improvement

	MSP-IMPROV						
	Baseline[16]	SVM	DNN	CNN-LSTM-DNN[19]	BiGRU+Att	ASRNN[20]	GA-GRU
Happiness	-	47.09	38.58	-	54.43	-	52.43
Neutrality	-	45.1	46.07	-	54.5	-	**59.83**
Sadness	-	61.13	61.13	-	47.68	-	52.54
Anger	-	57.95	62.37	-	52.4	-	**65.15**
WA	-	48.9	46.9	-	53.49	-	56.21
UA	41.4	52.82	52.04	52.43	52.25	55.7	**57.47**

Table 2: *The table shows the unweighted average recall (UA), weighted accuracy (WA) and each emotion class's recall rate obtained from the baseline models and our proposed model in the MSP-IMPRO.*



(a) The IEMOCAP (b) The MSP-IMPROV

Figure 2: *Edge connection between three different time partitions of an utterance for the IEMOCAP and the MSP-IMPROV. F represents the beginning 1/3 part, M is mid part, and B is last 1/3. FF means connection within F partition, and FB means connection occurs between F and B and so on.*

in the recognition rates come from similar emotion classes in both the IEMOCAP and the MSP-IMPROV database. Specifically, it improves 4.62, 4.45 for sadness and anger and 2.34 for neutrality for the IEMOCAP. For the MSP-IMPROV, sadness improves 4.86, and anger improves 12.75 and even 5.33 increase in neutrality by integrating such a graph structure in modulating the attention weights.

3.3.2. Attention and Graph Analysis

We provide an analysis on the learned graph-based attention weights. We first split each utterance into three equal partitions in time, i.e., front (F), middle (M), and back (B). We examine our within-utterance graph's edge weights specifically on those that are in the top 50-th percentile. We plot the ratio of those top 50-th percentile edges that result in a linked connection between the three partitions (FF, MM, BB, FM, FB, MB) for each of the four emotion classes (Figure 2). It is nature to see that the highest peaks exists within individual partition (FF, MM, BB), i.e., the local contextual information is inevitably the most important and similarly related to emotion. However, we observe a phenomenon that furthest cross segments (FB connections) tend to occur mostly in sadness and least likely in happiness. This effect is consistent over the two databases.

We can further analyze this phenomenon by plotting the *original* attention distribution versus *graph* modulated attention profile over an utterance in Figure 3 (in order to better visualize the effect, the profile is derived by summing those attention weights that are in the 70-th percentile only). Figure 3 shows a general trend across the four emotions in two databases that the beginning and and ending portion of the attention weights become larger after applying graphs. One interesting thing that we observe is that for the class of sadness across the two different databases, not only the values but even the shape of attention

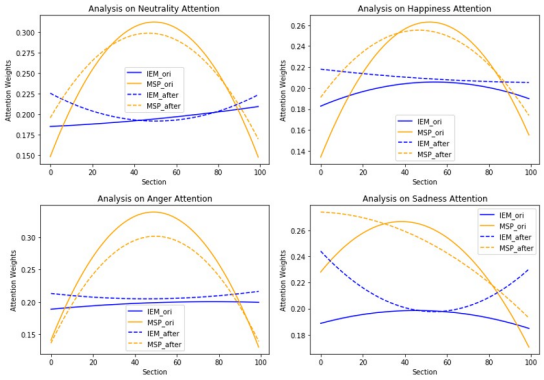


Figure 3: *Attention analysis of experiments: blue/yellow line is the IEMOCAP and the MSP-IMPROV, the solid/dash line indicates original/graph based attention weights respectively.*

distribution are different before and after the graph modulation. In summary, by applying graph based attention mechanism on times series GRU models, our proposed method is able to re-distribute (re-emphasize) those acoustic segments that are are emotionally-salient beyond left-to-right and/or vice versa over the course of an utterance to improve the overall performances.

4. Conclusion

In this work, we propose a graph based attention mechanism that is jointly trained with bidirectional GRUs to enhance the modeling capacity of time series model for SER. We observe an improvement in SER accuracy by imposing a graph structure for attention GRU in both the IEMOCAP and the MSP-IMPROV. When comparing to the existing state-of-art using the exact same experimental setting, our method outperforms all of them. Our analyses demonstrate that by joining a graph structure with attention mechanism, it would effectively re-distribute the attention weighs in handling the complex nature of acoustic encoding of emotion in speech - resulting in an improved recognition accuracy. One of our immediate future work is to integrate lexical information where graph structure would provide a better structure and flexibility in handling the intertwining effect of lexical and acoustic modality in carrying emotion information in speech. Furthermore, we would also explore other node-edge construction mechanism by inclusion of meta attributes, such as personality, to derive a multi-view graph in advancing the learning for SER applications and bring additional insights on the saliency profile of speech emotion expressions.

5. References

[1] Z.-T. Liu, F.-F. Pan, M. Wu, W.-H. Cao, L.-F. Chen, J.-P. Xu, R. Zhang, and M.-T. Zhou, "A multimodal emotional communication based humans-robots interaction system," in *2016 35th Chinese Control Conference (CCC)*. IEEE, 2016, pp. 6363–6368.

[2] J. Hauswald, M. A. Laurenzano, Y. Zhang, C. Li, A. Rovinski, A. Khurana, R. G. Dreslinski, T. Mudge, V. Petrucci, L. Tang *et al.*, "Sirius: An open end-to-end voice and vision personal assistant and its implications for future warehouse scale computers," in *ACM SIGPLAN Notices*, vol. 50, no. 4. ACM, 2015, pp. 223–238.

[3] T. Schwartz, I. Zinnikus, H.-U. Krieger, C. Bürckert, J. Folz, B. Kiefer, P. Hevesi, C. Lüth, G. Pirkl, T. Spieldenner *et al.*, "Hybrid teams: flexible collaboration between humans, robots and virtual agents," in *German Conference on Multiagent System Technologies*. Springer, 2016, pp. 131–146.

[4] 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.

[5] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2122–2132.

[7] S. Zhou, J. Jia, Q. Wang, Y. Dong, Y. Yin, and K. Lei, "Inferring emotion from conversational voice data: A semi-supervised multi-path generative neural network approach," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," *arXiv preprint arXiv:1712.08708*, 2017.

[9] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6685–6689.

[10] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[11] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, and B. W. Schuller, "Towards temporal modelling of categorical speech emotion recognition." in *Interspeech*, 2018, pp. 932–936.

[12] M. Tsakiris and H. De Preester, *The Interoceptive Mind: From Homeostasis to Awareness*. Oxford University Press, 2018.

[13] C. Breitenstein, D. V. Lancker, and I. Daum, "The contribution of speech rate and pitch variation to the perception of vocal emotions in a german and an american sample," *Cognition & Emotion*, vol. 15, no. 1, pp. 57–79, 2001.

[14] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 582–596, 2009.

[15] 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.

[16] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[17] 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*. ACM, 2010, pp. 1459–1462.

[18] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[19] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," *arXiv preprint arXiv:1904.03833*, 2019.

[20] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends," *IEEE Access*, vol. 8, pp. 16 560–16 572, 2020.

[21] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[22] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.