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

Using Speaker-Aligned Graph Memory Block in Multimodally Attentive Emotion Recognition Network

Jeng-Lin Li[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
cllee@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw

Abstract

Integrating multimodal emotion sensing modules in realizing human-centered technologies is rapidly growing. Despite recent advancement of deep architectures in improving recognition performances, inability to handle individual differences in the expressive cues creates a major hurdle for real world applications. In this work, we propose a Speaker-aligned Graph Memory Network (SaGMN) that leverages the use of speaker embedding learned from a large speaker verification network to characterize such an individualized personal difference across speakers. Specifically, the learning of the gated memory block is jointly optimized with a speaker graph encoder which aligns similar vocal characteristics samples together while effectively enlarge the discrimination across emotion classes. We evaluate our multimodal emotion recognition network on the CMU-MOSEI database and achieve a state-of-art accuracy of 65.1% UAR and 74.7% F1 score. Further visualization experiments demonstrate the effect of speaker space alignment with the use of graph memory blocks.

Index Terms: emotion recognition, speaker embedding, graph, memory network

1. Introduction

Emotion recognition technology has attracted substantial attention due to a surge of various human-centered applications, such as conversational AI systems, human robot interactions, and voice-based services. Advancements in deep learning architectures with the increasing availability of affective corpora have together brought a significant improvement in emotion recognition algorithms and also the generalizability of these technical frameworks across databases. However, a major challenge in obtaining a reliable emotion recognition for real world applications remains due to the natural individual differences that create a vast amount of variability in the multimodal expressive behaviors (e.g., speech, language, facial expressions). These individual differences are manifested as emotion expressed, elicited, and experienced and are characterized by each individual's unique life history [1]. Nowadays, collecting adequate amount of data from an individual user is no longer a bottleneck given the proliferation of media data, such as broadcast news, monologue vlogs, and political debates [2]. The ability to handle individual differences to further improve emotion recognition is becoming a crucial next direction of research.

Past literature have generally agreed that the underlying mechanism in emotion processing, i.e., from experiencing, regulating, to expressing, involves multi-component constructs with complex evolutionary, physiological, cognitive and social factors. These emotion-related constructs often intertwine with personal attributes such as age, gender, and personality, which jointly affect one's multimodal emotion expressions in a spontaneous manner [3, 4]. In fact, many past research studies have indeed shown the influences of these attributes and demonstrated the effectiveness in jointly considering their effect when developing emotion recognition algorithms [5]. Exemplary works include: Sagha et al. investigate technique of model selection by considering these personal attributes which yields improved valence recognition [6]; Zhang et al. perform feature space learning that encodes gender as distributional embedding instead of using a simple one-hot vector [7]; Li et al. estimate personality attributes of a target speaker and integrate this attribute for emotion recognition through an attention mechanism [8]; several research works cast this issue in terms of a speaker-dependent setup which can be dealt with by directly incorporating speaker identity to enhance the recognition performance [3, 9, 10].

While these approaches of integrating static personal attributes as additional input features to the recognition frameworks help improve the performances, they tend to ignore the interactive effect of these factors in the modulating of one's behavior. Moreover, exhausting all individual attributes or simply expanding speaker index to account for individual differences in emotion expressions is impractical in real world applications. Recently, there have been attempts to utilize speaker representation techniques from speaker recognition (SR) community in order to provide robust characterization of each subject for emotion recognition task. This idea presents an intriguing technical challenge where on one hand SR tasks focus on enlarging inter-speaker variability with minimal intra-speaker variability, and emotion recognition task tend to broaden intra-speaker affective distribution with minimal inter-speaker influence. Several recent works include: Bancroft et al. report improved results using a speaker verification framework to retrieve target speaker's emotion state [11]; Williams et al. present that style and emotion can be disentangled in speaker embedding [12]; Pappagari et al. devise a transfer learning scheme to examine mutual influences between the two tasks [13]; Li et al. attempt to eliminate speaker information to generalize SER across domains [14]; our previous work is one of the first works in constructing a speaker space via semantic word classes to encode individual variability in the emotion network learning strategy [15].

In this work, we follow a similar learning strategy to incorporate speaker embedding jointly in the learning of multimodal emotion recognition networks. Specifically, we propose a Speaker-aligned Graph Memory Network (SaGMN) consisting of an attentive memory network as the multimodal backbone architecture with its memory block jointly optimized with a speaker graph encoding network. The backbone multimodal network is an improved memory fusion network [16] with attention mechanism. We additionally impose a graph convolutional layer in the memory block whose adjacency matrix is constructed based on between speaker's embedding distances,

http://dx.doi.org/10.21437/Interspeech.2020-1688

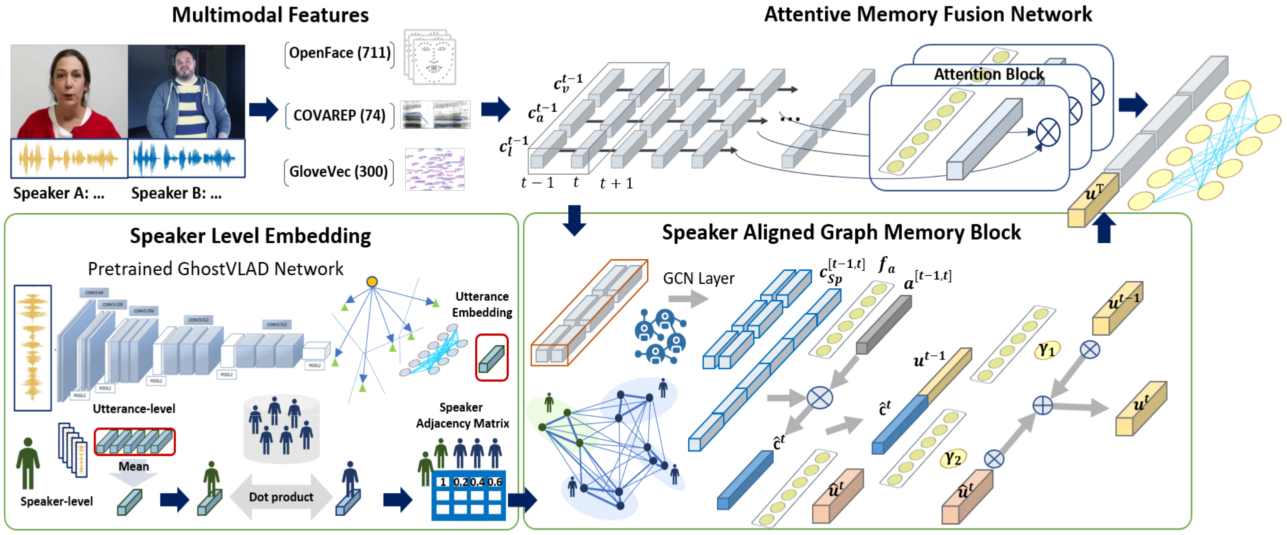

Figure 1: *Our framework SaGMN has a multimodal backbone network with a speaker aligned memory block. The similarity of speaker embeddings extracted from pre-trained speaker recognition network are used to derive adjacency matrix for the graph convolutional layer in the memory block. The resulting memory vector and multimodal attended vectors are used for the final emotion recognition.*

i.e., aligning (moving samples closer) similar speaker characteristics samples together, to achieve the desired effect of minimal inter-speaker differences while optimizing for emotion classification discrimination. We evaluate our framework on a large speaker set CMU-MOSEI database [17] and achieve a 74.7 F1 score and 65.1% UAR. Our analysis experiment illustrate the effect of speaker alignment via memory visualization.

2. Method

2.1. Database

2.1.1. Multimodal Emotion Database

In this work, we utilize a multimodal sentiment and emotion database, the CMU-MOSEI [17], which contains 3228 monologue movie review video clips selected from YouTube after manual quality inspection. Each video is around 40-50 seconds long, and there are a total of 23,453 sentences given by gender balanced speakers. Each sentence is manually annotated with a value in the range of [0,3] indicating the occurrence of six emotions (angry, disgust, fear, happiness, sadness, and surprise). We binarized the annotation by considering those emotion types with value greater than zero indicating the presence of that specific emotion, i.e., label equals to one and zero otherwise. Each sentence could have multiple emotion present.

2.1.2. Multimodal Features

We utilize the multimodal features provided by the CMU-multimodal SDK [18] as the original paper, which includes 74 dimensional COVAREP acoustic features [19], 711 dimensional OpenFace visual features [20], and 300 dimensional Glove textual features [21]. The acoustic features comprise 12 MFCC, pitch, voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients. The visual features contain facial landmarks, facial shape parameters, facial HOG features, head pose, head orientation and eye gaze. The average acoustic visual features are word aligned to word level using the P2FA tool [22]. We use these multimodal features as input to our framework.

2.2. Speaker Aligned Multimodal Emotion Recognition

Figure 1 shows overall architecture. It has a multimodal Attentive Memory Fusion Network (AMFN) as a backbone network to preform emotion recognition. We propose a novel mechanistic graph-gated memory block to include individual speaker characteristics, and the adjacency matrix of the graph is built on speaker embedding extracted from a pre-trained speaker recognition network. We will detail each component in the following sections.

2.2.1. Speaker Embedding Network

To represent each speaker, we use a pre-trained speaker recognition model on VoxCeleb2 database which is a large unconditioned audio recording database with over 1 million utterances from over 6000 speakers [23]. The pre-trained model is based on the GhostVLAD deep network with thin ResNet-34 front-end architecture [24]. The GhostVLAD layer has the ability to aggregate frame-level features into robust utterance-level speaker embedding. During training, an angular space margin softmax loss L_i optimizes for verification which computes the cost in assigning each sample to the correct speaker class y_i.

$$L_i = -\log \frac{\exp^{s(\cos\theta_{y_i} - m)}}{\exp^{s(\cos\theta_{y_i} - m)} + \sum_{j \neq y_i} \exp^{s(\cos\theta_{y_j})}} \quad (1)$$

where hyper-parameters s and m are temperature for softmax and angular margin which are set as 30 and 0.4, respectively. The extracted speaker embedding from this pre-trained network encodes fine-grained individual speaker's vocal characteristics that helps achieve the state-of-the-art speaker verification accuracy. Finally, since each speaker utters multiple sentences, we take the average of speaker embedding over each of the uttered sentences as the *speaker embedding* that will be used in constructing the speaker graph.

2.2.2. Speaker-aligned Graph Memory Network

Our backbone multimodal framework, i.e., termed as AMFN, is an improved memory fusion network (MFN) [16] which incor-

Table 1: *The results of multimodal binary emotion classification on CMU-MOSEI database*

	LF-LSTM			MFN			GMFN			AMFN			P-AMFN			SaGMN		
	WA	F1	UAR	WA	F1	UAR	WA	F1	UAR	WA	F1	UAR	WA	F1	UAR	WA	F1	UAR
Anger	0.648	0.672	0.619	0.518	0.548	0.607	0.580	0.613	0.590	0.613	0.643	0.613	0.637	0.663	0.622	**0.703**	**0.717**	**0.650**
Disgust	0.706	0.738	0.697	0.656	0.697	0.692	0.632	0.676	0.694	0.725	0.752	0.690	0.726	0.754	0.713	**0.761**	**0.782**	**0.728**
Fear	0.846	0.853	0.534	0.757	0.800	0.526	0.651	0.730	0.588	0.819	0.840	0.562	0.776	0.816	0.574	0.801	0.833	**0.587**
Happy	0.633	0.632	0.628	0.644	0.643	0.650	0.676	0.675	0.673	0.649	0.649	0.652	0.661	0.662	0.663	**0.701**	**0.701**	**0.699**
Sad	0.653	0.669	0.585	0.593	0.622	0.586	0.482	0.507	0.574	0.607	0.635	0.606	0.667	0.681	0.598	**0.698**	**0.707**	**0.627**
Surprise	0.639	0.712	0.564	0.652	0.723	0.580	0.735	0.781	0.572	0.757	0.796	0.585	0.754	0.791	0.546	0.680	0.744	**0.615**
AVG	0.688	0.713	0.605	0.637	0.672	0.607	0.626	0.664	0.615	0.695	0.719	0.618	0.704	0.728	0.619	**0.724**	**0.747**	**0.651**

porates additional modality specific attention blocks to models the temporal dependencies. AMFN has parallel modality specific attentive-LSTM systems with a gated memory block to extract cross modality dynamics in each time step. For $m = \{T, V, A\}$ denoting textual, visual, and acoustic modality, the LSTM cell state is written as $c_m = \{c_m^t : t \leq T, c_m^t \in \mathbb{R}^{D_m}\}$ and the output state is $h_m = \{h_m^t : t \leq T, h_m^t \in \mathbb{R}^{D_m}\}$ where D_m is the LSTM hidden dimension. The concatenation of current and previous cell states from the three modalities is denoted as $c^{[t-1,t]}$ serves as the input to the gated memory block.

In this work, we extend our original idea in capturing inter-speaker relation via similarity estimation to account for individual difference [8], we construct a graph that integrate the speaker similarity into a network layer. Specifically, we impose a graph convolutional (GC) layer [25], f_{GC}, to transform $c^{[t-1,t]}$ using the adjacency matrix A:

$$c_{Sp}^{[t-1,t]} = f_{GC}(c^{[t-1,t]}, A) = ReLU(Ac^{[t-1,t]}W) \qquad (2)$$

where W is the learnable weight of f_{GC}. Each element of the adjacency matrix A_{ij} indicates the dot product similarity computed between the i^{th} and j^{th} speaker pair using their speaker embedding (described in section 2.2.1). The use of GC layer links the utterances, i.e., the nodes on the graph, with other nodes that are similar in terms of speaker's embedding; this transforms the original multimodal feature space by aligning utterances to a speaker-aware space. We use a dense layer f_a followed by a softmax layer to learn attention coefficients from $c_{Sp}^{[t-1,t]}$, and multiply them together which results in the final attended vector $\hat{c}^t$.

$$a^{[t-1,t]} = f_a(c_{Sp}^{[t-1,t]}) \qquad (3)$$

$$\hat{c}^t = c_{Sp}^{[t-1,t]} \odot u^{[t-1,t]} \qquad (4)$$

We obtain a memory vector $\hat{u}^t$ by feeding $\hat{c}^t$ to a dense network. Concatenating $a^{[t-1,t]}$ and $\hat{c}^t$ through two neural networks f_{γ_1} and f_{γ_2} with sigmoid function, we derive update gates γ_1^t and γ_2^t for the stored memory u^{t-1} and $\hat{c}^t$, respectively.

$$\hat{u}^t = f_{\hat{u}}(\hat{c}^t) \qquad (5)$$

$$\gamma_1^t = f_{\gamma_1}([\hat{c}^t; u^{t-1}]), \gamma_2^t = f_{\gamma_2}([\hat{c}^t; u^{t-1}]) \qquad (6)$$

$$u^t = \gamma_1^t \odot u^{t-1} + \gamma_2^t \odot \hat{u}^t \qquad (7)$$

Note that γ_1^t and γ_2^t each controls how much the stored memory to keep and new memory to update. This gated process lasts until the final time step T which results in u^T. We devise this new gated memory block with inputs first transformed through the speaker aligned graph. We expect this provide a joint mechanism in consolidating useful multimodal behavior information in u^T by considering inter-speaker relations. In the backbone AMFN, the modality specific LSTMs are attentively re-weighted as context vectors using the learned attention vectors $a_m = softmax(\tanh(W_m^T h_m))$ from hidden vectors h_m. These context vectors are concatenated with the speaker aligned memory vector u^T to the following DNN classification layers.

3. Experiment

3.1. Experimental Setup

In this work, we train six different models to conduct a set of binary recognition tasks in classifying whether each of the six targeted emotion is present or not as our evaluation scheme. We compare our method to the following methods:

- **PAaAN**: The multimodal architecture without personalized attention proposed in [15]
- **MFN**: Memory Fusion Network proposed in [16]
- **GMFN**: Graph Memory Fusion Network in [17]
- **AMFN**: Our proposed Attentive Memory Fusion Network without speaker aligned graph learning
- **P-AMFN**: Using *PAaAN* approach [15] of integrating speaker embedding in the learning of memory block instead of graph convolution layer proposed in this work
- **SaGMN**: Our proposed speaker aligned graph memory multimodal framework

We first compare to a set of multimodal baseline architectures including *PAaAN*, *MFN*, *GMFN*, and our backbone *AMFN*. *PAaAN* indicates the multimodal attention learning architecture, i.e., without the personalized profile, previously used in [15]. *MFN* and *GMFN* represent two other multimodal learning frameworks recently proposed. *AMFN* is our modified attentive memory fusion network without speaker-aware learning. Furthermore, since our backbone learning is based on *AMFN*, we additionally compare to using the personalized profile integration learning strategy [15], i.e., by computing the distance of each individual speaker embedding derived from GhostVLAD to the background corpus, and concatenate this information into the learning of memory block, denoted as *P-AMFN*.

The data are split into training, validation, and testing sets followed the same experimental setting of the database paper [17] which corresponds to 14984, 1709, and 4321 samples in the three sets, respectively. In SaGMN, the parallel LSTM networks have [128,256,64] hidden dimensions for [T,V,A] features. f_{GC} has 256 output nodes and $f_{\hat{u}}$ has two layers with 128 and 32 nodes. Both f_{γ_1} and f_{γ_2} have a single 32-node layer and the memory size is 128. The final recognition network is a 2-layer fully connected feedforward neural network with 128 and 64 neurons. We use Adamax optimizer to train the model

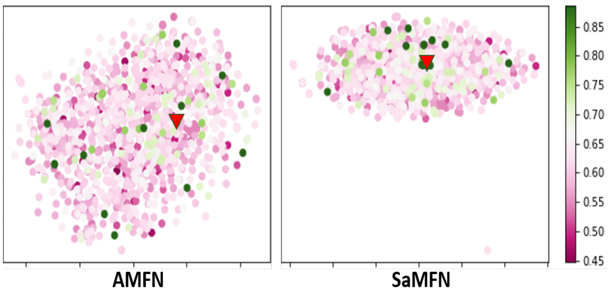

Figure 2: *This is the visualization of memory layer. The left and right figures are from AMFN and SaMFN, respectively. The targeted sample is represented in a red triangle and the other samples are sketched in different colors based on similarity.*

in 200 epochs with 64 batch size and 0.0001 learning rate. The architecture parameters are selected from $\{32,64,128,256,512\}$ based on validation performances while the testing set results are presented. Three different evaluation metrics are included, i.e., F1 score (F1), unweighted average recall (UAR) and accuracy (WA), in order to provide more complete results as compared to most prior works that report part of these metrics.

3.2. Experimental Results

The summary of the emotion recognition is demonstrated in Table 1. Our proposed *SaGMN* achieves 72.4% ACC, 74.7% F1, and 65.1% UAR in 6-class averaged emotion recognition performances which are 4.17%, 3.89%, and 5.34% relatively improved in terms of the model without speaker alignment, *AMFN*. When comparing multimodal architectures (without the speaker-aware memory block learning), *AMFN* obtains 69.5% WA, 71.9% F1, and 61.8% UAR which is generally higher than *LF-LSTM*, *MFN* and *GMFN*. Among these multimodal frameworks, the two renowned models, *MFN* and *GMFN* obtain comparable results with each other. *MFN* has higher WA and F1 (63.7% and 67.2%) while *GMFN* obtains 61.5% averaged UAR. We observe similar results in *LF-LSTM* demonstrating high WA and F1 though lower UAR than *MFN* and *GMFN*. This also shows the importance in reporting all of these metrics in comparing between models. *AMFN* utilizes modality specific attention blocks to re-weight important time step, and the memory block additionally accumulates key discriminatory multimodal information. The results of AMFN achieve the highest accuracy across all three metrics when compared with *LF-LSTM*, *MFN* and *GMFN*. The *AMFN* is used as our mutimodal backbone architecture in performing the emotion recognition learning.

When examining the emotion recognition accuracy obtained in learning with the extracted speaker embedding that aligns speaker space, *P-AMFN* outperforms *AMFN*, i.e., without speaker information, by 1.25% relative F1 improvement. Our proposed method, *SaGMN*, further obtains 2.84%, 2.61%, and 5.19% relative improvement over *P-AMFN* on ACC, F1, and UAR metric, respectively; except for rare class of fear and surprise, we observe a consistent improvement across all metrics. This results demonstrate that by modeling inter-speaker through a graph structure is much more effective in handling individual variability in improving emotion recognition accuracy than simply concatenating speaker embedding in the memory block [15]. The CMU-MOSEI database includes a large speaker set where each utters few utterances, it is encouraging to see our framework is able to obtain a state-of-the-art recognition rate in a scenario that is closer to real world applications.

3.3. Visualization Analysis

In this section, we analyze the effect of integrating speaker embedding in the learned multimodal memory. We visualize both the final memory vector u^T in *AMFN* and *SaGMN* by TSNE 2-dimensional projection from the angry classification model (the one that obtains the highest accuracy). Figure 2 shows results from *AMFN* in the left and *SaGMN* in the right. The targeted sample is represented with a red triangle; the color bar indicates the similarity degree computed between all samples in the database and this targeted sample, where darker green indicates higher similarity and purple indicates low similarity.

One thing to notice is the dark green samples in the *AMFN* are scattered over the memory vector space. We observe these dark green samples are more concentrated around the targeted sample in our proposed *SaGMN*. By examining these dark green samples, we also see that most of these green samples are in fact coming from the same speaker as the targeted sample. This provides an simple illustration that our proposed *SaGMN* using graph convolution layer on the speaker embedding adjacency matrix, has an effect on aligning speaker space, i.e., those with similar vocal characteristics would be memorized to have multimodal behavior representation closer to each other. Having this clustered speaker space effect in the memory block helps the backbone memory fusion network to enhance emotion discrimination. Another thing to note is that the robustness of the pre-trained speaker embedding extraction network is also important in the function of *SaGMN*. This pre-trained embedding achieves a fairly robust 5.60% EER on this CMU-MOSEI database.

4. Conclusion

Individual speaker's idiosyncratic behavior variability increases the complexity in the task of learning multimodal emotion recognition for real world applications. To address this issue, we propose a Speaker aligned Graph Memory Network (SaGMN) to embed inter-speaker vocal characteristics in a graph convolutional layer in the memory block for multimodal emotion recognition. Our framework uses robust speaker embedding extracted from a speaker recognition pre-trained network. The similarity based graph consequently aligns the samples in a latent speaker space. The speaker aligned multimodal information is accumulated over time and encoded in the memory block. Our framework obtains a state-of-the-art on a large-scale affective corpus across three evaluation metrics. Our analysis also shows that similar samples converge together after speaker alignment in the memory layer.

In this work, we integrate speaker's idiosyncratic information in a emotion recognition framework through the use of speaker embedding. There are several directions to extend this work. Firstly, each speaker has around one minute expressions in the CMU-MOSEI database. Since the total length requirement of each individual's data would determine the generalizability and usability of our framework in real world scenarios, we will investigate the emotion discriminatory effect of speaker space alignment as a function of data quantity. Secondly, we will explore additional graph algorithm to model the latent speaker relations, e.g., hypergraph for complex links and graph pooling for redundancy pruning. Finally, the current characterization of speaker space is through vocal characteristics, which may not capture the full spectrum of personal factors such as gender, age, personality, we will explore approaches in constructing the speaker space with multimodal signals to further improve the robustness of our framework.

5. References

[1] P. Kuppens, J. Stouten, and B. Mesquita, "Individual differences in emotion components and dynamics: Introduction to the special issue," *Cognition and Emotion*, vol. 23, no. 7, pp. 1249–1258, 2009.

[2] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer speech & language*, vol. 29, no. 1, pp. 186–202, 2015.

[3] C. Welch, V. Perez-Rosas, J. Kummerfeld, and R. Mihalcea, "Look who's talking: Inferring speaker attributes from personal longitudinal dialog," in *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, 2019.

[4] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[5] S. Rukavina, S. Gruss, H. Hoffmann, J.-W. Tan, S. Walter, and H. C. Traue, "Affective computing and the impact of gender and age," *PloS one*, vol. 11, no. 3, 2016.

[6] H. Sagha, J. Deng, and B. Schuller, "The effect of personality trait, age, and gender on the performance of automatic speech valence recognition," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 86–91.

[7] L. Zhang, L. Wang, J. Dang, L. Guo, and Q. Yu, *Gender-Aware CNN-BLSTM for Speech Emotion Recognition: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I*, 09 2018, pp. 782–790.

[8] J.-L. Li and C.-C. Lee, "Attention learning with retrievable acoustic embedding of personality for emotion recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 171–177.

[9] M. Sidorov, A. Schmitt, E. Semenkin, and W. Minker, "Could speaker, gender or age awareness be beneficial in speech-based emotion recognition?" in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 61–68.

[10] M. Sidorov, S. Ultes, and A. Schmitt, "Comparison of gender-and speaker-adaptive emotion recognition." in *LREC*, 2014, pp. 3476–3480.

[11] M. Bancroft, R. Lotfian, J. Hansen, and C. Busso, "Exploring the intersection between speaker verification and emotion recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2019, pp. 337–342.

[12] J. Williams and S. King, "Disentangling style factors from speaker representations," in *Proc. Interspeech*, vol. 2019, 2019, pp. 3945–3949.

[13] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7169–7173.

[14] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7144–7148.

[15] J.-L. Li and C.-C. Lee, "Attentive to individual: A multimodal emotion recognition network with personalized attention profile," *Proc. Interspeech 2019*, pp. 211–215, 2019.

[16] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[17] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[18] 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.

[19] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[20] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[21] R. JeffreyPennington and C. Manning, "Glove: Global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing*. Citeseer, 2014.

[22] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[23] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech 2018*, pp. 1086–1090, 2018.

[24] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

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