AN INTERACTION-AWARE ATTENTION NETWORK FOR SPEECH EMOTION RECOGNITION IN SPOKEN DIALOGS

Sung-Lin Yeh, Yun-Shao Lin, Chi-Chun Lee

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

ABSTRACT

Obtaining robust speech emotion recognition (SER) in scenarios of spoken interactions is critical to the developments of next generation human-machine interface. Previous research has largely focused on performing SER by modeling each utterance of the dialog in isolation without considering the transactional and dependent nature of the human-human conversation. In this work, we propose an interaction-aware attention network (IAAN) that incorporate contextual information in the learned vocal representation through a novel attention mechanism. Our proposed method achieves 66.3% accuracy (7.9% over baseline methods) in four class emotion recognition and is also the current state-of-art recognition rates obtained on the benchmark database.

Index Terms— speech emotion recognition, interaction, attention mechanism, spoken dialogs

1. INTRODUCTION

Emotion plays an important role in human-human interaction, it usually comes with intense and short-time responses expressed behaviorally in the form of facial expressions, gestures, and voice signals. Decade worth of research in speech emotion recognition (SER) have devoted into understanding acoustic manifestation of emotion and developing appropriate computational algorithms in achieving robust recognition performances (e.g., [1, 2, 3]). Due to the recent surge in deploying deep learning methodologies for machine intelligent tasks, several works have further demonstrated significantly improved speech emotion recognition rates; for example, Han et al. used deep neural networks to model the utterance-level emotion [4], Trigeorgis et al. combined convolutional neural networks (CNNs) with long short-term memory (LSTM) to learn better raw time representation [5], and Mirsamadi et al. used attention-based CNN to perform SER from frame-level characterization [6]. These developments of SER have not only enabled more personalized spoken dialog system [7] but also found its use in quantifying emotion in human-centered applications [8, 9].

While these works have achieved better recognition performances, their frameworks focus on modeling vocal information of target speech segments in isolation often without considering interaction context. Researches in psychology have emphasized the importance in characterizing the *transactional* dynamics of emotion during human-human interaction. These dynamics include not only transitions and co-occurrences of emotion states of a given speaker [10, 11] but also emotion contagion phenomenon [12], i.e., interacting partners are capable of affecting each other's emotion states and behaviors. Consequently, to obtain better characterize a target speaker's current emotion state, his/her own previous state and behaviors from his/her interacting partners are two prime contributions in this transactional aspect of emotion.

In this work, our aim is to further improve the speech emotion recognition in spoken dialogs by learning to embed these transactional aspect into vocal representation using attention network. A couple related works that have similarly taken advantage of contextual information for SER. For example, Hazarika et al. utilized a memory network to model the relevance of the current utterance and the history of utterances between the two speakers in dialogs to perform SER [13]. Ruo et al. proposed an interaction and transition model based on frame-level acoustics features, where each utterance's emotion probability is re-estimated by previous utterance and currently estimated posteriors using an additional LSTM [14]. While they both model the contextual information, however, the *emotionally-relevant* information embedded in the current utterance as a result of the transactional, i.e., transitional and contagious, effect is not explicitly learned and integrated in the representation of the current utterance.

To address this issue, we propose a complete architecture of interaction-aware attention network (IAAN), which is built based on attention-based gated recurrent units (GRUs) [15]. By including two contextual utterances as a unit of transactional frame, i.e., the previous utterance of the current speaker, and the previous utterance of the interlocutor, we devise an attention mechanism that embed the transactional information into the current utterance. Finally, we concatenate contextual representations and interaction-aware current utterance representation for emotion recognition. We evaluate our framework on the benchmark IEMOCAP corpus [16]. It obtains 66.3% accuracy, which 7.9% better than without using the contextual information. Our framework also outperforms the known state-of-art SER accuracy on the IEMOCAP.

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

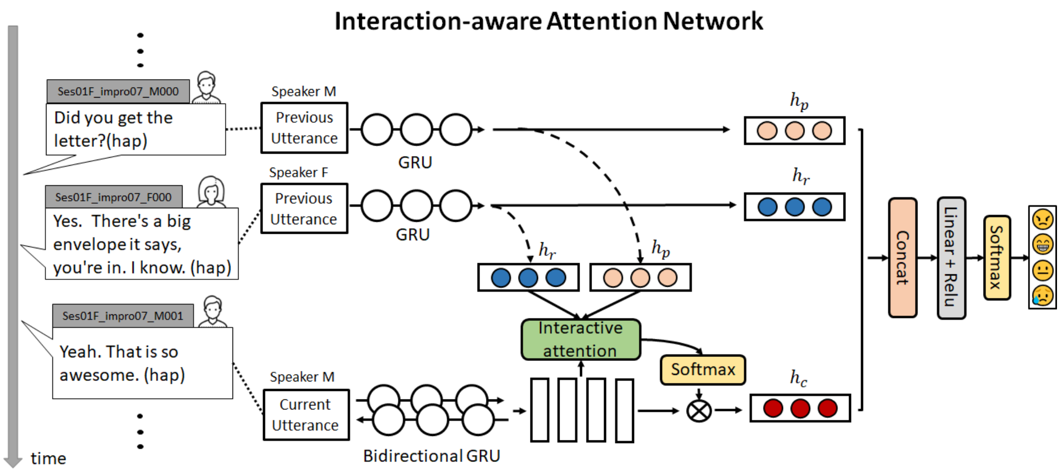

Fig. 1. An illustration of our proposed interaction-aware attention network (IAAN) for speech emotion recognition.

2. RESEARCH METHODOLOGY

In this section, we first describe the benchmark emotion dataset, acoustic feature extraction, and finally our proposed interaction-aware attention network (IAAN).

2.1. Dataset Description

We use the IEMOCAP dataset in this work [16]. It is a benchmark dataset that is widely used in speech emotion recognition research. It contains 10 speakers, each session consists of multiple conversational scenarios between two actors. In this work, in order to compare with the past state-of-art performances, we conduct four emotion class classifications, i.e., anger, happiness, sadness and neutrality, using a total of 5531 utterances, where happiness and excitement are considered together as happiness. The distributions of the four emotion classes in the 5531 utterances are: anger: 19.9%, happiness: 29.5%, neutrality: 30.8%, sadness: 19.5%

2.2. Acoustic Low-level Descriptors

We extract acoustic low-level descriptors (LLDs) based on Emobase 2010 Config using the openSMILE toolkit [17], including features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch and their statistics in each short frame of an utterance. We obtain a sequence of a total of 45 dimensional frame-level acoustic features for each utterance. We apply speaker-dependent z-normalization for each descriptor, and we further downsample the frame numbers by averaging feature values every five frames to reduce the computational cost.

2.3. Interaction-aware Attention Network (IAAN)

We propose an interaction-aware attention network (IAAN) to integrate influences of the contextual information between interlocutors within a *transactional* frame to perform emotion recognition. We will first define the transactional context, describe the gated recurrent unit that models the sequence of LLDs for each utterance, and finally details our interaction-aware attention framework.

2.3.1. Transactional Context

Our proposed IAAN extends beyond conventional framework that often relies solely on single utterance modeling by integrating influence of interlocutors' utterances within a defined transactional context. Consider a set of utterances in dialog, we aim to recognize emotion of a current utterance U_c. We define a transactional context by including the previous utterance of the current speaker and the previous utterance of the other speaker in the conversation as auxiliary utterances, denoted as U_p and U_r respectively. Consequently, each training data point includes a triple of (U_c, U_p, U_r) with the label of U_c. The goal of IAAN is to identify the emotion of U_c by simultaneously leveraging U_p and U_r. Note that labels of U_p and U_r are not used in the training procedure.

2.3.2. Interaction-aware Attention Representation

The basic building block of our IAAN is based on Gated Recurrent Unit (GRU) mainly due to its lower computational cost with comparable performance as compared to LSTM [15]. Within each frame of transactional context, we encode U_p and U_r to fixed-length utterance-level features h_p and h_r using GRU with Bahdanau attention mechanism [18]. Regarding the current utterance, U_c, we use bidirectional GRU (BiGRU). That is, given i^{th} current utterance $u_i \in U_c$ with $frame_{it}, t \in [1, T]$, BiGRU encodes u_i as follows:

$$\overrightarrow{h_{it}} = \overrightarrow{GRU}(frame_{it}), t \in [1, T], \tag{1}$$

$$\overleftarrow{h_{it}} = \overleftarrow{GRU}(frame_{it}), t \in [T, 1]. \tag{2}$$

The hidden states h_{it} of BiGRU are obtained through concatenation $h_{it} = [\overrightarrow{h_{it}}; \overleftarrow{h_{it}}]$.

Then, instead of using classical Bahdanau attention in Bi-GRU for U_c, we propose a novel interaction-aware attention. The interaction-aware attention is designed to capture the affective transition (previous utterance of the same target speaker) and affective influence (previous utterance of the

Model	Method	Recall(%)				WA(%)	UA(%)
		Anger	Happiness	Neutrality	Sadness		
SVM Trees	Rozgić et al.(2012)	-	-	-	-	60.8	60.9
BiLSTM+ATT	Mirsamadi et al.(2017)	-	-	-	-	63.5	58.8
CMN	Hazarika et al.(2018)	-	-	-	-	65.3	-
MDNN	Zhot et al.(2018)	-	-	-	-	61.8	62.7
BiGRU+ATT	Our method	56.6	59.4	48.4	71.6	57.6	58.4
BiGRU+IAA	Our method	65.3	61.0	51.7	73.0	60.7	62.9
RandIAAN	Our method	66.0	62.3	53.5	73.7	62.0	63.4
IAAN	Proposed method	72.1	65.4	53.1	74.6	64.7	**66.3**

Table 1. The performance of models in comparison with the state of the art (upper part) and different network variants (lower part). Note that Hazarika et al. only tested their model on Session 5 in IEMOCAP.

interlocutor) into the representation of current target utterance. Hence, while encoding the current utterance's attentive representation h_c, the previous utterance information in the h_p and h_r are integrated into current utterance encoder. We define the score function $e(\cdot)$ and attention weight α_t as:

$$e\left(h_{it}, h_p, h_r\right) = v_a^T \tanh\left(W_c h_{it} + W_p h_p + W_r h_r + b_a\right), \tag{3}$$

$$\alpha_t = \frac{\exp(e\left(h_{it}, h_p, h_r\right))}{\sum_{t=1}^{T} \exp(e\left(h_{it}, h_p, h_r\right))}, \tag{4}$$

where $v_a \in \mathbb{R}^d$ and $W_c, W_p, W_r \in \mathbb{R}^{d \times d}$ are weight matrices $b_a \in \mathbb{R}^{d \times 1}$ is a bias vector; these are all learnable parameters. The defined score function iteratively loop through every timestep of current utterance's hidden states h_{it} based on the contextual representations of h_p and h_r. With the obtained attentive weights, we perform weighted pooling over the output of BiGRU hidden states h_{it} to obtain the modified current utterance representation using the learned interaction-aware attention:

$$h_c = \sum_{t=1}^{T} \alpha_t h_{it}. \tag{5}$$

2.3.3. Emotion Classification Network

Hence, within every transactional frame, we assemble a joint representation collected from h_p and h_r, encoded for U_p and U_r, and the current utterance representation, h_c. The joint representation is passed to subsequent two projection layers with a ReLU activation in between,

$$R = [h_c; h_p; h_r], \tag{6}$$

$$LP(R) = relu\left(RW_1 + b_1\right) W_2 + b_2, \tag{7}$$

where LP stands for projection layer, W_1, W_2, are weight matrices, b_1, b_2 are bias vectors. The final emotion recognition are done using a $softmax$ function,

$$\hat{y} = softmax\left(LP\left(r\right)\right). \tag{8}$$

The complete IAAN is then trained by minimizing the cross-entropy loss to perform multi-class emotion recognition.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental setup

The exact structure of our IAAN is as follows: the hidden unit dimension is set to 512 for two GRUs and 256 for each direction of BiGRU. The learning rate is set as 0.0001 and a mini-batch size is set as 64. We apply 90% dropout to each GRU and BiGRU cells as well as the output of first projection layer. Moreover, we add a weight decay of 0.001 to all weights and biases in the projection layers. Regarding to activation function, we choose rectified linear unit. We train our model using Adam optimizer with cross-entropy loss as our objective. In experiments, we carry out early stopping by observing the performance on validation set in every 100 training epochs.

To evaluate the performance, we present both unweighted accuracy (UA) and weighted accuracy (WA). All hyperparameters are optimized based on results of 5-fold leave-one-session-out (LOSO) cross validation. Past works have used 10-fold leave-one-person-out (LOPO) cross validation [19, 20], which tend to give a higher accuracy number than LOSO. In order to evaluate under realistic scenarios of our performance for a new interacting dyad, all of our analyses are based on LOSO cross validation.

3.1.1. Baseline Methods

The following are the baseline methods of network variants and previous works that we use to compare with IAAN:
BiGRU+ATT: A BiGRU network with the classical attention (ATT) trained using current utterances only.
BiGRU+IAA: The framework and inputs are same as IAAN, but instead of the joint concatenated representation, the predictions only depend on current utterance's representation.
RandIAAN: The IAAN approach but trained using the randomly selected auxiliary utterances in the dialog as a transactional frame.
SVM Trees: A binary SVM based tree structure for emotion classification [19].
BiLSTM+ATT: A BiLSTM network that ultilized a attention-based pooling layer on frame-level features [6].
CMN: A conversational memory network that incorporated emotional context information into memory cells from history

utterances in conversations [13].

MDNN: A multi-path deep neural network composed of several local classifiers and a global classifier [20].

Table 1 summarizes the unweighted accuracy (UA) and weighted accuracy (WA) of the current state of the art methods (upper part), different network variants (lower part) and our proposed IAAN results. Also, the accuracies of each emotional category are demonstrated.

3.2. Result and Analysis

3.2.1. Recognition Performances

As shown in Table 1, the performance of our proposed IAAN reaches 66.3% UA in the four classes emotion recognition, which is the best accuracy among all baseline methods. To further analyze the effectiveness of various modules integrated in our IAAN, we compare different network variants in the lower part of Table 1.

Firstly, when comparing the performance of BiGRU+ATT and BiGRU+IAA, the proposed interaction-aware attention shows an improvement over using self-attention by +4%, which indicates that when learning to represent the current utterance's emotionally-relevant behavior, utilizing attention mechanism by jointly considering the past contextual information (previous utterance of the target speaker and the interlocutor) provides a substantial benefit. Secondly, we investigate the usefulness of auxiliary emotional contexts for final emotion predictions by comparing our proposed IAAN to BiGRU+IAA. With the joint representation concatenating representation within emotional contexts, IAAN obtain further +4% improvements over BiGRU+IAA. More interestingly, we compare RandIAAN with IAAN to evaluate the effectiveness of the *immediate* emotional contexts, and we observe that IAAN obtains +3% higher UA than RandIAAN, indicating the important interactive information should be embedded from the immediate context.

The upper part of Table 1 summarizes the comparison of our proposed IAAN to the existing methods on the same database. The existing methods include context-free and context-dependent frameworks. For context-free model [19, 6, 20], our method outperforms each of them by 5.3%, 7.5% and 3.6% in the UA measure. The context-dependent CMN proposed by Hazarika et al. [13] only presented their WA results on session 5 of the IEMOCAP, our method obtains 65.5% for that particular session.

In summary The comparison between BiGRU+ATT and BiGRU+IAA shows that the interaction-aware attention possess better ability in extracting emotionally-relevant information in a current utterance by integrating contextual information. The second comparison (IAAN vs BiGRU+IAA) indicates that concatenated representation demonstrates even further improved modeling power in recognizing emotion state of the current utterance. Furthermore, the comparison between IAAN and RandIAAN demonstrates the effectiveness of incorporating *immediate* emotional contexts. Lastly, our proposed IAAN, to the best of our knowledge, obtains the best emotion recognition performances among the known state-of-art methods on the IEMOCAP.

Scenario	Data points	BiGRU+ATT(%)		IAAN(%)	
		WA(%)	UA(%)	WA(%)	UA(%)
Case 1	27.1	72.3	69.2	76.0	74.4
Case 2	47.1	58.2	60.2	64.3	66.6
Case 3	25.8	42.0	42.4	53.6	54.8

Table 2. Analysis of IAAN predictions in three emotional context scenarios.

3.2.2. Analysis

In this section, we further investigate how does IAAN perform in different emotional context scenarios. For each transactional frame, we define three different emotional scenarios: (1) U_c shares the same emotion as U_p and U_r, (2) U_c shares the same emotion with one of U_p or U_r, (3) U_c has emotion different from of U_p and U_r. Table 2 demonstrates the accuracy of our IAAN under three different conditions.

Although no labels of previous utterances are given in whole training procedure, our framework obtains improves recognition rates in all three conditions when compared with method without interaction-aware attention. In Case 1 and Case 2, we observe that once U_c shares the same emotion (even partially) as their immediate preceding emotional contexts, IAAN achieves the best recognition rates. On the contrary, the condition where the previous utterances have completely different emotion from U_c (Case 3) results in lowest accuracies. More interesting, if we examine the type of emotions of U_c in Case 3, the emotions are dominated by neutrality that accounts for 37% of data, where angry, happiness, sadness are 24.8%, 22.2% and 14.9%, respectively. Furthermore, the UA of neutral category is only 47.6% in Case 3, which suggests that the emotional characteristics of neutrality has less relevance from their emotional contexts as compared to others. This result corroborates with several past works that have argued whether neutrality is considered as an *emotional* state or an simply a mixed state that is *absent* from apparent emotional expressions [21, 22].

4. CONCLUSIONS AND FUTURE WORKS

In this work, we propose an interaction-aware attention network, which effectively incorporates contextual information during dyadic conversations, to perform utterance-based emotion recognition. The contextual information is incorporated both at the learning of current utterances representation and the final prediction stage. Our method shows outstanding performance with unweighted accuracy of 66.3%, and outperforms the best-known state-of-the-art methods.

In the future, since we observe initially that neutrality seems to be more *context-free*, developing a strategy that simultaneously considered the nature of emotion classes will be an immediate future step. Also, we will evaluate the generality of IAAN in other conversational dataset, including dyadic to small group interactions, to further validate the robustness of IAAN in a variety of spoken interaction contexts.

5. REFERENCES

[1] B. Schuller, G. Rigoll, and M. Lang, “Hidden markov model-based speech emotion recognition,” in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03). 2003 IEEE International Conference on*, vol. 2. IEEE, 2003, pp. II–1.

[2] Y.-L. Lin and G. Wei, “Speech emotion recognition based on hmm and svm,” in *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, vol. 8. IEEE, 2005, pp. 4898–4901.

[3] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[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] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

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

[7] K. Jokinen and M. McTear, “Spoken dialogue systems,” *Synthesis Lectures on Human Language Technologies*, vol. 2, no. 1, pp. 1–151, 2009.

[8] S. Narayanan and P. G. Georgiou, “Behavioral signal processing: Deriving human behavioral informatics from speech and language,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[9] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, and S. Narayanan, “Signal processing and machine learning for mental health research and clinical applications [perspectives],” *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 196–195, 2017.

[10] M. A. Thornton and D. I. Tamir, “Mental models accurately predict emotion transitions,” *Proceedings of the National Academy of Sciences*, p. 201616056, 2017.

[11] S. Hareli, S. David, and U. Hess, “The role of emotion transition for the perception of social dominance and affiliation,” *Cognition and Emotion*, vol. 30, no. 7, pp. 1260–1270, 2016.

[12] S. G. Barsade, “The ripple effect: Emotional contagion and its influence on group behavior,” *Administrative Science Quarterly*, vol. 47, no. 4, pp. 644–675, 2002.

[13] 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)*, vol. 1, 2018, pp. 2122–2132.

[14] R. Zhang, A. Atsushi, S. Kobashikawa, and Y. Aono, “Interaction and transition model for speech emotion recognition in dialogue,” *Proc. Interspeech 2017*, pp. 1094–1097, 2017.

[15] 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.

[16] 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.

[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] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[19] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, “Ensemble of svm trees for multimodal emotion recognition,” in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–4.

[20] 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,” 2018.

[21] A. Metallinou, S. Lee, and S. Narayanan, “Decision level combination of multiple modalities for recognition and analysis of emotional expression,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010.

[22] A.-M. Laukkanen, E. Vilkman, P. Alku, and H. Oksanen, “On the perception of emotions in speech: the role of voice quality,” *Logopedics Phoniatrics Vocology*, vol. 22, no. 4, pp. 157–168, 1997.