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.
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 $frame_{c_{it}}, t \in [1, T]$, BiGRU encodes $u_i$ as follows:

$$h_{it}^{\rightarrow} = GRU_{\rightarrow}(frame_{c_{it}}), t \in [1, T],$$

$$h_{it}^{\leftarrow} = GRU_{\leftarrow}(frame_{c_{it}}), t \in [T, 1].$$

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

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
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 $\alpha_t$ as:

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

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

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,

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

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

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 $\text{softmax}$ function,

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

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-person-out (LOPO) 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 LOPO. In order to evaluate under realistic scenarios of our performance for a new interacting dyad, all of our analyses are based on LOPO 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 utilized 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.
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 possesses 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

### 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 generalization 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.

---

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

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Data points</th>
<th>BiGRU+ATT(%)</th>
<th>IAAN(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UA(%)</td>
<td>WA(%)</td>
<td>UA(%)</td>
</tr>
<tr>
<td>Case 1</td>
<td>27.1</td>
<td>72.3</td>
<td>69.2</td>
</tr>
<tr>
<td>Case 2</td>
<td>47.1</td>
<td>58.2</td>
<td>60.2</td>
</tr>
<tr>
<td>Case 3</td>
<td>25.8</td>
<td>42.0</td>
<td>42.4</td>
</tr>
</tbody>
</table>
5. REFERENCES


