2024 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3503-9015-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICME57554.2024.10687895

ENHANCED SPEECH EMOTION RECOGNITION INCORPORATING SPEAKER-SENSITIVE INTERACTIONS IN CONVERSATIONS

Jiachen Luo[⋆] *Huy Phan*[⋄] *Lin Wang*[⋆] *Joshua Reiss*[⋆]

[⋆] Centre for Digital Music, Queen Mary University of London, UK
[⋄] Amazon Alexa, Cambridge, MA, USA

ABSTRACT

Accurately detecting emotions in conversation is a necessary yet challenging task due to the complexity of emotions and dynamics in dialogues. The emotional state of a speaker can be influenced by many different factors, such as interlocutor stimulus, dialogue scene, and topic. In this work, we propose a conversational speech emotion recognition method to deal with capturing attentive contextual dependency and speaker-sensitive interactions. First, we use a pretrained WavLM model to extract frame-based audio representation in individual utterances. Second, an attentive bi-directional gated recurrent unit (*GRU*) models contextual-sensitive information and explores listener dependency and speaker influence jointly in a simple, fast, parameter-efficient way. The experiments conducted on the standard conversational dataset MELD demonstrate the effectiveness of the proposed method when compared against state-of-the-art methods.

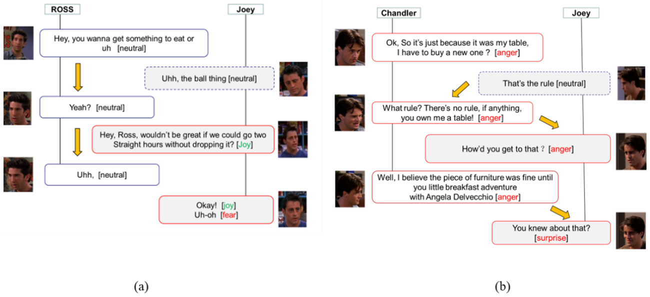
Index Terms— Emotion recognition, affective computing, speaker-sensitive

1. INTRODUCTION

Automatic recognition of human emotions has widespread applications in areas such as dialogue generation, and human computer interaction [1]. Unlike vanilla emotion recognition of utterances, emotion recognition in conversation (ERC) ideally relies on mining human emotions from conversations or dialogues having two or more interlocutors and requires context modeling of the individual utterances [2]. ERC aims to understand human emotions when the interlocutors are interacting with one another during conversations and classify each utterance into its associated emotional state.

Humans convey emotions through various modalities including speech, facial expression, body postures, etc [2-3]. The vast majority of ERC problems focus on extracting information from textual modalities. It has been highlighted that the text modality contains less noise compared to other modalities [3]. However, speech is the main communication medium in which people can clearly and intuitively feel emotional changes. Emotion perception from audio signals only is much easier to be obtained. How to capture emotion-relevant information from single audio data is a challenging task.

In this work, we focus on speech signals in interactive conversation. Speech signals naturally can carry the emotional characteristics. Conventionally, conversational emotion recognition usually requires a strong ability to model context-sensitive attributes, select crucial information, and capture speaker-sensitive dependencies [3]. Among all the factors, speaker information is important for tracking the emotional characteristics of conversations, especially listener and speaker state.


ROSS
Joey
Hey, you wanna get something to eat or uh [neutral]
Uhh, the ball thing [neutral]
Yeah? [neutral]
Hey, Ross, wouldn't be great if we could go two Straight hours without dropping it? [Joy]
Uhh, [neutral]
Okay! [joy]
Uh-oh [fear]
Chandler
Joey
Ok, So it's just because it was my table, I have to buy a new one ? [anger]
That's the rule [neutral]
What rule? There's no rule, if anything, you own me a table! [anger]
How'd you get to that ? [anger]
Well, I believe the piece of furniture was fine until you little breakfast adventure with Angela Delvecchio [anger]
You knew about that? [surprise]

(a) (b)

Fig. 1. Emotion dynamic of speakers in a dialogue in comparison.

In interactive conversations, these factors lead to diverse emotional dynamics. Fig. 1 presents some examples demonstrating such patterns from the Multi-modal EmotionLines Dataset (MELD) [4]. On the one hand, conversation (a) depicts the presence of emotional inertia which speakers influence on themselves. The character Ross maintains a neutral emotional state by not being influenced by the other speaker. On the other hand, conversation (b) refers to the interaction of participants (listener and speaker) that counterparts induce in a speaker. "I have to buy a new one?" shows negative attitude. With a particular voice shade of anger, it can affect the feeling of the addressee/listener. "How'd you get to that?" emphasizes to the listener that Joey is affected by the feeling of speaker Chandler's responses.

To model such conversations, an architecture would need to deal with these challenges: how to capture participant's state to govern emotional dynamics, and how to interpret latent emotions from its contextual information in the conversation flows. What's more, the raw emotion can be enhanced, weakened, or reversed based on the contextual information from neighboring utterances [5]. Further, the relevant features are to be extracted precisely by utilizing the speech signals, but the challenging task is to choose the appropriate features for the emotion recognition systems.

For utterance-level speech emotion recognition, an underlying issue is a loss of dynamic temporal information and short-term emotion dynamics by compressing speech into utterance-level features [6]. However, little progress has been made in analyzing the emotion estimation among frame-based feature representation in individual utterances, context-sensitive information and speaker influences in conversations. Devamanyu *et al.* used text modality features to model the contextual information into listener and speaker emotional in-

[⋄] The work was done when H. Phan was at School of Electronic Engineering and Computer Science, Queen Mary University of London, UK and The Alan Turing Institute, UK and prior to joining Amazon.

979-8-3503-9015-5/24/$31.00 ©2024 IEEE

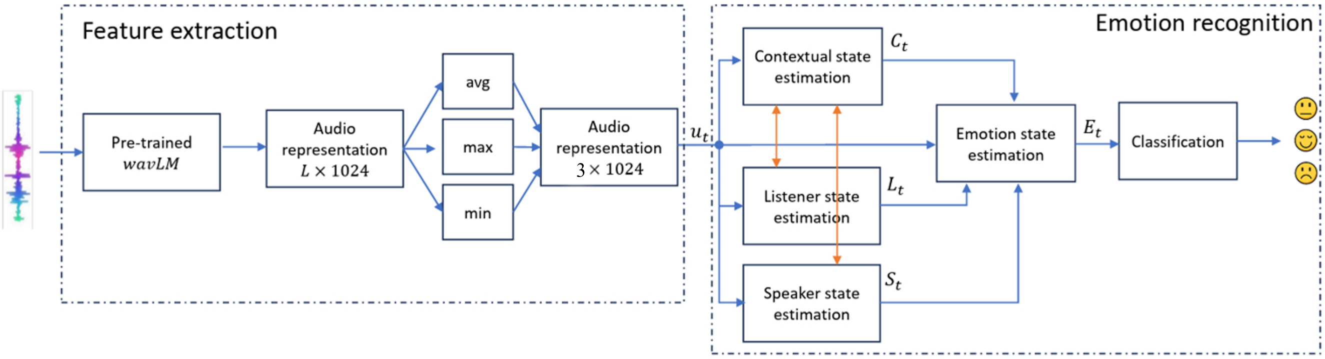
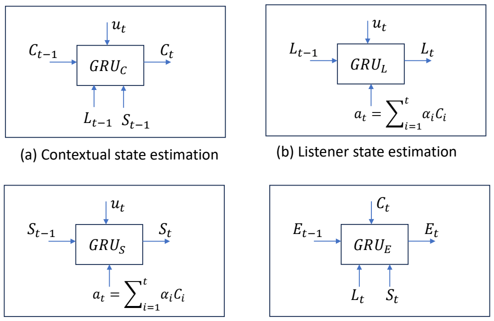


Fig. 2. Proposed emotion recognition method.



(a) Contextual state estimation

(b) Listener state estimation

(c) Speaker state estimation

(d) Emotion state estimation

Fig. 3. Update scheme of the four emotional states.

fluences in the ERC task [7]. Unfortunately, existing models were not designed to model complex dialogue context cues in different environments with arbitrary turns, listener dependency and speaker influences.

In this paper, we present an approach which can enable the co-evolution of the attentive contextual information among frames in utterances and accommodate speaker-sensitive interactions in the emotion-aware spoken dialog system. We first use a pre-trained WavLM model to extract frame-level audio representation in an utterance. Next, a statistical strategy is employed to determine emotional dynamics of utterance-level information in speech. To dynamically integrate attentive contextual information and listener and speaker state, we employ bi-directional *GRU* to model such relations in a simple, fast, parameter-efficient way. Overall, our contributions are summarized as follows:

• A frame-level enhanced speech feature extraction strategy which is able to empower a dialogue system with statistical strategy temporal emotional representation.

• We utilized the bi-directional *GRU* layer to capture attentive contextual information, and speaker-sensitive interactions, in combination with attention mechanism to highlight the important global contextual utterances.

• Our proposed approach was shown to be superior to state-of-the-art methods for conversational emotion recognition.

2. EXISTING LITERATURE

Global and local audio features of speech emotion recognition systems are typically classified into the following four categories: prosodic features, spectral features, voice quality features, and Teager Energy Operator-based features [8]. Traditionally, a number of spectral features are generally depicted using one of the cepstrum-based representations available. Commonly, Mel-frequency cepstral coefficients (MFCC) or Mel-scale spectrograms were used, and in some studies, formants were utilized as well [8]. Suraj *et al.* demonstrated the effectiveness of convolutional neural networks in emotion classification with MFCCs [5]. Besides, the direct use of Mel-scale spectrograms for ERC was proved successful as well [9]. In this work, we use a pretrained model to extract high-level acoustic features for emotion recognition.

ERC requires deep understanding of human interactions in conversations [10-13]. Some of the important works attribute emotional dynamics to be interactive phenomena [13,16], rather than being within-person. We utilize this trait in the design of our model that incorporates inter-speaker dynamic in a conversation. Since conversations have a natural temporal nature, context also plays a crucial role in emotion analysis [4,17]. Poria *et al.* employed a bi-directional LSTM to capture temporal context information of the same speaker to infer emotions [4].

However, there is no provision to model context and speaker interactive influences. We propose to capture this contextual-sensitive information via hierarchical recurrent networks. Additionally, our proposed approach adopts an interactive scheme that actively models listener and speaker emotional dynamics in conversations.

3. METHODOLOGY

As shown in Fig. 2 and 3, our proposed approach consists of two stages: 1) frame-level feature extraction using pre-trained audio models; and 2) conversation-aware emotion recognition that models the attentive context information and the interactions of participants (listener and speaker) in the dialogue.

3.1. Feature Extraction with Pre-trained Models

We adopt the pre-trained WavLM-Large model to extract frame-level audio representation. It is a recently introduced self-supervised model that improves on HUBERT [14]. WavLM-Large variant comprises of a convolutional feature encoder and 24 stacked Transformer encoders. We use the outputs of the feature encoder and all Transformer encoders as acoustic features.

To convert the frame-level representations given by WavLM into utterance-level representation, we use a statistical unit with three parallel one-dimensional statistics, average, max and min, along the sequence direction to reduce the sequence of *N* frame-wise embedding vectors and produce utterance-wise embedding vectors (see Fig.2). We concatenate them into one feature vector for utterance-wise representation, obtaining 3 × 1024-dimensional utterance-level acoustic features.

3.2. Conversation-aware Emotion recognition

The proposed module has four branches of bi-directional *GRU* cells to capture the attentive contextual information, listener, speaker and emotion state in conversations (see Fig.2 and 3).

3.2.1. Attentive Contextual State

In conversational emotion recognition, to determine the emotional state of an utterance at timestamp t, the preceding utterances can be considered as its cumulative context. The context state stores and propagates overall utterance-level information along the sequence of the conversation flow. The contextual state C_{t-1}, listener state L_{t-1} and speaker state S_{t-1} of the previous utterance, and audio representation u_t at timestamp t are used to update the contextual information from C_{t-1} to C_t (see Fig. 3(a)). The steps in the attentive contextual state update C_t are described using the following formula.

$$C_t = GRU_C(C_{t-1}, (L_{t-1} \oplus S_{t-1} \oplus u_t)) \tag{1}$$

where $\oplus$ represents concatenation. At the time step $t = 0$, the context state is randomly initialized.

In order to amplify the contribution of the context-rich information, we employ soft-attention from the history interactive context to combine long-context speaker interaction influences and conversational dependence [15]. We pool the attention vector a_t from the surrounding context history $[C_1, C_2, \ldots, C_{t-1}]$ using soft-attention. This contextual attention vector a_t can be computed as follows:

$$\begin{aligned} u_i &= \tanh(WC_i + b), \;\; 1 \leq i \leq t-1, \\ \alpha_i &= \frac{\exp(u_i^\mathsf{T})}{\sum_{i=1}^{t-1} \exp(u_i^\mathsf{T})}, \\ a_t &= \sum\nolimits_{i=1}^{t-1} \alpha_i C_i. \end{aligned} \tag{2}$$

3.2.2. Listener State

The listener state is conditioned on how the listeners tend to maintain emotions during the conversations. This state is also known as emotional inertia, as speakers may not always express explicitly their feeling or outlook through reactions. Concretely, listener state only involves listener himself/herself. Listener state refers to the emotional and psychological condition of the individual who is receiving and processing the information being communicated within a conversation. GRU_L attempts to memorize the emotional inertia of listener which represents the emotional dependency of the person with their own previous states. For time step t, the listener state is updated by the previous listener state of the person L_{t-1}, and the attentive contextual vector a_t, and the utterance u_t. At time step t, the listener state L_t can be computed as (see Fig. 3(b)):

$$L_t = GRU_L(L_{t-1}, (a_t \oplus u_t)) \tag{3}$$

3.2.3. Speaker State

The speaker state is easily observed, felt, and understood by the other participants. Speaker state refers to the emotional and psychological condition of a speaker who is engaged in a conversation. More Specifically, this state is usually about the expressions, reactions, and responses [3]. Since a speaker constantly interfere with the other participant (listener) in conversations, we construct an attentive interactive module called *Attention Interactive Dependency*. Soft-attention plays a crucial role in capturing the attentive contextual information for the interaction of participants. For the utterance at time t the speaker S_t is updated by the previous speaker state S_{t-1}, attentive contextual vector a_t, and utterance u_t. At time step t, the speaker state S_t can be computed as (see Fig. 3(c)):

$$S_t = GRU_S(S_{t-1}, (a_t \oplus u_t)) \tag{4}$$

3.2.4. Emotion State

The emotional state is reflected in the utterance's emotion and its corresponding emotional category. For the utterance at time t the emotion state E_t depends upon the previous emotion state E_{t-1} and the composite of the attentive contextual information C_t, listener state L_t, and speaker S_t. The emotion state E_t can be computed as (Fig. 3(d)):

$$E_t = GRU_E(E_{t-1}, (C_t \oplus L_t \oplus S_t)) \tag{5}$$

3.2.5. Classification

The attentive contextual state C_t, speaker state S_t and listener state L_t are input into emotion state E_t. The final output emotion state E_t is fed into two fully-connected layers with a residual connection. In conversations with two main participants, we identify the first person as the speaker and the second as the listener. When updating the speaker's state S_t, we ignore the listener's details, and vice versa for the listener's state L_t.

To train the model, categorical cross-entropy loss with softmax activation in the last layer is used as the loss function. To alleviate the problem of overfitting, we utilize L2 regularization with a weight of 0.0001 and apply dropout with a rate of p = 0.3.

4. EXPERIMENTS

4.1. Database and Metrics

We used the multi-modal and multi-speaker conversational dataset, namely Multi-modal EmotionLines Dataset (MELD) [4], in the experiments. The MELD dataset is closer to real-word conditions and contains more emotion categories than other existing benchmark emotion datasets. MELD contains acoustic, textual, and visual information from the TV series "Friends". There are seven emotion categories including: anger, disgust, sadness, joy, neutral, surprise and fear. The dataset was split into the training set, validation set and test set which contains 9989, 1109, and 2610 utterances, respectively [4]. Table 1 shows the distribution of MLED dataset.

In this work, we only used acoustic modality in related experiments. We mainly used weighted-average F1 (w-average F1) score as the evaluation metric since it is commonly used for unbalanced data and it is also adopted in previous work based on MELD dataset [4, 13, 16-19].

Table 1: Emotion distribution in the MELD dataset

Emotion	Train	Dev	Test
Anger	1109	153	345
Disgust	271	22	68
Fear	268	40	50
Joy	1743	163	402
Neutral	4710	470	1256
Sadness	683	111	208
Surprise	1205	150	281

4.2. Baselines and State-of-the-Art

Totally six state-of-the-art methods are compared in the experiments to verify the effectiveness of our proposed approach (see Table 2). bc-LSTM is the traditional method for context dependent sentiment analysis [4]. While CMN [13], FacialMMT[16] and DialogueRNN [17] are mainly to model speaker dynamic, M2FNet [18] and MMTr [19] are attention-based methods. The brief introductions of these 6 compared methods are presented below:

• bc-LSTM leverages an utterance-level LSTM to learn context dependency [4].

• CMN extracts utterance context from dialogue history information and leverages attention-based hops to capture inter-speaker dependencies [13].

• FacialMMT obtains the frame-level emotion distribution to help utterance-level emotion recognition [16].

• DialogueRNN utilizes *GRU* to capture the participant emotional states throughout conversation and the sentence-context representation between speakers [17].

• M2FNet employs a multi head attention-based fusion mechanism to learn emotion-rich latent information [18].

• MMTr acquires emotional cues at both levels of the speaker's self-context and contextual context and learns the information interactions [19].

4.3. Model Configuration

We implemented our proposed model using the Pytorch 1.11.0 framework. The model was trained with Adam optimizer with an initial learning rate of 1e-4 and a batch size of 32. Cross-entropy loss was used as the loss function for network training. To mitigate overfitting, the network was regularized by L2-norm of the model's parameters with a weight of 3e-4.

5. RESULTS AND DISCUSSION

5.1. Overall Results and Discussion

Overall results of our proposed model in comparison to the previous state-of-the-art results are presented in Table 2. Our proposed model reaches an F1 score of 45.5%, and is more effective in recognizing the seven different emotions compared to each of the other models [4,13,16-19]. In particular, it substantially improves performance on anger, joy, sadness and surprise (see Table 2), suggesting the positive effects of the attentive contextual information, listener dependency and speaker influence introduced to the proposed model for an enhanced speech emotion recognition.

Table 3: Ablation study on the MELD dataset

Method	w-average F1 (%)	Accuracy (%)
w/o statistical unit (SU)	42.5	45.9
w/o attentive contextual state	41.8	43.7
w/o listener state	42.6	45.8
w/o speaker state	42.3	45.6
Proposed model	**45.5**	**49.6**

Our proposed model infuses attentive contextual representation from surrounding utterance history and adds it to listener dependency and speaker influence to capture emotional dynamics on multi-turn conversations. To comprehensively study the impact of these four components, we removed them one at a time and evaluated their impact on the performance (see Table 3). Table 3 demonstrates that the full version of our approach achieves the best performance in both w-average F1-score and overall accuracy. The removal of either the statistical unit, attentive contextual module, listener, and speaker state adversely affected the model results, suggesting that these four modules are contributive to the overall performance of conversational emotion recognition (see Table 3).

Our method uses frame-based feature representation for utterance-level classification. The removal of statistical unit adversely affects the model results, resulting in a drop of 3.0%. This observation indicates that the statistical unit has an positive influence on capturing more emotion-relevant information in conversation. Emotions are brief in duration, most lasting only up to a few seconds. Thus, a frame-wise approach is beneficial for capturing temporal information and short-term emotion interaction. In addition, a pre-trained WavLM model retains a major portion of their prior knowledge for better high-level emotional audio features extraction. In particular, a pre-trained model is useful in the current situation where the dataset has limited size and is unbalanced. Our model excels in recognizing emotions like joy, anger, sadness, and surprise, as detailed in Table 3 and Fig. 4. Particularly with joy in the MELD dataset, its performance is notable, likely due to joy being expressed in a higher, more distinct voice tone.

As expected, the results in Table 3 imply that attentive contextual information is very important, as without its presence the performance falls by 3.7%. We suspect that attentive contextual state plays a crucial role in capturing emotion-relevant contextual information for emotional state of the participants. In this regard, it is more advantageous than prior approaches like bc-LSTM [4], which often loses the ability to determine this kind of situation. In addition, the interaction of listener and speaker are either synchronous (for example, cheer after speaking good news) or asynchronous (for example, laughter after speaking something funny). Specifically, listener and speaker state are also impactful, but less than attentive contextual state as its absence causes performance to fall by 2.9% and 3.2%, respectively. We believe the reason to be the lack of context flow from interactions of speaker and listener through the emotion representation of the preceding utterances.

Our framework is generally effective at recognizing emotions but struggles with fear and disgust (see Table 2). By analyzing the confusion matrix for the MELD dataset (see Fig. 4), we can better understand the causes of the model's inaccuracies. This issue has also been identified in previous studies [4,13,17-19] (see Table 2). In predicting emotions in conversations, our model's errors stem from a few key factors:

• Our system achieves good performance on dominant emotion (e.g.,neutral) and most of minor classes (e.g., anger, joy, sadness and surprise) than other published work [4, 13, 17-19]. However, the MELD dataset is unbalanced, and most of published work performed poorly for minor classes (e.g., fear and disgust). Unbalanced data

Table 2: Performance comparison with the state-of-the-art and baselines on MELD

Method	Anger	Disgust	Fear	Joy	Neural	Sadness	Surprise	w-average F1 (%)
bc-LSTM	21.9	0	0	0	66.1	0	16	36.4
CMN	29.6	0	0	11.8	67	0	2.8	38.3
DialogueRNN	32.1	5.1	0	11.2	53	8.3	15.6	34
FacialMMT	33.7	0	0	9.6	66.3	3.9	0	38.0
M2FNet	25.2	0	0	8	67.7	7.5	14.5	39.2
MMTr	27.3	0	0	15.8	66.9	8.2	0	38.8
Proposed method	**33.5**	**0**	**0**	**28.3**	**68.6**	**16.7**	**22.1**	**45.5**

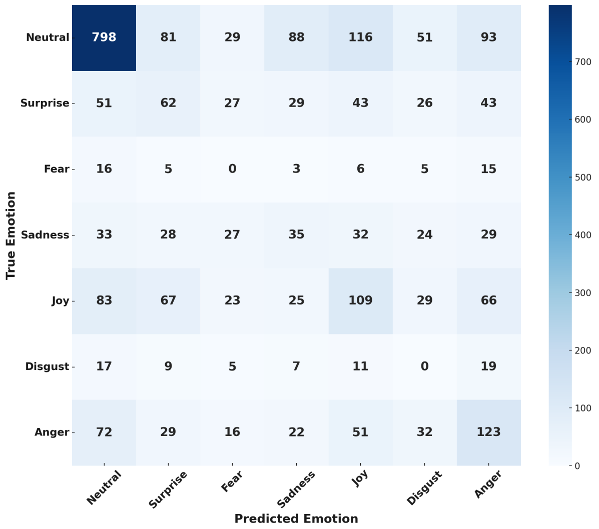


Fig. 4. Confusion matrix of the MELD dataset.

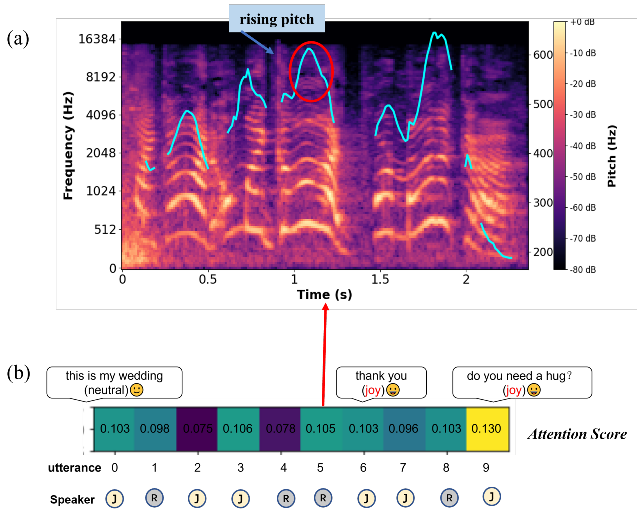


Fig. 5. Visualization of the attention weights of our model for a conversation in MELD. The spectrogram is from the 5th in a conversation snippet (Rachel (R) and Joey (J))

limits the model's ability to effectively discern fear and disgust (see Table 1 and Table 2). Delving into the datasets, we find that the dialogues in MELD dataset is relatively shorter. For speech emotion recognition task, contextual information plays a crucial role in prediction. However, a major challenge is that this task has longer input sequences with more noise and redundancy. The flow of noise information interferes with capturing long-term emotional contextual information, which leads to a significant negative effect on the model. To address this, data augmentation and transfer learning techniques will be utilized to improve the recognition of emotions that are not as well represented in the future.

● The MELD dataset contains an abundance of neutral emotions, which contribute to the frequent misclassification of other emotions as neutral. Moreover, emotions that share similar traits, such as anger and surprise, sadness and fear (see Fig. 4), often get confused due to common characteristics like high pitch and volume. Incorporating speech, facial expressions, and body language into emotion detection could enhance its accuracy, particularly in distinguishing similar emotions or interpreting unclear statements. This multimodal method gives a fuller picture of emotions, as each way of expression adds different and helpful information.

5.2. Case Studies

Fig. 5 illustrates a conversation snippet between two characters, Rachel (R) and Joey (J), with pre-defined labels from the MELD dataset classified by our proposed method. In this snippet, Joey serves as the initial speaker with a neutral state. The conversation develops with Rachel altering the topic. A notable change occurs when Rachel excitedly announces his upcoming marriage, marked by a joyful tone, high pitch, and increased frequency (see Fig. 5(a)). Joey's reaction is visible, transitioning from neutral to joy.

On the other hand, our method incorporates an attention mechanism that emphasizes key aspects within the broader conversational context, as demonstrated in Fig. 5(b). We provide a visualization of this attention process to clarify its role. In the specific instance depicted in Fig. 5(b), our system successfully identifies the emotional shift from 'neutral' to 'joy'. This exemplifies our model's proficiency in discerning and adapting to the subtle variations of emotional states in dialogues.

6. CONCLUSION

In this paper, we proposed the aggregation of frame-level speech representation in conversations. It capitalized on inferring the contextual information that incorporates dynamic listener and speaker state. An attention-based mechanism was employed to determine the important contextual-sensitive information from surrounding utterances history. The bi-directional *GRU* was used to capture contextual dependency, listener dependency and speaker influence in a simple, fast, parameter-efficient way. The experiment results on the MELD dataset demonstrate the effectiveness of the proposed model.

7. REFERENCES

[1] J. T. Wen, D. Z. Jiang, G. Tu, E. Cambira. "Dynamic interactive multiview memory network for emotion recognition in conversation,"*Information Fusion*, 2023, pp. 123-133.

[2] S. Mohammad, A. E. Sadjad, P. Maja, F. Yun. "Continuous emotion detection using EEG signals and facial expressions," *IEEE International Conference on Multimedia and Expo (ICME)*, 2014, pp. 223-229.

[3] K. Liu, J. Z. Hu, Y. T. Liu, J. Feng. "Speech emotion recognition based on discriminative features extraction," *IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 60-66.

[4] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea. "MELD: a multimodal multi-party dataset for emotion recognition in conversations,"*Proceedings of the 57th Conference of the Association for Computational Linguistics*, 2019, pp. 527-536.

[5] J. H. Yeh, T. L Pao, C. Y. Lin, Y. W. Tsai, Y. T. Chen. "Segment-based emotion recognition from continuous Mandarin Chinese speech," *Computers in Human Behavior*, 2011, vol. 27, pp. 1545-1552.

[6] S. Mao, P. C. Ching. "Deep learning of segment-Level feature representation with multiple instance learning for utterance-level speech emotion recognition," *Interspeech*, 2019, pp. 1686-1690.

[7] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, S. Poria. "COSMIC: commonSense knowledge for eMotion identification in conversations," *Findings of the Association for Computational Linguistics (EMNLP)*, 2020, pp. 1-12.

[8] M. B. Akçay, K. Oğuz. "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, 2020, vol. 116, pp. 56-76.

[9] S. Neil, K. Mikolaj, B. Pierre, C. Milos. "SERAB: a multilingual benchmark for speech emotion recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7697-7701.

[10] X. Y. Cai, J. H. Yuan, R. J. Zheng, L. Huang, K. Church. "Speech emotion recognition with multi-task learning," *Interspeech*, 2021, pp. 4508-4512.

[11] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L. P. Morency. "Context-dependent sentiment analysis in user-generated videos," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 873-883.

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

[13] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh. "DialogueGCN:a graph convolutional neural network for emotion recognition in conversation," *Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, pp. 154-164.

[14] S. Chen, C. Wang, Z. Chen Z, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022, vol. 16, pp. 1505-1518.

[15] D. K. Yang, Z. Y. Chen, Y. Z. Wang. "Context de-confounded emotion recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19005-19015.

[16] W. J. Zheng, J. F. Yu, R. Xia, S. J. Wang. "A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations" *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, vol. 1, pp. 15445-15459.

[17] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, E. Cambria. "DialogueRNN: an attentive RNN for emotion detection in conversations,"*Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 2019, pp. 6818-6825.

[18] C. Vishal, K. Purbayan, G. Ashish, S. Nirmesh, W. Pankaj, O. Naoyuki. "M2FNet: multi-modal fusion network for emotion recognition in conversation", *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 4652-4661.

[19] S. H. Zou, X. Y. Huang, X. D. Shen, H. K. Liu. "Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation", *Knowledge-Based Systems*, 2023, vol. 258, pp. 1-9.