INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Emotion Awareness in Multi-utterance Turn for Improving Emotion Prediction in Multi-Speaker Conversation

Xiaohan Shi[1], *Xingfeng Li*[2], *Tomoki Toda*[1]

[1]Nagoya University, Japan
[2]Hainan University, China

xiaohan.shi@g.sp.m.is.nagoya-u.ac.jp, lixingfeng@hainanu.edu.cn,
tomoki@icts.nagoya-u.ac.jp

Abstract

The aim of emotion prediction in conversation (EPC) is to predict the future emotional state of a speaker based on context information, which is essential for conducting a friendly human-computer conversation. Most EPC works only investigated context information by merging a speaker's multiple utterances into a single utterance per turn and focused on conversations in a dual-speaker scenario, which ignored the information in multi-utterance turn and a more complex and natural scenario of multi-speaker conversations. This paper introduces a context information modeling approach that considers potential emotional interactive information within a speaker's multi-utterance turn, which dominates his/her future emotions. Moreover, our approach advances emotion prediction in both dual- and multi-speaker conversations. Experimental results show that such an approach significantly enhances context information modeling and renders a higher accuracy in EPC than reported in the literature.

Index Terms: affective computing, emotion prediction in conversation, dialog management

1. Introduction

In recent years, conversational agents have played a particularly prominent role in our daily life. Humans can express and understand emotions of each other, while it is difficult for agents to understand the emotional changes and tendencies of the interlocutor in the conversation to continue the conversation more fluently [1]. Therefore, predicting the emotional variation is important for conducting a friendly human-machine conversation.

Emotion prediction in conversation is the task of predicting the coming emotional state of the speaker from previous context information [2]. In the prior studies, researchers have agreed that modeling conversational historical context information is vital for an accurate EPC [3, 4, 5]. Noroozi et al. [3] manually concatenated speaker turns as time series context information to predict the coming emotional state. Shahriar and Kim [2] built a model to predict coming emotions using historical information from a single speaker and also to predict emotion categories. Shi et al. [4] proposed a hierarchical gated recurrent unit (GRU) framework that individually models the speaker, interlocutor, and interaction information to perform emotion prediction. Wang et al. [5] proposed a surroundings-aware individual emotion prediction model (SAEP) based on individual and individual neighbors to predict individuals' future emotions. In most previous studies, a speaker's multiple utterances in each turn were merged into a single utterance to investigate the context information. However, different emotional states of a speaker's multi-utterance turn, which cause significant effects on his/her future emotional state, had been ignored [6, 7]. For this reason, it is better to consider a speaker's multi-utterance information per turn in predicting emotional changes instead of roughly merging the multiple utterances in each turn into a single utterance.

Moreover, multi-speaker conversations are common in daily life, including online group meetings, social gatherings, and house parties. It is vital to accurately model the interactive context information of multiple speakers for better performance in the multi-speaker scenario. This is mainly because each speaker has a specific personality and characteristic of uttering, which significantly impacts the emotional expression of each other [8]. However, how to model contextual information for such complex conversations in the EPC task is still unsolved.

To address these unexplored issues, we propose a novel multimodal emotion prediction model that examines the effects of the multi-utterance turn information on emotion prediction in multiple speakers' situations. To investigate the effect of multi-utterance turn information, we compare prediction performance characteristics using merged and full multi-utterance turn information. To investigate the effects of emotion prediction under the circumstances of multiple speakers, we utilize a dialog management module to model the contextual information of the multiple speakers, which is categorized into the speaker, interlocutor, and spectator information.

The contributions of this article are summarized as below.

- To utilize the context information effectively, we assume it is better to consider the full multi-utterance turn information rather than merge the multiple utterances per turn.
- We propose a novel structure to model the emotional information based on the speaker, his/her interlocutor, and the spectators' context information to infer the speaker's coming emotional state in multi-speaker conversations.
- Compared with the results of previous studies, our results indicate that the proposed framework can reach the improvement of 3.03% in speech, 2.34% in text, and 3.87% in multimodality when using the IEMOCAP database [9], and 1.71% in speech, 3.25% in text, and 2.82% in multimodality when using the MELD database [10].

2. Proposed Method

2.1. Multi-speaker Emotion Prediction in Conversation

Figure 1 shows the definition of a context sequence in a conversion. There are multiple speakers in the given conversation, $U = (U^A, U^B, U^C, ..., U^Z)$ represents the utterances of speaker A, his/her interlocutor B ,and spectators C ,..., Z in current conversation, respectively, and N is the number of the turns in the conversation. When predicting emotion in multi-speaker conversation, the past and current information from the

10.21437/Interspeech.2023-1236

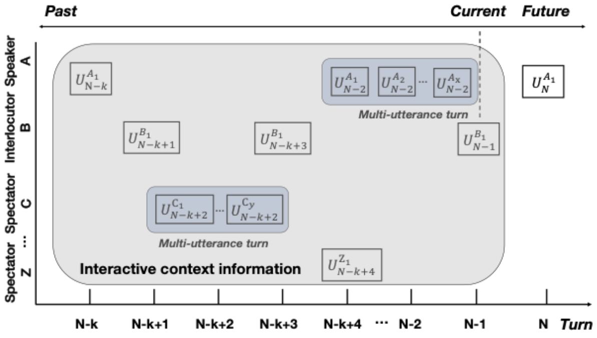

Figure 1: *Definition of context sequence in conversation*

speaker, interlocutor, and spectators are given, and the target is to predict the speaker's future emotion variation.

We take speaker A, interlocutor B, spectators C ,..., Z all into account. For instance, the information for the speaker can be denoted as $(U^{A1}_{N-k}, ..., U^{A1}_{N-2}, U^{A2}_{N-2}, ..., U^{A_x}_{N-2})$, The interlocutor information can be denoted as $(U^{B_1}_{N-k+1}, U^{B_1}_{N-k+3}, ..., U^{B_1}_{N-1})$, and the spectator information would only be considered in context information, as fellow: $(U^{A_1}_{N-k}, U^{B_1}_{N-k+1}, U^{C_1}_{N-k+2}, ..., U^{Z_1}_{N-k+4}, ..., U^{A_x}_{N-2}, U^{B_1}_{N-1})$.

In this study, we use category values to describe the emotional state; thus, the emotion label can be represented as $L_N = (L_{\text{Happy}}, L_{\text{Anger}}, L_{\text{Neutral}}, L_{\text{Sad}})$, where L_N denotes the label of the N-th turn.

2.2. Model Description

We propose a hierarchical model to predict the coming emotion using the interactive context in a conversation. As shown in Figure 2, the network mainly consists of three components, namely an encoder for the speaker and his/her interlocutor, an encoder for interactive context information, and a dialog management unit. These three components serve for different purposes in the encoding process. The encoder for the speaker and his/her interlocutor is set to capture the individual speaker's emotion information. The encoder for interactive context information is set to capture multi-speaker interaction context information. The dialog management unit is set to track the role of each speaker in the conversation.

Since each encoder deals with time series tasks and the gated recurrent unit (GRU) is suitable for capturing the temporal properties of the data, we use GRU for the each encoder in our model. The GRU network is an enhanced version of the recurrent neural network (RNN) that has advantage of being able to handle the vanishing gradient problem [11]. Each GRU cell consists of an update gate (z) and a reset gate (r) to control the flow of information. Let x be the input to the GRU network, W and U the weight parameters ,and b denote the bias of the GRU network. At time step t, the hidden state (h_t) can be computed as

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{1}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{2}$$

$$\begin{aligned} h_t =& (1 - z_t) \odot h_{t-1} + z_t \odot \tanh(W_h x_t \\ &+ U_t(h_{t-1} \odot r_t) + b_h). \end{aligned} \tag{3}$$

To capture the context information of multi-speaker, the utterance-level feature is conveyed to the interaction emotion GRU layer of the encoder for context information. Then, using the dialog management unit, the hidden state at each time step of the interaction emotion GRU layer is separately conveyed to the individual emotional layer. The individual layer can capture the context information of the speakers and his/her interlocutor. Finally, we concatenate the information from the speaker h_S, interlocutor h_I, and interaction h_A, using a fully-connected layer (FC) and softmax layer to predict the emotion label. The above process is summarized as follows:

$$h = \text{Concatenate}(h_S, h_I, h_A) \tag{4}$$

$$\hat{y} = \text{Softmax}(f_\theta(h)) \tag{5}$$

where $\hat{y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)^T$ is the predicted probability distribution, n is the number of emotion categories, f_θ is the fully connected layer with parameter θ.

In the encoder for the speaker and his/her interlocutor, to effectively capture the emotion information, following the previous study [12], we utilize the self-attention layer in the encoder. We first calculate each hidden layer information of the GRU, then calculate queries, keys, values of dimension d_k, d_k, d_v with linear projections. By packing them into matrices Q, K, V, the attention output is calculated as

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V. \tag{6}$$

3. Experiment

3.1. Database

We experiment with two publicly available multimodal emotion databases, i.e., Interactive Emotional Dyadic Motion Capture (IEMOCAP) [9] and the Multimodal EmotionLines Dataset (MELD) [10].

The IEMOCAP database is a widely used corpus in affective computing. It contains approximately 12 hours of audio-visual recordings and is designed for two-person dialogs. Each conversation in IEMOCAP has been segmented into utterances with the continuous label in the Valence-Arousal dimension and category label in categories such as anger, happiness, sadness, and neutrality.

MELD is a prevailing database for multi-speaker emotion tasks. It contains more than 1,400 conversations and 13,000 utterances from the TV series *Friends*. Emotion annotation in the dataset includes seven emotion labels (anger, disgust, sadness, joy, surprise, fear, and neutral).

The statistics of the databases are shown in Table 1. Each database contains both single and multi-utterance turn. Single utterance turn represents the speaker turn which only has one utterance per turn, and multi-utterance turn represents the speaker turn which has more than one utterance per turn.

Table 1: *The number of speaker turn in each database*

Turns	IEMOCAP	MELD
Single utterance Turn	7256	9075
Multi-utterance Turn	1166	1984

In the experiment, we choose four categories (happy, sad, neutral, and angry) and six turns of interactive context information in each database to predict the coming emotional state. Therefore, for the IEMOCAP database, the total training data in the experiment is 4043, and the numbers of emotional utterances of neutral, happy, angry, and sad are 1254, 1166, 929,

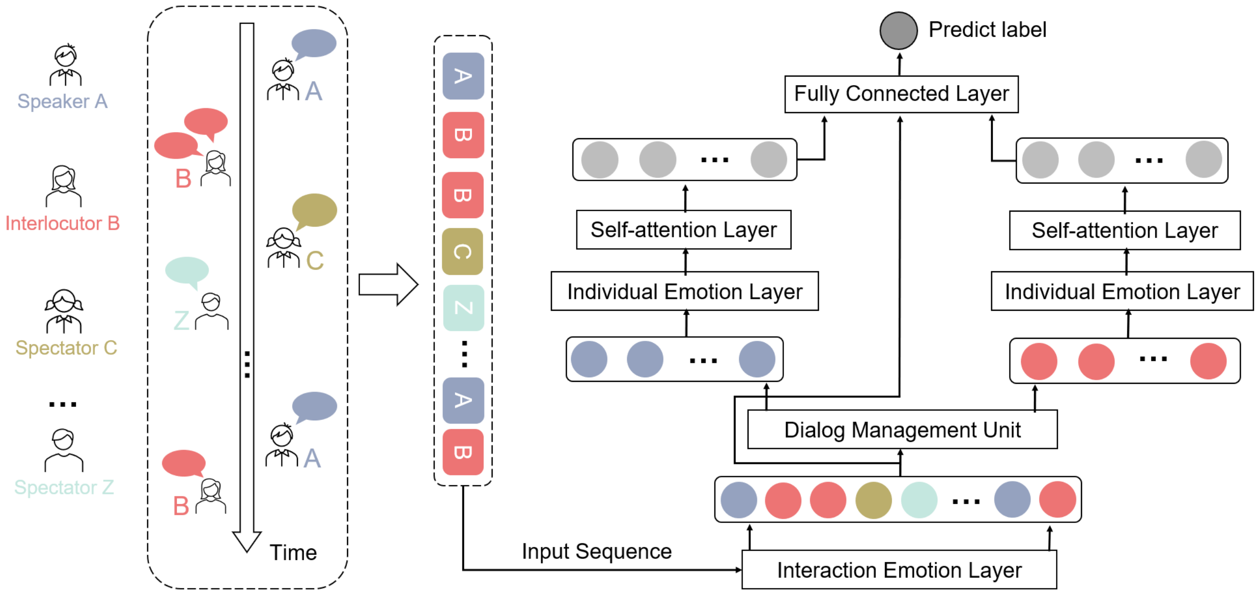

Figure 2: *Structure of multimodal emotion prediction model based on interactive context*

and 694, respectively. To keep with prior work [13], we use 10-fold cross-validation during the experiments to ensure that the experimental process is speaker-independent. For the MELD database, following the work of [14], we use predefined Train-data and Val-data to train and predefined Test-data to test the model. The numbers of training data is 4338, and the test data is 806.

3.2. Experimental Procedure

We conduct two experiments in this study. In Experiment 1, we investigate the effect of multi-utterance turn information on emotion prediction. In Experiment 2 we investigate the effects of multiple speakers' situation on emotion prediction.

In Experiment 1, we study different context information by comparing the performance of three approaches.

- Shahriar and Kim [2]: Only consider the timeline but ignore the speaker's turn information.
- Shi et al. [4]: Consider the speaker's turn, but use only merged multi-utterance turn information in each turn.
- Proposed model: Consider the speaker's turn and use full multi-utterance turn information in each turn.

In Experiment 2, we study the effects of a different method for multiple speakers by comparing the performance characteristics of three approaches.

- BLSTM: Model the multi-speaker's context information based on the timeline but ignore the speaker's turn information.
- Shi et al. [4]: Split the multi-speaker's context information into the speaker's and the others' information.
- Proposed model: Use dialog management to split the context information into the speaker's, interlocutor's, and spectator's information.

3.3. Features

In this study, we use acoustic and textual features to describe context information.

Speech: For alignment with previous works [15], two aspects of acoustic information are utilized to obtain acoustic features for the input utterances. We use openSMILE with the eGeMAPS feature set [16] for heuristic features and the Wav2Vec model for the pre-training representation features. [17]

- Wav2Vec is a self-supervised model which learns representations from raw audio data. It has shown impressive performance on many downstream tasks like automatic speech recognition (ASR) [18], and speech emotion recognition (SER) [19]. Therefore, we employ this unsupervised model as a feature extractor for raw audio samples. More specifically, we employ the Wav2Vec 2.0 (wav2vec2-base-960h)[1], which is trained on 960-hour LibriSpeech audio data [20].

Text: We choose the pre-trained model BERT to obtain the representation for the input text utterance [21].

- BERT is a self-supervised model that utilizes text data to learn representations. It showed good performance in representing semantic meanings in vector space and has demonstrated impressive efficacy in text emotion recognition [22]. More specifically, we employ the bert-as-service toolbox[2] in this study.

In the data pre-processing stage, we extract 88-dimensional heuristic features with the openSMILE toolbox and 768-dimensional representation features with the pre-trained model. The pre-trained model BERT we used to extract 378-dimensional textual features is from bert-as-service [23].

3.4. Implementation

All our deep learning models are implemented with Python 3.7 and Pytorch 1.11.0. In the training process, each GRU for the model is set as two layers with 256 units and a dropout rate of 0.5. The optimizer is set as the Adam optimizer with a learning

[1] https://huggingface.co/facebook/wav2vec2-base-960h

[2] https://github.com/hanxiao/bert-as-service

rate of 0.0001 and the cross-entropy loss function.

3.5. Evaluation

Because of the unbalanced data distribution in each database, we utilize unweighted Average Recall (UAR) and F1 which has been widely used in unbalanced data experiments to evaluate the performance [24].

4. Results and Evaluation

To analyze the multi-utterance turn information effect, we compare the prediction performance characteristics of merged multi-utterance turn information with full multi-utterance turn information in the dual-speaker situation, as shown in Table 2.

Table 2: *The performance using multi-utterance turn information*

Modality	Model	IEMOCAP	
		UAR	MacroF1
Speech	Shahriaret al., [2]	56.78%	55.11%
	Shi et al., [4]	61.98%	60.21%
	Proposed	**65.01**%	**65.91**%
Text	Shahriaret al., [2]	71.19%	70.65%
	Shi et al., [4]	74.96%	74.54%
	Proposed	**77.30**%	**76.67**%
Text + Speech	Shahriaret al., [2]	74.61%	73.62%
	Shi et al., [4]	76.31%	75.50%
	Proposed	**80.18**%	**80.01**%

Apparently, using full multi-utterance turn information achieves higher performance than that using merged multi-utterance turn information alone. The observed UAR improvements are 3.03% for speech modality, 2.34% for text modality, and 3.87% for multimodality. This implies that it is better to consider complete full multi-utterance turn information for emotional prediction rather than to only consider merged multi-utterance turn information alone. Moreover, the UAR improvements are 8.23%, 6.11%, and 5.57%, respectively, in comparison with the result of the previous work [2]. This implies that speaker turn information is effectively represented by context information.

Table 3: *Confusion matrix of results of our proposed model in multimodality*

Ground Truth	Prediction			
	Neutral	Happy	Angry	Sad
Neutral	867	184	110	93
Happy	118	1018	17	13
Angry	95	22	786	26
Sad	80	25	36	553

Table 3 shows the confusion matrix of the results of our proposed model on the IEMOCAP dataset. The classification of the neutral emotion has a low performance of around 69.14%; however, all of the other three emotions have a UAR of over 80%. It is easy to misclassify the neutral emotion to the happy emotion, and it is also easy to misclassify the happy emotion to the neutral emotion.

Table 4: *Comparison of performance characteristics in the multi-speaker situation*

Modality	Model	MELD	
		UAR	MacroF1
Speech	BLSTM	25.22%	20.66%
	Shi et al., [4]	26.96%	25.13%
	Proposed	**28.67**%	**25.97**%
Text	BLSTM	41.41%	41.45%
	Shi et al., [4]	42.19%	42.67%
	Proposed	**45.44**%	**45.13**%
Text + Speech	BLSTM	42.31%	42.46%
	Shi et al., [4]	42.39%	42.51%
	Proposed	**45.21**%	**44.36**%

Table 4 shows the performance results obtained using the proposed model in the multi-speaker situation. Apparently, achieving good performance in multi-speaker situations is a challenging task. Compared with the previous work [4], the proposed model shows an improvement in emotion prediction model performance in multi-speaker situations while using the dialog management unit. The observed UAR improvements are 1.71% for speech modality, 3.25% for text modality, and 2.82% for multimodality. This shows that using the dialog management unit to take the multi-speaker situation into account improves the EPC performance. This implies that context information should be modeled on the basis of speaker, interlocutor, and spectators turns in multi-speaker situations.

5. Conclusions and Future Work

In this paper, we took multi-utterance turn information and multi-speaker situation effects into consideration and proposed an emotion prediction model using interactive context information. Our results showed that the performance is better when considering full multi-utterance turn information rather than only merged multi-utterance turn information. Furthermore, We introduced a dialog management module in multi-speaker conversations, which achieved higher accuracy in emotion prediction. In future efforts, we recommend further investigation of alternative methods and models for interaction information in situations involving multiple speakers.

6. Acknowledgements

This work was financially supported by JST SPRING, Grant Number JPMJSP2125, and in part by JSPS KAKENHI Grant Number 21H05054. The author would like to take this opportunity to thank the Interdisciplinary Frontier Next-Generation Researcher Program of the Tokai Higher Education and Research System.

7. References

[1] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, “Emotion recognition in conversation: Research challenges, datasets, and recent advances,” *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[2] S. Shahriar and Y. Kim, “Audio-visual emotion forecasting: Char-

acterizing and predicting future emotion using deep learning," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–7.

[3] F. Noroozi, N. Akrami, and G. Anbarjafari, "Speech-based emotion recognition and next reaction prediction," in *2017 25th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2017, pp. 1–4.

[4] X. Shi, S. Li, and J. Dang, "Dimensional emotion prediction based on interactive context in conversation." in *INTERSPEECH*, 2020, pp. 4193–4197.

[5] Y. Wang, Y. Du, J. Hu, X. Li, and X. Chen, "Saep: A surrounding-aware individual emotion prediction model combined with t-lstm and memory attention mechanism," *Applied Sciences*, vol. 11, no. 23, p. 11111, 2021.

[6] X. Lu, B. Di Eugenio, T. C. Kershaw, S. Ohlsson, and A. Corrigan-Halpern, "Expert vs. non-expert tutoring: Dialogue moves, interaction patterns and multi-utterance turns," in *Computational Linguistics and Intelligent Text Processing: 8th International Conference, CICLing 2007, Mexico City, Mexico, February 18-24, 2007. Proceedings 8*. Springer, 2007, pp. 456–467.

[7] M. Atterer, T. Baumann, and D. Schlangen, "Towards incremental end-of-utterance detection in dialogue systems," in *Coling 2008: Companion volume: Posters*, 2008, pp. 11–14.

[8] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations." in *IJCAI*, 2019, pp. 5415–5421.

[9] 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, pp. 335–359, 2008.

[10] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[11] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[12] Y. Li, T. Zhao, T. Kawahara *et al.*, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning." in *Interspeech*, 2019, pp. 2803–2807.

[13] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "Cnn+ lstm architecture for speech emotion recognition with data augmentation," *arXiv preprint arXiv:1802.05630*, 2018.

[14] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "Cosmic: Commonsense knowledge for emotion identification in conversations," *arXiv preprint arXiv:2010.02795*, 2020.

[15] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[16] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[17] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[18] J. Li, V. Manohar, P. Chitkara, A. Tjandra, M. Picheny, F. Zhang, X. Zhang, and Y. Saraf, "Accent-robust automatic speech recognition using supervised and unsupervised wav2vec embeddings," *arXiv preprint arXiv:2110.03520*, 2021.

[19] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[22] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of bert-based approaches," *Artificial Intelligence Review*, pp. 1–41, 2021.

[23] H. Xiao, "bert-as-service," https://github.com/hanxiao/bert-as-service, 2018.

[24] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.