SPEECH EMOTION RECOGNITION WITH CO-ATTENTION ON ACOUSTIC AND LINGUISTIC FEATURES FROM PAST UTTERANCES

Bella Godiva, Jeongsoo Choi, Jeong Hun Yeo, Yong Man Ro

Integrated Vision and Language Lab., School of Electrical Engineering, KAIST
{bellagodiva, jeongsoo.choi, sedne246, ymro}@kaist.ac.kr

ABSTRACT

Speech Emotion Recognition (SER) is a task focused on categorizing a speaker's emotional state from their speech. Existing approaches presented networks that focus solely on analyzing acoustic information like pitch and rhythm in current utterance and often neglect the broader context in which the utterance is spoken. To mitigate this issue, we propose a novel SER framework that is designed to comprehensively understand the context from past utterances while also jointly utilizing the acoustic information from the current utterance. We construct the proposed method as a two-stream network. One stream integrates acoustic features from the current speech, while the other incorporates linguistic context from both past and present utterances. We achieve state-of-the-art performance and validate the effectiveness of incorporating linguistic context through extensive experiments.

***Index Terms*—** Speech Emotion Recognition in Conversation (SERC), ASR, HuBERT, RoBERTa

1. INTRODUCTION

Speech Emotion Recognition (SER) is a task to classify each user's utterance into one of a fixed set of emotion categories like happy, sad, angry, or fearful. This technology can be used for interactive intelligent systems in domains like customer service, marketing, and healthcare. Based on these points, SER technologies have received increasing attention.

The main research stream for SER based on deep learning focuses on modeling how well the acoustic information such as prosody, pitch, and rhythm from a speech utterance predicts the emotional state. From this perspective, some works have been presented. Recent advances use methods such as attention mechanism on low-level features (MFCC) [5, 19, 15], fine-tuning of SSL features (HuBERT, Wav2Vec) [18, 6, 10], or a mix of both [12, 14]. However, these previous methods focus only on analyzing acoustic information of the current utterance. Research in multimodal emotion recognition [13, 9] has highlighted the importance of context in which the utterance is spoken. For instance, when the underlying context revolves around a joyful subject, a greater presence of positive emotions like happiness is expected. Therefore, a novel approach must be devised to enable contextual emotion recognition when only speech information is available.

In this paper, we focus on developing a novel SER method that comprehensively understand the context from past utterances while also jointly utilizing the acoustic information from the current utterance. According to the results of [13, 9], text-based emotion recognition achieves the highest accuracy compared to speech and visual modalities. Adding speech and visual modality on top of text modality only boosts the system performance by an insignificant value of around 1-2%. Speech contains speaker bias where each person has a different loudness or pitch of voice while expressing the same emotion. Utilizing only visual modality is also challenging because the size and quality of picture are not appropriate for facial emotion features to be extracted well. Nonetheless, in numerous real-world scenarios, text transcriptions are frequently unavailable. On the other hand, speech stands as one of the fundamental and instinctive forms of human interaction. Taking inspiration from these observations, our objective is to harness text transcriptions obtained via the Automatic Speech Recognition (ASR) system from speech modality to supplement acoustic features of speech in recognizing the emotion of the current utterance. With the transcribed text obtained through Automatic Speech Recognition (ASR), we can now integrate contextual information from previous utterances. This was a challenge not achievable with earlier methods of speech emotion recognition.

To address the challenges mentioned above, here we summarize the contribution of this study:

- Incorporates linguistic context from past utterances into speech emotion recognition through ASR (Automatic Speech Recognition) and language model.
- Achieves SoTA (State-of-The-Art) on speech emotion recognition in conversation on IEMOCAP dataset for speaker-independent setting.

2. RELATED WORKS

Early end-to-end methods combine a convolutional neural network (CNN) and a long short-term memory (LSTM) [17].

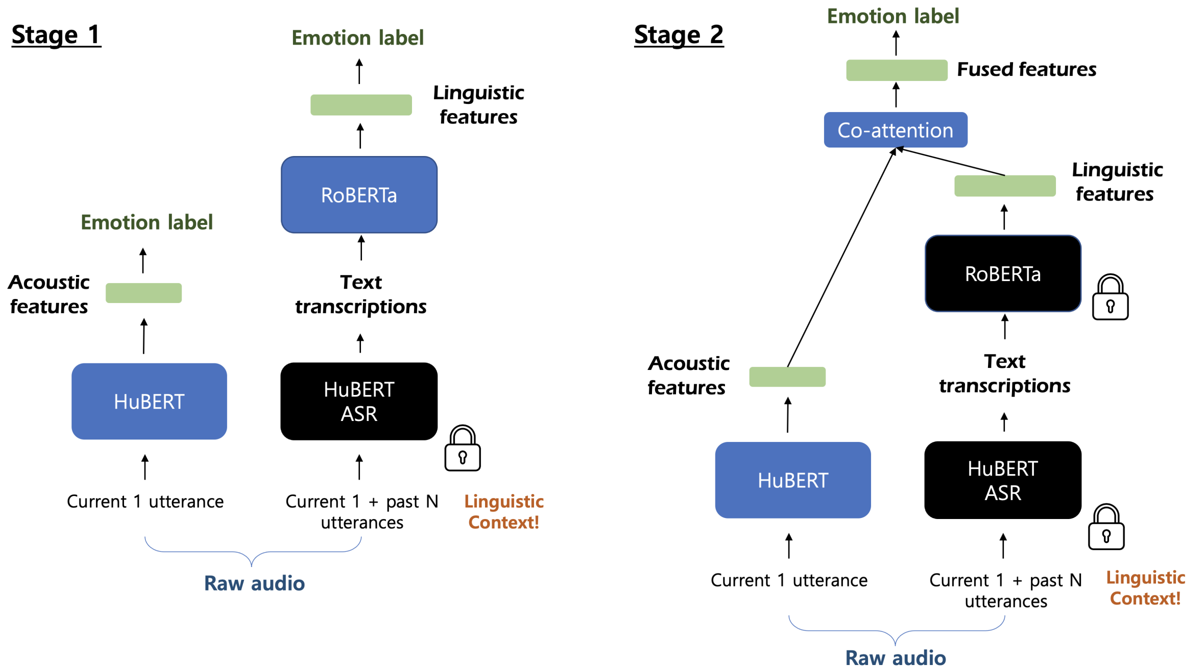

Fig. 1: Proposed Model Architecture and Two-Stage Freeze training (Section 3.3). Blue blocks represent trainable blocks while Black blocks represent non-trainable blocks. Stage 1: RoBERTa and HuBERT are trained separately on emotion classification. Stage 2: The pre-trained RoBERTa from Stage 1 is frozen, the pre-trained HuBERT from the first stage is fine-tuned, and the co-attention transformer and linear classification are trained from scratch on the emotion classification task.

This traditional LSTM or GRU-based methods suffer from vanishing gradients in longer sequences that hinder the learning of long-term dependency. Later, transformer-based models with variations of attention mechanisms such as local, hierarchical, or windowed attention outperformed the CNN and LSTM combination. [4, 5]. In recent years, self-supervised learning models like Wav2Vec and HuBERT have generated significant interest in speech processing research due to their ability to learn high-quality speech representation. Yang et al. [18] fine-tunes such models on speech emotion recognition. For transformer and self-supervised learning models based methods, a longer temporal view of speech features requires higher computational memory which is often infeasible. Li et al. [12] has the most similar approach to us where they fuse ASR outputs with MFCC features in joint training. However, existing work does not yet explore the linguistic context from past utterances.

3. PROPOSED METHOD

3.1. Linguistic Context Features from Past Utterances through ASR and Language Model

Large pre-trained speech models like Wav2Vec [2] and HuBERT [8] produce high-quality speech representation and have been proven to achieve one of the best results on ASR. In our work, we are using a pre-trained HuBERT model (see Figure 1) to extract text transcriptions from raw audio. HuBERT's convolutional encoder is frozen and the rest of the weights are fine-tuned with CTC [7] loss. The projection layer(s) of HuBERT is removed and replaced with a randomly initialized softmax layer. The CTC target vocabulary includes 26 English characters, a space token, an apostrophe, and a special CTC blank symbol.

After the transcripts of current and past utterances are available, we can now extract emotional linguistic context features by fine-tuning a large pre-trained language model on the emotional classification task. In this work, RoBERTa [16] is chosen as the pre-trained language model. RoBERTa uses two $\langle /s \rangle$ tokens consecutively as the [SEP] token, which separates the first and the second segments. We use the first segment to store word embedding tokens of past utterances and the second segment to store word embedding tokens of the current utterance of a dialog. This segmentation helps the model to distinguish between past utterances and the current utterance and attend to corresponding parts that contribute the most to the final classification [11]. RoBERTa uses $\langle s \rangle$ and $\langle /s \rangle$ as [CLS] and [EOS] tokens respectively.

3.2. Acoustic Features and Modality Fusion

Acoustic features from speech can be extracted through diverse methods, such as utilizing self-supervised learning representation [2][8], prosody features like pitch and energy

[1], or MFCC [19]. In our approach, we input raw audio of the current utterance into HuBERT and finetune HuBERT on emotion classification task. The hidden state output serves as our emotional acoustic feature. This acoustic feature is then fused with the linguistic context features from past utterances through the co-attention transformer layer and fed into a linear layer that functions as the emotion classifier.

3.3. Training Strategies

To achieve optimal results, selecting an appropriate training strategy is crucial. The architecture consists of four components that require training: HuBERT which extracts emotional acoustic features, RoBERTa which extracts emotional linguistic context features, co-attention transformer which fuses the extracted emotional features, and the classification layer which predicts the emotion class. Various approaches exist for training these components, including:

- One-Stage Joint training: RoBERTa, HuBERT, co-attention transformer, and linear classification layers are trained in parallel on the emotion classification task.

- Two-Stage Fine-tune training: RoBERTa and HuBERT are trained separately on emotion classification in the first stage. In the second stage, the pre-trained RoBERTa and HuBERT from the first stage are fine-tuned while the co-attention transformer and linear classification are trained from scratch on the emotion classification task.

- Two-Stage Freeze training: RoBERTa and HuBERT are trained separately on emotion classification in the first stage. In the second stage, the pre-trained RoBERTa from the first stage is frozen, the pre-trained HuBERT from the first stage is fine-tuned, and the co-attention transformer and linear classification are trained from scratch on the emotion classification task.

We opt for the two-stage freeze training because training two large models concurrently (RoBERTa and HuBERT) requires a huge amount of memory which sometimes cannot be afforded by a regular GPU. Furthermore, this strategy allows us to maximize the potential of both RoBERTa and HuBERT. During the first stage, each model can concentrate solely on extracting linguistic and acoustic features, respectively, without interference from one another. This ensures the models reach their optimal feature extraction capabilities. Subsequently, in the second stage, these high-quality features are fused, capitalizing on their respective strengths to enhance each other through the co-attention layer. The effectiveness of this approach is demonstrated through the experiments conducted in Section 4.3 of the paper.

4. EXPERIMENTS

4.1. Dataset and Training Settings

In these experiments, we use the IEMOCAP corpus [3] as it is the most used dataset in SER research. It consists of five dyadic sessions with ten actors. We use three sessions for training, one session for validation, and one session for testing. The corpus contains approximately 12 hours of speech where each utterance is annotated with 1 out of 9 emotions (angry, excited, fearful, sad, surprised, frustrated, happy, disappointed, and neutral). Based on prior research (Table 4), we combine happy and excited and use four categories: angry, happy, neutral, and sad. We removed utterances that did not have audio, bringing the total number of utterances used in this study to 5500.

We trained speech emotion recognition task in speaker-independent settings. For the ASR system, we use the weight of HuBERT Large model pre-trained with 60k hours of Libri-Light and fine-tuned on 960 hours of Librispeech dataset available on Fairseq Git Hub. RoBERTa Large is used for all experiments. The linear classification layer consists of one linear layer without an activation function (we tried using more layers with an activation function on each layer but it performed worse). Training is done on a single NVIDIA Titan RTX GPU. We use AdamW optimizer, gradient clip of 0.8, and learning rate ranging between $1e^{-5}$ to $1e^{-6}$.

4.2. Effect of Numbers of Past Utterances Used for Context and Effect of WER from ASR

Table 1 illustrates the impact of various numbers of past utterances incorporated for context. Experiment shows that incorporating more past utterances ($0 \rightarrow 5 \rightarrow 10$) into context enhances the performance of emotion recognition. However, after a certain threshold, adding more past utterances ($10 \rightarrow 15 \rightarrow 30$) begins to detrimentally affect the performance. Using too few past utterances might not provide the model with enough information to grasp the linguistic context of the present utterance. Conversely, employing an excessive number of past utterances could introduce noise into the model. Utterances too distant in the past might become irrelevant to the current utterance's emotion, especially considering how emotions tend to fluctuate during a conversation.

Word Error Rate(WER) measures the difference between the transcribed output generated by the ASR system and the reference (ground truth) transcript. This transcription errors can negatively impact the accuracy of speech emotion recognition. From Table 1, we can notice that the model trained with transcriptions from ASR (ASR 0past) performs 8.8% worse than the model trained on ground truth text (GT 0past). Incorporating lingusitic context from past utterances reduces the WER-induced degradation from ASR. With 10 past utterances, the gap between ASR-trained (ASR 10past) and ground truth-trained models (GT 10past) decreases to just 1.6

Table 1: Effect of WER and numbers of past utterances used for context on emotion classification F1 score.

WER	24.2	23.9	23.8	16.7	27.8
Model	**Overall**	**Neut**	**Sad**	**Ang**	**Hap**
GT 0past	67.3	65.5	68.2	66.9	68.7
GT 5past	80.7	75.9	83.9	76.8	84.7
GT 10past	**81.7**	**77.3**	**85.5**	77.6	85.0
GT 15past	80.2	72.1	80.4	**77.9**	**88.1**
GT 30past	80.4	73.6	82.7	75.6	86.7
ASR 0past	58.5	58.0	58.1	60.4	58.3
ASR 5past	77.6	73.2	80.5	72.0	82.1
ASR 10past	**80.1**	**75.9**	80.7	**76.4**	84.7
ASR 15past	78.7	70.6	81.4	72.2	**86.7**
ASR 30past	78.4	71.8	**81.9**	73.1	84.1

4.3. Effects of Training Strategies

As HuBERT and RoBERTa models are too large for the one-stage and two-stage finetune to be carried out on a single 24GB GPU, we use prosody features (pitch and energy) to implement the experiments. Pitch and energy are concatenated and projected into a dimension of 1024 using a linear layer and fed into a transformer encoder consisting of 2 attention heads with 4 layers each and topped with an average pooling layer. These prosody features are then concatenated with the linguistic context features before being fed into the classification layer. Based on the experiment results, the two-stage freeze strategy emerges as the most effective approach for fusing prosody and linguistic features (Table 2). In joint training, the model's focus gets divided between learning parameters for both models. RoBERTa, due to its larger parameter size, demands more comprehensive training compared to the prosody transformer encoder. By enabling RoBERTa to undergo isolated training for emotion classification in the initial stage, it is allowed to reach its maximum learning potential in extracting emotional linguistic features. Subsequently, in the second stage, these emotional linguistic features are concatenated with emotional prosody features, mutually enhancing each other's contributions to the model's ability to differentiate emotions in an utterance. These findings lead us to adopt the two-stage freeze training strategy for the HuBERT/RoBERTa architecture depicted in Figure 1 for optimized results and efficient memory utilization.

Table 2: Effect of training strategies on emotion classification F1 score.

Model	WA
ASR 10past + Prosody one-stage	77.1
ASR 10past + Prosody two-stage finetune	76.6
ASR 10past + Prosody two-stage freeze	**80.0**

4.4. Importance of Linguistic Context from Past Utterances and Co-attention Layer

All experiments are done in 10 past utterances and two-stage freeze training settings, both of which were determined as optimized configurations based on previous experimentations. We can see that incorporating linguistic context from past utterances boosts the performance by 12% (Table 3) compared to only using HuBERT. Adding a co-attention transformer on top of the concatenated emotional and linguistic features further boosts the performance by 3.6% (Table 3). This shows the effectiveness of a co-attention transformer in fusing the two modalities to complement one another in learning emotional features that helps in emotion classification. We conducted experiments across all configurations of HuBERT and RoBERTa, yet this HuBERTb+RoBERTaL setting demonstrated the most superior performance.

Table 3: Importance of linguistic context from past utterances and co-attention layer

Model	Overall
HuBERTb (acoustic)	66.2
HuBERTb + RoBERTaL (linguistic context)	78.3
HuBERTb + RoBERTaL (linguistic context) + co-attention	**81.9**

Table 4: Comparison with SER methods

Paper	Method	Year	WA
[12]	MFCC+W2V2+ASR	2022	63.4
[5]	DSTransformer	2023	71.8
[19]	MFCC+TIM-Net	2023	72.5
[14]	MFCC+W2V2+A-MLF	2023	72.9
[18]	HuBERT	2022	73.1
[6]	HuBERT+SpkNorm	2022	74.2
[10]	HuBERT+LAM	2023	75.4
[15]	MFCC+TFA+TFW	2023	81.6
Proposed	**HuBERT+ASR +RoBERTa(ling context)**	**2023**	**81.9**

5. CONCLUSION

Our proposed method achieves SoTA outcome on speech emotion recognition in speech conversation on IEMOCAP dataset for speaker-independent setting. Table 4 provides comparisons with the latest models. Through extensive experiments, co-attention layer and linguistic context from past utterances through ASR has proven their effectiveness in aiding acoustic features (HuBERT features). It outperforms other HuBERT-based methods and ASR-based methods.

6. REFERENCES

[1] Starlet Ben Alex, Leena Mary, and Ben P. Babu. Attention and feature selection for automatic speech emotion recognition using utterance and syllable-level prosodic features. 39(11):5681–5709, nov 2020.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. volume 42, pages 335–359, December 2008.

[4] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. Speechformer: A hierarchical efficient framework incorporating the characteristics of speech, 2022.

[5] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. Dst: Deformable speech transformer for emotion recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[6] Itai Gat, Hagai Aronowitz, Weizhong Zhu, Edmilson da Silva Morais, and Ron Hoory. Speaker normalization for self-supervised speech emotion recognition. *CoRR*, abs/2202.01252, 2022.

[7] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks. volume 2006, pages 369–376, 01 2006.

[8] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *CoRR*, abs/2106.07447, 2021.

[9] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164. Association for Computational Linguistics, July 2022.

[10] Lei Kang, Lichao Zhang, and Dazhi Jiang. Learning robust self-attention features for speech emotion recognition with label-adaptive mixup, 2023.

[11] Taewoon Kim and Piek Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *CoRR*, abs/2108.12009, 2021.

[12] Yuanchao Li, Peter Bell, and Catherine Lai. Fusing asr outputs in joint training for speech emotion recognition, 2022.

[13] Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1610–1618, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[14] Ke Liu, Jingzhao Hu, and Jun Feng. Speech emotion recognition based on low-level auto-extracted time-frequency features. In *ICASSP 2023*, pages 1–5, 2023.

[15] Ke Liu, DongYa Wu, Dekui Wang, and Jun Feng. Speech emotion recognition via heterogeneous feature learning. In *ICASSP 2023*, pages 1–5, 2023.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[17] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, 2016.

[18] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. 2022.

[19] Jiaxin Ye, Xincheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In *ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5, 2023.