Takumasa Kaneko

University of Electro-Communications
Chofu, Tokyo
Japan

k2440004@gl.cc.uec.ac.jp

1 Research interests

Emotion recognition is vital for improving the quality of human–human and human–machine inter- actions. Especially in spoken dialogue systems (SDSs), responses can be generated by predicting users' emotions and considering their intentions. The response content can change depending on the user's emotion for similar utterances, allowing for natural communication. However, emotion recognition remains a challenging task, and responses based on incorrect emotion predictions can significantly impair the user experience.

One of my research interests is multimodal emotion recognition. Studies have proposed several emotion recognition models using multiple modalities to improve recognition accuracy (Sun et al., 2023; Wu et al., 2023; Yang et al., 2023). Multimodal emotion recognition models perform better than unimodal ones.

Another area that caught my attention is **personalization** in the emotion recognition task. Specifically, methods for personalization without fine-tuning have been recently proposed (Tran et al., 2023). Personalization without fine-tuning can greatly improve the utility and user experience of dialogue systems.

1.1 Speech Emotion Recognition

Speech emotion recognition models have improved year by year through various efforts. Zou et al. (2022) proposed an emotion recognition model that uses three types of acoustic information as input: raw waveform data, Mel-Frequency Cepstrum Coefficient (MFCC), and spectrogram. This model extracts features from the three acoustic data types and fuses the extracted features with a coattention mechanism. Kim et al. (2022) developed a model that combines the focus attention mechanism and the calibration attention mechanism. This proposed attention mechanism allows us to focus more on the important regions in the feature space of speech data. Pan et al. (2024) proposed a model that uses contrastive learning and gender information. Using information other than speech for prediction, such as gender information, is important for improving accuracy. Hence, many proposed models use text and video in addition to speech.

Although several multimodal models have been proposed that use video, text, and speech as input (Sun et al., 2023; Wu et al., 2023; Yang et al., 2023), their use is currently limited to video analysis and other applications that do not consider real-time performance because of the high computational complexity of handling video. In an SDS, it is difficult to use all the user's video images during speech because the speed of emotion recognition is important.

Furthermore, a multimodal dataset is more expensive to create than a unimodal dataset. Therefore, multimodal emotion recognition datasets are not available, but many cases have dealt with emotion recognition datasets with facial expressions and speech. Against this background, I aim to construct a multimodal emotion recognition model with facial expressions and speech using emotion recognition datasets for each modality. I construct a multimodal emotion recognition dataset by pairing similar labels from each dataset and trained a multimodal model. I compare the difference in performance between the multimodal model trained on the constructed dataset and the model trained on the unimodal dataset to confirm the effectiveness of the method.

1.2 Personalization

Typical speech emotion recognition tasks aim to predict emotion labels such as happiness, sadness, anger, and neutral. However, these labels often fall short of capturing the complexity of human emotions. An alternative approach is to use emotion attributes as suggested by core affect theory (Russell, 2003). Emotion attributes are represented as continuous scores in dimensions such as arousal (calm versus active), valence (unpleasant versus pleasant), and dominance (weak versus strong) for more nuanced expressions of emotions.

The prediction of valence is known to heavily depend on the speech characteristics of individual speakers, making it more challenging than predicting the other two attributes (Sridhar et al., 2018). However, adopting personalization techniques can address the variability in expression among different speakers. This approach allows for accurate predictions by considering each speaker's unique features.

Sridhar and Busso (2022); Tran et al. (2023) showed that prediction accuracy can be improved by embedding

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 5–7
September 16–17, 2024. ©2024 Association for Computational Linguistics

speaker characteristics from training data and identifying speakers with similar characteristics in test data. This method necessitates that the training data include a diverse array of speakers. If the training data does not contain a sufficient number of speakers, adding new trainable speaker data may be necessary. In such cases, existing methods require retraining not only the speaker embedding module but also the speech encoder weights after each data addition.

To address this, I introduce the concept of continuous prompt tuning, in which speaker prompts are added to the inputs of each speech encoder layer. In this approach, the weights of the speech encoder are frozen, and only the speaker prompts are updated to learn the speaker's characteristics. This allows for the addition of new speaker data without retraining the weights of the speech encoder.

2 Spoken dialogue system (SDS) research

Recently, the development of large language models has made it possible to perform tasks that had been constrained by technical limitations and costs and has allowed us to achieve high performance in demanding tasks. For example, in natural language processing, tasks such as document summarization, translation, and question–answering systems, which were considered challenging, can now be executed with high accuracy. This advancement has allowed for various practical applications such as information retrieval and customer support.

However, many unresolved challenges remain. One significant difficulty is the integration and consistent processing of multiple modality information (e.g., text, images, audio). Effective methods to model the interactions between these modalities are not yet fully established.

Additionally, there is a demand for the rapid processing of such diverse information. Current models consume a vast computational resources, making real-time response challenging. In speech dialogue systems, understanding the user's intent quickly and accurately is paramount; any latency can degrade the user experience. Hence, improving computational efficiency remains a critical research area.

Moreover, current models have limitations in accurately understanding human intent and emotions. While many language models are trained on large datasets and are adept at understanding general patterns and contexts, they continue to struggle with grasping subtle nuances and emotional changes in specific situations. For instance, they may misinterpret sarcasm or the use of polysemous words.

Future research will resolve these issues and further advance SDSs. If technology can be established to manage multiple modalities of information in an integrated manner and process it at high speed, a system that can understand the user's intentions more accurately will be realized. In addition, if emotions and intentions can be accurately captured, more natural and humanlike dialogue will be possible. As a result, SDSs are expected to find applications in a wide range of fields, including medicine, education, and entertainment.

3 Suggested topics for discussion

I suggest discussing the following:

- How can we improve the accuracy of emotion recognition in SDSs?
- What are the challenges in multimodal emotion recognition?
- Can personalization be applied to tasks other than speech emotion recognition tasks?
- How can reinforcement learning be used in dialogue systems?

References

Junghun Kim, Yoojin An, and Jihie Kim. 2022. Improving Speech Emotion Recognition Through Focus and Calibration Attention Mechanisms. In *Proc. Interspeech 2022*. pages 136–140. https://doi.org/10.21437/Interspeech.2022-299.

Yu Pan, Yanni Hu, Yuguang Yang, Wen Fei, Jixun Yao, Heng Lu, Lei Ma, and Jianjun Zhao. 2024. Gemo-clap: Gender-attribute-enhanced contrastive language-audio pretraining for accurate speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 10021–10025.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110(1):145.

Kusha Sridhar and Carlos Busso. 2022. Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech. *IEEE Transactions on Affective Computing* 13(4):1959–1972.

Kusha Sridhar, Srinivas Parthasarathy, and Carlos Busso. 2018. Role of regularization in the prediction of valence from speech. *Interspeech 2018* .

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. 2023. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 658–670.

Minh Tran, Yufeng Yin, and Mohammad Soleymani. 2023. Personalized adaptation with pre-trained speech encoders for continuous emotion recognition. *arXiv preprint arXiv:2309.02418* .

Shaoxiang Wu, Damai Dai, Ziwei Qin, Tianyu Liu, Binghuai Lin, Yunbo Cao, and Zhifang Sui. 2023. Denoising bottleneck with mutual information maximization for video multimodal fusion. *arXiv preprint arXiv:2305.14652* .

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 7617–7630.

Heqing Zou, Yuke Si, Chen Chen, Deepu Rajan, and Eng Siong Chng. 2022. Speech emotion recognition with co-attention based multi-level acoustic information. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 7367–7371.

Biographical sketch

Takumasa Kaneko is a PhD student in the Department of Informatics at the University of Electro- Communications. His master's research focused on developing a speech dialogue system that considers emotions.