Breaking the Silence: Whisper-Driven Emotion Recognition in AI Mental Support Models

Xinghua Qu
Tianqiao and Chrissy Chen Institute
Singapore
quxinghua17@gmail.com

Zhu Sun, Shanshan Feng, Caishun Chen
*CFAR, IHPC, A*STAR*
Singapore
{feng_shanshan,chen_caishun}@cfar.a-star.edu.sg

Tian Tian
University of California, Davis, CA
United States
liltian@ucdavis.edu

***Abstract*—Most emotional support conversations (ESCs) currently rely on text-based interfaces, which may not be user-friendly, especially for individuals with visual impairments or those who struggle with reading and writing. Thus, we present a personalized voice-based ESC system powered by large language models (LLMs). It can analyze emotional status from vocal user inputs, which provides deep insights that text-based methods cannot, enabling the LLM-driven chatbot to offer more tailored and effective emotional support to its users. Our code is available at https://github.com/xinghua-qu/speech_emotion_recognition**

***Index Terms*—Mental Healthcare, Voice-based Emotional Support Conversation, Large Language Models, AI for Social Good**

I. Introduction

Mental health concerns have intensified due to increased rates of mental illnesses. It was revealed that the prevalence is 20% for depression, 35% for anxiety and 53% for stress in a combined study population of 113,285 individuals from Dec 2019 to Jun 2020 [1]. These factors affect everyday life quality, underscoring the need for efficient and reachable mental health support (MHS) systems, e.g., emotional support conversations (ESCs), to mitigate the risk of mental health issues.

Recently, the breakthrough made by large language models (LLMs) has spurred their extensive integration across diverse domains [2], [3], without mental healthcare (e.g., ESCs) being an exception [4]. Despite the success, most ESCs use text-based interfaces, which are less user-friendly and present challenges for users with visual impairments or limited reading and writing abilities. Contrarily, we posit that the emotional status derived from user voice input can markedly enhance the efficacy of response generation in LLMs. Although prior studies [5] disclose that fusing emotional stimuli into user prompts can augment LLM performance, they primarily focused on emotional stimuli embedded within textual information. We argue that a more direct and potent avenue for capturing human emotional context lies in voice-based inputs, which can offer a richer and more nuanced source of emotional data for LLMs.

To validate this, we design a Speech Emotion Recognition (SER) model, seamlessly amalgamated with a frozen Whisper encoder, for the streamlined extraction of vocal characteristics. It synergistically converges with specialized downstream transformer layers, meticulously fine-tuned for classification tasks. Our model exhibits an exceptional accuracy rate (95%), a substantial enhancement over SOTAs' accuracy (85%) as documented on Hugging Face, showcasing the efficacy of vocal data integration into LLMs for more effective ESCs.

II. Related Works

Early methods for MHS have applied the psychological theory of empathy, e.g., affective empathy [4]. Later, some works attempt to provide emotional support in interactions [6] to improve users' mental state. To further free users from emotional distress, positive emotion elicitation has been studied [7]. Recent studies have adopted LLMs in the medical domain [8] thanks to their prowess in a spectrum of healthcare tasks [9], [10]. With the prevalence of stress-related concerns, MHS, particularly focusing on emotional well-being, has garnered significant attention within the research community. For instance, Liu et al. [11] design task-adaptive tokenization to enhance long-form text generation efficacy in MHS. Zheng et al. [12] use LLMs for dialogue augmentation in the task of ESCs. Despite the success of traditional methods and LLMs for MHS, most ESCs primarily use text-based interfaces, which are less user-friendly and might present challenges for users with visual impairments or limited reading and writing abilities. In contrast, we introduce a personalized voice-based ESC, which extracts nuanced contextual cues, e.g., sentiment, age, and gender, from voice inputs, capturing insights inaccessible through text-based input.

III. The Proposed Framework

Our SER system employs a pretraining and fine-tuning paradigm, using a Whisper encoder as the pre-trained base. During fine-tuning, it is augmented with two additional Transformer layers, followed by a classification projection layer to enhance its ability to discern emotional nuances in speech.

The OpenAI Whisper model [13] is engineered to transcribe speech across multiple languages, employing a novel approach termed 'large-scale weak supervision'. This paradigm diverges from the conventional supervised learning framework, which predominantly depends on extensively annotated datasets, by utilizing an expansive corpus of data accompanied by 'weak' or imprecise labels. In our research, we exclusively harness the Whisper encoder for deriving audio representations. These representations are then employed for classification tasks. A pivotal aspect of the Whisper model is its resilience to a spectrum of speech attributes, encompassing diverse accents,

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00063

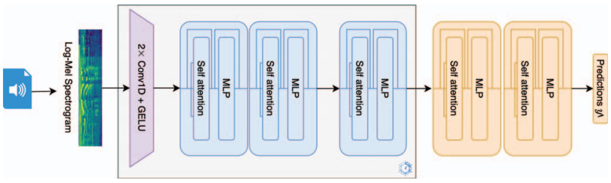

Fig. 1. The network structure of SER-Whisper model.

ambient noise interference, and varying audio fidelity. We posit that these robust representations will substantially augment the downstream emotion classification tasks.

The training architecture is depicted in Figure 1. It incorporates two convolutional 1D (Conv1D) layers, succeeded by transformer layers (highlighted in blue), adapted from the OpenAI Whisper model and maintained in a frozen state. Our fine-tuning process concentrates on the subsequent two transformer layers and the projection layer, which are optimized to predict the target variable y'. The model receives speech input x, with the aim of identifying the corresponding emotional state, denoted as y. The training employs a cross-entropy loss function. Optimization is achieved through the Adam optimizer, configured with a learning rate of 0.000025 and a decay rate of $1 \cdot 10^{-7}$. To cater to diverse computational and performance needs, we offer the model in five different scales: tiny, base, small, medium, large, large-v2 and large-v3.

IV. Experiments and Analysis

We adopt the dataset of Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [14] for evluation. RAVDESS comprises 7,356 audio files, aggregating to a total size of 24.8 GB. It features (a) 24 professional actors (12 female and 12 male), who perform two lexically-matched statements in a neutral North American accent; and (b) a range of expressed emotions, including calm, happiness, sadness, anger, fear, surprise, and disgust, each articulated in two levels of emotional intensity: normal and strong, along with a neutral expression. We exclusively employ the audio component of RAVDESS, re-sampling it to 16,000 Hz to maintain consistency with the Whisper model's settings.

The experimental results are shown in Table I. In general, our SER-Whisper models outperform all baselines significantly. In particular, across all variants from 'tiny' to 'large-v3', they consistently achieve promising accuracy, with SER-Whisper-medium and its subsequent versions attaining an impressive 94.44%. This is a notable improvement over the runner-up model, XLSR-Wav2Vec2 with an accuracy of 86.7%. The key to this performance disparity is the use of pretraining, which all SER-Whisper models employ, in contrast to some of the lower-performing methods like IAF and ERANN. This underscores the efficacy of pretraining in enhancing model accuracy. The evolution of SER-Whisper models, from 'tiny' to 'large-v3', further demonstrates a significant advancement in model architecture and optimization. These findings not only establish the dominance of SER-Whisper models in SER but also open up new avenues for future research in AI, particularly in refining model architectures and exploring their practical applications in real-world scenarios.

TABLE I
The results comparison on speech emotion recognition.

Method	Accuracy	Is Pretrained
XLSR-Wav2Vec2 [15]	86.7%	Yes
IAF [16]	81.58%	No
CNN+biLSTM [17]	80.08%	Yes
ERANN [18]	74.8%	No
SER-Whisper-tiny	83.83%	Yes
SER-Whisper-base	88.89%	Yes
SER-Whisper-small	88.89%	Yes
SER-Whisper-medium	94.44%	Yes
SER-Whisper-large	94.44%	Yes
SER-Whisper-large-v2	94.44%	Yes
SER-Whisper-large-v3	94.44%	Yes

V. Conclusion

We present a novel SER model, termed SER-Whisper, which is predicated on the Whisper architecture and is designed to augment LLMs in facilitating ESC. Our empirical results demonstrate that SER-Whisper achieves a significant performance enhancement over SOTA models in the domain of SER. In the future, we plan to extend SER-Whisper to distill more contextual cues, e.g., age and gender, from the voice inputs for more personalized ESC.

References

[1] R. Lakhan *et al.*, "Prevalence of depression, anxiety, and stress during covid-19 pandemic," *Journal of Neurosciences in Rural Practice*, 2020.

[2] Z. Sun *et al.*, "Dynamic in-context learning from nearest neighbors for bundle generation," in *SIGIR*, 2024.

[3] ——, "Large language models for intent-driven session recommendations," in *SIGIR*, 2024.

[4] Q. Li *et al.*, "Knowledge bridging for empathetic dialogue generation," in *AAAI*, 2022.

[5] C. Li *et al.*, "Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus," *arXiv e-prints*, 2023.

[6] Q. Tu *et al.*, "Misc: a mixed strategy-aware model integrating comet for emotional support conversation," in *ACL*, 2022.

[7] J. Zhou *et al.*, "Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach," *arXiv preprint arXiv:2307.07994*, 2023.

[8] M. Moor *et al.*, "Foundation models for generalist medical artificial intelligence," *Nature*, 2023.

[9] C. E. Haupt and M. Marks, "Ai-generated medical advice—gpt and beyond," *Jama*, 2023.

[10] K. Yang *et al.*, "Towards interpretable mental health analysis with large language models," in *EMNLP*, 2023.

[11] S. Liu *et al.*, "Task-adaptive tokenization: Enhancing long-form text generation efficacy in mental health and beyond," in *EMNLP*, 2023.

[12] C. Zheng *et al.*, "Augesc: Dialogue augmentation with large language models for emotional support conversation," in *ACL*, 2023.

[13] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[14] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, 2018.

[15] C. Luna-Jiménez and other, "A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset," *Applied Sciences*, 2021.

[16] K. Chumachenko *et al.*, "Self-attention fusion for audiovisual emotion recognition with incomplete data," in *ICPR*, 2022.

[17] C. Luna-Jiménez and other, "Multimodal emotion recognition on ravdess dataset using transfer learning," *Sensors*, 2021.

[18] S. Verbitskiy *et al.*, "Eranns: Efficient residual audio neural networks for audio pattern recognition," *Pattern Recognition Letters*, 2022.