INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Promoting Mental Self-Disclosure in a Spoken Dialogue System

M. Rohmatillah[1], B. Aditya[1], L.-J. Yang[2], B. G. Ngo[1], W. Sulaiman[2], J.-T. Chien[2]

[1]EECS International Graduate Program [2]Institute of Electrical and Computer Engineering
National Yang Ming Chiao Tung University, Taiwan

{mahdin.ee08, bobbiaditya.ee10, jtchien}@nycu.edu.tw

Abstract

This paper proposes a mental health spoken dialogue to relax mental illness for university students by acting as an active listener to promote self-disclosure. The proposed system is designed for Mandarin with the specific accent and lexicon in Taiwan which is known as one of the underrepresented spoken languages. To achieve the objective, this work considers three key factors which are high quality speech components including automatic speech recognition and text-to-speech models, and the personalized responses while keeping the trustworthiness and seamless integration among dialogue system components.
Index Terms: mental health dialogue system, automatic speech recognition, text-to-speech, active listener

1. Introduction

The concern of mental health related problems in the student campus life has been crucial for many years. This issue is basically affected by several factors, like academic pressure, job crisis, social disharmony and even sleeping problems. Recent surveys have shown that the number of students taking psychotropic medication is increasing considerably, from only 9% in 1994 to around 24% in recent years [1]. Furthermore, the unpredictable situations, for example pandemic due to Covid-19 also cause a huge impact to the student mental health stability.

While most of the universities and schools offer free mental counseling sessions to the students, such a counseling service is likely insufficient due to the imbalance between the number of counselors and students which obstructs immediate treatments. As a consequence, an effective and reliable treatment should be designed to promptly alleviate the student mental discomforts.

Due to the advancement of deep learning methods, a lot of alternative solutions have been offered to alleviate the mental health problems, for example Woebot [2] and Wysa [3] which were developed to reduce the mental discomforts by a mean of promoting the user self-disclosure. Unfortunately, most of the existing methods were designed in the text-based format and only allowed the users to select a response from the provided options. Furthermore, most of them generated the responses based on a specific designed rule without considering the personalized and diverse responses which might restrict some potential benefit of deep learning solution to cure mental illness.

In order to exploit the speech technologies for mental health, this paper proposes a spoken dialogue system that acts as an active listener to promote self-disclosure for students so as to relax their mental illness. The role of active listener is set in the dialogue manager (DM) component [4] of a dialogue system according to some pre-defined rules following the sessions of journaling [5] and small talk. To enrich user experiences, both speech components and text components are optimized carefully along with the appropriate system integration to allow seamless communication among various dialogue components.

2. System Architecture

Figure 1 illustrates NYCUKA as a mental health spoken dialogue system for university students. FLASK 2.2.2 is utilized as a backend to allow seamless integration of system interface with various text and speech components as described below.

2.1. Text Components

There are three main modules or components that handle text data in the proposed dialogue system. Those are natural language understanding (NLU), DM and natural language generation (NLG). All of the text components are designed to make the system able to generate the diverse responses while maintaining reliability. Diverse responses are essential to attract user interests to interact with the dialogue system, while trustworthiness is an important factor that must be considered to avoid the unwanted negative effects in the interaction between humans and the dialogue system. NLU aims to classify the text input from the user into a specific intent or semantic slot. For example, if a user answers "我很好" (I am fine) given a question "你今天過得很好嗎？" ("Are you having a good day?"), then the system will bring the conversations into the conversation flow which corresponds to "positive" intent. Due to the data scarcity to train traditional Chinese NLU, then GPT-3 model is utilized and converted to be an NLU by using the pre-defined prompts. DM is the brain of the system that will determine the system response in every conversation turn. The proposed DM contains several rules that govern the conversation flow with the users. The rule is developed by following the active listener theory to avoid any trustworthy issues. In order to provide diverse and personalized responses, NLG based on the GPT-3 model is implemented. GPT-3 model will receive the specific prompts that ask NLG to paraphrase output from the rule defined in DM conditioned on the dialogue context which contains the problems faced by the users. However, it is not required that all DM outputs should be passed to the NLG. Some of the rules are employed to maintain the conversation flow to be controlled within the topic. This task is performed by a controllability function.

2.2. Speech Components

Speech components play an important role to establish an active listener for dialogue system. The automatic speech recognition (ASR) must transcribe correctly the user long utterances into text sequences. Meanwhile, text-to-speech (TTS) must provide high quality speech signals to convey the message clearly. However, it is very challenging to build ASR and TTS that works for

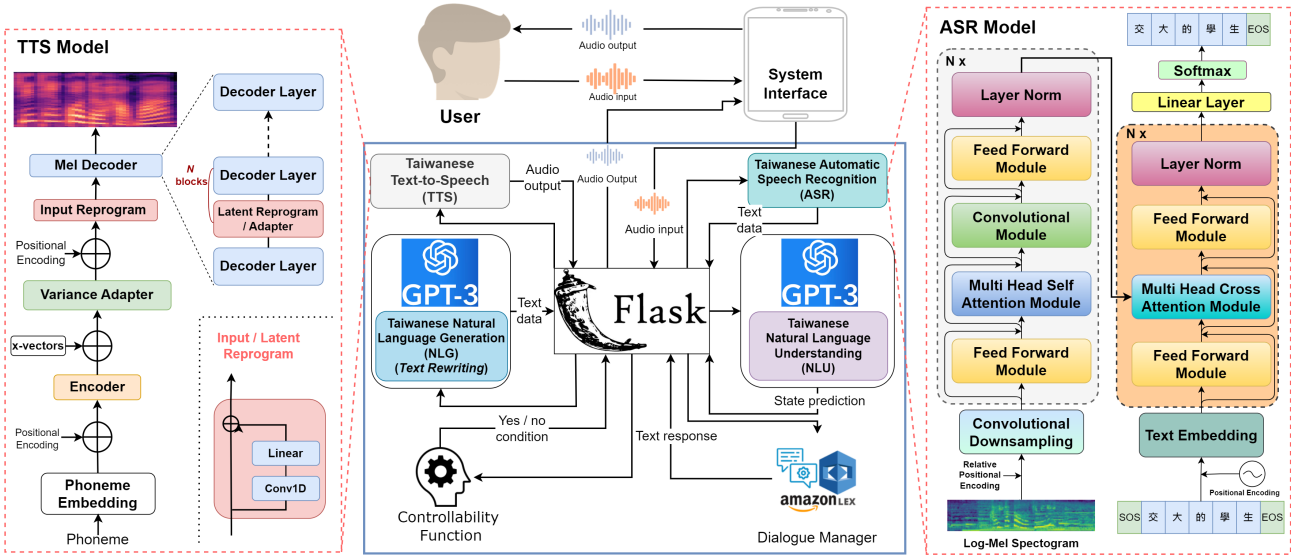

Figure 1: *Model architectures for TTS (Left) and ASR (Right). (Middle) System integration and data flow of NYCUKA dialogue system.*

Mandarin in Taiwan (zh-TW) due to the limited public dataset to train the models. Therefore, data collection and parameter efficient learning should be considered carefully. For the ASR learning, data crawling from YouTube videos of news channels by taking only the audios and the corresponding captions is carried out. This process collects 208 hours of speech data from various speakers in different backgrounds to ensure that the model is robust to a variety of accents and speaking styles in Taiwan. Next, the pre-processing techniques are applied to clean and downsample the voice signals. Data cleaning involves removing the empty signal and too long speech to ensure consistency. Signal downsampling improves the memory efficiency during data processing. The process data were then used to train ASR model consisting of a Conformer as encoder and a transformer as decoder. The adaptation strategy from the Mandarin in China (zh-CN) model to zh-TW model is carried out by initializing the ASR model parameters from the pre-trained weights based on AISHELL 2 [6] which contains 1000 hours of speech data. The model is then fine-tuned with the collected data under the shared tokenization as shown in Figure 1 (right).

TTS model is also trained by using the adaptation strategy from zh-CN to zh-TW. Different from the ASR training that fully fine-tunes the pre-trained model [7], this work develops the parameter efficient learning and model regularization for low-resource accent adaptation in TTS model. The Conformer-fastspeech2 [8] model from ESPNet [9] is utilized as the backbone of TTS model. The speech spectrogram from phoneme input is synthesized through a stack of components consisting of phoneme embedding, encoder, variance adapter and Mel decoder where the additional position embeddings and x-vectors [10] are added during the stack processing as depicted in Figure 1 (left). The parameter efficient learning is performed by only reshaping the architecture of Mel decoder in the frozen Conformer-fastspeech2 model through input reprogramming layer and latent reprogramming layer. The input reprogramming layer is introduced in conjunction with the Mel decoder of a frozen TTS model which aims to address the challenge of re-deploying accent voices under a fixed Mel decoder. Meanwhile, the latent reprogramming layer is employed to reprogram for feature adaptation within Mel decoder.

3. Conclusions

This paper has presented a mental health spoken dialogue system which acts as an active listener to cure mental discomfort for students by mean of promoting the self-disclosure. The speech-driven dialogue system was designed to work for Mandarin with specific accent and lexicon in Taiwan which is seen as one of the low-resource spoken languages. The solutions to learning machine and system design were proposed to cope with the issue of data scarcity for seamless system integration.

4. References

[1] D. Iarovici, *Mental Health Issues and the University Student*. JHU Press, 2014.

[2] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial," *JMIR Mental Health*, vol. 4, no. 2, pp. 1–11, 2017.

[3] B. Inkster, S. Sarda, V. Subramanian *et al.*, "An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study," *JMIR Mhealth Uhealth*, vol. 6, no. 11, pp. 1–14, 2018.

[4] M. Rohmatillah and J.-T. Chien, "Causal confusion reduction for robust multi-domain dialogue policy," in *Proc. of INTERSPEECH*, 2021, pp. 3221–3225.

[5] M. Kawasaki, N. Yamashita, Y.-C. Lee, and K. Nohara, "Assessing users' mental status from their journaling behavior through chatbots," in *Proc. of ACM IVA*, 2020, pp. 1–8.

[6] C.-H. Leong, Y.-H. Huang, and J.-T. Chien, "Online compressive transformer for end-to-end speech recognition," in *Proc. of INTERSPEECH*, 2021, pp. 2082–2086.

[7] L.-J. Yang, I.-P. Yeh, and J.-T. Chien, "Low-resource speech synthesis with speaker-aware embedding," in *Proc. of ISCSLP*, 2022, pp. 235–239.

[8] Y. Ren, C. Hu, X. Tan, T. Qin *et al.*, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. of ICLR*, 2018.

[9] S. Watanabe, T. Hori *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. of INTERSPEECH*, 2018, pp. 2207–2211.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. of ICASSP*, 2018, pp. 5329–5333.