TOWARDS JOINT MODELING OF DIALOGUE RESPONSE AND SPEECH SYNTHESIS BASED ON LARGE LANGUAGE MODEL

Xinyu Zhou (Communication University of China), Delong Chen (Hong Kong University of Science and Technology)
xinyuzhou@cuc.edu.cn

This paper explores the potential of constructing an AI spoken dialogue system that *"thinks how to respond"* and *"thinks how to speak"* simultaneously, which more closely aligns with the human speech production process compared to the current cascade pipeline of independent chatbot and Text-to-Speech (TTS) modules. We hypothesize that Large Language Models (LLMs) with billions of parameters possess significant speech understanding capabilities and can jointly model dialogue responses and linguistic features.

We conduct two sets of experiments. Firstly, we showcase the speech understanding ability of LLMs by performing prosodic structure prediction, which is a typical task within the TTS text analysis front-end. Results show that both prompting-based ChatGPT and fine-tuning based ChatGLM model achieve competitive performance against traditional methods. We also show that LLM can utilize linguistic knowledge to improve prediction accuracy.

Secondly, we aim to further integrate a wide array of linguistic features into the model, and maintain LLM's dialogue capability at the same time. To address the lack of a parallel dataset of dialogue response and linguistic annotations, we employ an automated dialogue context generation approach inspired by LongForm, then train an LLM to produce both dialogue response speech features at the same time. Our results indicate that the LLM-based approach is a promising direction for building unified spoken dialogue systems.