Generative Speech Models for Human–AI Interaction: Advancing Personalization, Cross-Lingual Speech, and Multimodal Generation

Name: Tran Quang Chung

Laboratory' s name: Human-AI Interaction Laboratory

Supervisor' s name: Sakriani Sakti

Abstract (should be within 1st page)

Seamless Human–AI Interaction (HAI) depends critically on an AI agent to communicate as naturally, adaptively, and intuitively as a human partner. In HAI context, speech generation plays a crucial role in communication and has seen significant advancements, especially in single-language systems developed for high-resource majority languages. However, current systems still lack the flexibility necessary for truly immersive interactions, especially in terms of voice personalization, support for low-resource languages, cross-lingual accessibility, and multimodal context awareness. This thesis aims to bridge these gaps by developing novel solutions that enhance the adaptability and inclusivity of AI agents.

First, this thesis introduces an adaptation text-to-speech (TTS) model that uses a style-enhanced diffusion technique to clone voices just a few seconds of reference audio. This capability enables AI agents to instantly generate speech in a specific user's voice, fostering intimacy, trust, and personalization in long-term interactions. Second, building upon this, the thesis presents novel training techniques to develop robust TTS models using limited paired data, thereby enabling broader language support. Third, this work tackles the challenge of cross-lingual translation by proposing the SAM-Translator, which uses **S**elf-P**A**ced Learning and a **M**ixture-of-Experts approach. SAM-Translator, which directly converts source text (e.g., Japanese) into target speech (e.g., English), achieves state-of-the-art results on standard datasets for cross-lingual translation. Fourth, to expand interaction, this thesis introduces the Image-to-Speech model with dual guidance, which decodes visual inputs and generates descriptive speech. Finally, the thesis presents OmniVIVO, a unified model for generating high-fidelity **VI**sual and **VO**ice modalities together. This development paves the way for next-generation unified AI models that can perceive inputs and generate coherent responses across multiple modalities.