SEMINAR ANNOUNCEMENT

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
Faculty of Engineering
Website: https://www.eng.nus.edu.sg/ece/

Area: Signal Analysis & Machine Intelligence

Host: Prof Li Haizhou

Co-organized by

Chinese and Oriental Languages Information Processing Society, Singapore
IEEE Singapore Systems, Man and Cybernetics Chapter

TOPIC	**:**	**Waveform loss-based acoustic modeling for text-to-speech synthesis and speech-to-musical sound transferring**
SPEAKER		**Dr. Yi Zhao,** **Postdoctoral Researcher,** **Yamagishi Lab, National Institute of Informatics,** **Japan**
DATE	**:**	**6 November 2019, Wednesday**
TIME	**:**	**2pm to 3pm**
VENUE	**:**	**E3-06-01, Engineering Block E3, Faculty of Engineering, NUS**

ABSTRACT

Recent neural waveform synthesizers such as WaveNet, WaveGlow, and the neural-source-filter (NSF) model that learn directly from speech waveform samples have achieved very high-quality synthetic speech. Such neural networks are being used as an alternative to vocoders and hence they are often called neural vocoders. The neural vocoder uses acoustic features as local condition parameters, and these parameters need to be accurately predicted by another acoustic model. However, it is not yet clear how to train this acoustic model, which is problematic because the final quality of synthetic speech is significantly affected by the performance of the acoustic model. Significant degradation happens, especially when predicted acoustic features have mismatched characteristics compared to natural ones. Also, the similarity between speech and music audio synthesis techniques suggest interesting avenues to explore in terms of the best way to apply speech synthesizers in the music domain.

I will firstly introduce a study on multi-speaker Text-to-Speech Synthesis (TTS) Systems that involves generative adversarial network (GAN) as well as waveform loss. The idea is to use the loss of a well-trained WaveNet in addition to mean squared error and adversarial losses as parts of objective functions of the acoustic model, with the aim of reducing the mismatched characteristics between natural and generated acoustic feature. I will also show a work on comparing three neural synthesizers used for musical instrument sounds generation under three scenarios: training from scratch on music data, zero-shot learning from the speech domain, and fine-tuning-based adaptation from the speech to the music domain. The results of a large-scale perceptual test demonstrated that the performance of three synthesizers improved when they were pre-trained on speech data and fine-tuned on music data, which indicates the usefulness of knowledge from speech data for music audio generation.

BIOGRAPHY

Dr. Yi Zhao is currently a post-doctoral researcher in National Institute of Informatics, Japan. She received the Ph.D. degree from University of Tokyo in 2018. Before that, she received the M.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2014, and B.E. degree in communication engineering from Beijing University of Posts and Telecommunications, China, in 2011. Her current research interests include speech synthesis, audio signal processing, and statistical machine learning.