Self-adaptive and Incremental Machine Speech Chain

Name : Sashi Novitasari
Laboratory : Augmented Human Communication
Supervisor : Satoshi Nakamura

Abstract

In human spoken communication, speech production and perception are inseparable. It is reflected in the human speech chain mechanism, showing that humans speak while listening. This mechanism allows them to monitor and improve their speech performance in various situations. It is also important for language acquisition.

Inspired by the human speech chain mechanism, a machine speech chain framework based on deep learning was recently proposed for a semi-supervised development of TTS and ASR. However, the basic framework was aimed only for non-incremental TTS and ASR training, in which the systems require a long delay when encountering a long input sequence. Moreover, the TTS and ASR still perform separately during inference. They could not do self-adaptation or change the speech by considering environmental situations. By contrast, humans can listen to what they speak in real-time and enhance the intelligibility of their speech, which is called the Lombard effect. If there is a delay in the hearing, they won't be able to continue speaking and adapting to the environment appropriately.

In this thesis, we propose self-adaptive and incremental machine speech chain frameworks for training and inference by mimicking the human speech chain closely. To achieve this, first, we reduce the delay in the basic machine speech chain by replacing the components with an incremental TTS (ITTS) and an incremental ASR (ISR). During speech chain training, we let these systems improve together through a short-term loop. Second, we design a self-adaptation framework focusing on speech synthesis in noisy environments through a speech chain mechanism. It synthesizes the speech not only by taking text input but also the auditory feedback representing the current system performance and the environmental situation. This mechanism allows the TTS to speak in a Lombard effect automatically according to real situations. Finally, we perform experiments of self-adaptive incremental speech synthesis with a low adaptation delay in noisy environments. A low delay is critical in the adaptation process, so the system can perform optimally in dynamic situations. All this contribution shows that the feedback mechanism is not only essential for the human speech chain but also for machines to dynamically adapt and improve themselves in various situations.