Japanese incremental text-to-speech synthesis based on accent phrase unit

Name: Tomoya Yanagita
Laboratory's name: Augmented Human Communication Laboratory
Supervisor's name: Satoshi Nakamura

Abstract (should be within 1st page)

Speech-to-speech translation (S2ST) is an innovative technology that translates speech signals in a source language to another language. The systems consist of three components: automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS). In conventional systems, the process is done sentence by sentence. The conventional S2ST will produces translated speech with significant delays for longer speech in lectures and meetings.

On the other hand, simultaneous speech translation systems consist of incremental components: incremental ASR, incremental MT (iMT), and incremental TTS (iTTS). The process of incremental systems is done chunk by chunk. Each chunk is a smaller unit than a sentence. Existing iTTS technologies based on hidden Markov model (HMM) and neural end-to-end frameworks generate a speech from a word unit as a synthesis chunk. An English end-to-end iTTS which allows waiting look-ahead a word or two produced good speech quality. Moreover, the look-ahead approach needs to wait for a look-ahead length. However, Japanese TTS needs an accent phrase that is a longer unit than a word to improve speech quality. To construct simultaneous speech interpretation for the Japanese language, we proposed a Japanese incremental text-to-speech synthesis based on an accent phrase unit.

In this thesis, we first proposed Japanese iTTS based on HMM and showed that an accent phrase unit longer than a word unit improved speech quality. Second, we proposed an end-to-end iTTS which used an accent phrase as a unit without any look-ahead. The experiment results showed that the proposed end-to-end iTTS can synthesize speech in various input lengths without look-ahead length.