論文内容要旨

A Study on Separation and Integration of Speech Information for Speech Synthesis

(音声合成のための音声情報の分離・統合に関する研究)

氏名: 齋藤 大輔

For output of information, speaking is one of the most important and fundamental abilities of human beings. On the other hand, synthesizing of voices is a sound modality for computers. That is to say, speech synthesis can be and should be considered as an artificial process of speech generation of human beings. Nowadays, most of the speech synthesizers are text-to-speech (TTS) converters, which take a phoneme sequence as input and generate speech sounds corresponding to the sequence. To build a synthesizer, symbol-to-sound mapping is learned from a speech corpus. If a speech corpus of speaker A is used, the synthesizer learns A's voices and can read text out for him/her. A very good synthesizer may be able to deceive speaker verification systems. With the preparation of large amounts of speech corpora and the development of statistical learning theories and approaches, the quality of synthesized speech samples produced by these speech synthesizers is improving.

Let us think about the case of humans' learning of speech. Developmental psychology tells that infants acquire spoken language through imitating the utterances from their parents, called vocal imitation. However, they never imitate the voices of their parents. It is impossible for infants to create their parents' voices due to a difference in the shape of vocal tubes. To enable the vocal imitation in this situation, some abstract representation of utterances should exist between infants and their parents. One may claim that they communicate orally via phonemic representation as similar to the framework of TTS, but researchers of infant study deny this claim. This is because their phonemic awareness is very immature and it is difficult for them to decompose an utterance into sequence of phonemes. Hence phonemic representation is not necessarily required to abstract representation of utterances. What makes the vocal imitation possible?

For media information processing, implementation of media processing of human beings, which is very robust and flexible is one of the most essential problems. From this viewpoint, media information processing should deal with vocal imitation of infants correctly. For this

purpose, some abstract representation of utterances which infants imitate in the case of vocal imitation should be defined physically and acoustically. Researchers answer that infants extract the holistic sound pattern from word utterances, called word Gestalt and they reproduce it with their short vocal tubes. Here, we can say that the Gestalt has to be speaker-invariant because, whoever speaks a specific word to infants using different voices, it seems that infants always extract the same Gestalt.

Recently, a candidate answer for the word Gestalt is showed mathematically and the validity of the answer was verified experimentally. The proposed method of extracting the Gestalt from an input utterance was used successfully for Automatic Speech Recognition (ASR), and Computer Aided Language Learning (CALL). This study is trying to implement speech synthesis framework based on this implementation of the word Gestalt.

The objective of this study is to realize the framework of speech information processing, which is robust and flexible as human speech processing. For this purpose, this study is trying to implement the similar process to the vocal imitation of infants based on the implementation of the word Gestalt mentioned above. From the viewpoint of the treatment of information in speech, this process includes two operations; to separate speech information into linguistic and speaker information, and to integrate them into produced speech again. In order to realize the former operation, we focus on representation for the difference of speakers and that for speaker-invariant linguistic information. In this study, the difference of speakers is represented as geometrical properties of a feature space. Speaker-invariant linguistic information is represented as acoustic word Gestalt mentioned above. For more flexible operation, a smaller acoustic unit based on word Gestalt is proposed. For the latter operation, we propose two frameworks of conversion to speech; structure to speech conversion and model-integration-based voice conversion. Structure to speech conversion can be regarded as an implementation of the vocal imitation of infants inspired by the viewpoint of developmental psychology. Model-integration-based voice conversion is a preliminary approach to integrate linguistic and speaker information by a probabilistic manner.

Compared with our proposed framework, modeling of speech by the conventional speech synthesizers can be regarded as "simultaneous" modeling of linguistic and speaker information. On the other hand, recognition processes in speech applications model focused and unfocused information "separately." Then synthesis and recognition processes of computers do not share the models of each other perfectly. Then synthesis and recognition processes of computers do not share the models of each other perfectly. However, speech processing of human is called a "speech chain", which integrates both the perception and production processes flexibly. From this viewpoint, this study can be regarded as an optimization of the whole process of speech processing, i.e. "speech chain."