論文の内容の要旨

Gesture Design for a Real-time Gesture-to-Speech Conversion System
Based on Space Mapping Between a Gesture Space and an Acoustic Space
(音響空間からジェスチャ空間への写像に基づく
リアルタイム音声生成系におけるジェスチャ設計)

37-097093 國越 晶

These days, most speech synthesizers such as TTS (Text to Speech) converters require symbol inputs. The quality of synthesized speech sample produced by the speech synthesizers is improving. However, this approach still has some drawbacks, for example, in emotional speech synthesis or in expressive pitch control. On the other hand, synthesis methods which do not require symbol inputs, such as articulatory synthesis, are effective for continuous speech synthesis and pitch control based on dynamic body motion. Therefore these alternatives also attract research interest and several applications have been proposed.

A dysarthric engineer, Ken-ichiro Yabu, developed a unique speech generator that relied on a pen tablet. The F1-F2 plane is embedded in the tablet. The pen position controls the F1 and F2 of vowel sounds and the pen pressure controls their energy. Another example of speech generation from body motions is Glove Talk proposed by Sidney Fels. With two data gloves and some additional devices equipped to the user, body motions are transformed into parameters for a formant speech synthesizer. In this study, we consider the process of speech production as media conversion from body motions to sound motions.

Recently, GMM-based speaker conversion techniques have been intensively studied, where the voice spaces of two speakers are mapped to each other and the mapping function is estimated based on a GMM. This technique was directly and successfully applied to estimate a mapping function between a space of tongue gestures and other speech sounds. This result naturally makes us expect that a mapping function between hand gestures and speech can be estimated well. People usually use tongue gesture transitions to generate a speech stream. But previous works showed that tongue gestures, which are inherently mapped to speech sounds, are not always required to speak. What is needed is a voluntarily movable part of the body whose gestures can be technically mapped to speech sounds. However, Yabu and Fels use classical synthesizers, i.e. formant synthesizers. Partly inspired by the remarkable progress of voice conversion techniques and voice morphing techniques in this decade, we are developing a GMM-based Hand-to-Speech conversion system (H2S system). Unlike the current techniques, our new synthesis method does not limit

the input media. Therefore, our technique would be useful in assistive technology, in which devices are tuned for person to person, and in performative field, in which people pursue the human capability of expression.

In this study, we focus attention on the design of the system. As an initial trial, a mapping between hand gestures and Japanese vowel sounds is estimated so that topological features of the selected gestures in a feature space and those of the five Japanese vowels in a cepstrum space are equalized. Experiments show that the special glove can generate good Japanese vowel transitions with voluntary control of duration and articulation.

We also discuss how to extend this framework to consonants. The challenge here is to figure out appropriate gestures for consonant sounds when the gesture design for vowels is given. We reported that inappropriate gesture designs for consonants result in a lack of smoothness in transitional segments of synthesized speech. We have considered the reason to be: (1) the positional relation between vowels and consonants in the gesture space and that in the speech space were not equivalent, (2) parallel data for transition parts from consonants to vowels did not correspond well. In order to solve those problems, we have developed a Speech-to-Hand conversion system (S2H system, the inverse system of H2S system) trained from parallel data for vowels only to infer the gestures corresponding to consonants. Listeners evaluated that an H2S system, which exploits gesture data for consonants derived from an S2H system, can generate more natural sounds than those trained with heuristic gesture design for consonants.

Natural speech generated by an H2S system trained exploiting data generated by S2H system are, however, obtained only when input gestures are the same as the one which generated by S2H system. S2H system sometimes outputs gestures whose dynamic range is too large or which is not smooth enough. In those cases, it is difficult for users to form those gestures in realistic time. In this thesis, we compensate those problems with two ways: (1) reduce the dynamic range by setting the optimal weight for the gesture model (2) smooth the gesture trajectories by considering delta features. Exploiting parallel data for consonants derived from a S2H system, we also implemented a real-time Hand-to-Speech conversion system and evaluated the effectiveness. Subjective user evaluations showed that almost a half of the phonemes, which are generated by our H2S system are perceived correctly and that this system is effective enough to generate emotional speech.