Abstract

This thesis deals with sound synthesis method for environmental sounds, which are any sounds that are not limited to speech or music. Environmental sounds such as “sound of wind” play an extremely important role in the production of contents such as movies and cartoon animations, as they explain the situation and symbolize the characters’ mental images. In the past, many studies have been conducted on method of synthesis and conversion for generating new sounds for speech and music. On the other hand, there have been very few studies on method of synthesis and conversion for environmental sounds. It is also unclear how the characteristics of diverse environmental sounds can be controlled. This thesis proposes methods of environmental sound synthesis based on a statistical method using various input informations in order to flexibly control the characteristics of diverse environmental sounds. The realization of an method of environmental sound synthesis that enables flexible control of diverse sounds is expected to be applied to various applications, such as the creation of background sounds and sound effects for movies, video content, games, etc., and data augmentation for environmental sound analysis.

The thesis constructs a dataset for environmental sound synthesis from onomatopoeic words corresponding to environmental sounds. It is very important to construct a dataset to be used for training of environmental sound synthesis model using statistical approach. In addition, while there are many datasets with sound event labels for environmental sounds, there are very few environmental sound datasets with other label information. Therefore, this thesis constructs a dataset in which environmental sounds are assigned onomatopoeic words, which is considered to be effective in expressing the temporal-change characteristics of environmental sounds. When assigning onomatopoeic words, this thesis collects the score of confidence level by the person who described the onomatopoeic words and the score of acceptance level for the onomatopoeic words from others. By analyzing the collected confidence and acceptance scores, this thesis shows that the collected onomatopoeic words is appropriately assigned to the environmental sound.

The thesis also organizes the evaluation methods for environmental sound syn-

thesis and discusses how the generated environmental sounds should be evaluated. This thesis proposes subjective evaluation methods for environmental sound synthesis. Through evaluation experiments, this thesis compares the subjective and objective evaluation methods used in environmental sound synthesis, and indicates how the synthesized environmental sounds should be evaluated.

Furthermore, this thesis discusses methods for environmental sound synthesis using various input informations. First, this thesis proposes a method of environmental sound synthesis from sound event labels such as sound of "rain" and "wind." Sound event labels can express the sound events of the environmental sound to be generated. Therefore, by using sound event labels as input information for environmental sound synthesis, it can be expected to control the sound event of the synthesized sounds. From the experimental evaluation of the quality of the synthesized sound shows that the use of sound event labels as input information is effective. This thesis also shows that nearly half of the sound events have the same level of naturalness as the natural sounds in the dataset.

Second, this thesis proposes a method of environmental sound synthesis from onomatopoeic words using the dataset constructed in this thesis. Since onomatopoeic words are effective for expressing the temporal-change characteristics of sounds, it can be expected to control temporal-change characteristics such as the number of repetitions of synthesized sounds by using onomatopoeic words as input information for environmental sound synthesis. In addition, by using onomatopoeic words and sound event labels simultaneously as input information for environmental sound synthesis, the temporal-change characteristics of the synthesized sound and the sound event of the synthesized sound can be controlled simultaneously. From the evaluation experiments on the synthesized sounds, this thesis shows that the use of onomatopoeic words is effective in controlling the temporal-change features of the synthesized sounds. This thesis also shows that the simultaneous input of sound event labels and onomatopoeic words can generate a wider diversity of environmental sounds than when only sound event labels are input.

Finally, this thesis proposes a method of environmental sound synthesis using vocal imitation that imitating environmental sounds using. One of the methods to express the pitch and rhythm of environmental sounds is to imitate environmental sounds by human voices. Vocal imitations can intuitively express the pitch and rhythm of environmental sounds. Therefore, it is possible to control the pitch

and rhythm of the synthesized sound by using a vocal imitation that imitates the environmental sound for environmental sound synthesis. From the evaluation experiment of the synthesized sound, this thesis shows that the vocal imitation of the environmental sound is effective in controlling the pitch and rhythm of the synthesized sound. In addition, when the pitch and rhythm of the vocal imitations used for input is changed, the pitch and rhythm of the synthesized sound changes in accordance with the change in pitch and rhythm.