Abstract of Doctoral Dissertation

Title：Advanced Speech Emotion Recognition for Understanding Various Emotion Expressions

Doctoral Program in Advanced Information Science and Engineering
Graduate School of Information Science and Engineering
Ritsumeikan University

ナガセ　リョウタロウ
NAGASE Ryotaro

This thesis tackles the study of speech emotion recognition (SER), which predicts emotions conveyed by speech. SER has been applied to a wide range of services and products.

This technology is often implemented using machine learning methods. In particular, SER based on deep learning methods has been actively studied, leading to steady improvements in the accuracy of SER over the years. Previous studies have examined methods to predict basic emotions from the whole utterance using only acoustic information as input. Therefore, it is difficult to recognize emotions that are often misrecognized using only acoustic information, emotions that change constantly, and emotions that cannot be represented by only basic emotions. To recognize these emotions, this paper presents the following three studies.

The study of “SER using both acoustic and linguistic information,” investigated the methods combined acoustic and linguistic information. Moreover, these methods were compared under the same conditions. These experiments clarified the effective combination to improve the performance of SER. In addition, the experiments using the transcripts by automatic speech recognition as linguistic information clarified the effect of this on the combining methods.

The study of “SER with emotion label sequence” proposed new methods to train the model to predict a sequence of emotion labels using phoneme information. The results of these methods indicated that phoneme information is effective to recognize fine-grained emotions. In addition, the experiment using utterances in which the emotion changes demonstrated that the proposed methods are effective to recognize emotions that change constantly.

The study of “SER using emotion captions” used an emotion caption which describes emotions as a form of prediction and proposed the method to predict emotion captions from speech or freely define classes by emotion captions during the predicting phase. The results showed that models of predicting emotion captions can recognize emotions which cannot be represented by basic emotions.

Through these studies, this thesis aims to realize the advanced SER which expands the range of the emotion recognized by previous methods.