ACOUSTIC MODELING FOR RECOGNITION AND CONVERSION

OF SPEECH CONTENT BY DYSARTHRIC PEOPLE

No. 2181210 INO Shunsuke

(Supervised by Prof.KYOSO Masaki)

ABSTRACT

Speech signal processing technologies have so far been mostly targeted at able-bodied people. However, with the implementation of the "Law for Elimination of Discrimination against Persons with Disabilities" in Japan in 2016, the importance of information technology support in various fields, both in terms of hardware and software, has increased greatly. In this study, we examine speech recognition for people with dysarthria caused by paralysis due to cerebral infarction. The speech style of dysarthric people with paraplegia differs significantly from that of normal people due to muscle relaxation caused by nerve paralysis, and is unstable, thus speech recognition with a specific speaker model is limited. To improve the accuracy of speech recognition for dysarthric people, this study examines a method that uses lip movements from videos as features in addition to speech features.

First, in setting the recognition rate target for the conceptual system in this study, we referred to a previous study that investigated the tendency of phoneme recognition errors by persons with dysarthria. In this study, phoneme recognition experiments with three dysarthric subjects revealed that phonemes with low correct recognition rates were similar for both vowels and consonants, with particularly low correct recognition rates for the vowels /a/, /i/, /u/, /e/, and /o/. The average correct recognition rate for the five vowels for the three subjects was 86.78%. In order to improve the speech recognition system, the goal of the recognition rate in the conceptual system for this study was set at 90%, which can be said to be a significant difference.

Next, a lip-reading system was created and the data was analyzed to analyze the characteristics of the way the lips of dysarthric people move during speech and to classify the five types of vowels. Support vector machines and neural networks were used for the models. The accuracy of class identification using the neural network for persons with dysarthria was 88%, which exceeded the target value, albeit for five types of vowels.

In the evaluation of the accuracy of the learning model, for normal subjects, the accuracy was higher for the combined features, speech-only features, and lipreading-only features, in that order. For dysarthric subjects, accuracy was higher for the combined features, lipreading-only features, and speech-only features, in that order. Although the accuracy of speech was poor, high accuracy was achieved by combining it with lip-reading features.

Finally, performance evaluation experiments were conducted on the conceptual system of this study. For dysarthria, the system succeeded in producing higher performance when combined with lip-reading features compared to texting done with speech alone.