様式８

論　文　内　容　要　旨

報告番号	甲　先　第　444　号	氏　名	福田　芽衣子
学位論文題目	Construction and evaluation of a new speech corpus of Japanese super-elderly speech recognition (日本人超高齢者音声認識の為の音声コーパス構築)		

内容要旨

Operation of computer and smartphone devices by hand often presents challenges for the elderly, due to aging-related impairment of their vision and motor skills. In theory, the development of easy-to-use speech recognition technology should allow the elderly to access digital communication and electronically stored information with less difficulty. However, due to the unique features of elderly speech, which tend to degrade the accuracy of automated speech recognition (ASR), the necessary level of recognition performance has not yet been achieved when using conventional speech recognition systems. Therefore, we have created a new speech corpus named EARS (Elderly Adults Read Speech), consisting of the recorded read speech of 123 super-elderly Japanese people (average age: 83.1), as a resource for training automated speech recognition models for the elderly.

While investigating the acoustic features of super-elderly Japanese speech using our new speech corpus, we observed that, in comparison to the speech of less elderly Japanese speakers, super-elderly speakers have slower speech rates, extended vowel duration. Males exhibit a slight increase in fundamental frequency, while females exhibit a slight decrease in fundamental frequency.

For this dissertation, we trained various acoustic models using three existing Japanese speech corpora (JNAS, S-JNAS, CSJ) as a baseline, and then adapted them for elderly speech using our super-elderly EARS speech corpus. We also used a combination of the three existing corpora for acoustic modeling, with and without our speech data, and conducted speech recognition experiments. The acoustic models trained with EARS speech achieve word error rates (WER) as low as 13.38%, exceeding the results of our previous study, in which we proposed a CSJ-trained acoustic model adapted for elderly speech (WER = 17.4%).

We also conducted speech recognition experiments using DNN-HMM and transformer based acoustic models trained with a combination of speech data from our EARS corpus and from the three conventional Japanese speech corpora. The models were trained with varying amounts

of EARS data using a simple data expansion method, and were also trained for varying numbers of epochs, without any modifications. When using the DNN-HMMbased model trained with all four corpora for 2 epochs, the word error rate (WER) was reduced to 9.08%, compared to 16.9% when using only the three baseline training corpora without EARS. When using the transformer-based end-to-end speech recognizer, the CER fell from 13.4% to 11.4% when the EARS speech in the four corpora training data was simply doubled.