論文の内容の要旨

Improvements in Pronunciation Evaluation for Reading-Aloud and Shadowing Speech Based on Speech Technology
（音声情報処理に基づく音読・シャドーイング音声の自動評価の改良手法）

氏　名　　羅　徳安

The main goal of this research is to improve automatic pronunciation evaluation of reading-aloud and shadowing based on speech technology for Computer-Assisted Language Learning (CALL) systems. One of the biggest challenges in CALL development based on speech processing is the mismatches between learners' speech and the native speech data that is used to train acoustic model. In Automatic Speech Recognition (ASR), speech adaptation techniques such as Maximum Likelihood Linear Regression (MLLR) have been used to reduce these mismatches by using small amount of the target speaker's speech as adaptation data. However, in the case of CALL, learners' pronunciations often contain errors. Conventional speaker adaptation techniques that use learners' imperfect pronunciations as adaptation data can cause the over-adaptation problem, in which case errors can be transformed into good pronunciations after adaptation. Although some studies use MLLR adaptation (with only one transformation for all pronunciations) to keep the main characteristic of speaker while ignoring the pronunciation details, to the best of the authors' knowledge, no quantitative analysis has been reported to investigate the adverse effects of conventional speaker adaptation techniques.

To address the over-adaptation problems, we first analyze the effects and side effects of conventional MLLR adaptation for pronunciation evaluation in terms of automatic scoring and error detection. Evaluation experiments show that: a) although global adaption with only one transformation for all pronunciations indeed improves performances, when more transformations are used for different pronunciations, over-adaption occurs. b) In automatic scoring, when the number of regression tree is larger than 4, the correlation between automatic scores and manual scores is worse than the original models. c) In error detection, the performance of recall rate decreases due to over-adaptation but the performance of precision rate increases even with over-adaptation.

In order to better benefit from speaker adaption and prevent over-adaption at the same time, this thesis presents a novel idea that uses a group of teachers' perfect

pronunciations to regularize learners' transformation so that over-adaptation problems can be prevented. We name this method Regularized Maximum Likelihood Linear Regression (Regularized-MLLR) and implement it in two ways: one is using the average of the teachers' transformations as constraints adding to conventional MLLR to prevent radical pronunciation transformation, and the other is using linear combinations of teachers' transformation matrices to represent learners' transformations. We refer to the formal implementation as R-MLLR1 and the latter as R-MLLR2. We compare R-MLLR1 and R-MLLR2 with conventional MLLR by conducting experiments on the same conditions as we investigate the adverse effects of MLLR. Automatic scoring and error detection experiments show that the proposed methods outperform conventional MLLR. By adding constraints to MLLR, R-MLLR1 indeed reduces the adverse effects of MLLR, yet performances still drop due to over-adaptation. R-MLLR2 not only out-performs MLLR global adaption, which is widely use for CALL, but also prevents over-adaptation by using linear combinations of teachers' matrices instead of using learners' directly. The proposed methods can better utilize speaker adaptation and prevent adverse effects, and thus more suitable for CALL systems.

Automatic evaluation methods for shadowing are also proposed. Shadowing is a kind of "repeat-after-me" type exercise, but rather than waiting until the end of the phrase heard, learners are required to reproduce nearly at the same time. Recently, shadowing has attracted much attention in the field of teaching and learning foreign languages for its effects of improving both listening and speaking skills. Since learners have to follow the speaking rate of the presented utterance, their pronunciation often becomes very inarticulate and unintelligible. These features of shadowing make it very difficult to build a reliable scoring system for shadowing speech.

Three techniques are proposed for evaluating shadowing speech. One is using Goodness of Pronunciation (GOP) scores calculated through HMM-based forced alignment. In this method, for automatic scoring, the transcription of the presented utterance and the acoustic models of the target language are required. Another is based on continuous phoneme recognition, in which the acoustic models are also needed, but no transcription is required. The third method is using a time-constrained bottom-up clustering technique. Here, only the presented utterance and the shadowed response are required. The transcription and the acoustic models are not needed. Correlations between automatic scores and manual scores, and correlations between automatic scores and learners' TOEIC scores have been investigated and very good results have been obtained.

We also compare the evaluation performances of shadowing and reading-aloud with different cognitive loads posed on learners. Experimental results prove that shadowing can better reflect learners' true proficiency than reading-aloud by posing an adequate level of cognitive load on learners. Therefore, our proposed shadowing evaluation methods can be used to predict learners' over-all language proficiency. A shadowing scoring system has been developed based on these methods. The system is being used for English classes in several universities in Japan and has received very positive feedbacks from teachers and students.

Finally, automatic prosodic evaluation has also been proposed for learners' personal-best shadowing. Experimental results show that rather high correlation with manual prosodic scores has been found. Automatic prosodic scores and segmental ineligibility scores are combined together by using a multiple regression model and the combined scores further improve the performance of automatic scoring that predicts learners' over-all language proficiency.