A pipeline for automatic assessment of foreign language pronunciation

Aku Rouhe[1], *Reima Karhila*[1], *Heini Kallio*[2], *Mikko Kurimo*[1]

[1]Aalto University, Finland
[2]University of Helsinki, Finland

firstname.lastname@aalto.fi

Abstract

We illustrate our prototype pipeline for automatic analysis and assessment of foreign language speech. It uses automatic speech recognition as a preprocessing step for phonetic analysis and predicts the grade that human experts would give for the utterance. The work is applied in two educational research projects: DigiTala for L2 Swedish for Finnish-speaking high school students, and SIAK ”Say it again, kid!” for L2 English for Finnish-speaking 6-9 year old children.

Index Terms: speech recognition, pronunciation grading, language learning, phonetics, speech analysis

1. Introduction

We demonstrate a pipeline for automatic assessment of foreign language pronunciation. First we describe our work in recognising, validating and segmenting read and spoken prompts into phonemes. Then we present our work in automatic phonetic analysis and its application to predict human experts' assessments. The pipeline is applied in two demonstrations that have been developed in two projects for two L2 languages.

In the DigiTala project we are currently developing a high stakes spoken language testing process for L2 Swedish of native Finnish students in matriculation examinations. Since expert human reviewing time is scarce and expensive, assistive technology is needed. After our initial work to establish the use of computerised testing environment [1], we have used the pilot systems to collect data and experiences in real use scenarios.

The SIAK ”Say it again, kid!” project approaches foreign language acquisition in early primary education from an interdisciplinary point of view. In the project we have created a L2 pronunciation learning game and tested it in primary schools.

2. Speech recognition for prompted content

The speech processing pipeline at the server starts with an automatic speech recognition. The rest of the system depends on the recognizer to provide an accurate phone-level segmentation, which requires a successful decoding of the input speech. For L2 learners and particularly children, the decoding is a hard task. Thus, our initial work focuses on read or repeated words, where the content of the recordings can be hypothesised. We demonstrate a system which follows reading, validates that the speech matches our expectation and segments the result into phonemes.

We have built acoustic models for L2 Swedish spoken by Finnish-speaking high school students and L2 English spoken by young Finnish-speaking children. Many of these speakers' data match poorly to the native target language speech corpora, so realistic in-domain data is immensely valuable. Furthermore, most speakers often do not speak fluently, so the conventional language models do not match well, either. We have implemented a heuristic model which tolerates miscues in reading, resembling to the work done in Project LISTEN[2]. An additional constraint is that in the L2 learning game all speech processing operates in realtime.

3. Automatic pronunciation rating

The pronunciation skill level of the speaker is estimated from phoneme recognition results. Individual phoneme segments are classified with bilingual LSTM-RNNs [3] that are trained with native speech data from both the target language and the native language of the users. Scores are computed as a weighted linear sum or other simple regression of phoneme recognition confusion matrices. The regression system is trained in a supervised manner with in-domain utterances of varying skill level annotated by experts.

The English children's system works on very sparse data, returning a score from 1 to 5 for each utterance. The demonstrated Finnish Swedish takes in 8 utterances consisting of around 200 phonemes, and it gives a rough quess of skill level on a 10-step scale.

4. Conclusions

We have demonstrated an automatic speech processing pipeline which provides assessments of spoken language skill. We have educational research projects using the pipeline in both L2 English and L2 Swedish in schools. In the projects we are gathering large in-domain datasets, which can be used to further improve the speech recognition performance and test and analyse the L2 pronunciation learning. We will also expand the systems to other, more open-ended speaking tasks.

5. Acknowledgements

This work has been supported by the Matriculation Examination Board of Finland, Svenska Folkskolans Vänner Foundation, Tekes Challenge Finland programme and the Academy of Finland's grants 251170 and 274075. We acknowledge the computational resources provided by the Aalto Science-IT project.

6. References

[1] R. Karhila, A. Rouhe, P. Smit, A. Mansikkaniemi, H. Kallio, E. Lindroos, R. Hildén, M. Vainio, M. Kurimo *et al.*, “Digitala: An augmented test and review process prototype for high-stakes spoken foreign language examination,” in *INTERSPEECH*. International Speech Communication Association, 2016.

[2] J. Mostow, “Why and how our automated reading tutor listens,” in *International Symposium on Automatic Detection of Errors in Pronunciation Training*, 2012.

[3] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735