Interspeech 2018
2-6 September 2018, Hyderabad

Captaina: Integrated pronunciation practice and data collection portal

Aku Rouhe[1], *Reima Karhila*[1], *Aija Elg*[1], *Minnaleena Toivola*[2], *Mayank Khandelwal*[1], *Peter Smit*[1], *Anna-Riikka Smolander*[2], *Mikko Kurimo*[1]

[1]Aalto University
[2]Helsinki University

aku.rouhe@aalto.fi

Abstract

We demonstrate Captaina, computer assisted pronunciation training portal. It is aimed at university students, who read passages aloud and receive automatic feedback based on speech recognition and phoneme classification. Later their teacher can provide more accurate feedback and comments through the portal.

The system enables better independent practice. It also acts as a data collection method. We aim to gather both good quality second language speech data with segmentations, and the teacher given evaluations of pronunciation.

Index Terms: Automatic pronuciation rating, data collection

1. Introduction

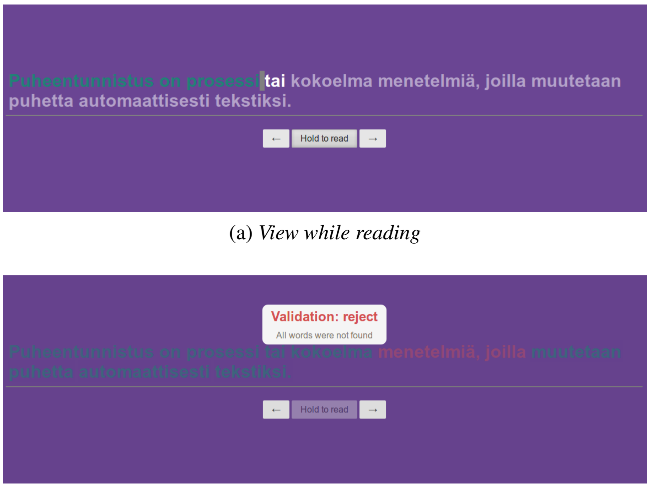
In second language instruction, learning and improving learners pronunciation is typically a very time consuming endeavor. Human teachers rarely have enough resources to provide the individual attention and feedback that learners need and the learners are often too shy to practice and to receive corrective feedback in class in front of peers. The nature of pronunciation learning makes automated pronunciation rating system a suitable tool for private learning and practicing[1]. Furthermore, computerized pronunciation practice tools free teachers to focus on interactive, communicative skills in the classroom.

In this demo we show our work in developing a pronunciation practice portal, which provides automatic feedback and a human evaluation interface. This portal has a dual role: it serves both the needs of language education and research data collection. Previously we have demonstrated automatic pronunciation evaluation of Swedish and English for Finnish speakers in [2] and [3]. The demo portal is being developed in cooperation with Aalto University language courses. The first prototype focuses on Finnish spoken by Arabic origin language speakers. We have bootstrapped an automatic pronunciation rating system for the Arabic-Finnish language pair with a manual data collection phase.

The main problems in computer assisted pronunciation training identified in [4] are first language dependency and integration of the various existing tools. We hope to address the latter concern with the demonstrated system and use it to gather multi-lingual language learning data to enable further research into better pronunciation teaching tools, including first language independency. More recently multiple deep neural network based pronunciation methods have been proposed [5, 6], and these type of models benefit from larger training datasets.

2. Practice portal

The student logs into the practice portal, which is a web based application. They select a batch of phrases to read. The phrases are read aloud, one by one. Figure 1 shows how automatic speech recognition follows the reader in real time to allow self-monitoring. Each utterance is automatically validated immediately, to ensure that the data collected is suitable for automatic processing. Though the goal of this validation is not to judge the reader, it acts as a preliminary level of feedback, and hopefully encourages the student to continue. The speech recogniser also produces a segmentation into words and phonemes, which are stored as well.



Puheentunnistus on prosessi tai kokoelma menetelmiä, joilla muutetaan puhetta automaattisesti tekstiksi.

(a) *View while reading*

Validation: reject

All words were not found

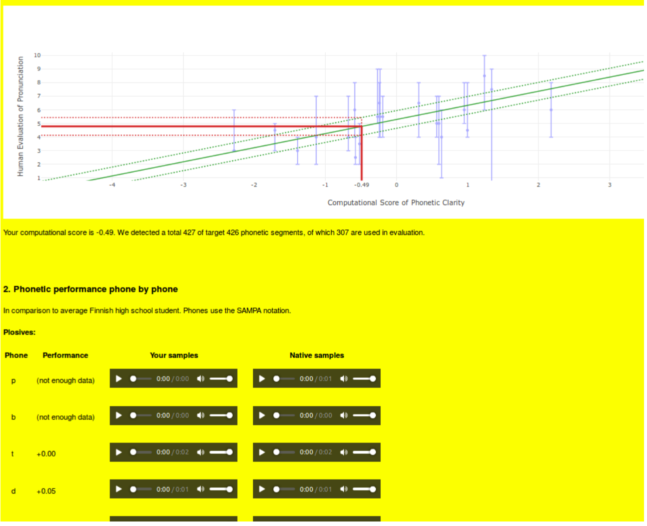
Puheentunnistus on prosessi tai kokoelma menetelmiä, joilla muutetaan puhetta automaattisesti tekstiksi.

(b) *Validation result*

Figure 1: *The automatic speech recognition follows reading in real-time. The speech recognition result is validated: if not all words are found, either the speaker or the speech recogniser has made an error.*

After the selected phrases have been read, the automatic pronunciation evaluation is performed. The student then gets various feedback: a general score and individual phoneme scores, as shown in figure 2. The student may also listen to the recordings and compare with native speaker references.

This demo is not intended for repetitive practice of the pronunciation of single words, but rather to practice full sentences. Our earlier work[2] has indicated that when the training data is limited the pronunciation statistics over a batch of sentences becomes more reliable than that of a single word. However a more recent study[7] shows that after collecting more human evaluation data, the automatic evaluation can be trained accurately enough to provide useful feedback even on individual words. Therefore we are hoping to use this system to gather more data and develop the individual word evaluation capability.


Human Evaluation of Pronunciation
10
9
8
7
6
5
4
3
2
1
-4
-3
-2
-1
-0.49
0
1
2
3
Computational Score of Phonetic Clarity

Your computational score is -0.49. We detected a total 427 of target 426 phonetic segments, of which 307 are used in evaluation.

2. Phonetic performance phone by phone

In comparison to average Finnish high school student. Phones use the SAMPA notation.

Plosives:

Phone	Performance	Your samples	Native samples
p	(not enough data)	0:00 / 0:00	0:00 / 0:01
b	(not enough data)	0:00 / 0:00	0:00 / 0:00
t	+0.00	0:00 / 0:02	0:00 / 0:02
d	+0.05	0:00 / 0:01	0:00 / 0:01

Figure 2: *A general goodness of pronunciation score is given in the chart, which maps computer scores to human rater scores. Individual phoneme scores are also presented.*

2.1. Technical implementation

The realtime automatic speech recognition is built with Kaldi[8] and the server framework from [9]. The training data for speech recognition is purely from native speakers. We have created a system that generates phrase-specific decoding graphs, which are very small and fast, and recognize deviations from the text.

The phonetic segments, extracted with the speech recognition segmentation, are passed through a bi-directional recurrent neural network phoneme classifier which is also trained purely on native speaker data[2, 7]. Finally, a statistical regressor is trained to map the phoneme classification probabilities to corresponding human evaluations. This regressor is then used in the demo to calculate an automatic pronunciation rating for the non-native user.

3. Evaluation interface

To prepare the trainin sessions, teacher can create batches of text to read in the portal. Phrase-specific decoding graphs are then created automatically on the speech recognition backend.

To provide evaluations, feedback and comments to the students, the teachers utilize an evaluation interface that displays the students' data and passes the human evaluations both to the students and the researchers to accumulate the anonymized training data for improving the pronunciation rating performance and enabling further pedagogical research.

This demonstration version gathers human evaluations for each uttered word separately. The granularity of evaluations is somewhat important; during model training in [2], we had low inter-rater agreement on ratings based on longer continuous sections.

4. Data collection

All of the successfully validated utterance recordings, their segmentation, and all human evaluations are transferred to storage. Firstly the data is needed to further develop the practice portal. Secondly if the portal is helpful enough, it may be incorporated into various language courses. The language courses at Aalto University have a steady supply of new students, who are very motivated to learn. Furthermore the teachers are also motivated to provide their evaluations in exchange for being able to use this portal in their courses. This will enable a relatively large-scale, multi-lingual second language audio and pronunciation evaluation data collection effort.

The collected data also has a longitudal aspect, as the courses last multiple weeks and students may take multiple courses. Another interesting research area might be the same students taking courses of different languages.

To allow sharing the data for speech research, the students and teachers will be asked for the proper permissions. The data will be made available through The Language Bank of Finland. Also the tools that we have developed for the data collection can be distributed to the community.

5. Conclusions

We demonstrate a second language pronunciation practice portal. It will enable students to better practice their pronunciation skills, teachers to spend more time in classroom on other skills and a multi-lingual second language pronunciation data collection effort.

Our current system works for Finnish of Arabic origin language speakers, but given suitable training data it can be extended for other languages and language pairs. The data collection effort is the key for improving the quality of the feedback for the student and for further research.

6. References

[1] M. Eskenazi, "Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype," 1999.

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

[3] R. Karhila, S. Ylinen, S. Enarvi, K. Palomäki, A. Nikulin, O. Rantula, V. Viitanen, K. Dhinakaran, A.-R. Smolander, H. Kallio, K. Junttila, M. Uther, P. Hämäläinen, and M. Kurimo, "SIAK–a game for foreign language pronunciation learning," *INTERSPEECH*, 2017.

[4] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," *Proc. IS ADEPT*, vol. 6, 2012.

[5] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[6] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving mispronunciation detection for non-native learners with multisource information and lstm-based deep models," *INTERSPEECH*, 2017.

[7] R. Karhila, A.-R. Smolander, S. Enarvi, K. Palomäki, M. Uther, A. Rouhe, P. Smit, S. Ylinen, and M. Kurimo, "Automatic pronunciation scoring in a language learning game for children," *In review*.

[8] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[9] T. Alumäe, "Full-duplex speech-to-text system for Estonian," in *Baltic HLT 2014*, Kaunas, Lithuania, 2014.