Karhila, Reima; Ylinen, Sari; Enarvi, Seppo; Palomäki, Kalle; Nikulin, Aleksander; Rantula, Olli; Viitanen, Vertti; Dhinakaran, Krupakar; Smolander, Anna-Riikka; Kallio, Heini; Junttila, Katja; Uther, Maria; Hämäläinen, Perttu; Kurimo, Mikko

SIAK — A Game for Foreign Language Pronunciation Learning

Published in:
Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH

Published: 01/08/2017

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Karhila, R., Ylinen, S., Enarvi, S., Palomäki, K., Nikulin, A., Rantula, O., Viitanen, V., Dhinakaran, K., Smolander, A-R., Kallio, H., Junttila, K., Uther, M., Hämäläinen, P., & Kurimo, M. (2017). SIAK — A Game for Foreign Language Pronunciation Learning. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (Vol. 2017-August, pp. 3429-3430). (Interspeech: Annual Conference of the International Speech Communication Association). International Speech Communication Association. http://www.isca-speech.org/archive/Interspeech_2017/pdfs/2046.PDF

INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

SIAK – A Game for Foreign Language Pronunciation Learning

Reima Karhila[1], Sari Ylinen[2], Seppo Enarvi[1], Kalle Palomäki[1], Aleksander Nikulin[1], Olli Rantula[1], Vertti Viitanen, Krupakar Dhinakaran[1], Anna-Riikka Smolander[2], Heini Kallio[2], Katja Junttila[2], Maria Uther[3], Perttu Hämäläinen[1], Mikko Kurimo[1]

[1]Aalto University, Finland
[2]University of Helsinki, Finland
[3]University of Winchester, UK

sari.ylinen@helsinki.fi, mikko.kurimo@aalto.fi

Abstract

We introduce a digital game for children's foreign-language learning that uses automatic speech recognition (ASR) for evaluating children's utterances. Our first prototype focuses on the learning of English words and their pronunciation. The game connects to a network server, which handles the recognition and pronunciation grading of children's foreign-language speech. The server is reusable for different applications. Given suitable acoustic models, it can be used for grading pronunciations in any language.

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

1. Introduction

Language and communication skills play an important role in our society. The benefits of early foreign language learning are acknowledged, but often the learning of foreign languages starts in school so late that the optimal period for language learning has already passed [1]. If foreign-language learning was based on speech rather than written textbooks, it could be started before school-age. Speech-based exposure to foreign language could also result in more native-like brain representations for foreign speech sounds and words compared to exposure to text. Speech-based learning applications and games have potential for this kind of language teaching, particularly as deep learning has opened new possibilities for misprounciation detection in computer assisted pronunciation training [2, 3].

We have designed a game for computers and tablets called Say it again, kid! (SIAK). It uses ASR for the assessment of children's speech in a foreign language. The game teaches foreign words and their pronunciation to children by encouraging them to listen and produce speech. By eliciting speech in a foreign language, the game is expected to gradually establish phonetic categories and word representations in the brain that resemble those of native speakers. Importantly, category learning utilizes subcortical areas that are responsive to feedback [4], which in our game comes in the form of a score computed using speech recognition technology after each utterance. In the project, we aim to verify the activation and plastic changes in these brain ares from playing the game, comparing to a group with a less gamified pronunciation learning environment.

SIAK is implemented as a computer board game. In the current version, the target group is Finnish children with little or no experience in English. The game presents the player a board, where the player can move. There are many boards in the game world. After collecting enough points and working their way through the board, the players can move to the next board and jump between unlocked boards (see Figure 1). A board contains a number of cards that the player can open. Each card introduces a new English word. Later in the game, cards may contain sentences consisting of a few words. Upon opening a card, the player hears the word in Finnish and in English (produced by different native English speakers) and sees a related picture. The child's task is to imitate the word aloud.

Figure 1: *SIAK game records player's utterances using a headset. The game allows the player to move between unlocked boards.*

The player's computer is only responsible for the game mechanics. Speech processing and utterance scoring is done over the network on a server. The game sends a recorded word to the server, and the server returns a numerical score. Then the child's own and native English speaker's utterances are played again for comparison, and the player receives one to five points based on the utterance score for any proper attemps. Increasing scores open new routes on the game board for exploration. In order to keep the game motivating, the players are expected to perform various gimmicks before they can finish a board. In addition to imitation, children are encouraged to test their word learning in test cards where they should recall and say the English word without the model pronunciation.

2. School trials

We are currently starting trials for evaluating the game in classrooms. At this stage, participants are 9-year-old Finnish chil-

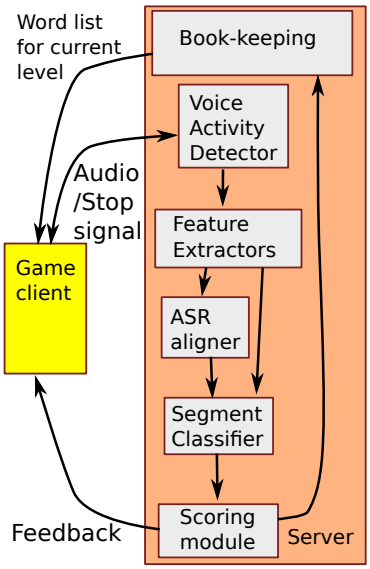

Figure 2: *Block diagram of scoring system.*

dren who have recently started learning English, but later also younger children with no foreign-language experience will be of interest. During gameplay, each player has an Android tablet or a Windows PC and a headset. Children will play approximately 15 minutes per day, 5 days a week for six weeks. During this period, we will collect a database of children's recorded utterances, part of which will be analyzed acoustically. We will also follow up players' progress and different paths of learning from assessment scores provided by the grading system.

In addition, to demonstrate plastic changes in the brain induced by gaming, we will measure childrens brain responses to foreign words with electroencephalography (EEG) before and after the gaming period. Brain responses will also be used to investigate how the accuracy of feedback given by the ASR system affects learning. Specifically, we will compare learning in child groups that receive either accurate realistic feedback or feedback that is not dependent on children's performance.

3. System architecture

The game has been developed using Unity game engine. It can be exported to over 25 platforms with relative ease. Currently we have created Windows and Android builds. When the game is started, it asks the player to sign in using a username and password. The player is identified in order to keep track of the progress of each player and possible in the future adapt the models to each player's voice.

Another component of the game is the network server. It is used for grading speech and storing the game state between sessions. When the player moves to a card on the game board, native English pronunciation of the word in question is played through the headset, the player repeats the word, and the player's pronunciation is recorded. The recording is streamed to the server, and the server starts to analyze the audio as soon as it receives the first packet.

As shown in Figure 2, the server contains several components, most importantly a speech aligner and a phoneme classifier. First features are extracted from raw audio waveform and then forced-alignment is performed using Aalto ASR, a GMM-HMM speech recognizer. As a result of the alignment, the

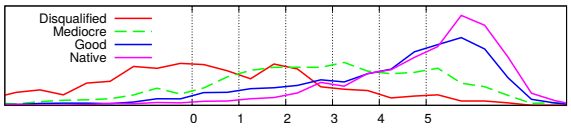

Figure 3: *Performance of the scoring system at the start of experiment: Distribution of points awarded to samples categorised by a human evaluator. Any score above 5 is rounded down to 5, any score under 0 means disqualification. The scores are computed from a leave-one-out validation version of the system. The running system uses all data to build the scoring system.*

system gets a mapping of time frames to phonemes. The second part of the evaluation consists of classifying each phoneme segment individually using a bilingual DNN classifier that has been trained with speech data from both the target language and the native language of the players. The pronunciation score is computed by comparing the phonemes obtained using the DNN classifier to the phonemes obtained using forced-alignment.

A voice activity detector is utilized in the server to detect when the player stops talking. The components have been optimised for speed, and so the score is sent back to the game client most of the time within one second after the server has received the entire word. Using segmentation instead of decoding has its downsides. The segmenter is very hard to tweak to accept a wide variety of speakers in different acoustic conditions, while still rejecting utterances that do not contain the target word. Sometimes, though quite rarely, a player will say something not related to the game and still get a good score.

Figure 3 shows the current performance of the scoring system. The task of evaluating individual, sometimes very short utterances from new players is hard, and as the players are young, sometimes encouragement is more important than precision. The scoring is based on approximately 5000 utterances of varying quality, including failed attempts, collected during the developing of the game, and broadly categorised by one evaluator. The scoring module can be easily improved as more game data is collected.

What still remains to be implemented is a decoder running with a small, restricted grammar running parallel to other speech processing tasks. The decoder results will be used to confirm that there is real effort in the utterance, even if other components would be fooled to award points for the utterance.

4. References

[1] J. Johnson and E. Newport, "Critical period effects in second-language learning: The influence of maturational state on the acquisition of english as a second language," *Cognitive Psychology*, vol. 21, pp. 60–99, 1989.

[2] A. Lee and J. R. Glass, "Mispronunciation detection without non-native training data," in *INTERSPEECH*, 2015, pp. 643–647.

[3] S. Joshi, N. Deo, and P. Rao, "Vowel mispronunciation detection using dnn acoustic models with cross-lingual training," in *INTERSPEECH*, 2015, pp. 697–701.

[4] H.-G. Yi, W. T. Maddox, J. A. Mumford, and B. Chandrasekaran, "The role of corticostriatal systems in speech category learning," *Cerebral Cortex*, vol. 26, no. 4, pp. 1409–1420, 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4785939/