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

Reading validation for pronunciation evaluation in the Digitala project

Aku Rouhe[1], *Reima Karhila*[1], *Peter Smit*[1], *Mikko Kurimo*[1]

[1]Aalto University School of Electrical Engineering, Finland

aku.rouhe@aalto.fi

Abstract

We describe a recognition, validation and segmentation system as an intelligent preprocessor for automatic pronunciation evaluation. The system is developed for large-scale high stake foreign language tests, where it is necessary to reduce human workload and ensure fair evaluation.

Index Terms: L2 learning, speech recognition, reading validation

1. Introduction

The Digitala project aims to create a spoken language test process for national high stakes matriculation examinations, initially for L2 Swedish for native Finnish upper secondary school students. We have previously established the use of computerised testing environments and the need to use all available technologies to reduce the human reviewer workload in [1].

In this demonstration we show our work on automatically recognising, validating and segmenting read text *prompts*. We describe how we tolerate human reading mistakes, or *miscues*. Then we justify a method to reject *utterances*, which we take to mean the audio recording of reading one prompt. Finally we outline the post-processing step and conclude.

To try out the demonstration system the user is invited to read some prompts aloud for the computer. The detected miscues and computed statistics are exposed to the user, along with the acceptance verdict. The user can then listen to and verify the truncated output. A block diagram of the system shows the structure of the demonstrated system in figure 1. We also exhibit our automatic pronunciation assessment system, where this system is a preprocessor. Figure 2 shows a screen capture of the automatic pronunciation assessment system in use.

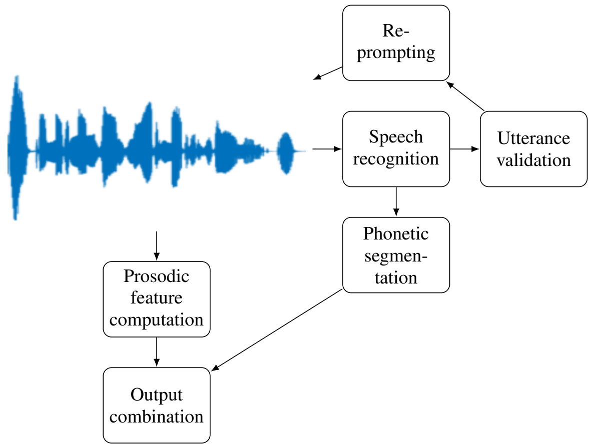
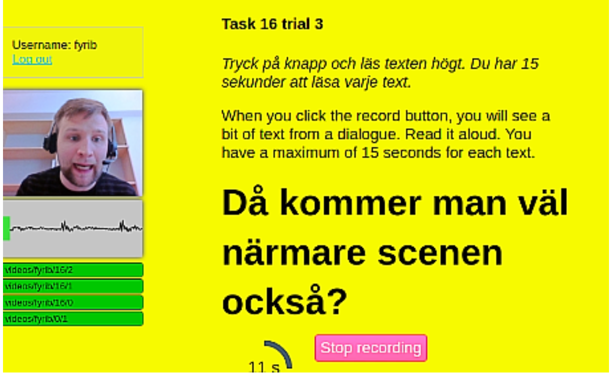


Figure 1: *Block diagram of the recogniser and validator system. Prosodic features are also computed as output.*



Task 16 trial 3

Tryck på knapp och läs texten högt. Du har 15 sekunder att läsa varje text.

When you click the record button, you will see a bit of text from a dialogue. Read it aloud. You have a maximum of 15 seconds for each text.

Då kommer man väl närmare scenen också?

Figure 2: *Screen capture from our automatic pronunciation assessment system in use. It is a browser based system.*

2. Miscue tolerant recognition

The demostration system first decodes the user's reading of the given prompt. The language model is very restricted, but tolerates miscues. The most common reading miscues include repetition, partial or wrong pronunciation, skipping words and adding words which do not appear in the original prompt. Other human noises such as coughing and verbal filler sounds might not be best described as miscues, but in practice we prepare for them as they present the same type of challenge.

The speech recognition task is quite close to forced alignment: we know the words the readers are supposed to say, and their order. This leads to a relatively small search space, which entails fast and accurate recognition. However, the speakers are tested precisely because they are not fluent, and will produce more miscues than native readers.

We use the Kaldi speech recognition toolkit, which is built from the start to use finite state transducers [2]. We use a finite state acceptor to model reading and build the grammar transducer for each prompt separately based on hand-crafted rules. For designing the finite state acceptors we have closely followed the work done in Carnegie Mellon University's Project LISTEN, described in [3]. Our new implementation is available online[1]; it includes some intricacies involving homophones and integration with the Kaldi toolkit. Our recogniser leverages up-to-date deep neural network acoustic models.

3. Utterance and recognition validation

After an initial recognition pass, the demonstration system validates acceptability of the utterance for further automatic assessment. Being able to reject a reading attempt as unfit for automatic assessment can help in building a fair system, both when it is the fault of the system and when the reader can be blamed.

[1] https://github.com/Gastron/miscue-tolerant-lm-fst

If the reading attempt is not well described by the presumptions of the system, then any later assessment will be somewhat unjustified.

Our demonstration system may ask the test taker to repeat a prompt, which is a big potential advantage. Of course, this affects the nature of the test somewhat: the test taker is given a second chance. We stress that the alternative may be performing automatic assessment on false premises.

The task of choosing which attempts to reject is performed using machine learning methods, which will have its own errors. Our demonstration system also supports hand-crafted rules. For example, the test preparer may have used a specific difficult word and may wish to require it to be present for automatic assessment.

3.1. Learned rejection boundaries

The feature set is gathered from automatic speech recognition output. This holds an assumption that an attempt's suitability for automatic assessment is correlated with the attempt's suitability for automatic speech recognition methods. We expect this to be true since we use similar methods and similar data in building the systems, but this remains to be tested when larger data sets are acquired.

Two types of features are included: statistics describing the output, such as percentage of target phonemes that were found and number of miscues detected and automatic speech recognition confidences, which indicate how well the speech recognition model seems to describe the utterance. There are two reasons we would want to reject an utterance: that our tools do not work for the utterance or that the utterance does not work for our tools. The difference is in the source of errors, but we expect the symptoms to be similar.

In practice we train a binary classifier on a data set which we have manually inspected and labeled. We expect to have a large majority of accepted utterances. A support vector machine has been our initial choice. Its prediction rules can be interpreted easily, which is important for a transparent test procedure.

4. Segmentation and post-processing

In the post-processing step, we gather the outputs to be passed to the automatic assessment system. Our current automatic pronunciation rating system is based on phoneme-level statistics. Therefore we retrieve phoneme-level segmentation. We also deduce which spoken words correspond to which prompt words. Figure 3 shows example output waveform annotated. We are also developing prosodic feature extraction to enable automatic prosodic analysis.

An open question remains as to what to do if there are multiple attempts at pronouncing a given word. In our system, the last attempt is assumed to be the one the reader was happy with; but we could also score each pronunciation attempt and choose to keep the best one or perhaps the average. We also compute a variety of other parameters such as number of miscues, acoustic confidences and speaking rate.

5. Conclusion

We have described a system which recognises, validates and segments read prompts, and we have justified its place in a nation-wide computerised spoken L2 language examination process. Larger data sets are needed for developing and testing performance, but the technical implementation is demonstrated. The effect on human workload and test fairness remain to be quantified.

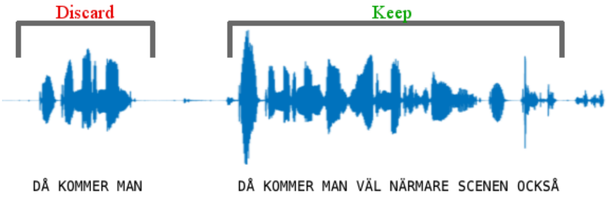


Figure 3: *Recognised output, the first part is discarded as it is repeated in the second part.*

6. Acknowledgements

This work is funded by the Finnish Matriculation Examination board.

7. 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] 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.

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