INTERSPEECH 2016: Show & Tell Contribution
September 8–12, 2016, San Francisco, USA

Digitala: An augmented test and review process prototype for high-stakes spoken foreign language examination

Reima Karhila[1], *Aku Rouhe*[1], *Peter Smit*[1], *André Mansikkaniemi*[1], *Heini Kallio*[2], *Erik Lindroos*[2], *Raili Hildén*[2], *Martti Vainio*[2], *Mikko Kurimo*[1]

[1]Aalto University School of Electrical Engineering
[2]Helsinki University
reima.karhila@aalto.fi, heini.h.kallio@helsinki.fi

Abstract

This paper introduces the first prototype for a computerised examination procedure of spoken foreign languages in Finland, intended for national scale upper secondary school final examinations. Speech technology and profiling of reviewers are used to minimise the otherwise massive reviewing effort.
Index Terms: Spoken language skill assessment, speech recognition, phonetics, speech analysis

1. Introduction

The Digitala project designs a testing and reviewing process for a national-scale, high stakes upper secondary school final examination in spoken foreign languages. In Finland, a country with a population of 5.5 million, each year around 35 000 upper secondary school students take part in the national final examinations, which include among other subjects two mandatory second languages, and many will also do an exam in one or more optional languages. The exams are taken seriously: performance affects access to higher level education. Currently the foreign languages exams test reading and writing skills as well as listening comprehension. However, the Matriculation Examination board, which operates under the Ministry of Education, has stated that a test in spoken skill will become a new mandatory part, starting with Swedish in 2020. With the current effort of computerising the written parts of the final exams, the lack of history in testing spoken skill, the large scale of the tests and the lack of extra resources for reviewing, it is necessary to computerise the test and use all available technologies to reduce the workload of the test organisers.

In this paper we will quickly describe the test from the point of the testee, give an overview of the reviewing process, and describe the system architecture of the complete test and review setup. Finally, we will finish by presenting an obvious conclusion.

This paper accompanies a demonstration system that includes a small subset of test tasks from a complete prototype system for Finnish Swedish. Guests are invited to try the role of a tested student, do the test with their own devices, and get an estimate how their speech parameters compare to the Finnish secondary school students' speech parameter distribution.

2. Test Procedure

The secondary school students are expected to use their own devices in the final examinations. Although some specifications are given for students to do their laptop and tablet shopping, there will be a myriad of hardware and software configurations present. To avoid compatibility problems, the prototype pilot test works in a web-browser with a minimalistic interface, of which Figure 1 shows an example. The test itself consists of a series of tasks, ranging from reading tongue-twisting sentences to reacting to described situations, or to a simulated video conversation. These tasks have a pool of trials, from which a random set is given to each examinee. The test proceeds with a quick pace. The clock starts ticking as soon as a question or task is shown. Answering times are limited to 10-30 seconds and no retakes are allowed. Due to the randomisation of the trials, the test can be conducted simultaneously for a group of students in a classroom using headset microphones given that the acoustic properties of the room keep the volume on a reasonably low level.

Figure 1: *An examinee's view of the test system.*

The testees do not get immediate feedback from the system. They will get their scores in due time, and they will have an official channel for filing complaints and requesting re-reviews of their audio and video materials.

3. Grading Procedure

The written final exams are first graded by local teachers, and the review is checked and corrected by a censor of the matriculation examination board. Using the same procedure for spoken exams poses two major challenges. First, There are no extra resources allocated to the grading process, and the teachers and censors need to find extra time for this work. The second problem is the lack of tradition in grading spoken skills in foreign languages in Finland. The majority of teachers have done their pedagogical and language studies with an emphasis on evaluating written performance. There are no guarantees for consistency of evaluation from one school to another. The simple solution to the second problem is to collect more reviews for each testee, which only makes the first problem worse.

The role of speech technology in the project is to reduce the amount of work the teachers need to put into grading. Speech recognition can give statistics on vocabulary size, and produce segmentation for computing phonetic features like speech rate

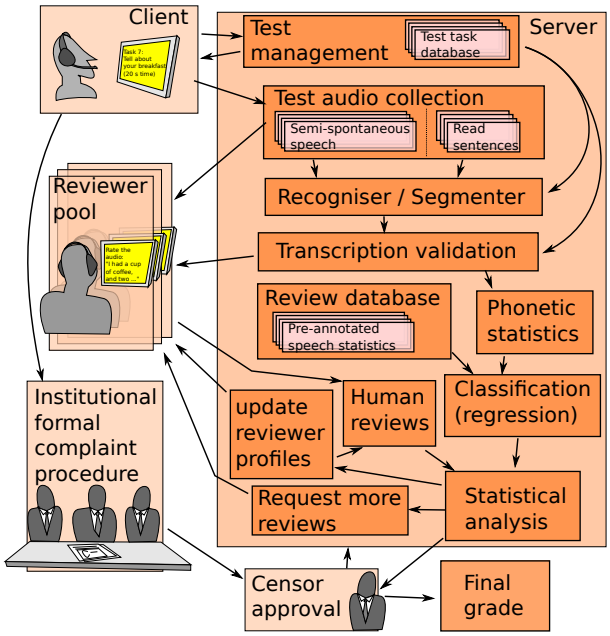

Figure 2: *Schematic of the proposed computer-assisted system.*

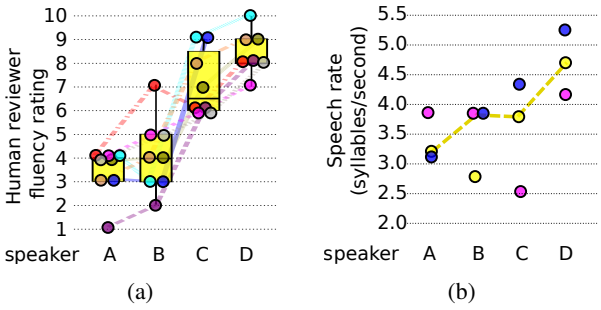

Figure 3: *(a) A sample from review collection: 7-8 expert reviewers have evaluated the fluency of Swedish spoken by four upper secondary school students. (b) Same students' speech rates for three sentences with median plotted.*

and prosodic chunking [1] (see Figures 3b and 4), which in preliminary experiments seem to correlate well with reviewers opinions. Together the various statistics are used to form a crude estimate of the speech quality. Whether this estimate will be used as a substitute for a human reviewer or as a guide to human reviewers is to be decided experimentally.

As shown in Figure 3a, the current consistency in evaluating spoken foreign languages leaves a lot to be desired. The large scale of the national exam allows profiling the reviewers and statistically evaluating the quality of the reviews using approaches typical in crowd-sourcing tasks [2]. Building the components for minimizing the number of reviews is taking baby steps, as gathering enough review data to robustly test the different available algorithms is slow.

4. Prototype System Architecture

Figure 2 shows a schematic of the system. Running the test for data collection requires the publicly available server code[1] and a database of exam tasks with associated media files. The tes-

[1]https://github.com/rkarhila/pilot_test_for_spoken_foreign_language

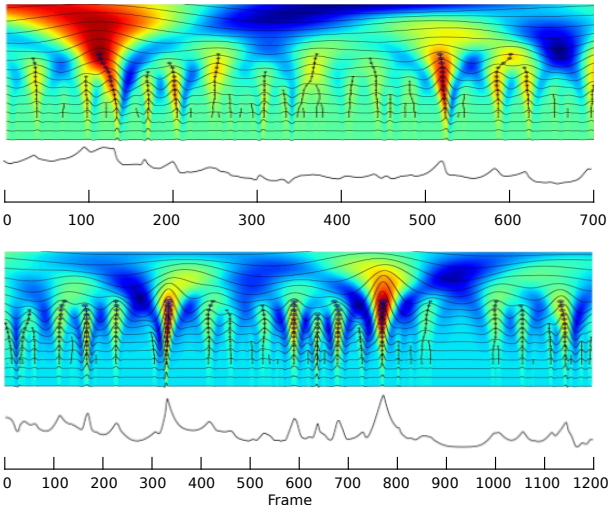

Figure 4: *Continuous wavelet transform analysis and F0 contours of a fluent speech sample (above) and a non-fluent speech sample for the utterance "Jag tycker att det är konstigt att människor håller på med allt möjligt hela tiden." In the fluent speech sample the prosodic chunking is more coherent.*

tees need a computer or tablet with a modern web browser, microphone and camera. The media devices are accessed through WebRTC functionality[2], with some experimental components[3]. The audio and video data is transferred to the test server via encrypted connections. The speech technology components work offline, and are based on AaltoASR[4] and Keras[5].

We are developing every aspect of the pilot test in an agile manner, making changes with feedback from all user groups and improvements with new, accumulated data. The first prototype test has been tested in schools around the country and we are currently working on developing the components related to grading.

5. Conclusions

We have introduced our prototype for a computerised, large-scale examination procedure of spoken foreign languages. The system is still under heavy development, but most remaining work lies in grading interface design and improving algorithms and models as more real-life data becomes available. User feedback has been encouraging despite the deficiencies.

6. Acknowledgements

This work was funded by Svenska Folkskolans Vänner.

7. References

[1] M. Vainio, A. Suni, and D. Aalto, *In Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*. Springer, 2015, ch. Emphasis, Word Prominence, and Continuous Wavelet Transform in the Control of HMM-Based Synthesis.

[2] P. Welinder and P. Perona, "Online crowdsourcing: Rating annotators and obtaining cost-effective labels," in *proc. CVPRW*, 2010.

[2]RFC 7478 https://tools.ietf.org/html/rfc7478
[3]https://www.webrtc-experiment.com/RecordRTC/
[4]https://github.com/aalto-speech/AaltoASR
[5]http://keras.io/