

TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Martin Väljaots 163097IAPM

COMPUTER AIDED PRONUNCIATION TRAINING TOOL FOR ESTONIAN

Master's thesis

Supervisor: Einar Meister
PhD

Tallinn 2018

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Martin Väljaots 163097IAPM

EESTI KEELE HÄÄLDUSTREENINGU RAKENDUS

Magistritöö

Juhendaja: Einar Meister
PhD

Tallinn 2018

Author's declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Martin Väljaots

07.05.2018

Abstract

The goal of this thesis was to develop a computer aided pronunciation training tool for Estonian vowels, using free signal processing components as much as possible. The application, featuring pronunciation exercises for Estonian vowels and a listening exercise for Estonian vowel quantity degrees, was tested with 26 subjects, approximately a third of whom had first languages that were not Estonian. Based on the test subjects' pronunciation results and feedback, as well as the author's remarks, the application was analyzed. The application was improved where possible at the current time, with some ideas remaining as potential future improvements.

This thesis is written in English and is 41 pages long, including 6 chapters, 17 figures and 2 tables.

Annotatsioon

Eesti keele hääldustreeningu rakendus

Töö eesmärgiks oli arendada hääldustreeningu rakendus eesti keele vokaalide jaoks, kasutades selleks võimalikult palju tasuta signaalitöötlusvahendeid. Töö jooksul valmis Java rakendus koos JavaFX kasutajaliidesega, mis sisaldab hääldusharjutusi kõigi eesti keele vokaalide jaoks ning üht kuulamisharjutust eesti keele völdete eristama õppimiseks.

Töö jooksul valminud rakendust testiti 26 katsealusega, kellest ligikaudu kolmandikul ei olnud eesti keel emakeeleks. Võttes arvesse katsealuste hääldustulemusi, katsealuste poolt esitatud tagasisidet ning autori tähelepanekuid, analüüsiiti rakendust, et näha, millistes aspektides oleks vaja muudatusi teha.

Eesti vokaalide referentsväärtused ja vokaalipiirid määratleti esmalt vastavates uuringutes esitatud andmete põhjal, hääldustulemuste analüüsi järel kohandati mõnede vokaalide esialgselt määratletud meeskõnelejate vokaalipiire. Katsealuste tagasiside tulemusena muudeti kuulamisharjutusele vastamist lihtsamini mõistetavaks. Hääldusharjutustesse lisati juhendeid oma kõne salvestamiseks ning võimalus enda häälduskatsete progressi jälgimiseks.

Rakendus vastab ülesandepüstitusele ning seda võib pidada edukaks. Tulevikus saaks rakendusele lisada harjutusi ning funktsionaalsuse, mis võimaldaks õpetajatel ise harjutusi koostada. Samuti saaks rakenduse kasutajaliidest muuta modernsemaks ning rakenduses võiksid olla esindatud ka naissoost emakeelerääkija hääldusnäited.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 41 leheküljel, 6 peatükki, 17 joonist, 2 tabelit.

List of abbreviations and terms

| | |
|-----------------|--|
| ASR | Automatic speech recognition |
| CAPT | Computer aided pronunciation training |
| Formant | An acoustic resonance of the human vocal tract |
| JSTK | Java Speech Toolkit |
| LPC | Linear predictive coding |
| Quantity degree | How long a vowel or consonant is pronounced |

Table of contents

| | |
|---|----|
| 1 Introduction | 10 |
| 2 Background and related works | 13 |
| 2.1 A Visual Feedback Tool For German Vowel Production – Carroll, Trouvain, Zimmerer 2015 | 16 |
| 2.2 An Audiovisual Feedback System for Pronunciation Tutoring – Mandarin Chinese Learners of German – Ding, Jokisch, Hoffmann 2011 | 17 |
| 3 Application overview | 19 |
| 3.1 User interface..... | 22 |
| 3.1.1 Login screen | 22 |
| 3.1.2 Registration screen | 22 |
| 3.1.3 Microphone volume adjustment screen..... | 24 |
| 3.1.4 Exercise selection screen | 25 |
| 3.1.5 Pronunciation exercise screen | 26 |
| 3.1.6 Listening exercise screen..... | 29 |
| 4 Testing | 32 |
| 4.1 Testing results..... | 33 |
| 4.2 Analysis of pronunciation results | 36 |
| 4.3 Test subject feedback..... | 40 |
| 4.4 Revisions based on test results and test subject feedback | 42 |
| 5 Assessment and future improvements | 48 |
| 6 Summary..... | 50 |
| References | 51 |
| Appendix 1 – GitHub link to the application | 53 |

List of figures

| | |
|---|----|
| Figure 1. Login screen. | 22 |
| Figure 2. Registration screen. | 23 |
| Figure 3. Successful account registration confirmation pop-up window. | 23 |
| Figure 4. Microphone volume adjustment screen. | 24 |
| Figure 5. Microphone volume level saving confirmation window. | 25 |
| Figure 6. Exercise selection screen. | 26 |
| Figure 7. Example of a pronunciation exercise screen for a male user. | 27 |
| Figure 8. Pronunciation exercise results graph tooltip. | 27 |
| Figure 9. Example of a pronunciation exercise screen with one user result. | 28 |
| Figure 10. Listening exercise screen. | 29 |
| Figure 11. Listening exercise screen with the answer partially filled. | 30 |
| Figure 12. Listening exercise pop-up window for a correct answer. | 30 |
| Figure 13. Listening exercise pop-up window for an incorrect answer. | 31 |
| Figure 14. Listening exercise results pop-up window. | 31 |
| Figure 15. Revised microphone volume adjustment screen. | 44 |
| Figure 16. Revised pronunciation exercise screen. | 46 |
| Figure 17. Revised listening exercise screen. | 47 |

List of tables

| | |
|--|----|
| Table 1. Testing mean results for female subjects..... | 37 |
| Table 2. Testing mean results for male subjects..... | 39 |

1 Introduction

When learning a foreign language it is important to acquire not only the vocabulary and grammar, but also the correct pronunciation. To acquire proper pronunciation, one must first learn to perceptually distinguish the acoustic difference between the phonological categories (vowels and consonants) of the learner's native language and the target language and only then is one able to use pronunciation exercises to acquire proper pronunciation in the foreign language. In learning pronunciation, it is important to get feedback on your pronunciation and assess how close your pronunciation is to the foreign language's correct pronunciation. Acoustic analysis allows us to analyze different characteristics of speech signal and give feedback about the learner's pronunciation. These kinds of pronunciation training applications exist for some widespread languages such as English, Spanish, French and German, but a successful one has yet to be developed for Estonian. Since being able to speak the national language is an important part of obtaining Estonian citizenship, an Estonian pronunciation training application would be very helpful to a language learner. This application could also be useful for children with logopedical problems and adults in need of speech rehabilitation, such as people whose speech is inhibited as a result of a stroke.

The main aim of the thesis is the development of a CAPT prototype which helps learners of Estonian to acquire the production and perception of Estonian vowels. In this thesis the pronunciation training and evaluating will be limited to Estonian vowel quality, as consonants are more difficult to accurately evaluate. Signal processing methods will be used to extract relevant acoustic characteristics from the recordings of a learner, which will be visualized to give feedback on the correctness of and how to improve the learner's pronunciation.

Main functionalities of the application:

- Users can create new accounts and sign in to existing accounts
- Users can set up their microphone for accurately evaluating their pronunciation

- Users can save their microphone volume settings to their account
- Pronunciation exercises for each vowel
- Perceptual exercises involving vowels in different quantity degrees
- Visualization of pronunciation results
- Assessment of pronunciation results and instructions to improve the production
- Instructions for use – how to set microphone volume for accurate pronunciation recording, how to complete exercises and how to interpret feedback

The technical specifications that the application will be developed under are as follows:

- Desktop application to work on Windows 10 Home
- Java programming language
- Use of free signal processing components
- Requires a microphone (for pronunciation exercises)

To achieve the goal and provide the functionality, the following research questions will be addressed:

1. Which signal processing methods should be used to extract different acoustic characteristics?
2. How should different acoustic characteristics be visualized?
3. Which acoustic characteristics would be the most practical and useful to visualize?
4. What kinds of instructions should the application give to improve pronunciation?
5. Should the application work with just a laptop microphone or should it require a separate, higher quality microphone?

To validate the results, the application will be continuously tested during development by the author, as well as tested by both native speakers and people with other first languages to gather feedback about the application as a whole and especially about how the application gives feedback to the user to validate the chosen methods for giving feedback and relaying information to the user. Results from the pronunciation exercises will be used to analyse whether the application needs vowel range recalibration. If the application is able to perform its functions, it should be considered as successfully developed.

2 Background and related works

For a person to learn a foreign language, there are four main language skills they must acquire: reading, listening to, writing in and speaking the language [1]. Reading, listening to and writing in a foreign language are activities that can perfectly and in a fairly straight-forward manner be implemented to be learned using a computer, especially in a world where everyone is connected to the Internet with a plethora of media in any language to consume and people speaking any language to chat to [1]. However, teaching a student the speaking aspect with the help of a computer proves to be a more complex task [1]. The task of implementing computer assisted speech training can be difficult for multiple reasons: on the technical side, for instance, the learner requires access to an adequate microphone and a quiet environment to use a computer in; on the linguistic side, the implementation of speech processing and making computer assisted pronunciation training (CAPT) useful to a learner is made difficult by the steps required for creating meaningful feedback out of a learner's speech signal [1].

Although feedback in a CAPT tool can be provided in a variety of ways, the more common methods are feedback through visualization and feedback through automatic speech recognition (ASR) [2].

Visualization makes use of “the graphical display of a native speaker's face... [and] the vocal tract“, spectrograms, waveforms and pitch tracings [2]. Of the displays named, spectrograms and waveforms, if they are used as the only method for feedback, are often seen as not worth being used due to not being understandable to the layman out of the box, requiring additional time solely for training the learner to interpret their feedback [2]. However, studies like [3], [4] suggest that they can be of use when learning a language, and Coniam, in 2002 [5], successfully used spectrograms to educate native Cantonese speaking English teachers in Hong Kong on the distinctions between the local Hong Kong English and American English [2].

Pitch tracings, showing how the speaker's pitch changes during speech, seem to be a more straight-forward and easier to understand method of feedback, although still

needing some instruction before use [2]. Pitch tracings are widely recommended and can be used to explain intonation on both sentence level (like in [4]) and word level [2].

Another visual feedback method is formant plotting, used, for example, by Carroll, Trouvain, Zimmerer in 2015 [6], which will be examined further in subchapter 2.1. A formant is “a concentration of acoustic energy around a particular frequency in the speech wave” [7]. Fundamental frequency (F0) correlates to the speaker’s pitch, the first formant (F1) frequency correlates to how high the tongue is positioned in the mouth during pronunciation and the second formant (F2) correlates to tongue position front-to-back [8]. The frequencies are found from the learner’s speech recording, then, formants 1 and 2 are correlated on an x-y graph, which has phonemes placed on it, to show the learner how close they are to pronouncing the target phoneme. An example of this can be found in Kay Sona-Match, which Carey, 2004, used to study whether computer-based visual feedback systems could improve the vowel quality of people who speak English as a second language [9]. Although formant plots can be useful in inspiring a learner to reach a target, formant frequencies for each phoneme vary from person to person, and when phonemes have similar formant frequencies, the frequencies found from a learner’s pronunciation may be closer to those of a different phoneme. This can cause possible confusion or distrust in the formant finding functionality of the CAPT tool. In addition, formant plots would require additional training for the learner to understand, as presenting a graph of correlated frequencies without explaining how they are related to phoneme pronunciation is sure to cause confusion and hinder improvement.

As for automatic speech recognition, whether it is able to “effectively provide immediate feedback” remains a core problem in CAPT [2]. If the non-native speaker’s pronunciations aren’t close enough to those of a native speaker, should the feedback on the errors be more general (example from [2]: “Your score is 62%. This means that you often have words that are mispronounced.”) or more specific, like feedback on the pronunciations of specific words or phonemes? And after this question is answered, another one, a general problem for CAPT design, arises: how will the learner be instructed to improve their pronunciation and fix their errors [2]?

With ASR systems, the issue of providing incorrect feedback is an enormous problem, and it’s one that can lead to immense frustration of a student. Because of this, some

studies recommend giving feedback in an implicit manner – less precise (example from [2] : “the word *system* was not pronounced correctly”) instead of giving it in an explicit manner – more precise and on a deeper level (example from [2], “the first vowel in *system* was pronounced wrongly”) [2]. Neri et al, in 2003, stated that feedback using ASR should “keep the recognition task as simple and as limited as possible, by carefully designing the learning activities” [2].

When done correctly, minimizing learner frustrations and guaranteeing a high level of consistency, CAPT can be a valuable tool for language learning, giving a learner individualized instructions and feedback faster than a teacher could, allowing the learner to practice more frequently and in a more focused, repeated manner, and providing automatic visual feedback for how accurate the learner’s pronunciation is compared to example pronunciations [2]. In addition, CAPT tools can help with languages for which a properly trained teacher may be hard to find, and also in situations where practicing pronunciation is given little time in class [2].

Hardison, 2004, argued that a CAPT application could give more specific feedback than a teacher, and that a CAPT tool would remain more accurate and consistent than a teacher is likely to be with regards to evaluating each learner [2], [4]. CAPT can drive a learner to become more independent in practicing their pronunciation, and here, Hardison is not dismissing a teacher’s guidance as unimportant, as both lessons by a teacher and using a CAPT tool lead to an improvement in pronunciation – instead, CAPT can be used to help with the evidence-backed problems of pronunciation teaching not being prevalent enough in classrooms and with issues of teachers feeling that they are not prepared to teach pronunciation. A quote from [2] sums up the notions of Hardison and the idea of CAPT as being a tool for teachers, instead of being their replacement: “rather than a false fight over an already too small piece of the language teaching pie, CAPT is a way to expand the pie so that more teachers and learners can enjoy their own use of spoken language.”

2.1 A Visual Feedback Tool For German Vowel Production – Carroll, Trouvain, Zimmerer 2015

An example of a recent CAPT application is a visual feedback tool for German vowel production by Carroll, Trouvain and Zimmerer, 2015 [6]. The tool is written as a script for Praat, a computer program by Boersma & Weenink for analyzing, synthesizing and manipulating speech [10]. The tool allows the learner to listen to native German recordings of consonant-vowel-consonant nonsense words and record their own pronunciations of these words. When a learner records themselves, they are given feedback on how long they pronounced the vowel part compared to the recording in the form of a bar representing time, without displaying the duration in seconds. An acoustic feedback space tells the learner how close the first and second formants of their pronunciation are to the recording. The learner can record themselves multiple times - the last five recordings are saved and can be replayed.

As the tool is written as a script for Praat, all of the speech signal processing is done using Praat. To determine the vowel part of the learner's recording, the script first finds where the recording's intensity curve exceeds a certain threshold, then, provided that the peak in intensity is longer than 40 milliseconds in duration, considers it as being the vowel segment. Formant analysis is done at the midpoint of this considered vowel segment using Praat's built-in find formant feature.

According to the authors, this method for vowel detection is “naive”, and one can see how: as only sound intensity and duration are considered for determining the vowel segment of the recording, a loud noise with a long enough duration, for example, could be detected as a vowel, which may result in completely incorrect formant values being found and incorrect feedback being given to the learner. The authors suggest that the tool could be improved in the future by improving the current vowel detection system to also be suitable for use with multi-syllabic words and even phrases, implementing adjustments of vowel targets to a learner's personalized acoustic space, and testing how effective the tool is for actually improving a learner's perception and production of German vowels. In addition to these three points, the authors also propose collecting user feedback on how intuitive and easy to use their chosen visual feedback method is.

2.2 An Audiovisual Feedback System for Pronunciation Tutoring – Mandarin Chinese Learners of German – Ding, Jokisch, Hoffmann 2011

The article explores the possibility of creating a German pronunciation tutoring system for Chinese speakers, using a German learning tutoring system named EURONOUNCE [11]. EURONOUNCE is “a corpus-based learning system, which integrates large speech corpora and multilingual speech databases”, which needs special speech corpora for each pair of German and another language. In this case, to create a German/Chinese pair, the authors collected speech data and analyzed Chinese students learning German, and reported on their attempt at building a prototype tutoring system.

An analysis of Chinese speakers learning German determined three main points for why CAPT is applicable in China. Firstly, in China, foreign language classes put more emphasis on reading and writing the language, while pronunciation is often neglected due to the amount of students one teacher has to divide their time for. What is more, the “teachers are embarrassed because of the lack of phonetic instruction strategies”. CAPT could give a chance for students to practice their pronunciation and receive feedback and suggestions for improvement.

Secondly, due to most students not learning German until at a university, the learner’s “perceptual discrimination of phonetic sounds is not as good as that of a child; their learning of pronunciation should be enhanced by informative visual feedbacks” to help them learn the language at their more mature age. Lastly, the logographic orthography of Chinese means that learners will process the material they read more visually than phonologically, due to their non-alphabetic first language background. The learners will have a very hard time understanding the pronunciation deviations between the two languages, making “tutoring systems with feedback information best fit the requirements of Chinese German learners”.

The EURONOUNCE system was installed in the language lab of CDHK (Chinesisch-Deutsches Hochschulkolleg). The authors worked with language teachers to find where students were having difficulty and tried to create a suitable curriculum.

The speech data needed for a special corpora was mostly collected when the EURONOUNCE system was used by students in the lab. The students were presented

with audiovisual feedback of how accurate their pronunciation is, they could find the areas they had most difficulty in and try to imitate the standard speaker's pronunciation. Questionnaires were designed to collect data for determining the efficiency of the tutoring system, which were carried out before, during and after the use of the tutoring system.

Acoustic and perpetual investigations of the students recordings were carried out to find common difficulties for students. The two that the authors pointed out were, verbatim:

- 1) Inaccurate production of those German vowels and consonants which are nonexistent in Chinese
- 2) Incorrect placement of tonal categories and wrong phonetic realization of a phonological category

An example of using CAPT feedback to enhance a student's learning is brought out by the authors about using an intonation curve to show how the student should be raising their pitch at the end of a question. A point is made that this curve should be smoothed out somewhat, as automatic pitch tracking algorithms often show many small pitch changes, which can confuse the learners.

The authors state that after testing the tutoring system, some obstacles still remain. For instance, when multiple students use the system in the same language lab, the accuracy of speech recognition will be affected by the fact that the room will be filled with multiple people speaking at the same time. On the other hand, if the students were to use the system at home, data collection would be much more difficult, if possible at all. Another problem is that most language teachers have difficulty in understanding the feedback that the system provides, due to having little knowledge of acoustic phonetics, meaning that introductory courses would be absolutely necessary.

The authors conclude that audiovisual feedback information can obviously, over time, improve the learner's pronunciation, but that "a faithful imitation of isolated words or sentences with visual aids cannot guarantee a good pronunciation in ordinary speech".

3 Application overview

The application is written in Java. Java was chosen mainly because the author of this thesis works as a Java developer. Although C++ was originally considered due to audEERING's openSMILE, which is a “versatile and fast open-source audio feature extractor” and seemed like the perfect tool for extracting formants, the author has no experience with C++ [12]. Praat was also considered due to Carroll, Trouvain and Zimmerer's work [6]. Praat has its own scripting language and has built-in formant finding, among other features [10]. Ultimately, Praat was not chosen due to Java being more widely used, which means that Java is more likely to have more readily available information and tutorials than Praat is. What is more, Praat was said by Carroll, Trouvain and Zimmerer to provide limited control over the application's visual layout, which threatened to be a problem when trying to create intuitive and meaningful visual feedback [6]. The GitHub link to the application is available in Appendix 1.

As the CAPT tool would be used to evaluate Estonian vowel pronunciation, the user's pronunciation would have to be analyzed. JSTK, or the Java Speech Toolkit, was capable of performing formant analysis on a recorded sound file [13]. In addition to this, the user should be able to listen to their own pronunciation and compare it to a native pronunciation, which can be used to establish a difference between pronunciations in the learner's native language and Estonian. This means that the user pronunciation should be recorded either way.

In developing the CAPT tool, one of the main problems would be how to separate the user's speech into segments to understand where the vowel pronunciation is, so that it could be recorded into a separate soundfile and then be processed by the formant finding functionality in JSTK [13]. An attempt was first made to improve on Carroll, Trouvain and Zimmerer's method, which they themselves called a “naive” method [6]. Their method considered the vowel pronunciation to be in a part of the user's recording where the sound intensity peaks above a threshold of 10 decibels below maximum intensity and which lasts at least 40 milliseconds [6]. The attempted method would use

both a sound intensity threshold and pitch detection to determine the vowel segment. Pitch detection is useful because unvoiced consonants cannot be assigned pitch, therefore, if the word being pronounced starts and ends with an unvoiced consonant (for instance, /k/, /p/, /s/) and has a vowel in the middle, where pitch is detected, there is a vowel [14].

Unfortunately, this method ultimately failed to fortify vowel segment determination because the pitch detection of TarsosDSP was often too slow and couldn't detect pitch even as a long vowel was being pronounced in a word [15]. The implementation of this method also caused issues with pronunciation recordings, leaving the users to listen to a crackling version of their pronunciation. Ultimately, it was decided that a similar method to Carroll, Trouvain and Zimmerer's was going to be used, except instead of deriving the threshold by subtracting a value from the maximum recorded sound intensity, the user would set the threshold themselves to a level where only the vowel pronunciation exceeds it. In a way, this method is a little better, because if a loud noise is 10 or more decibels louder than the rest of the learner's speaking, no vowel part may be detected. On the other hand, having the user make an extra step to use the application is something that should generally be avoided, if possible. What is more, as all sound that is above the threshold is recorded to be part of the vowel segment (even after the vowel pronunciation has ended), noises may interfere with formant analysis.

The vowel segment would then have to be analyzed and for this, formant values of the vowel segment are determined. For determining the formant values, the application uses JSTK, which has a formant finding feature [13]. JSTK applies a Hamming window to create frames from an audio source, feeds the frames through autocorrelation and uses a LPC (Linear Predictive Coding) spectrum, which formant values can be extracted from. These formant values are then filtered to be within a certain range of a native speaker's mean value for that vowel. This range is 200Hz for the first formant and 400Hz for the second formant. The filtering helps eliminate noise that might be in the recording. An average of the remaining values is then found for each formant.

To evaluate the correctness of the pronunciation, the average formant values are compared to formant values from [16]. The study from which these formant values originate has not been published at the time of writing this thesis and were provided by the supervisor of this thesis, who is one of the authors of the study. These formant

values were found from continuous speech and will have to be analysed after testing to see whether they are accurate for this application or whether they should be adjusted. To provide feedback on the learner's pronunciation, the formant values are depicted on a formant graph, like in [9].

For each user to have the ability to set their own volume threshold for vowel detection and not have to set that threshold each time they use the application, the application allows users to register an account. The username is used to name the folder where the threshold value file, log file and user pronunciation recordings will be kept. The user saves their threshold to their account and can later log in and start using the application without the need to set up their microphone again. If it is necessary, the threshold can later be adjusted. A log file is kept of user actions such as their exercise results and their moving between screens to aid in analyzing test results and user behavior.

The application features pronunciation exercises for Estonian vowels and one listening exercise for Estonian vowel quantity degrees. Both types of exercises feature native Estonian recordings of words. All of the utterances were recorded by the author of this thesis. In the listening exercise, Estonian vowels occur in two-syllable words representing the three Estonian quantity degrees. Listening to the vowels with different duration and in different word contexts contributes to the better production of the Estonian vowels.

Originally, the application was planned to have a pronunciation exercise for quantity degrees. The idea was to have the user record their pronunciation of a word like "*saada*" – consonant, vowel, consonant, vowel – and compare the lengths of the vowels in the recording to determine whether the user's pronunciation was in the correct quantity degree. Unfortunately, as there is no pitch detection backing up vowel detection, the author does not find the current implementation of vowel detection to be robust enough to attempt quantity degree evaluation. It is one thing to record one vowel segment in a consonant-vowel-consonant word and perform formant analysis on this one segment, but it is much more difficult to determine two vowel segments in a consonant-vowel-consonant-vowel word and compare their duration. The current implementation of vowel detection cannot, with a sufficient degree of certainty, claim to have recorded the full length of each vowel. This functionality could be added in the future.

3.1 User interface

JavaFX was used for prototyping the user interface [17]. As the application is written in Java, JavaFX can easily be used to create a GUI. In each of the following subchapters, a screen in the application will be described, along with the functionality behind it.

3.1.1 Login screen

The first screen that the user sees when the application is started is the login screen, as seen in Figure 1. On this screen, registered users can log in to their existing accounts and new users can click the “register new user” button to create a new account. This button takes the user to the registration screen. When the user logs in, they are taken to the exercise selection screen.

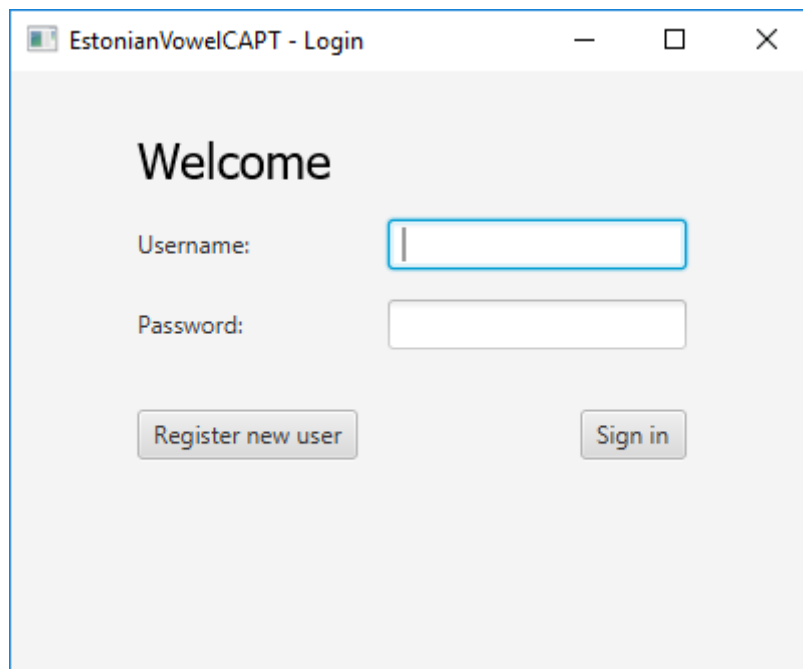


Figure 1. Login screen.

3.1.2 Registration screen

The registration screen is shown when the user clicks on “Register new user” in the login screen. The layout of the registration screen is as seen in Figure 2. The user is asked to enter a username, enter a password and repeat it, and select their gender. The only strength measures for the password are that the password needs to be at least one

character long and the entered passwords need to match. Gender selection has two options (male or female) and is mandatory. A tooltip is displayed when hovering the cursor over the gender selection box, explaining that gender is used to evaluate the user's pronunciation. In pronunciation exercises, male and female users have different native pronunciation ranges values for formants.

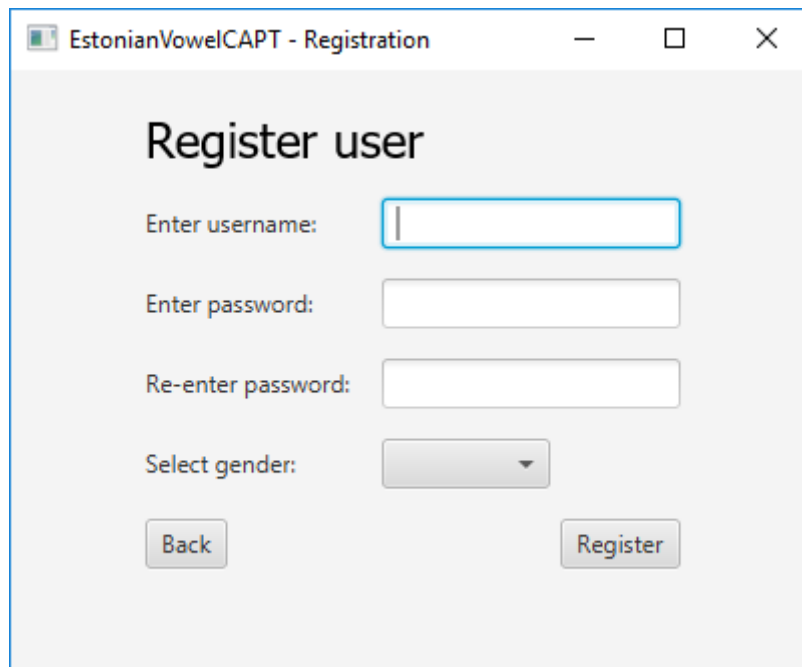


Figure 2. Registration screen.

When the user is done, they click the register button. Users can also move back to the login screen by clicking the back button. Once the user has registered successfully, meaning they have entered something in each of the fields and the entered passwords match, the user is shown a pop-up window confirming that they have successfully registered an account and next, they will be guided through adjusting their microphone volume. This pop-up window can be seen in Figure 3.

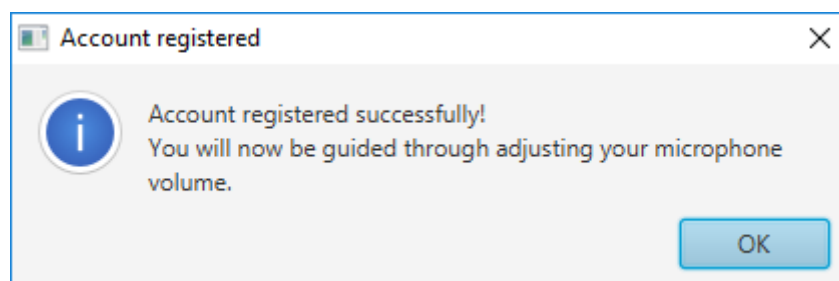


Figure 3. Successful account registration confirmation pop-up window.

3.1.3 Microphone volume adjustment screen

The microphone volume adjustment screen is where the user adjusts their microphone volume to set a threshold for vowel detection. The layout of this screen is seen in Figure 4. For setting the vowel detection threshold, the user is guided to first listen to the pronunciation of the Estonian word “*kook*” and then replicate that pronunciation. The user is instructed to move the slider on the screen to set the microphone volume so that only the vowel part “*oo*” lights up as green. This is necessary for accurately measuring your pronunciations. The red bar shows the current highest recorded volume. When the user is done, they can press the save button. When this screen is reached from the exercise selection screen, a cancel button also appears on screen and the user can choose not to save their threshold.

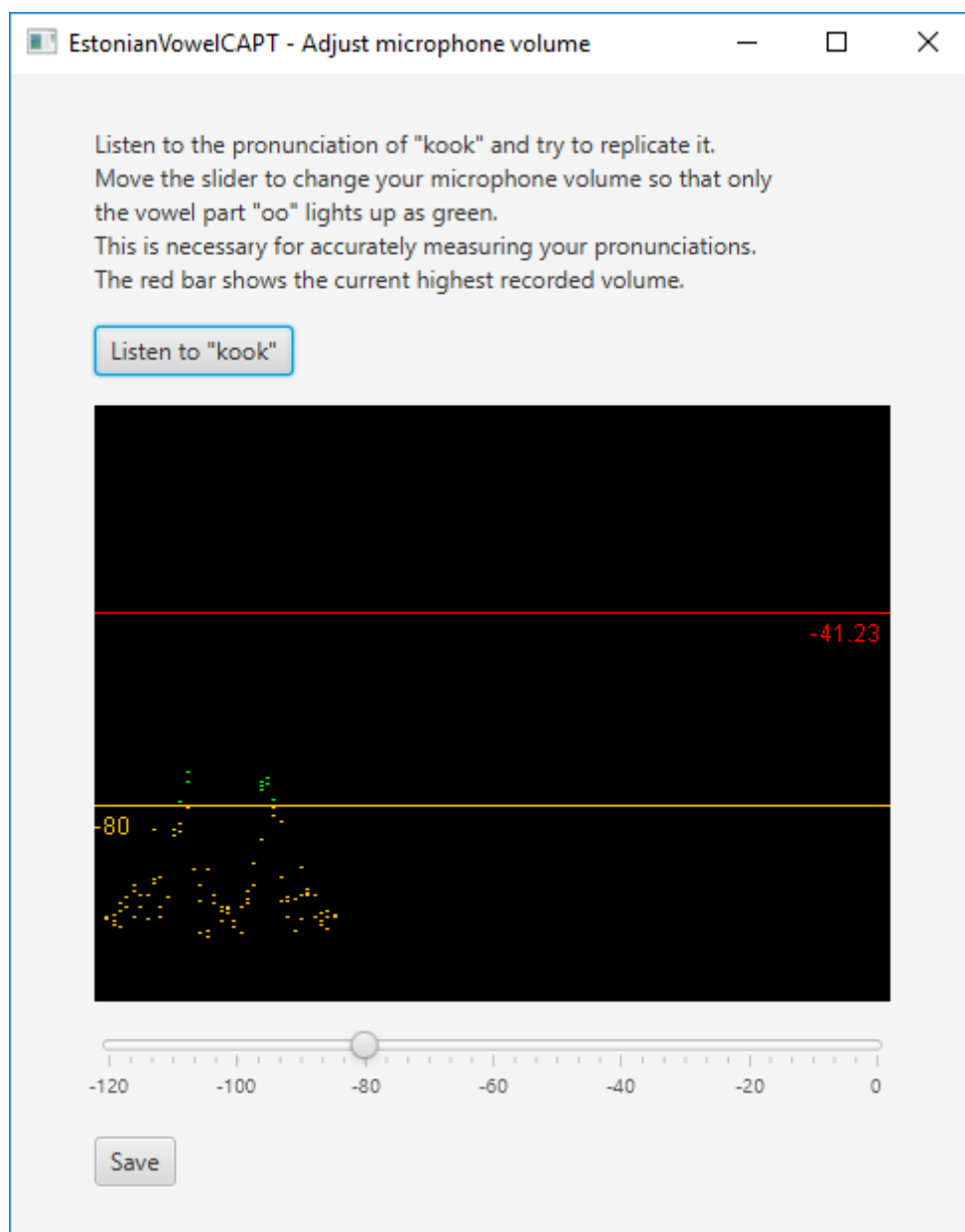


Figure 4. Microphone volume adjustment screen.

When the user clicks the save button, a pop-up window appears asking for confirmation. This window is seen in Figure 5. If the user chooses to click OK, the microphone volume threshold is saved in the user's threshold file. If the user cancels, they have another chance to adjust their microphone volume level.

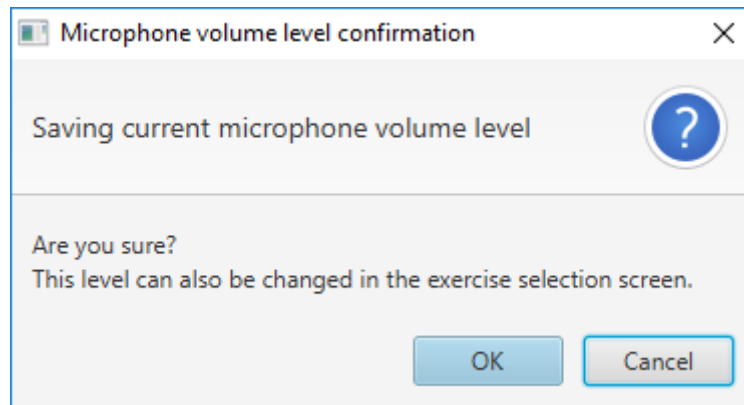


Figure 5. Microphone volume level saving confirmation window.

3.1.4 Exercise selection screen

When the user finishes registration or when the user logs in, they are greeted with the exercise selection screen. This screen is seen in Figure 6. The exercise selection screen shows which user is currently logged in and contains buttons to enter pronunciation and listening exercises or adjust microphone volume. The application has one pronunciation exercise for each vowel in the Estonian language (9 total) and one listening exercise. The exercise selection screen acts as the central screen for a logged in user. When they click on an exercise or go to adjust their microphone volume, the exercise selection screen is the screen they come back to when they're done.

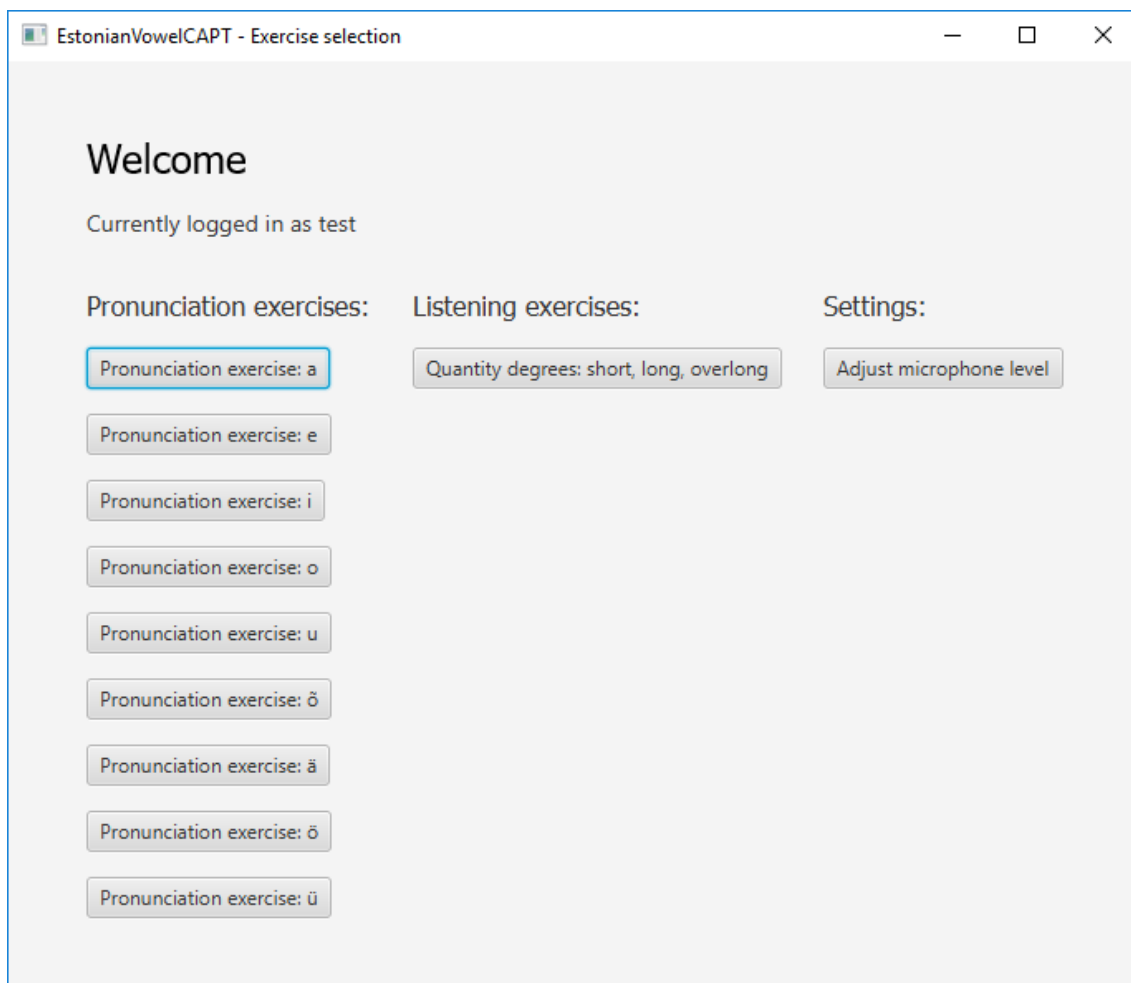


Figure 6. Exercise selection screen.

3.1.5 Pronunciation exercise screen

The pronunciation exercise screen contains various means to guide the user to improving their Estonian vowel pronunciations. An example of this screen for an /e/ pronunciation exercise is seen in Figure 7. The user is given a word (consonant, double vowel, consonant format) in Estonian, along with its English translation. The user can view an animation of the mouth and tongue position for pronouncing that vowel, and listen to a native speaker’s recording of the word. The animations for vowel pronunciations were made by Jürgen Lasn, based on roentgenograms published by Georg Liiv in 1961, for the application “Õpime hääldama!” developed by Priit Penjam [18] [19].

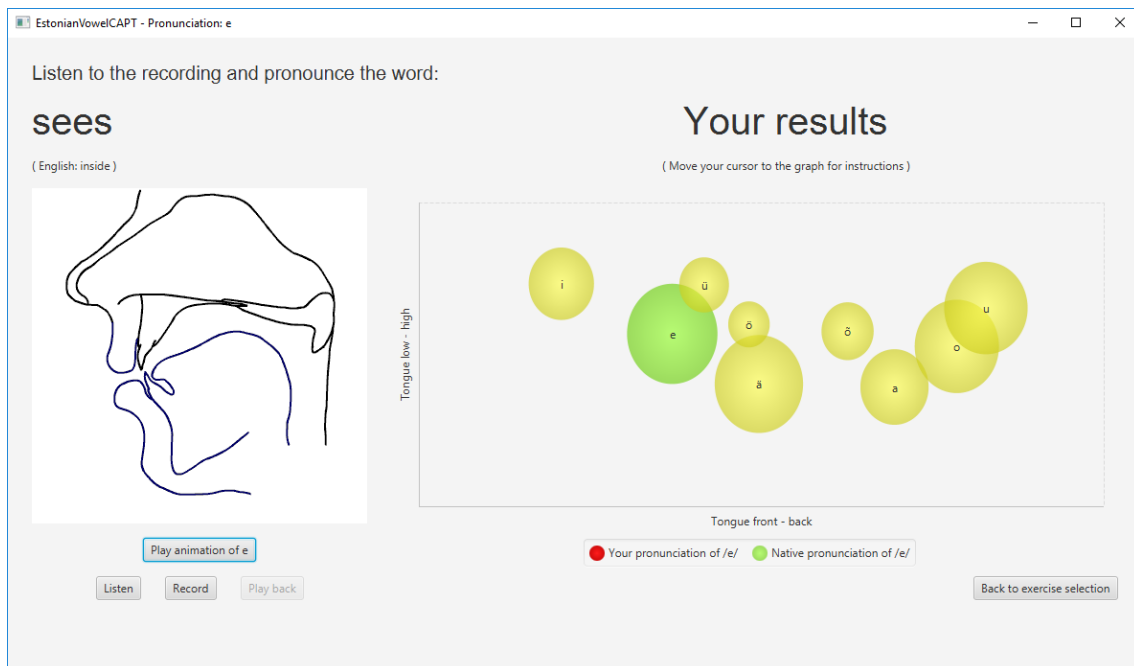


Figure 7. Example of a pronunciation exercise screen for a male user.

The results graph (on the right in Figure 7) shows native pronunciation ranges, which have been additionally widened by 20% so as not to discourage new learners when they are off the mark. When the user hovers their cursor over the graph, a tooltip with additional information explaining the meaning behind the graph and a guide to the exercise is shown. This tooltip is seen in Figure 8.

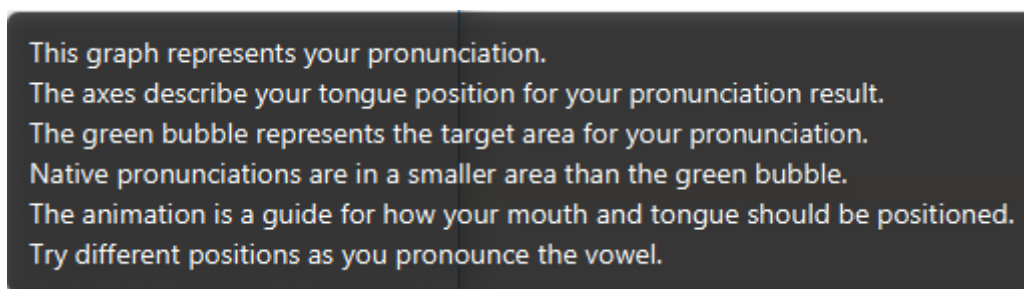


Figure 8. Pronunciation exercise results graph tooltip.

Once the speaker is ready to record their pronunciation, they can click the record button, pronounce the word, and click the button again to stop recording themselves. After this, formant analysis is performed on the user's pronunciation. Provided that the formant analysis found a result, the user's result appears on the results graph as a red dot. An example of how a user's result is represented on the graph is seen in Figure 9.

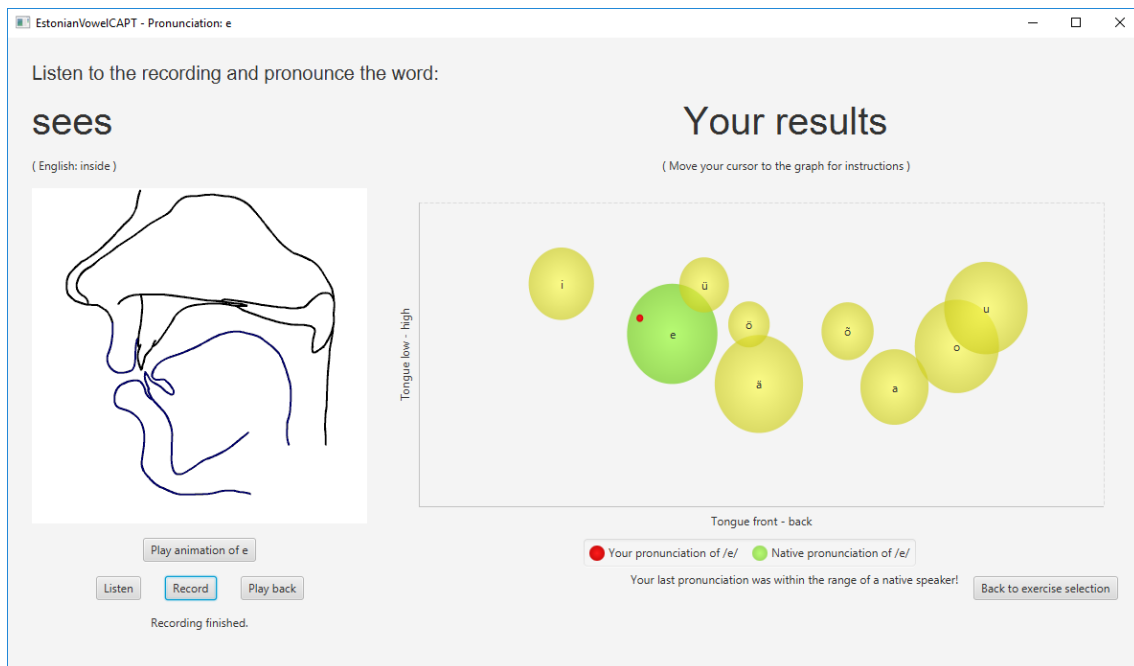


Figure 9. Example of a pronunciation exercise screen with one user result.

The user can listen to their recording and compare it to the native pronunciation. For each user recording, a message explaining their result is shown underneath the results graph. In the example in Figure 9, the user's pronunciation was within the range of a native speaker. If the user's recording would've been outside the range of a native speaker, a message suggesting that the user should look at the results graph and try positioning their tongue in a different way would've been displayed, in reference to the graph. A third kind of message would be displayed if the system wasn't able to detect a vowel, meaning that no values close enough to one or both formants were found.

The user can attempt the pronunciation as many times as they'd like. Each new attempt adds another red dot to the results graph. When the user is finished with the exercise, they can go back to the exercise selection screen by clicking the "back to exercise selection" button.

3.1.6 Listening exercise screen

The listening exercise is meant to help users learn to distinguish between the three Estonian quantity degrees – short, long and overlong – in vowels. The exercise screen is seen in Figure 10.

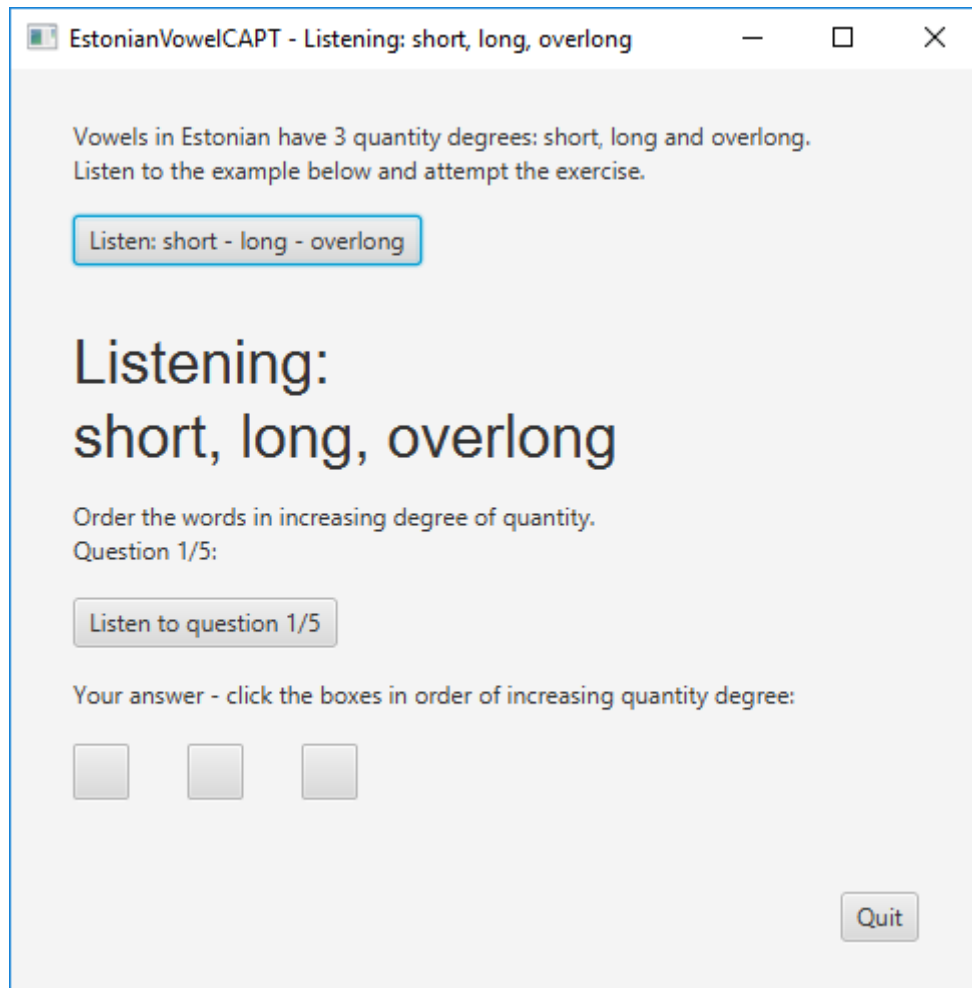


Figure 10. Listening exercise screen.

The user is first instructed to listen to an example of three words, all having the same vowel but ordered in an increasing degree of quantity. Then, the user is presented with a listening exercise, where three words, each with the same vowel but with a different quantity degree, are presented for listening in varying order of quantity degree. The user has three checkboxes representing the words that they heard. The user has to first tick the box for the word where they heard the short vowel, then tick the box for where they heard the long vowel and lastly, tick the box for where they heard the overlong vowel. An example of the user having chosen the short and long vowel words is seen in Figure 11.

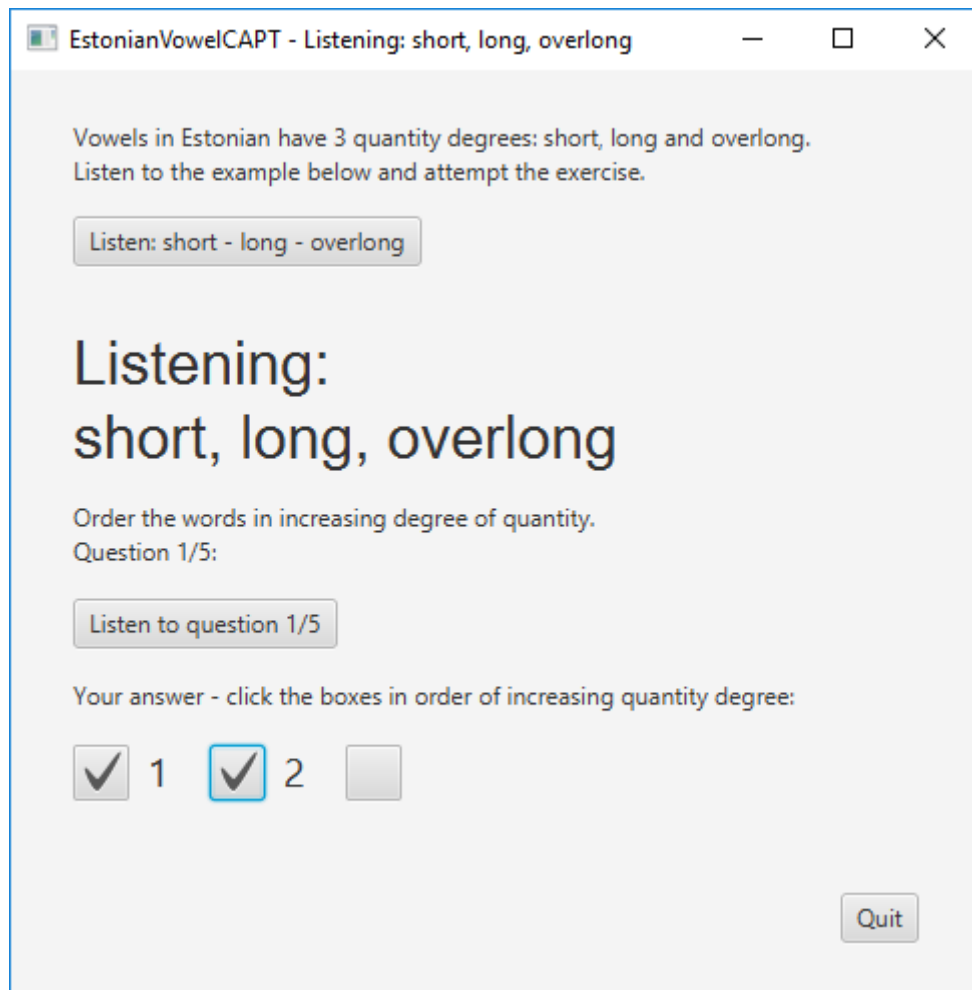


Figure 11. Listening exercise screen with the answer partially filled.

The user can change their answer by unticking the boxes. Once the user ticks all the boxes, the answer is evaluated. If the answer is correct, a pop-up window seen in Figure 12 is shown.

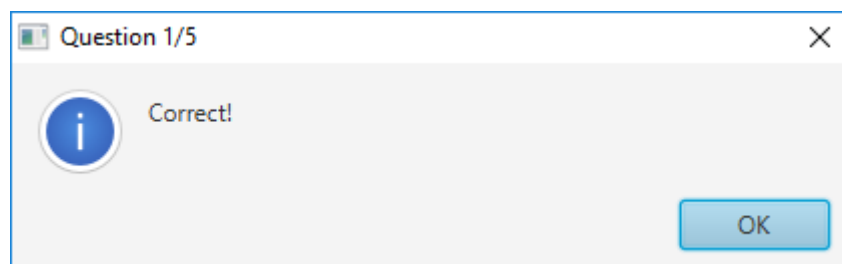


Figure 12. Listening exercise pop-up window for a correct answer.

If the answer is incorrect, a pop-up window seen in Figure 13 is shown. This window lets the user know what the correct answer would have been.

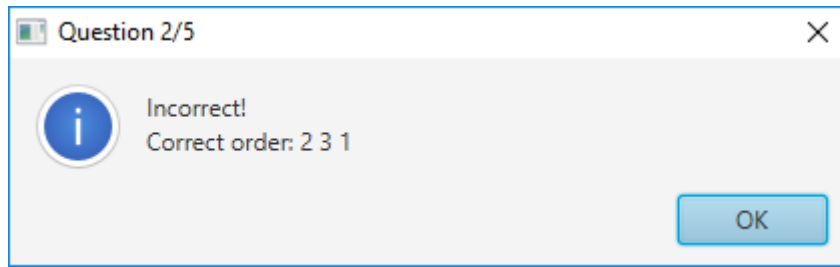


Figure 13. Listening exercise pop-up window for an incorrect answer.

From the second question onward, the question is first played once without the user clicking the “listen to question” button. This isn’t done for the first question because the user is expected to first listen to the example and then attempt the exercise, and is done from the second question onward to require less clicks from the user.

Once the user has answered all the questions, a final pop-up window with their result is shown. This window is seen in Figure 14. When the user clicks OK, they are taken back to the exercise selection screen.

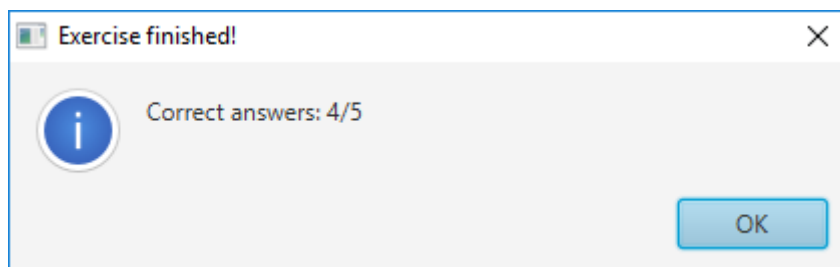


Figure 14. Listening exercise results pop-up window.

4 Testing

Test subjects were mostly native Estonian speakers, with slightly less than a third of the subjects speaking various other languages as a first language. Subjects were asked to use the application as though they were learning a language and to use the application independently as much as possible, but were also assured that if they had any questions or comments, they could express them. Each subject created an account, set the microphone threshold and were asked to get three results for each of the pronunciation exercises, as well as complete the listening exercise once. The author wrote down any comments, questions and feedback from the subjects, and provided instructions where necessary. The application was running on the author's personal laptop. All testing was carried out in an individual study room in the library of the Tallinn University of Technology and all subjects used the same headset for pronunciations. An individual study room in the library was chosen as the author wished to oversee all testing in case there were any questions, but also because the individual study room was quiet, minimizing potential noise recordings. Although the application could be used with a laptop microphone in theory, an average laptop microphone is sensitive to the laptop being tapped, which creates a spike in sound intensity that could peak above the vowel detection threshold. As some subjects could wish to use the laptop's touchpad instead of the mouse, the tapping could have interfered with the calculation of their results.

Afterwards, all test subjects filled out a feedback form with the following questions, which were translated to Estonian for native Estonian speaking subjects:

1. What did you think of the application?
2. How useful would you say the listening exercise implemented in the application is for learning to differentiate Estonian quantity degrees? (graded on a scale of 1-5, 1 being not useful at all, 5 being very useful, along with an elaboration on the answer)

3. How useful would you say the pronunciation exercises implemented in the application are for learning to pronounce Estonian vowels? (graded on a scale of 1-5, 1 being not useful at all, 5 being very useful, along with an elaboration on the answer)
4. In your opinion, was the visual feedback (animations, graphs, messages) provided in the application easy to interpret or was there something confusing?
5. How could the application be improved? Should something be removed or added?
6. Would you use an application like this for learning a language?

4.1 Testing results

In total, 26 subjects participated in testing, 18 of whom were native Estonian speakers and 8 of whom had various other first languages: Azerbaijani, Bengali, Georgian, Hindi, Japanese, Tamil, Turkish and Urdu. Of the native Estonian speakers, 10 were male and 8 female, and of the test subjects who had other first languages, 6 were male and 2 female. In all cases but one, the subject completed all three attempts of each pronunciation exercise, along with completing the listening exercise. In the one exception, the subject failed to receive results for the /u/ pronunciation exercise due to a lack of oversight by the author.

Most test subjects had no issue creating an account. A few subjects did, however, attempt to create accounts with a space in the username, which reminded the author that the testable application had no checking for special characters. Special characters can cause issues with creating files and folders and it is generally a good idea to avoid them in the username and password. The registration process should be strengthened with additional checks in the future.

Most subjects had severe difficulty understanding the microphone threshold setting instructions and the author had to explain this step on most occasions. Some of the subjects did not understand what to do at all, while most of the subjects listened to the example, then replicated the pronunciation and then asked for further instructions. Only

a few subjects listened to the example, replicated the pronunciation, set the threshold to a suitable level using the slider and then asked for the author's confirmation to continue. One subject commented that instead of having a save button at the bottom of the screen, the button could instead read "next".

For the most part, the subjects completed the pronunciation exercise without much difficulty. A few subjects asked whether the pronunciation was set up to be recorded as "push-to-talk", meaning that they would have to hold down the button to record their pronunciation. In one instance, a subject did not click the record button when attempting their first pronunciation. A few subjects commented that the native pronunciation stopped playing randomly, but that when they clicked the button again, the full pronunciation played without a problem. In some cases, a crash would occur in a pronunciation exercise after the subject stopped recording themselves. The application would eventually recover and give a result, but in the event of a crash, the subject was asked to make an additional attempt. The majority of the native Estonian speaking subjects did not pay attention to the animation, but this makes sense, as a native speaker might feel that they have already obtained this knowledge. Most non-native subjects did, however, use the animations to gain additional knowledge for vowel pronunciation. Some subjects wouldn't listen to their own recordings and when they were reminded that they could listen to their pronunciations after they had recorded them, some subjects replied that listening to one's own voice feels odd. A couple of subjects became quite captivated with how their results were represented on the graph and wished to make additional attempts to see how much they could improve their result. Sometimes, the subjects would make comments that would reveal that they had not noticed some of the information that they were being presented, such as asking for the meanings of words or what the graph represented.

On some occasions, the test subject became confused as to why their result was being displayed as being within the range of a native speaker, although the message beneath the graph said that they were outside the range. This is caused by the fact that the ranges on the graph are widened by 20%. On some other occasions, the result of a subject would be displayed as being outside the range of the native speaker, although the message beneath the graph would indicate that they were within the range. This is caused by the fact that the ranges are circles and only have one radius, whereas in actuality, the range would be an ellipse with a separate radius for each of the formants.

In the current implementation, the radius of the vowel range that is depicted on the graph is calculated by finding the average of the standard deviations of the formants and then adding 20% to it to be more forgiving and encouraging for new learners. This widening of the ranges is disclosed in the mouseover text of the results graph. At the same time, the actual calculations of whether the pronunciation was within range or not are done for each formant separately, taking into account their respective mean and standard deviation values. If either of the formants is not within range, the result is considered as not being within range, even though it may appear to be within range on the results graph. After testing, it can be concluded that this implementation can cause a lot of confusion in the user and a better, more consistent implementation should be added.

Most subjects initially experienced difficulty in answering the listening exercise. The subjects were confused as to what the boxes represented and what they had to do once they had listened to the question. Once the author explained the meaning behind the boxes and how the user should use them, the subject completed the exercise without issue. A bug was discovered in the second question where the correct answer would prompt a pop-up window stating that the user was incorrect, and one incorrect answer would prompt a pop-up window stating that the user was correct. When this occurred, the subject was informed of whether their answer had actually been correct or not.

Taking the testing results into account, the application has points that need addressing. The main focus should be to fix the bug in the listening exercise and reduce or eliminate the crashes in the pronunciation exercise, but it is also apparent that the information being presented to the user is, at times, either hard to notice or understand. The problem with how the native pronunciation ranges and user results are displayed on the results graph, which can be inconsistent with what the message below the results graph states, can cause a lot of confusion. This issue could be mended, for instance, by either removing the message entirely, displaying the ranges in a more accurate way, having a better way to tell the user that the ranges displayed on the graph are wider than the native ranges for vowel pronunciation or by taking the widened target area into account when calculating the user result. The case where the result is displayed to be outside of the range and the message below the graph states that the user was within the range of a native speaker has to definitely be fixed.

4.2 Analysis of pronunciation results

In this subchapter, the pronunciation results obtained during testing are analyzed. Whether or not the result is within a native speaker's range is calculated based on the native ranges set within the application, without the additional 20% range extension displayed on the results graph. Additionally, the native speaker's results are compared to the native speaker formant ranges to see whether the ranges should be adjusted. If the native speaker ranges currently implemented in the application are very different from the average results of native subjects, the native speaker formant mean values used in the application, which were obtained from continuous speech, are compared to native speaker formant mean values for isolated speech to see whether the values in the application should be adjusted [16]. The formant mean values for isolated speech are from [20].

The mean results for female test subjects are seen in Table 1. The table contains the mean values of the results of both native speaking female test subjects and female test subjects with other first languages, along with formant values for continuous and isolated speech. For continuous speech, the formant value's standard deviation is noted after the "+-". For isolated speech, the standard deviations weren't available in the source material.

In the case of native Estonian speaking females, the results for the first formant values of /õ/ were often lower than the supposed lowest value for a native female speaker, with two subjects out of 8 not achieving a single result that was within the native speaker's range for the first formant. In many cases, the results for the first formant of /i/ were lower than the supposed lowest value for a native female speaker, however, all but one subject achieved at least one result within the native female speaker's range for /i/. All 8 subjects achieved two or more results within the native female speaker's range for the /a/, /e/, /o/, /ä/ and /u/ vowels and one or more results within the native female speaker's range for /ü/. With these results for female native speakers, with all mean values other than the mean value for the average results for the second formant of /ü/ being within range, the adjustment of native female formant ranges does not seem necessary.

Table 1. Testing mean results for female subjects.

| Vowel | Continuous speech | | Isolated speech | | Native subjects mean | | Non-native subjects mean | |
|-------|-------------------|----------------|-----------------|--------|----------------------|--------|--------------------------|--------|
| | F1, Hz | F2, Hz | F1, Hz | F2, Hz | F1, Hz | F2, Hz | F1, Hz | F2, Hz |
| a | 700 +- 103 | 1362 +- 215 | 810 | 1121 | 714 | 1355 | 735 | 1305 |
| e | 523 +- 72 | 2187 +- 252 | 341 | 2636 | 487 | 2245 | 488 | 2172 |
| i | 380 +- 59 | 2618 +- 168 | 263 | 2797 | 332 | 2615 | 345 | 2598 |
| o | 509 +- 63 | 1046 +- 184 | 435 | 778 | 507 | 1070 | 498 | 1161 |
| u | 408 +- 66 | 1001 +- 219 | 295 | 615 | 395 | 1104 | 383 | 1118 |
| ö | 474 +- 56 | 1447 +- 152 | 347 | 1335 | 422 | 1397 | 442 | 1385 |
| ä | 762 +- 103 | 1803 +- 175 | 855 | 1700 | 824 | 1828 | 803 | 1824 |
| ö | 482 +- 39 | 1853 +- 119 | 385 | 1891 | 473 | 1862 | 482 | 1896 |
| ü | 390 +- 62 | 1903 +- 207 | 253 | 2159 | 347 | 1982 | 403 | 1964 |

The two foreign first language speaking female test subjects performed very well overall. The subjects achieved all three results within a native female speaker's range for /a/, /u/, /ä/ and /ü/. One of the subjects achieved all three results within a native female speaker's range for all vowels except for /ö/, where they achieved two results within a native female speaker's range. The other subject managed at least one result for each vowel that was within a native female speaker's range.

The average results for male test subjects are seen in Table 2. The table contains the mean values of the results of both native speaking male test subjects and male test subjects with other first languages, along with formant values for continuous and isolated speech. For continuous speech, the formant value's standard deviation is noted after the "+-". For isolated speech, the standard deviations weren't available in the source material.

Native Estonian speaking males had particular difficulty achieving a result within the native male speaker's range for /õ/ due to the first formant value, with only 2 out of 10 subjects achieving a result within the native male speaker's range. Another problematic vowel was /i/, with only 3 out of 10 subjects achieving a result within the native speaker's range. 5 subjects achieved a result within the native speaker's range for /ü/. All but one subject achieved one or more results within the native speaker's range for /ä/.

In native Estonian speaking males, the second formant for /a/ seems to have caused difficulty for the test subjects to achieve a result within the native speaker's range, as the mean value of the results is higher than the supposed highest value for the second formant of /a/. Since the mean value of results for the second formant is not closer to the second formant value of isolated speech and is actually further away, the application should be changed to instead consider the mean value of the results for the second formant as the new mean value. For /i/, /õ/ and /ü/, the mean value of the results for the first formant was lower than the supposed lowest value for a native speaker and is closer to the first formant value of isolated speech, so for those vowels, the application should be changed to instead consider the first formant value for isolated speech as the mean value. For /u/, the mean value of the results for the first formant was lower than the supposed lowest value for a native speaker, however, the mean value of the results is almost the mean value of the first formant values for isolated and continuous speech, so the application should instead consider the mean found during testing as the new mean value for the first formant for /u/. For /i/, the mean value of the results for the second formant was higher than the supposed highest value for a native speaker and is closer to the mean value of the second formant value of isolated speech, therefore, the second formant value of isolated speech should be used as the new mean value.

Table 2. Testing mean results for male subjects.

| Vowel | Continuous speech | | Isolated speech | | Native subjects mean | | Non-native subjects mean | |
|-------|-------------------|----------------|-----------------|--------|----------------------|--------|--------------------------|--------|
| | F1, Hz | F2, Hz | F1, Hz | F2, Hz | F1, Hz | F2, Hz | F1, Hz | F2, Hz |
| a | 586 +- 61 | 1111 +- 105 | 666 | 1001 | 598 | 1222 | 597 | 1239 |
| e | 446 +- 51 | 1760 +- 169 | 381 | 2074 | 421 | 1837 | 426 | 1825 |
| i | 314 +- 37 | 2084 +- 122 | 239 | 2279 | 251 | 2211 | 291 | 2200 |
| o | 479 +- 41 | 929 +- 164 | 454 | 799 | 450 | 998 | 485 | 1053 |
| u | 378 +- 42 | 844 +- 161 | 304 | 666 | 335 | 930 | 323 | 993 |
| õ | 439 +- 33 | 1248 +- 94 | 361 | 1225 | 376 | 1262 | 388 | 1284 |
| ä | 578 +- 70 | 1507 +- 145 | 684 | 1575 | 624 | 1523 | 609 | 1532 |
| ö | 421 +- 30 | 1536 +- 71 | 386 | 1613 | 411 | 1526 | 427 | 1430 |
| ü | 317 +- 29 | 1667 +- 92 | 255 | 1813 | 271 | 1648 | 283 | 1557 |

Foreign first language speaking male test subjects seemed to have the most difficulty with the /ü/ and /õ/ vowels, with only one subject being able to attain a result within the range of a native speaker for either vowel. For /ü/, if the result was outside the range of a native speaker, both formants' values were usually lower than the lowest value in the range of a native speaker. For /õ/, if the result was outside the range of a native speaker, the first formant's value was mostly below the lowest value in a native speaker's range and the second formant's value was mostly above the highest value in a native speaker's range. The /ä/ vowel was pronounced the most successfully, with 5 subjects out of 6 achieving all three results within the range of a native speaker. The /e/ vowel was a close second, with 5 subjects out of 6 achieving two or more results within the range of a native speaker.

Analyzing the results for native male speakers, it can be seen that the formant mean values for males should be adjusted for some vowels. Formant values for female speakers can stay the same.

4.3 Test subject feedback

All test subjects filled out the feedback form described in chapter 4. The current subchapter summarizes the feedback to see what the subjects' opinion of the application is as a whole, how they rated the exercises in the application, how intuitive the subjects found the feedback provided in the application and what could be improved.

The feedback was mostly very positive. To the question of what they thought of the application, subjects wrote that they found the application to be an interesting and useful approach to learning to pronounce Estonian vowels, and that the idea of the application will help improve language skills and increase interest in learning Estonian. A few subjects did mention that the application's design could be improved.

The pronunciation exercise was rated an average of 4.5 out of 5 for being useful for learning to pronounce Estonian vowels. It was noted that the pronunciation exercise clearly shows the correctness of one's pronunciation and that the results graph is a good indication of what the target is for the learner's pronunciation. One test subject noted that it is quite useful that they could see the difference between native and non-native pronunciation graphically. The animations were mentioned multiple times as being a good guide to pronouncing the vowel, however, on two occasions, it was also said that a recording of the vowel could be played as the learner views the animation.

The listening exercise was rated an average of 4.4 out of 5 for being useful for learning to differentiate Estonian quantity degrees. The comments for the listening exercise were mostly positive with some more negative remarks. On one hand, test subjects noted that the exercise helps new learners to learn to tell the difference between Estonian quantity degrees, saying that the examples chosen for questions were clear. On the other hand, one subject said that the answering process was unintuitive, and suggested that the answering could be done as some sort of a drag-and-drop. More than one subject also suggested that the words being pronounced in each question could also be shown as

text, and one subject suggested that the exercise could include instructions as to which vowel they should pay attention to, as the words in the questions had multiple vowels in them (for instance, one question had the words *koli*, *kooli* and *kooli*, and it seems that the subject thought that they should perhaps consider the /i/ vowel when answering).

The visual feedback provided in the application was overall found easy to interpret, with the exception of one subject saying that it was “not good” and a few subjects mentioning issues with the results graph. One test subject said that sometimes, their result would be outside the target range on the graph, although the message would tell them that they were within the range of a native speaker. One subject brought out that they didn’t initially understand that the graph is an indication of the position of their tongue. One subject also noted that visually, it wasn’t clear what they had to do in the listening exercise.

Multiple suggestions were made as to how the application could be improved. The most frequent suggestion was that the interface could be improved, particularly to be more modern and more stylish. The crashes and bugs should be fixed. The listening exercise could be improved to have a better answering system and/or better guidance for how to answer. The listening exercise could shuffle the questions so that the user might better learn to tell the difference between the quantity degrees, instead of learning the correct sequences for answers. The native pronunciations could also feature a female voice. The application could show the user their overall progress, meaning that the user could be shown some kind of identifier for which exercises they’ve completed, as well as having numbers accompany the result dots on the graph in the pronunciation exercises. The instructional texts were said to be too long by one subject, who also brought out that long text should be considered a bad practice. The instructional texts could be brought out more to the user, as some subjects noted that they did not notice many of the explanatory messages. The application should also feature more exercises in general. One subject pointed out that the application should feature more challenging listening exercises.

On the question of whether they would personally use an application like this when learning a language, almost all test subjects said that they would, with a few exceptions. One subject brought out that for them to definitely use the application when learning a language, the application would have to be usable with any random words which feature

the vowel that they are practicing in their current exercise, and that in the current state, the exercises have too little content to be enough for a language learner. Two subjects wrote that they are not sure whether or not they would use the application, with one of them saying that they would actually rather feel the need for an application for learning consonant pronunciation.

Taking all of the test subject feedback into account, it is clear that the application has plenty of room for improvement. Particular trouble spots seem to be the crashes and bugs that occur, along with making instructions clearer and explanatory messages stand out more. In the future, more exercises should definitely be added, along with native female pronunciation examples. The listening exercise could be improved to have the questions show up in a random order each time the user starts the exercise. The user interface could overall be more modern and easy to use.

4.4 Revisions based on test results and test subject feedback

Testing and subject feedback revealed multiple improvements that could be made to the application. Based on some of the suggestions made by users and some of the author's remarks during testing, the application was revised and improved. The improvements to be made were chosen according to what the author thought to be the most important or most useful to the learner. Some suggestions and ideas, while useful, will have to be implemented in the future due to time and resource constraints.

During testing, the author realized that the user registration process had no checking for special characters in neither the username nor the password. This could cause issues with file and directory creation, rendering the application, at times, unusable. User registration was revised to only allow letters and numbers in usernames and passwords. If the user tries to register an account with special characters, the registration will not be completed and the user will receive an error message stating that neither the username nor the password can contain special characters.

Both testing and test subject feedback revealed that the microphone volume adjustment was confusing and hard for the subject to follow. As the correct volume threshold is crucial to detecting and evaluating the user's vowel pronunciation, the adjustment

screen was revised and is now as seen in Figure 15. The microphone volume adjustment screen was changed to have a more instructive explanation text explaining the on-screen graph and what it represents, along with what the slider underneath it represents and what the user should do with it. The instructions now also guide the user to click the button at the bottom of the screen after they are finished with the microphone set-up. Taking one test subject's comment into consideration, the text of the save button now reads "Next" for the screen of the initial microphone volume set-up, but still reads "Save" for when the user goes to readjust their microphone volume threshold. Although one test subject stated that the instructional texts in the application were already perhaps too long and that long text should be considered bad practice and avoided if possible, the microphone volume adjustment screen received longer instructions, firstly, to make sure that the user understands what they have to do and secondly, to somewhat space out the information, making it easier to understand. The red line representing the highest recorded volume was removed as it has no relevance to the user and most likely only served to distract the user from the instructions.

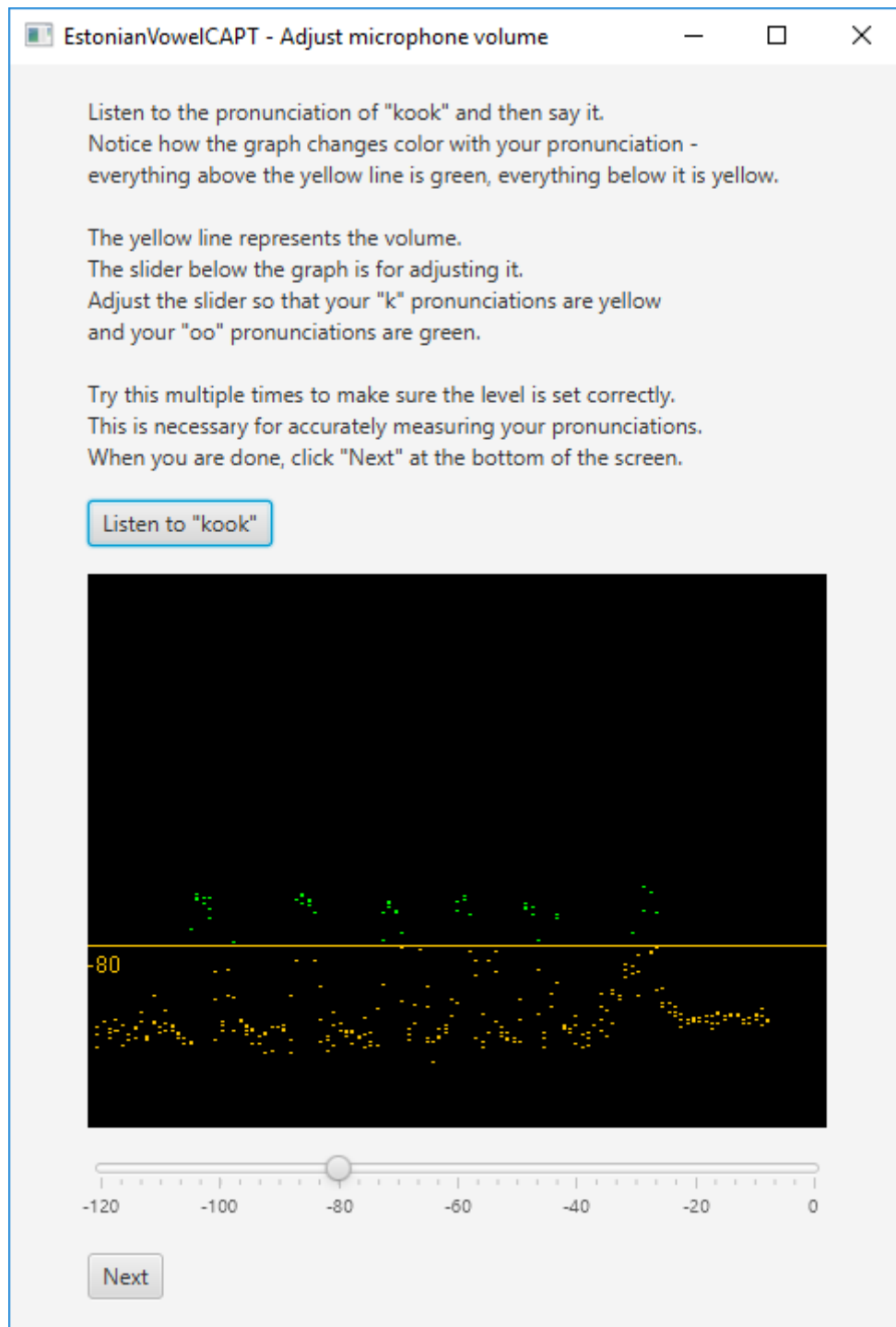


Figure 15. Revised microphone volume adjustment screen.

The native speaker ranges of some vowels were adjusted for males where it was deemed necessary, based on the analysis of pronunciation results in subchapter 4.2. When calculating whether a pronunciation result is within the native speaker's range, the application now considers the additional 20% range extension as part of the native speaker's range. This additional range was previously shown on the results graph of the pronunciation exercise, but was not considered when calculating the result, which

caused a lot of confusion and was a glaring inconsistency in feedback. This extension of the vowel range is more forgiving to new learners, and the extension was added to calculations to create consistency between the calculations and the results graph. In the text shown when the user moves their cursor onto the results graph, a line disclosing the widening of the vowel ranges was removed. The widening of vowel ranges is not necessary knowledge for the user.

The pronunciation exercise's main problem had to do with the subjects not understanding how to record themselves and with the exercise missing visualization of pronunciation result progress. The revised pronunciation exercise is seen in Figure 16. To better explain the recording process, instructions were added below the record button explaining that the user must click the record button, then pronounce the word and then click the same button to stop recording themselves and receive their result. The color of the record button was changed to red to draw more attention to it. To track progress, the red dots on the results graph were improved to have numbers and letters on them so the user can see which result they received last. The results are first numbered 1-9, then lettered A-Z and after that, if any user should reach it, simply labelled with three dots. The lettering after numbers 1-9 was added due to the dots being too small to contain double-digit numbers. The text shown when the user moves their cursor onto the results graph was changed to include a sentence explaining that the graph is where their results would appear. The issue with the native pronunciation occasionally not playing was also fixed. The vowel ranges were revised to accurately reflect first and second formant values, meaning that the ranges are now represented with two separate radiuses, one for each formant, instead of having one radius which was averaged from the standard deviation values for each formant. The previous implementation caused severe issues with results being shown as though they were within the target range, although in reality they were not, or the opposite – being shown as though they were outside the target range, although in reality the subject had been within range.

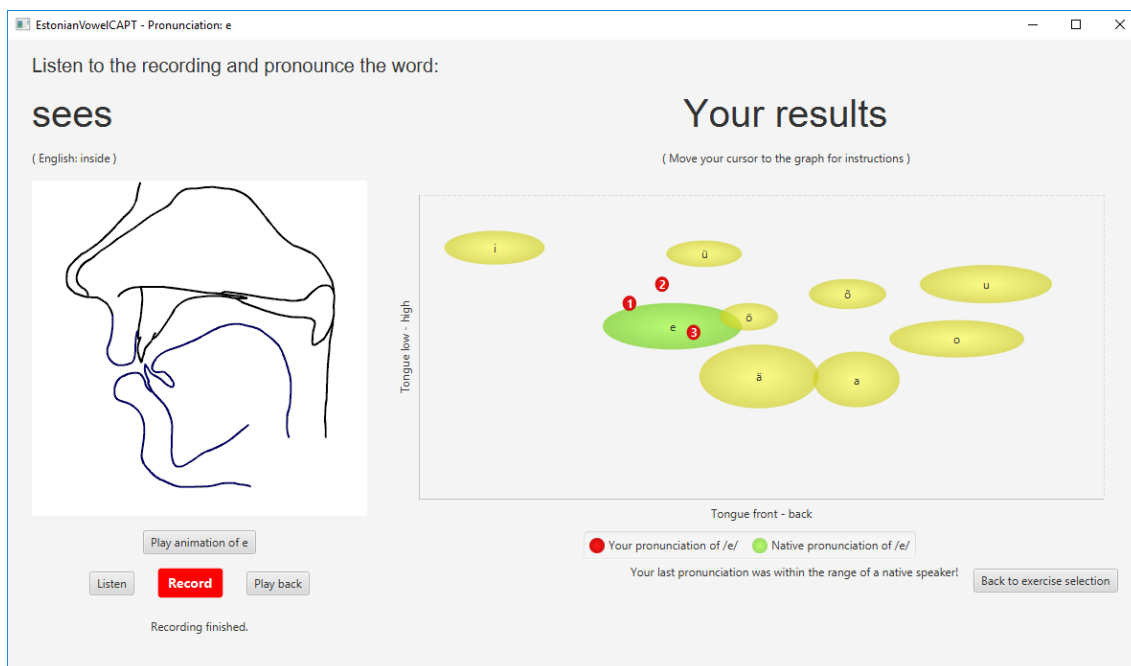


Figure 16. Revised pronunciation exercise screen.

The listening exercise was very confusing to the majority of test subjects. The answering system was quite complicated and not intuitive at all, requiring the user to enter a sequence with checkboxes to signify where they heard the short, long and overlong vowel. There was also a bug with one of the questions where the correct answer would result in an incorrect result and one specific incorrect answer would result in a correct result. To simplify the exercise, it was decided that instead of comparing three words at once, the user would hear one word at a time and then would have to choose which quantity degree the vowel they heard was in. The revised listening exercise is seen in Figure 17. The instructions for the exercise now also include an explanation that the user must choose an answer based on the first vowel they hear in the word. When the user answers incorrectly, the pop-up window now states the correct quantity degree. The spelling of the word was not added to the exercise, however, as the purpose of a listening exercise is to listen and because the spelling is a hint. For instance, for the words *koli*, *kooli* and *kooli* (first, second and third degree of quantity), the single /o/ is a giveaway that the word has a short vowel, while the double /oo/ is either long or overlong. The bug in the second question was fixed when the answering system was altered.

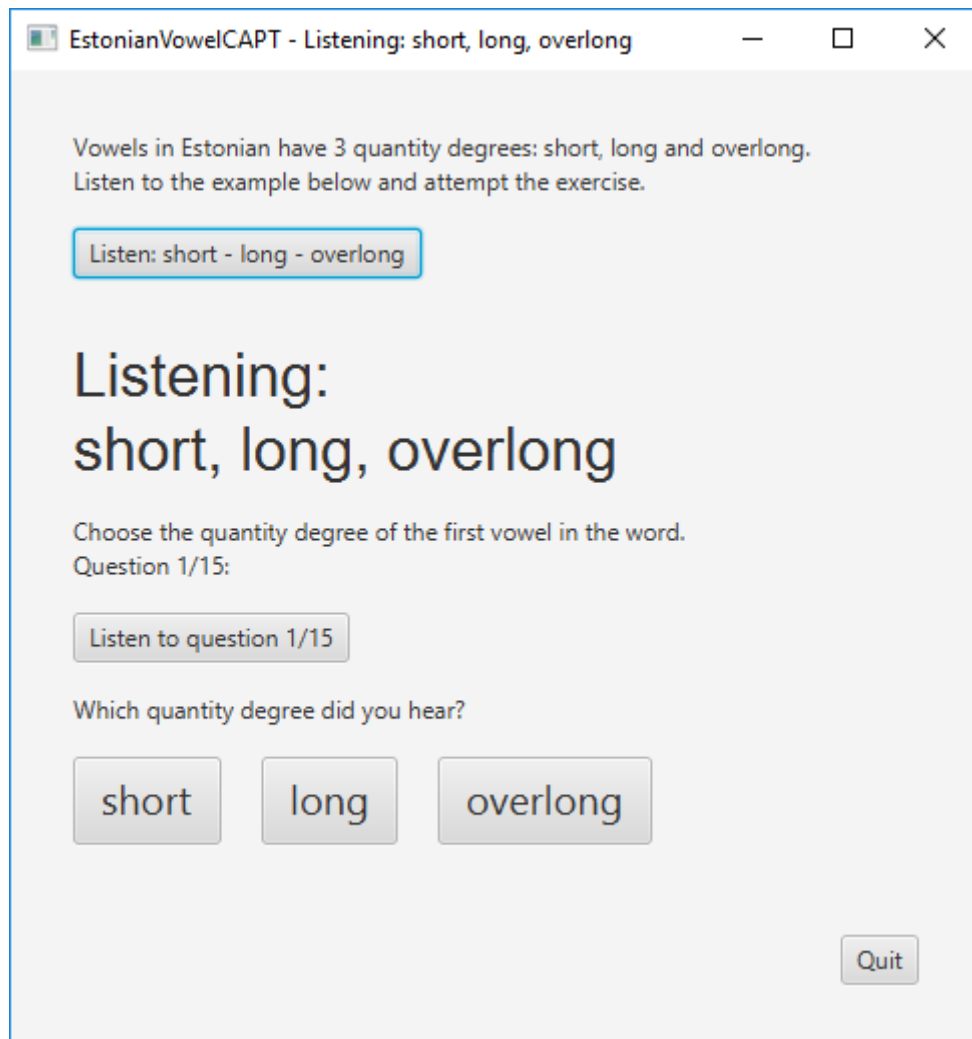


Figure 17. Revised listening exercise screen.

Although multiple revisions were made to the application, more remain. Some of the ideas would have required too much time and resources to complete at the present time. These revisions are listed as future improvements in chapter 5.

5 Assessment and future improvements

All in all, the application can be considered successful. The goal of this thesis was not to create a complete CAPT product that learners could start using immediately, but to create a prototype to evaluate whether the chosen methods and solutions would work.

The application has vowel detection and gives feedback about a learner's pronunciation. The application has instructions to enable the user to use it independently, or at least that is the hope after the listening exercise and microphone volume adjustment were revised. The user is able to create an account and complete pronunciation exercises for each Estonian vowel as well as a listening exercise for Estonian vowel quantity degrees. The formant graph chosen for feedback was easy for most users to understand, as the green target on the graph is simple and to the point.

Although the application can be considered successful, the application has aspects that could be improved in the future. While perhaps not detrimental to the application's functionality, they would be clear improvements and would aid learners in understanding and obtaining the correct Estonian vowel pronunciation.

Firstly, a pronunciation exercise for quantity degrees could be added. Although currently not possible to implement with a sufficient degree of accuracy due to the vowel detection system not being accurate enough to record the entire durations of vowel segments, quantity degrees are an important part of the Estonian language. A pronunciation exercise would, in the author's opinion, be a better teaching tool than an exercise where the user, for example, just listens to a native speaker's pronunciation to determine the quantity degree, although this exercise is useful in the beginning. This could be achieved with a more robust vowel detection system. It is unfortunate that the attempts in this thesis to implement pitch detection ultimately failed, however, it may still be possible.

Secondly, a native female speaker's recordings should be added to the application to add variety and to create a better comparison for female users. A native female speaker's voice could also be used to create native pronunciations for the learner to

listen to when the animation of the mouth and tongue is played. Of course, these pronunciations should also include male recordings for male users.

Thirdly, the user interface design of the application could still be improved, among other things, to be more modern. This was currently not done as the author is not a designer and recruiting a designer would have taken too much time and resources. And although progress is displayed for the pronunciation attempts on the results graph of the pronunciation exercise, no progress is saved permanently, apart from the volume threshold value. Progress could be saved on which exercises the user has completed, whether they have reached a native speaker's pronunciation range, what their best result is in a listening exercise and so on.

Moreover, more listening exercises should be added, as the application's content currently runs out fairly fast. The pronunciation exercises could also be improved to include multiple consonant-vowel-consonant words for each vowel so the user learns vocabulary as well as improves their pronunciation. However, this will have to be done in collaboration with a language teacher. The application should then also have a system in place for language teachers to add new exercises without the help of a software developer.

Lastly, the crashes which would sometimes occur in the pronunciation exercise should be investigated to see if they could be removed completely. The author's guess is that the laptop that was used for development and testing could be too weak for the application, however, no specific reason for the crashes has been found thus far.

6 Summary

The application developed during this thesis can be considered successful. The application performs the functions specified in the introduction, allowing the user to create an account that they can log back in to and use to keep track of their personal microphone volume settings, having pronunciation exercises where the user can record their pronunciation and receive feedback on how accurate their pronunciation was, and a listening exercise where the user can learn to differentiate between the vowel quantity degrees of Estonian.

The application was tested with a total of 26 test subjects, 8 of whom had first languages that were not Estonian. After testing was conducted, the subjects' results from pronunciation exercises were used to evaluate whether the formant values used in the application should be modified, and for males, the formant values for some vowels were changed. The feedback from the test subjects was analyzed to see what the application could improve on, and some of the more important or useful ideas were added to the application. Some of the ideas, while useful, will have to be implemented in the future due to time and resource constraints.

According to the test subjects, the feedback in the application was mostly easy to understand. The feedback in the pronunciation exercise was improved to be more precise. The instructions in both the listening exercise and microphone volume adjustment were revised to be more easy for the user to understand. The answering system and the questions in the listening exercise were simplified significantly.

In the future, the application could include more exercises, such as a pronunciation exercise for vowel quantity degrees, which would require a more robust vowel detection system to implement. A native female speaker's pronunciations should be added to the application to provide a better comparison for female learners. With the help of a designer, the interface of the application could be improved to be more modern and pleasant.

References

- [1] R. Hincks, "Speech technologies for pronunciation feedback and evaluation," *ReCALL*, vol. 15, no. 1, pp. 3-20, 2003.
- [2] J. Levis, "Computer Technology In Teaching And Researching Pronunciation," *Annual Review of Applied Linguistics*, vol. 27, pp. 184-202, 2007.
- [3] D. M. Chun, "Come Ride the Wave: But Where Is It Taking Us?," *CALICO Journal*, vol. 24, no. 2, pp. 239-252, 2007.
- [4] D. M. Hardison, "Generalization Of Computer-Assisted Prosody Training: Quantitative And Qualitative Findings," *Language Learning & Technology*, vol. 8, no. 1, pp. 34-52, 2004.
- [5] D. Coniam, "Technology as an Awareness-Raising Tool for Sensitizing Teachers to Features of Stress and Rhythm in English," *Language Awareness*, vol. 11, no. 1, pp. 30-42, 2002.
- [6] P. Carroll, J. Trouvain and F. Zimmerer, "A Visual Feedback Tool for German Vowel Production," in *Elektronische Sprachsignalverarbeitung*, Eichstätt, 2015.
- [7] S. Wood, "What are formants?," 15 January 2005. [Online]. Available: <http://person2.sol.lu.se/SidneyWood/praat/whatform.html>. [Accessed 24 March 2018].
- [8] "Chapter 1 Introduction," [Online]. Available: http://ec-concord.ied.edu.hk/phonetics_and_phonology/wordpress/learning_website/chapter_2_vowels_new.htm. [Accessed 24 March 2018].
- [9] M. D. Carey, "CALL Visual Feedback for Pronunciation of Vowels: Kay Sona-Match," *CALICO Journal*, vol. 21, no. 3, pp. 571-601, 2004.
- [10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," [Online]. Available: <http://www.fon.hum.uva.nl/praat/>. [Accessed 24 March 2018].
- [11] H. Ding, O. Jokisch and R. Hoffmann, "An Audiovisual Feedback System for Pronunciation Tutoring – Mandarin Chinese Learners of German," in *Cognitive Behavioural Systems*, Dresden, Springer-Verlag Berlin Heidelberg, 2011, pp. 191-197.
- [12] audEERING GmbH, "audEERING | Intelligent Audio Engineering - openSMILE," [Online]. Available: <https://audeering.com/technology/opensmile/>. [Accessed 26 April 2018].
- [13] "sikoried/jstc: Automatically exported from code.google.com/p/jstc," [Online]. Available: <https://github.com/sikoried/jstc>. [Accessed 19 April 2018].
- [14] "5. The Voice," 24 11 2014. [Online]. Available: <http://msp.ucsd.edu/syllabi/170.13f/course-notes/node5.html>. [Accessed 23 April 2018].
- [15] "JorenSix/TarsosDSP: A Real-Time Audio Processing Framework In Java," [Online]. Available: <https://github.com/JorenSix/TarsosDSP>. [Accessed 19 April 2018].

- 2018].
- [16] E. Meister and L. Meister, "(Work in progress) Production of Estonian vowels by Finnish speakers," Tallinn, 2018.
 - [17] Oracle, "What Is JavaFX? | JavaFX 2 Tutorials and Documentation," Oracle, [Online]. Available: <https://docs.oracle.com/javafx/2/overview/jfxpub-overview.htm>. [Accessed 1 May 2018].
 - [18] G. Liiv, "On qualitative features of Estonian stressed monophthongs of three phonological degrees of length," *Eesti NSV Teaduste Akadeemia Toimetised*, vol. 10, no. 1, pp. 41-66, 1961.
 - [19] P. Penjam, "Eesti keele hääldamise ja kõnetaju interaktiivne õpivahend "Õpime hääldama!,"" Tallinn, 2005.
 - [20] A. Eek, *Eesti keele foneetika I*, Tallinn: Tallinna Tehnikaülikooli Kirjastus, 2008.

Appendix 1 – GitHub link to the application

<https://github.com/martinvaljaots/Estonian-vowel-CAPT>