Can audio-visual instructions help learners improve their articulation?
- an ultrasound study of short term changes

Olov Engwall

1 Centre for Speech Technology, CSC, KTH, Stockholm, Sweden
engwall@kth.se

Abstract

This paper describes how seven French subjects change their pronunciation and articulation when practising Swedish words with a computer-animated virtual teacher. The teacher gives feedback on the user’s pronunciation with audiovisual instructions suggesting how the articulation should be changed. A wizard-of-Oz set-up was used for the training session, in which a human listener choose the adequate pre-generated feedback based on the user’s pronunciation. The subjects’ change of the articulation was monitored during the practice session with a hand-held ultrasound probe. The perceptual analysis indicates that the subjects improved their pronunciation during the training and the ultrasound measurements suggest that the improvement was made by following the articulatory instructions given by the computer-animated teacher.

Index Terms: talking heads, pronunciation training, computer-assisted language learning, articulatory changes

1. Introduction

Commercial state-of-the-art Computer Assisted Pronunciation Training (CAPT) programs typically merely indicate to the user if a pronunciation error has been made, without any information on the type of error or how the error could be corrected. Using a computer-animated talking head that acts as a language teacher could improve CAPT immensely, for several reasons: 1) A virtual teacher can use computer animations to illustrate how the pronunciation should be made [1] or changed [2]. 2) The interaction with a computer-animated tutor is more similar to that between the teacher and students in the classroom, and the feedback may hence be more effective than if pronunciation scores are used [3]. 3) Users who interact with an animated agent spend more time with the system, think that it performed better and enjoy the interaction more [4, 5].

We are therefore developing a computer-animated articulation tutor, ARTUR [2], who should assist hearing- or language-impaired children and second language learners with their pronunciation of Swedish. The key feature of ARTUR is that audio-visual feedback is given on what articulatory changes the user should make in order to correct pronunciation errors. Sensory feedback on articulatory features, such as the touch between the tongue tip and the palate; and comparisons to words with a similar articulation in the mother tongue are also used.

One important question is if the audio-visual articulatory instructions really can help the users improve their pronunciation, since this type of feedback is unfamiliar. Most language learners are not used to seeing the tongue move or to actively think about the articulatory production, as most feedback in the language classroom focus on the acoustic targets instead.

This paper investigates the short term changes that the users of ARTUR make when they are given audio-visual articulatory feedback on their pronunciation of two Swedish phonemes [i] and [r]. The palatovelar fricative [r] is particularly troublesome and is often mispronounced as one of [i, ə, χ] depending on the student’s first language. The Swedish [r] may also be difficult for speakers of languages where a rhotic [r] is used instead. We focus on the practice of Swedish, but the general scope is valid for other languages as well.

2. ARTUR – An ARticulation TUtoR

The goal of the work on ARTUR is to create a computer-animated language teacher, who presents the user with appropriate pronunciation exercises, detects errors and gives feedback that helps the user improve the pronunciation [2]. ARTUR should automatically detect pronunciation errors from the speech signal, supported by video images of the user’s face [6].

Our current main interest, in previous [2, 3] and this study, is focused on ensuring that the multimodal feedback from ARTUR is as efficient as possible.

2.1. The user interface

The graphical user interface of ARTUR, shown in Fig. 1a), consists of five information windows and one user control frame. The main window displays the face of the virtual teacher Artur and articulatory animations. These show either how the tongue should be moved in the target word or the articulatory changes that the user should attempt. Pre-recorded natural speech and time-aligned articulation movements generated from a text-to-visual-speech synthesizer [7] are used. The two progress windows show how far into the training session the user has come, with one window for the Swedish words and one for translations. In the sub-title window, all instructions spoken by Artur are given in writing. The target word is displayed and the user is prompted to speak when the window’s background colour changes to green. The user control frame allows the user to request that Artur repeats the word either at normal speed or slower, to see the difference between the own and the correct articulation and to listen to the own previous attempt.

The user’s and the correct articulations are synthesized using the KTH models of the face [7] and vocal tract [8]. The tongue model is articulatorily correct since its shape for different phonemes is based on statistical analysis of Magnetic Resonance Imaging data and the articulatory movements are modelled on Electromagnetic Articulography (EMA) measurements, both for the same subject.

The feedback to the user can be [9]: (P) positive for a correct pronunciation. (C) corrective the first time the user mispronounces a word. (A) augmented instructions for repeated errors. (V) vague if the articulatory cause of the mispronunciation...


3. Data acquisition

3.1. The training session

A French-speaking Artur was used to give instructions and feedback. He initiated the training by explaining the task and encouraging the user to request additional help with control buttons. The training then lasted five to ten minutes.

3.2. Subjects & Training material

The subjects were seven speakers of French. One had Tunisian and one Persian as mother tongue, while French was the first language for the remaining five. None had received any previous pronunciation training in Swedish.

The training material was the Swedish words rik [ʁik:] (rich), rak [ʁak:] (straight), kor[kura] (select as), kir [kɪr] (the French aperitif), karikerar [kɑrkɛːrɑr] (making a caricature of), schack [ʃɑːk:] (chess), sjuk [ʃʊk:] (sick), skick [ʃɪkːː] (condition), chock [ʃʊkː] (shock). The words were chosen so that they contained the phonemes [ɾ] and [ɻ] in different vowel contexts (front vs. back, open vs. closed), and at the same time [k], since its palatal closure made temporal alignment between the acoustic and articulatory data possible.

3.3. Acquisition set-up

The acquisition of acoustic and articulatory data was made using a set-up including audio and video recordings, an ultrasound scanner and an electromagnetic tracking system.

The acoustic of each of the subjects’ attempts was recorded separately on computer disc, using the Snack [10] plug-in, with which the recording could be controlled automatically by the ARTUR software.

The articulatory data was collected at around 66 Hz using a Logiq5 ultrasound (US) machine from GE Healthcare. The acquisition frequency depends on the area that is imaged, the depth of penetration and the resolution of the image. These properties were set individually for each subject prior to the training session to ensure that relevant parts of the tongue were captured with the highest possible image quality and frame rate. The resulting image frequency hence varied somewhat for different subjects, due to differences in tongue size and shape.

The transducer was a microconvex 8C probe, producing ultrasound between 5MHz and 9MHz. The transducer probe was held under the chin by the subjects themselves, rather than being attached to a helmet [11] or being fixated together with the subject’s head. It is customary in quantitative speech production measurements with ultrasound to restrict the relative movements between transducer and the head, since the data may be lost or erroneous if the angle between the transducer and the chin is changed. In particular, if the transducer is tilted outside the midsagittal plane, a lateral part of the tongue is measured instead and the data can not be compared to midsagittal images.

A hand-held transducer was used to give the subjects the maximal freedom and create a training situation that was as close to a real practise session as possible. For the current study it is further not crucial to acquire continuous, quantitative data of tongue movements. What is relevant is instead to identify image frames that show qualitative changes in the articulation between the user’s different attempts. It is important, on the other hand, to ensure that the image frames that are selected for analysis can be trusted, i.e., that the user held the transducer correctly in the midsagittal plane, at an adequate angle for these frames. A tracking system was therefore used to monitor the transducer’s orientation relative the subject’s head.

The tracking system was made with an Aurora miniature electromagnetic system. The system consists of a magnetic field generator, a system control unit and miniature coils. Two coils were fastened to the ultrasound transducer and one to a point behind the subject’s ear that could be assumed to only display rigid movements, even when the subject spoke. The system of three coils permits to monitor the transducer’s relative movements in three spatial coordinates and two angles.
4. Data analysis

The data was analyzed using the acoustic recordings as starting point. For each subject and each training word, the first and last attempts on the target phoneme ([r] or [l]) were identified. The tongue contour was then extracted manually from corresponding frames in the ultrasound data using ImageJ [12].

The aim of the analysis is to compare the tongue contours from the first and last attempts to determine if the subjects changed their articulation and if the changes are consistent with Artur’s instructions. The delay and transfer of articulatory changes are also of interest. Delay refers to if the subjects change the articulation directly following a feedback instruction or if several attempts are needed. Transfer refers to if the subjects use the new articulation learned in one word in following training words with different phonetic contexts, either directly or after fewer attempts than for the previous word.

Since the ultrasound technique relies on the reflection of ultrasound waves at air-tissue boundaries, the tongue tip will not be visible if it raised from the mouth floor, as for [r]. The actual shape of the tip can hence not be measured for [r], but a raised tongue tip is nevertheless signalled by the angle of the tongue body posterior to the tip, as shown in Fig. 2. It is hence still possible to investigate qualitative changes from [r] to [r].

Table 1 summarizes the subjects’ changes for each training word. An empty slot signifies that the target phoneme was reached acoustically, whereas phonetic signs indicate phonemes that replaced the target. * indicates that the erroneous pronunciation was judged to be unchanged throughout the attempts.

5. Results

No formal subjective evaluation of the users impression was made, since this has been the scope of previous studies [2, 3]. Several users nevertheless spontaneously commented that they believed that they were able to change the articulation according to the feedback. They further stated that the articulatory instructions with the most important features highlighted by animations were more helpful than animations of the entire training word, since the tongue movements were too fast.

5.1. Acoustic change

The recordings of all subject attempts were analyzed to determine how the subject changed the pronunciation.

Table 1 summarizes the subjects’ changes for each training word, in terms of number of attempts, and if the target was reached, or which phoneme that was produced instead. There are several reasons that can explain why the training continued with the next word, even if the target was not reached: 1) a maximum number of attempts had been made (the limit was imposed to avoid repetitions of the same feedback), 2) the user did not make any improvements compared to the previous attempts and it was judged that it would not be fruitful to continue prac-

Figure 2: The midsagittal tongue contour is extracted manually from the ultrasound images. The tongue tip (to the left) is lifted for an [r].
5.2. Articulatory change

For each subject and training word, the frames corresponding to the practised sound were analyzed in the first and last attempts, to monitor the articulatory changes made.

The analysis was qualitative and focused on intra-subject changes, as exemplified in Fig. 3. Since the probe orientation was not fixed and no reference frame, such as the palate, was defined, no quantitative measures were made. The ultrasound image quality further differed substantially. While the images were quite clear for some subjects, they could sometimes be analyzed only with difficulty for others, either due to probe orientation or subject anatomy or articulation. Since the subjects differed in anatomy and the probe orientation was not fixed, it is impossible to make any inter-subject comparisons, but it is the intra-subject changes that are of interest in this study.

For ‘r’, the subjects made a clear articulatory change, to a front [r] with lifted tongue tip. and several image sequences shows the tongue tip vibration of [r]. Figs. 3a-c) show the change made by three subjects in the articulation of ‘r’ when followed by [i] or [e]. This articulation was transferred to other vowel context. For [i], the changes were not as prototypic and differed between subjects. The most clearly observable articulatory change was that when subjects changed the articulation from a palatal plosive to a fricative, as shown in Fig. 3d). All the subjects’ articulation of [i] approached the correct one, even if they were more or less successful in reaching it.

6. Conclusions

The acoustic analysis showed that the subjects, to a large extent, were able to change the pronunciation of ‘r’ from the initial [r] to an [r]. The articulatory analysis showed that this change in pronunciation was indeed the effect of changing the articulation to that indicated by the virtual teacher.

For [i], the change was less clear, both acoustically and in the articulation. The pronunciation was nevertheless judged to have been improved by following the articulatory instructions.

This small scale short-term study hence shows that audiovisual articulatory instructions are beneficial. The subjects do improve their pronunciation and they do it by changing the articulation according to the instructions.

To prove the long-term efficiency and retention certainly requires a larger study over several training sessions, but these results are a positive indication that subjects are able to transfer articulatory instructions to the own articulation.

7. Acknowledgements

The user test in this study was performed while the author was a visiting researcher at the Speech group at LORIA, INRIA, France, as part of the collaboration within the ASPI (Audiovisual speech inversion) project. The leader of the LORIA speech group, Yves Laprie, is thankfully acknowledged for this opportunity. The data acquisition was made using equipment from the ASPI project, financially supported of the Future and Emerging Technologies (FET) programme within the Sixth Framework Programme for Research of the European Commission, under FET-Open contract no. 021324.

8. References


