Using ultrasound tongue imaging to improve L2 English pronunciation in Dutch students

Lisanne de Jong[1], Teja Rebernik[1], Sonja Vaziri[1], Martijn Wieling[1,2]

[1]*University of Groningen, The Netherlands*
[2]*Haskins Laboratories, United States of America*
lisannemereldejong@gmail.com, t.rebernik@rug.nl, s.vaziri@student.rug.nl, m.b.wieling@rug.nl

Abstract

This study set out to investigate whether visual feedback using ultrasound tongue imaging could help Dutch learners to improve their pronunciation of the English sound contrasts /æ/-/ɛ/ and /k/-/g/. Thirty-seven high school students took part in our experiment which consisted of a perception task and a production task (pre-test, phonetic training session and post-test). During the training session, half of the group received UTI visual feedback, whereas the other half only received auditory feedback. Based on subjective ratings by native speakers of English, our results showed that the pronunciation of the target sounds significantly improved after the training session, but that there was no significant difference in improvement between the group that received visual feedback and the group that did not. We furthermore did not find any statistically significant differences in the actual pronunciation of the target vowel contrast, measured acoustically.

Keywords: ultrasound tongue imaging; visual feedback; L2 speech

1. Introduction

The pronunciation of non-native sounds is typically considered one of the most difficult skills to master when learning a second language (L2). Even though it has been suggested that receiving explicit training on phonetic differences between similar sounds is correlated with performance on L2 pronunciation (e.g., Bongaerts, 1999), pronunciation still tends to receive little attention in the language learning classroom. Recently, interest in the application of speech-production-based technologies in pronunciation training for second language learners has increased. Similar to the application in clinical settings where studies show that the visualization of articulators (such as the tongue) can facilitate speakers in producing target sounds (e.g., Preston, Brick & Landi, 2013), this type of bio-visual feedback might also aid L2 learners in producing non-native sounds. Ultrasound tongue imaging (UTI) is a non-invasive technique that can be used to visualize tongue movements in a way that is relatively easy to interpret. Up until now, several studies have shown beneficial effects of a pronunciation training using UTI-based visual feedback on the production of non-native sounds (e.g., Ouni, 2014; Cleland *et al*., 2015).

Building onto this research, our study investigated whether a short training using UTI could improve Dutch high school students' pronunciation of two English target contrasts, /æ/-/ɛ/ (e.g., in *bat* - *bet*) and /k/-/g/ (e.g., in *pick* - *pig*). These two contrasts were chosen as Dutch learners of English tend to find them difficult (Broersma, 2005; Cutler *et al*., 2004). In our study, we compared pre- and post-test recordings of a group of students that received UTI-based visual feedback to those of another group that only received auditory feedback. In line with previous studies, it was hypothesized that students who received visual feedback would show more improvement in the pronunciation of target sounds after the training than the group that was not exposed to the visual information provided by the ultrasound images.

We were also interested in assessing whether the potential beneficial effect of providing UTI feedback would be restricted to the sound contrast /æ/-/ɛ/. Only for this sound contrast, the difference is related to the tongue shape. For the sound contrast /k/-/g/, the difference is related to voicing and thus UTI feedback was expected to be less informative.

2. Methodology

2.1. Participants

The data was collected at the RSG Ter Apel, a high school located in Ter Apel, a village in the north of the Netherlands. The 37 participants (24 female, 13 male) consisted of first- and second-year students with ages ranging between 12 and 15 (with an average age of 12.7). Since participants were under the age of 18, parent(s)/guardian(s) of the participants were asked to provide their written consent ahead of the experiment and to provide background information about their child through a survey. These background questions concerned demographic factors as well as specific questions about language learning, such as the student's motivation to learn English. A requirement for participation was that participants did not have any other native language(s) other than Dutch. Considering that English is a mandatory subject in high schools in the Netherlands and English lessons are also part of most primary school curricula, participants had been learning English for at least a few years. No severe language, speech or hearing disorders that could influence linguistic performance were reported. Students received 10 euros for participating. Ethical approval for the study was obtained through the University of Groningen, Faculty of Arts' Central Ethical Testing Organization.

2.2. Materials

The stimuli in this experiment consisted of minimal pairs containing the target contrasts /æ/-/ɛ/ and /k/-/g/. The vowel contrast was chosen as Dutch learners of English tend to perceive both vowels as an exemplar of the Dutch /ɛ/ and as a result typically assimilate the two English sounds (Broersma, 2005; Wester *et al.*, 2007). The /k/-/g/ contrast is another example of two English sounds that Dutch speakers tend to merge to one sound. Since /g/ is not present in the Dutch language (excluding loan words), Dutch speakers tend to map both sounds to /k/ (Cutler *et al*., 2004).

The overview of the 16 minimal pairs used in the study can be found in Table 1. All words contained one syllable. We tried to control for phonetic environment as much as possible, for instance by having the target consonant in the /k/-/g/ items appear in word-initial position for half of the items and in word-final position for the other half of the items. All 32 target

words appeared in the pre- and post-test, but only half of the minimal pairs (indicated with an asterisk in Table 1) were part of the training session. By only training some of the items, we wanted to find out whether a possible beneficial effect of the training session was limited to the practiced items, or whether it would be generalized over different phonetic contexts.

Table 1: *Minimal pairs used in the study. Words with an asterisk (*) were part of the training session.*

/æ/	/ɛ/
band*	bend*
pat*	pet*
axe*	ex*
tan*	ten*
pan	pen
sat	set
bat	bet
and	end

/k/	/g/
crease*	grease*
came*	game*
coat	goat
kill	gill
pick*	pig*
clock*	clog*
wick	wig
buck	bug

Prior to the experiment, audio and UTI recordings of the target words were made in the Articulate Assistant Advanced software (Articulate Instruments Ltd) by two adult native speakers of American English (one male, one female). Pronunciations were recorded with a microphone (Shure WH20) attached to the ultrasound headset (Articulate Instruments Ltd). These native speaker recordings were then loaded into SonoSpeech (Articulate Instruments Ltd), the program that was used to show the UTI images (see Fig. 1 for an example image) during the training part of the experiment.

2.3. Procedure

After welcoming the participants, the first part of the experiment consisted of a perception task. Participants heard items (as pronounced by the native speakers) from minimal pairs containing either the /æ/-/ɛ/ or the /k/-/g/ contrast, after which they were asked to click on the word they thought they had heard (e.g., *bat* or *bet*, as presented in written form on the screen). The perception task took 5 minutes.

Next, participants were led to another area where the production experiment took place. After explaining the procedure and attaching the ultrasound headset to the participant, the pre-test recordings were made. Participants were asked to read a list of words (presented in randomized order), containing either the /æ/ or /ɛ/ sound or the /k/ or /g/ sound. The order of the items was created in such a way that two words from the same minimal pair would never follow each other (e.g., *bat* would never follow *bet*). Similar to the procedure for the model speakers, both audio and UTI recordings were made during the pre- and post-test.

Following the pre-test, in a session of roughly 20 minutes, the researcher trained the participants on the articulatory differences between /æ/-/ɛ/ and /k/-/g/. Participants could practice the target words and listen to the previously recorded pronunciations produced by a gender-matched native speaker of English. The researcher encouraged participants to practice the target words and would answer their questions, but made sure to keep the training sessions as similar as possible for all participants. The participants were divided into two groups: audio-only (n = 17) and audiovisual (n = 20). During the training session, the audio-only group only received auditory feedback. This meant that they received the articulatory instructions for the target sounds, but did not see their own UTI image nor that of the model speaker. Participants in the audiovisual condition, on the other hand, did receive this visual feedback, as they saw their own UTI image (presented in real-time) on the screen and were therefore able to watch their tongue movements while they practiced the target words. Moreover, they had access to the UTI videos of the pronunciation of the native speaker. In these videos, the tongue position for the target sound was indicated with a colored line, which made it easier to compare the tongue positions for the two contrasting sounds (see Fig. 1). After the training session, participants read the list of the (newly randomized) words again in the post-test. The entire experimental session took around 40 minutes, equivalent to the duration of one class at school (which students were allowed to miss to participate in this experiment).

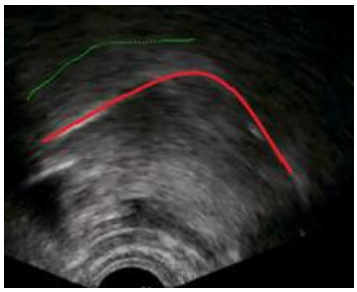

Figure 1: *Example UTI image that was provided in the training session.*

2.4. Analysis

Even though UTI data was collected in this experiment, the analysis in the current study only focused on the audio recordings. Our analysis contained two parts, namely a perceptual judgment task and acoustic measurements.

2.4.1. Perceptual judgment task

The goal of the perceptual judgment task was to determine whether the participants changed their pronunciation in a way that could be detected by a naïve listener. After excluding files that contained severe mispronunciations or other errors, an online survey was set up to collect native speaker judgments of the pronunciations by the Dutch speakers. A total of 248 native speakers of English (60 female, 179 male, 9 who indicated 'other' or preferred not to indicate their gender) with a mean age of 49.8 participated in this part of the study. The participants were recruited via Language Log.[1] The participants were given a set of recordings (in randomized order) and for each recording, they were asked to indicate which word out of two (i.e. *bat* or *bet*) they heard. Since this was an online survey, we asked participants to rate at least 10 recordings, but they could rate as many as they liked. The average number of recordings rated per participant was 44. None of the raters reported severe hearing issues.

2.4.2. Acoustic measurements

The goal of the acoustic measurements was to determine whether participants changed their pronunciation of vowels /æ/ and /ɛ/, and whether the type of feedback that participants received (audio-only versus audiovisual) played a role. We manually measured the first and second vowel formant of these

[1] See https://languagelog.ldc.upenn.edu/nll/?p=43095 for recruitment text.

two vowels in PRAAT (Boersma & Weenink, 2020), at the approximate midpoint of the vowel. We calculated Euclidean distances (EDs) between the minimal pairs in the pre-test and post-test using the formula in (1) as a way to measure vowel contrast.

$$\sqrt{(F1æ - F1ɛ)^2 + (F2æ - F2ɛ)^2} \qquad (1)$$

An increase in the Euclidean distance between two vowels in the post-test versus the pre-test would indicate that participants started making a larger distinction between the vowels. This would indirectly mean an improvement in pronunciation.

3. Results

3.1. Perceptual judgment task

To assess the effect of the training session on the pronunciation, we performed a mixed-effects logistic regression analysis with the dependent variable being whether or not the target word was recognized correctly by the rater. In the model with the optimal random-effects structure, a significant effect of test phase was found (β = .21, p < .05), meaning that target words recorded in the post-test were significantly more likely (0.2 logits, corresponding to an increase of about 5% in recognition probability) to be recognized correctly by the raters than words recorded in the pre-test. However, the experimental condition (audio-only versus audiovisual) did not have a significant effect on recognition, either by itself or in interaction with test phase. Figure 2 visualizes this result.

Regarding the target contrasts, target words in the /k/-/g/ category were more likely to be recognized correctly than target words in the /æ/-/ɛ/ category (β = .66 , p < .01; approximately a 15% increase in recognition probability). An interaction with test phase did not improve the model and was therefore not included. However, we did find a significant effect of participants' score in the perception task and the extent to which target words in the /k/-/g/ category were correctly recognized (β = .27, p < .001; approximately a 7% increase in recognition probability). Participants who had better perception, were better at producing the /k/-/g/ contrast (but not significantly better at producing the /æ/-/ɛ/ contrast: β = .08, p = .09). No other significant influences of personal characteristics on recognition were found.

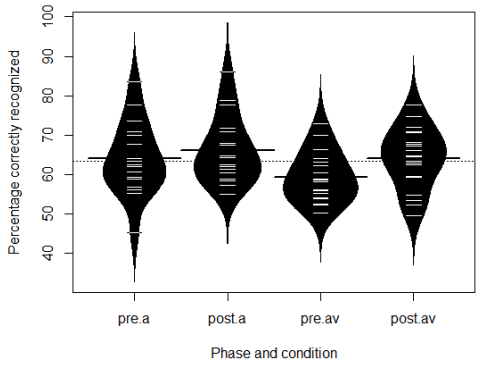

Figure 2: *Effect of training phase (pre-test versus post-test) and condition (audio-only versus audiovisual) on the percentage of correctly recognized items*

3.2. Acoustic measurements

To assess the general effect of training on the /æ/-/ɛ/ vowel contrast, we averaged the ED of minimal pairs in the pre-test and the post-test. This resulted in two average EDs per participant. We performed a simple linear regression, with the dependent variable being the Euclidean distance and the independent variables being the test phase (pre- versus post-test) and group (audio-only versus audiovisual). We found no significant effect for neither test phase (β = 2.6, p = .9) nor for condition (β = -39.9, p = .08). Figure 3 visualizes the effect of training and condition on the vowel contrast /æ/-/ɛ/ (as expressed in ED). While not significant, the EDs were somewhat larger after training for both conditions.

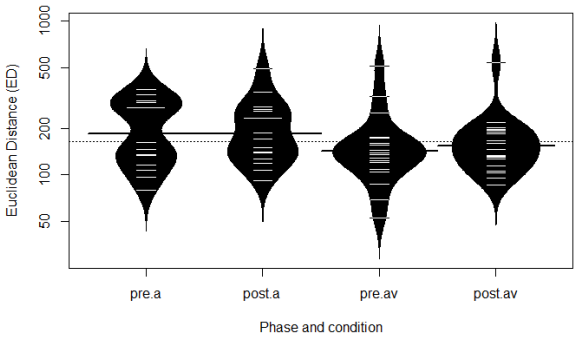

Figure 3: *Effect of training phase (pre-test versus post-test) and condition (audio-only versus audiovisual) on Euclidean distances.*

4. Discussion

This study investigated whether a short training session using UTI-based visual feedback would improve Dutch learners' pronunciation of the English sound contrasts /æ/-/ɛ/ and /k/-/g/. Looking at the native speaker ratings, we found that words pronounced after the training were more likely to be recognized correctly, indicating that even in a short session, explicit instruction on the articulatory differences between target sounds could help speakers in improving their pronunciation of non-native sounds. However, we did not find any significant differences between the participants in the audio-only and the audiovisual condition, whereas this was observed in several other studies (e.g., Ouni, 2014; Cleland *et al.*, 2015).

For the acoustic measurements, we found no significant effect of neither condition nor training. However, this could also be due to our choice of acoustic measure (i.e., vowel contrast). Specifically, target words in the perceptual judgment task included both the /k/-/g/ minimal pair contrast as well as the /æ/-/ɛ/ minimal pair contrast, with the consonantal pair contrast being more likely to be recognized correctly. As for the acoustic analysis, we only focused on the Euclidean distances in the vowel pairs. Further analysis is needed to determine whether an acoustic difference can be found in the velar context (e.g., a difference in Voice Onset Time, VOT). It is also possible that a different vowel measure (for example, vowel duration) would show a greater change. Nevertheless, even for the present vowels and measure, the result was close to significance (p = .08) and the direction of the effect was in line with the perceptual results.

Several reasons could be named for the absence of a specific training condition effect. One shortcoming of the study might be the length of the training session, which was only

twenty minutes. In this short time, the students had to learn about the phonetic differences between the target sounds, learn to interpret the UTI signal for the first time and practice the target words. In order to observe a beneficial effect of visual feedback, students might need more time to familiarize themselves with the interpretation of the UTI signal, especially for target contrasts like /æ/-/ɛ/ where the tongue shape differences are subtle (also given that the jaw was not fixed, and therefore the location of the UTI probe shifted relative to the hard palate).

Moreover, it is possible that students were less motivated to ask for clarification or ask questions to an unfamiliar researcher than they might be if their own teacher had provided the training. Although most participants in the audiovisual condition indicated that they found it interesting to work with UTI, more practice sessions could increase their engagement, which might in turn lead to different results. Finally, an important point concerns the age of our participants. Whereas many of the previous studies on this topic focused on older, usually college-aged learners, the participants in our study were in an early stage of second language learning and had received little to no phonetic training in their curriculum so far. Future research could look at whether a beneficial effect of visual feedback might be linked to language learning stage (i.e., early-stage versus more advanced L2 learners).

5. Acknowledgements

We thank the Groningen University Fund for the financial support that made this research project possible. Moreover, we thank RSG Ter Apel for recruiting participants and for allowing us to collect data at their school.

6. Note

Our original submission to ISSP 2020 only contained the analysis of the perceptual judgment task, which was carried out by Lisanne de Jong and Martijn Wieling. We later added the acoustic analysis by Teja Rebernik and Sonja Vaziri, hence the extended list of authors for this proceedings paper.

7. References

Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners. In D. Birdsong (Ed.), Second language acquisition and the critical period hypothesis (pp. 133 159). Mahwah, NJ: Erlbaum.

Broersma, M. (2005). Perception of familiar contrasts in unfamiliar positions. The Journal of the Acoustical Society of America, 117(6), 3890–3901.

Boersma, P., & Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.15, retrieved from http://www.praat.org

Cleland, J., J. M. Scobbie, S. Nakai & A. Wrench (2015). Helping children learn non-native articulations: The implications for ultrasound-based clinical intervention. Paper presented at the 2015 International Conference of Phonetic Sciences, Glasgow, Scotland.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. The Journal of the Acoustical Society of America, 116(6), 3668-3678.

Ouni, S. (2014). Tongue control and its implication in pronunciation training. Computer Assisted Language Learning, 27(5), 439-453.

Preston, J., Brick, N. & Landi, N. (2013). Ultrasound Biofeedback Treatment for Persisting Childhood Apraxia of Speech. American Journal of Speech Language Pathology, 22, 27-643.

Wester, F., Gilbers, D., & Lowie, W. (2007). Substitution of dental fricatives in English by Dutch L2 speakers. Language Sciences, 29(2-3), 477-491.