Abstract

Pronunciation training based on speech production techniques illustrating tongue movements is gaining popularity. However, there is not sufficient evidence that learners can imitate some tongue animation. In this paper, we investigated human awareness of controlling their tongue body gestures. In a first experiment, participants were asked to perform some tongue movements. This task was evaluated by observing ultrasound imaging of the tongue recorded during the experiment. No feedback was provided. In a second experiment, a short session of training was added where participants can observe ultrasound imaging in real-time of their own tongue movements. The goal was to increase their awareness of their tongue gestures. A pre-test and post-test were carried out without any feedback. The results suggest that without a priori knowledge, it is not easy to finely control tongue body gestures; and that we gain in performance after a short training session which suggests that providing visual feedback, even a short one, improves tongue gesture awareness.

1. Introduction

In the field of second language learning, speech technologies such as speech signal visualization (i.e. spectrum and F0 for instance), speech signal modification, speech synthesis, and speech recognition are often used as training tools. Most research focuses on the training of acoustic features [1, 2]. Even though some research exploits visual cues, this line of research focuses mostly on perception rather than production [3]. Only recently, interest has increased among researchers to apply speech-production-based techniques for pronunciation training in second language learning. One approach is to use virtual embodied conversational agents (ECAs) as tutors in the pronunciation training. In this context, an ECA has a 3D tongue of a talking head is accurately animated, with a display of a palate, and a velum [4, 5]. In addition, animating the talking head takes into account advanced coarticulation model [6, 7]. Typical pronunciation training consists of showing the articulation of the critical phonemes, either in isolation or placed in words or sentences. The specialized ECA is able to show the articulation of each sound from different views. For example, from front view, or a transparent midsagittal view, where a 2D view of the tongue, palate and velum are displayed, or a semi transparent 3D view, where it is possible to see through the skin, and therefore to observe the inner articulators as the tongue and the velum. The animations can be slowed down or repeated as many times as wanted. Moreover, additional instructions can be given explicitly or implicitly to learners.

During recent years, some studies examined the benefit of using ECAs in the pronunciation training to improve the production of second language learners. Of special interest was to assess whether seeing ECAs in views that cannot be provided by seeing a natural speaker are helpful for improving pronunciation.

For example, Massaro et al. [8] designed a lesson to teach native English speakers phoneme pairs of a second language, where one phoneme in a pair was highly similar to a native phoneme, while the second phoneme was not. The idea behind this approach was that learners may benefit in their pronunciation training of the novel phoneme when they can see how it is produced differently from a phoneme they know how to produce. Two phoneme pairs were trained: one in Arabic (/k/, /q/) and one in Mandarin (/i/, /y/). For Arabic, a midsagittal view was used as the phonemes were velar and uvular, in comparison with a front view of the face. For Mandarin a front view of the talking face was used as the phonemes were bilabials, in comparison with audio only. Although learners showed some improvements in the different conditions, the benefit was larger for the full face than when showing the tongue. These production results are also in line with perceptual data [9].

The pronunciation training results that although seeing inner articulators may help pronunciation, it does not seem to provide an additional benefit. The general assumption behind showing articulations to learners is that they will imitate or implicitly improve their perception of the to-be-learned phonemes and thus their production. However, in the case of second language learning, this perception-production link seems less useful when the relevant phoneme that is to be learned is absent in the learners first language. The existence of a highly confusable native phoneme will provide additional difficulties. One remedy here could be to provide instructions to learners on how to reach a target from a position of a phoneme in their native language by, for example, moving their tongue in some direction. However, there is little evidence that learners can follow correctly such instructions. In other words, pronunciation trainings based on illustrating tongue movement for some phonemes cannot be successful if learners are not able to reproduce those movements, even if they understand the instructions.

To investigate the human self awareness of tongue gestures, we designed an experimental study where tongue movements were recorded by ultrasound imaging. In a first experiment, participants were asked to reproduce some simple tongue displacements without any visual feedback. Their realizations have been evaluated by observing how well they succeeded in achieving the different gestures. In a second experiment, we additionally investigated whether a short training where participants observed their tongue movements in real-time would improve their awareness of their own tongue gestures (i.e. a correspondence between the intended gesture and executed one). Identically as in the first experiment, this training session was followed by an evaluation, and the performance with a control group was compared.
2. Tongue Control Study

2.1. Tongue Gestures

In our study, we considered 12 tongue gestures. The choice of tongue movement directions was based on the kinematics of the most important muscles of the tongue. We consider this set of gestures as simple, as they did not involve a deformation of the shape of the tongue. That is, the tongue movements here are not as complex as found in speech. In fact, speech sound articulations are based on a finely tuned deformation of the tongue shape, in addition to the displacement of the tongue.

There are two parts of the tongue that are mainly involved in producing speech: the tip of the tongue and the tongue body. The tip of the tongue participates in the production of sounds that are dental, alveolar, and post-alveolar. In the contrast, the body of the tongue participates in the production of almost all speech sounds where the articulation is taking place in the vocal tract. It seems easy to control the tip of the tongue and move it almost in any direction, but this seems less likely to be the case for the body of the tongue. For these reasons, we focused only on the tongue body movements.

The 12 tongue movements are presented in Fig. 1 and can be organized in two sets. The purpose of the first set of gestures was to observe how humans can control the motion of their tongue in various directions. The second set allowed assessing to what degree it is possible for speakers to move the body of their tongue from a known position of a phoneme of their native language to a new position. This study addresses therefore the question to what extent speakers can control the movement of the body of their tongue and the degree to which each set of gestures is easy or difficult to accomplish. Note that the starting position of the tongue is a neutral position, i.e. the tongue lying on the jaw and not touching the palate.

![Gesture Set I (GSI)](image1)

![Gesture Set II (GSII)](image2)

Figure 1: The different directions used for the two sets of gestures. The vocal tract is just for illustration to show the tongue movement.

**Gesture Set I (GSI)** The first set of gestures is not physically difficult to produce by the tongue and follow the major muscle movements. In addition, these gestures are independent of any speech or language knowledge. The purpose of such a choice is to limit as much as possible the influence of the complexity of the articulation on the realization of a given gesture. Some of these movements correspond to the articular form an existing phoneme (as in gesture 5 that corresponds to some extent a French /i/), but this information was not provided to participants.

**Gesture Set II (GSII)** This set was based on the articulation of three phonemes: /a/, /i/ and /k/ representing three distinct articulatory regions in the vocal tract (approx. back and down; up and front; up and back). From each of these positions, participants were asked to move their tongue slightly in some direction. Participants were asked to start from the position of /a/ and /i/ and for the phoneme /k/ without releasing it. For the phonemes /a/ and /i/ they were asked to pronounce the sound out loud to make sure that they started from the correct position. This situation can be similar to teaching non-native phonetic contrasts to learners. In this case, we usually show the known phoneme and provide instructions on how to reach the new sound from that position. The place of articulation of the new sound is usually in the vicinity of the known one.

2.2. Tongue movement observation

In this study, we observed the motion of the tongue using ultrasound imaging. An ultrasound transducer (probe) was placed against the chin produces a beam across the tissues of the tongue and is reflected at the surface of the tongue when it makes contact with air. The placement of the probe allows obtaining a midsagittal view of the vocal tract at a frame rate of 66 images/second. While the tongue is moving, it is possible to interpret the movement and make sense of the tongue gestures. The interpretation needs however some efforts and some reference gestures are needed to help with interpretation.

2.3. Experimental Design

**Pre-test and post-test sessions** Participants were asked to move their tongue (mainly the body of the tongue) in a certain direction, or from a given position towards a target position. As explained above, we were solely concerned with the movement of the body of the tongue. Thus, participants were asked to reproduce the 12 gestures described above. Each gesture, was repeated twice for better interpretation while evaluating by human judges. Their production was observed and recorded using an ultrasound machine. Participants did not see the ultrasound images during the pre-test and post-test sessions. The recording was analyzed off-line. Instructions were given to participants by showing the direction of the tongue movement by hand. Before starting the experiment, participants were explained what the task was, and what was meant by the body of the tongue, making it clear that the body of the tongue gestures are different from those of the tip of the tongue. In addition, the production of few phonemes was recorded that served as reference in the data analyses.

During the different experiments, participants did not receive any visual feedback. The ultrasound probe was handled by participant themselves. They were explained how to handle it correctly, in addition, the experimenter checks the orientation of the probe on the screen, before asking a participant to do a given gesture. However, they were not given any information about the aim of the study before and during the experiment.

**Training session** A group of participants were involved in a very short observation and practice session (about 15 to 20 minutes). During this session, participants had the possibility to observe the movement of their tongue as displayed in real time by the ultrasound machine. They were able to practice the 12 predefined gestures. The experimenter provided them with a description of the different gestures and explained how to read an ultrasound image by showing the palate, the tongue and the overall shape of the vocal tract. In this study, it was not our purpose to provide any training on how to control the tongue. The aim was to investigate whether a visual feedback session can be
sufficient to improve the awareness of tongue gestures.

Two experiments In this study, we present two experiments tightly related. In the first experiment, we investigate whether humans are aware of their tongue gestures, and whether there is a performance difference in reproducing the set of GSI compared to the set GSII. In the second experiment, we investigate whether a short training session would improve participants awareness of their own tongue gestures. The experimental design was as follows: (1) a pre-test session; (2) a training session for one group, and nothing of the second group; (3) a post-test session. For the first experiment, we consider only the pre-test sessions, and thus to increase the reliability of the result, the two groups were pooled. For the second experiment, we considered the three sessions, where one group has a training session, and the other group has no training. During the two test sessions, no feedback was provided. The only difference between pre-test and post-test is that the sequences of the presented gestures were randomized.

2.4. Participants
We considered two groups of participants. We call the first group, the control group. They had a pre-test and post-test sessions, separated by a pause for about 15 minutes. This group did not get any visual feedback. They were ten native speakers of French, all between 24 and 38 years. They reported no history of a speech or language disorder. They did not have any particular training in phonetics.

We call the second group Ultrasound group. They had a pre-test and post-test sessions separated by a training session of about 15-20 minutes. During this training session they have got a visual feedback as explained above. They were fourteen native speakers of French, all between 24 and 36 years old. Participants reported no history of a speech or language disorder. They did not have any particular training in phonetics.

2.5. Results
The ultrasound data were examined and the different gestures were evaluated. The evaluation was performed by two judges who verified the adequacy of the different gestures. Theoretically, it is better to measure the displacement of the body of the tongue and the orientation in degrees from the first image to the last one. Unfortunately and in practice, the nature of the ultrasound imaging made it very difficult. In fact, the ultrasound images were evidently noisy and it was very hard to delimit the tongue and the palate. In addition, the dynamic is important to have an idea of the movement. Considering other acquisition techniques as Electromagnetic Articulography (EMA) is not very practical, takes a lot of time to set up, and it will not be possible to find many participants who will accept such acquisition technique. For these reasons, we preferred to use human judge evaluation in our study. We considered two judges to verify the reliability of the results.

The realization of each gesture was repeated twice by each participant. The evaluation of each gesture was a global score for the two realizations. A 10-point scale was used as a rating scale for the evaluation: (10) completely correct and (0) incorrect. A gesture was rated as completely correct, if a displacement of the tongue toward the correct direction was observed across the two repetitions. The more the overall gesture was correct the higher the score. Completely incorrect gestures, where those for which a gesture was performed in an incorrect direction (for instance, lowering the tongue, instead of backing it, or advancing the tongue instead of lowering it). In this study, the total number of the evaluated gestures was 552. In the following, we present the results for the two experiments.

Experiment I For the first experiment, the results were pooled across all participants of the two groups (Control group and Ultrasound group), to increase the reliability of the ratings. The results showed that reproducing a specific gesture was not easy or obvious ($M = 5.77$, $SD = 2.24$). There was no participant among the 23 that was able to reproduce all the gestures correctly. Although the selected gestures did not present a particular difficulty and are physically easy to produce by the different muscles, the participants were not able to control their tongue body movement and succeed in reproducing the different gestures. Fig. 2, present the mean score ($\pm$ standard deviation) for every gesture. The gestures 2 and 8 obtained the highest score ($2: M = 6.78$, $SD = 2.81$; $8: M = 6.7$, $SD = 2.36$). However, they are barely 1 point above the average of all the gestures. These two gestures have on common the backing of the tongue. The two gestures that were the most difficult to reproduce were gestures 1 and 6 ($1: M = 4.65$, $SD = 2.22$; $6: M = 4.74$, $SD = 2.28$). These two gestures have on common the advancement of the tongue. When grouping the results by the type of gestures, we obtained almost the same results for the first set of gestures (GSI) ($M = 5.9$, $SD = 2.10$) and for the second set (GSII) ($M = 5.68$, $SD = 2.34$). In addition, the difference was not significant ($F(1, 44) = 0.35$, $p = 0.55$). For the second set (GSII), we should note that the articulation of the starting phoneme was very accurate across all participants. However, the execution of the following gesture was not successful in the majority of the cases. This shows that starting from a very well-known position does not help that much, as participants did not seem to be aware of the place of articulation of this gesture, but just executing a "pre-recorded" movement.

![Figure 2: First experiment result. Mean score (± standard deviation) for every gesture. The twelve gestures are presented in two sets : GSI and GSII.](image)

Experiment II The goal of the second experiment is to see the effect of a very short observation and practice session in improving participants’ realization of the tongue gestures. This improvement is measured by comparing the performance of the ultrasound group with the performance of the control group.

Fig. 3 shows the progress gained from pre-test to Post-test, for the two groups. It is clear that almost all the 12 gestures (except gesture 9 and 10) gained improvement across ultrasound group participants. However, the gestures realized by the
control group did not gain much improvement. Moreover, the scores of the set GSI actually were deteriorated.

Figure 3: Mean score differences of the 12 gestures across Pre-test Post-test. Results are presented for Ultrasound group and Control group.

Fig. 4 presents the average ratings for the different gestures for the ultrasound group and the control group, before and after training, according to the gesture type (GSI and GSII). The ultrasound group presented similar ratings for GSI and GSII, before (GSI: $M = 5.6$, GSII: $M = 5.85$) and after training (GSI: $M = 6.23$, GSII: $M = 6.45$). For the control group, while gesture set GSI means were almost the same during the pre-test ($M = 5.44$) and post-test ($M = 5.48$), the performance of the gesture set GSII decreases from pre-test ($M = 6.28$) to post-test ($M = 4.52$). There were no interaction between group and test, and no interaction between group and gesture type.

Figure 4: Average ratings for the two sets of gestures (GSI and GSII) across Ultrasound group (US) and Control group (CTRL) before (Pre-test) and after training (Post-test).

### 3. Discussion

The main finding of the tongue control study is that controlling the movement of the tongue body is to some extent difficult. Producing specific tongue gestures that were not learned during native language acquisition or second language learning processes is not an easy task. Humans are not really aware of the mechanism of articulating a given segment of speech as soon as they succeed in reaching the needed target. The coordination of the different tongue muscles to produce a given phoneme or word was acquired during early stages after several repetitions and the retention became permanent. Moreover, the gestures starting from places where the articulation is well-known, as it is in their phonetic repertoire, did not help to reach a given target position. This implies that it is very likely that pronouncing phonemes during continuous speech follows an already learnt articulatory path, and that it is not very easy to decompose this path in elementary gestures. This suggests that pronunciation training methods based on contrasts would not be very effective, if the purpose is to transfer the tongue movement illustrated by a talking head to the learner by imitation. It seems that a visual feedback step is unavoidable for these training methods.

In fact, this study showed that learners benefit from having visual feedback available, even though during an extremely short session of practice. During this error and trial process, participants were capable of increasing their awareness of their tongue gestures. Learners can visualize their own tongue movement and readjust a particular gesture based on the observation and the given instructions. We should notice that the effect of the session of practice lasts even when the feedback was removed.

Pronunciation training based on illustrating speech articulation should take into account the awareness of learners of the used gestures. We recommend that pronunciation training should be either based on some explicit or implicit real-time visual feedback or preceded by an awareness task of the articulation gestures. As this is not an easy task, future studies should focus on how to integrate visual feedback techniques efficiently in the learning process.

### 4. References


