Tongue control and its implication in pronunciation training

Slim Ouni

Tongue Control and its Implication in Pronunciation Training

Slim Ouni

Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-les-Nancy, F-54506, France

Slim.Ouni@loria.fr

Tongue Control and its Implication in Pronunciation Training

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 argue that although controlling tongue movement related to speech is not such an easy task, training with visual feedback improves its control. We investigated human awareness of controlling their tongue body gestures. In a first experiment, participants were asked to perform some tongue movements composed of two sets of gestures. 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. The second experiment showed that we gain in performance after a short training session and this suggests that providing visual feedback, even a short one, improves tongue gesture awareness.

Keywords: pronunciation training; tongue movement; visual feedback; speech production

Introduction

In the field of second language learning, speech technologies such as speech signal visualization (e.g. spectrum and F0), speech signal modification, speech synthesis, and speech recognition are often used as training tools (Cucchiarini, Strik & Boves, 2000; Neri, Mich, Gerosa & Giuliani, 2008; Strik, Neri & Cucchiarini, 2008; Eskenazi, 2009). Most research focuses on the training of acoustic features (Pisoni Lively & Logan, 1994; Lambacher, 1999; Hazan & Simpson, 2000; Colotte, Laprie & Bonneau, 2001; Probst, Ke & Eskenazi, 2002). Even though some research uses visual cues, this line of research focuses mostly on perception rather than production; see for instance Hazan, Sennema, Faulkner, Ortega-Llebaria, Iba & Chung (2006). Actually, in second language learning literature, researchers usually make the link between perception and production in how to discriminate non-native phones from native phonemes. In these cases, articulatory configurations of these different phonemes are usually discussed (Best, McRoberts & Goodell, 2001; Flege, Munro &

MacKay, 1995; Kuhl, 1991). However, they naturally focus on perception as they consider that we cannot obtain a good production of non-native phonemes if perception is not accurate (Flege, Bohn & Jang, 1997). Only recently, interest has increased among researchers to apply speech-production-based techniques for pronunciation training in second language learning. The main idea behind these techniques is that training learners on how to articulate non-native phonemes, or showing them the articulatory differences between native and non-native phonemes, can help to improve their production, and perhaps, their perception of these sounds. The articulatory improvement can be assessed either directly by measuring the learner's articulation using articulatory visualization techniques, or indirectly through the evaluation of the learner's acoustic realization.

A first approach of speech-production-based techniques is to provide a real-time visual feedback of the actual articulation of the learners. The visual feedback can be delivered using machines that are traditionally used in the medical field, such as electropalatographs, electromagnetic articulographs, or ultrasound machines. However, these real-time visual feedback techniques are not yet sufficiently convenient or practical to be widely used for language learning or at large scale for speech therapy. In fact, most of these techniques are invasive and require a substantial amount of preparation and tuning by an expert before one can use them. For these reasons, researchers have become interested in alternative approaches that are more accessible to the general public. Over a decade, virtual embodied conversational agents (ECAs) have been used as tutors in pronunciation training (Bosseler & Massaro, 2003; Wik, 2009; Massaro, Liu, Chen, & Perfetti, 2006; Wang, Qian, Scott, Chen & Soong, 2012). This ECA specialized in pronunciation must be based on an accurately animated talking head, where the animation is synchronized with auditory

speech. Many talking heads are highly developed and include improvements that are principally related to the realism of the articulation. For instance, in some of the developed systems, the 3D tongue of the talking head is accurately animated, with a display of a palate, and a velum (Cohen, Massaro & Clark, 2002; Engwall, 2003; Massaro, 2003; Badin, Elisei, Bailly & Tarabalka, 2008). In addition, animating the talking head takes into account advanced coarticulation modelling (Cohen & Massaro, 1993; Cosi, Caldognetto, Perin & Zmarich, 2002); see Beskow (2004) for an overview. Typical pronunciation training consists of showing the articulation of the critical phonemes, either in isolation or placed in words or sentences.

In the following, we present some case studies and examples where the two approaches based on speech production techniques were used for language learning, or more widely for speech therapy. The first approach is the use of machines that can provide real-time visual feedback, and the second is the use of ECA as a pronunciation tutor.

Real-time visual feedback

Real-time visual feedback techniques can be considered an interesting way to show learners in real-time their own articulation. An example of such a technique is electropalatography (EPG). This technique is used mainly in speech therapy. An artificial palate implanted with electrodes is worn by a patient and allows the visualization in real time of the contact between the tongue and the palate during speech production (Hardcastle & Gibbon, 1997). Typically, the therapist shows an articulatory target of a sound, which the patient tries to reach. For example, Crawford (1995) used EPG in teaching two profoundly deaf children to produce initial voiced velar stops. The two participants made a significant improvement in their articulatory production of that category of sounds. Dent, Gibbon and Hardcastle (1995) presented

a case study of a patient who improved the production of lingual stops and fricatives using EPG therapy. In another case study, Pantelemidou, Herman and Thomas (2003) showed that using EPG improved significantly the articulation of voiced and voiceless velar plosives in a cochlear implanted child. The effect of the learning was persistent after 5 weeks from the end of the therapy. Similarly, electromagnetic articulography (EMA) can be used to provide real-time articulatory visual feedback. EMA is a tracking system that dynamically provides the positions of sensors glued on the tongue using electromagnetic field variations. Katz et al. (2007) used EMA to provide visual feedback in the training of consonant production for patients with apraxia and aphasia. EMA-based therapy improved the production for some articulatory targets. These results are very promising regarding the use of EMA as feedback for articulatory training. As mentioned above, these techniques are not easy to use widely. Hopefully, technological improvements will make the machines more accessible, easily transportable and user friendly. Nevertheless, these techniques will still be invasive.

ECA as a visual pronunciation training tutor

An ECA specialized in pronunciation training can be capable of showing the articulation of each sound from different views. For example, a 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 such as the tongue and the velum. In some systems, the animations can be slowed down or repeated as many times as wanted. Moreover, additional instructions can be given to learners. During recent years, some studies examined the benefits of using ECAs in 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, French participants were asked to pronounce some Swedish words by observing a talking head with a view of the tongue (Engwall, 2012). During training, instructions were given to explain the articulation presented by the animation of the talking head. Pronunciation improvement was assessed through ultrasound by measuring articulation. Participants' articulation improved through the training. The contribution of seeing the tongue and receiving instructions can, however, not be assessed here. Massaro and Light (2003) assessed the additional contribution of seeing the internal articulatory processes compared to simply seeing the face of a talking head for teaching non-native phonetic contrasts to Japanese learners of American English. The task was to identify and produce /r/ and /l/ in American English. Training either involved showing the face or also included showing the internal articulatory processes of the oral cavity. A pre-test/post-test design was used. The participants' realizations were scored; where each acoustically produced word was scored by a human judge as correct or incorrect, without any knowledge of the experimental conditions. Although both speech identification and production improved, showing the internal articulators did not show an additional benefit. Similar results were obtained by using a contrastive teaching approach. Massaro, Bigler, Chen, Perlman & Ouni (2008) 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, for the Arabic talking head the benefit was larger for the full face than when showing the tongue. These production results are also in line with perceptual data. Badin, Tarabalka, Elisei and Bailly (2010) assessed the benefit from seeing the tongue to better perceive sounds. They performed audiovisual perception experiments in a noisy environment with different viewing conditions. The main presentation conditions were the following: Participants either saw a midsagittal view of the jaw, vocal tract walls, palate and pharynx with the tongue or received the same view but without seeing the tongue. This was compared to a condition where a profile view of the full face was provided during training. An auditory-only condition where no speaker was visible was also added. One of the results of this study is that the full face appears to provide the best perception results. Similarly, Grauwinkel, Dewitt and Fagel (2007) found that showing the tongue and other articulators in comparison to not showing them did not provide a significant benefit in a consonant identification task. However, a short training where articulatory gestures were explained improved recognition. In another experiment, Grauwinkel & Fagel (2007) used the visualization of inner vocal tract in a learning lesson addressed to three children with Sigmatismus interdentalis to improve their sibilant production. In an experiment presented in (Kröger, Graf-Bortscheller, & Lowit, 2008) the visual recognition of mute uttered phonemes by children (five to eight years old) when presenting a 2D- or 3D-model was very low (19% to 22%).

Seeing the tongue and controlling it

In summary, the pronunciation training results show that although seeing inner articulators may help pronunciation, it does not seem to provide an additional benefit *a priori*. 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. In fact, there is evidence for a strong link between perception and production in the motor regions of the brain. Fadiga, Craighero, Buccino and Rizzolatti (2002) showed that while listening to speech, listeners show an increase of motor activities of tongue muscles when the heard words involved tongue movements. Watkins, Strafella and Paus (2003) found similar results for lip muscles, when lip movements were observed in addition to listening to speech. 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 (Best et al., 2001; Iverson et al., 2003). 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, as shown to some extent in the studies discussed above, there is little evidence that learners can correctly follow such instructions. In other words, we think that learners cannot easily reproduce some tongue movements just by illustrating or describing the gestures. In fact, 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 animations. For instance, a French /r/, an English /r/, an Arabic /q/ or a German /ç/ are not easy to pronounce just by showing how to do so for learners for whom these phonemes do not exist in their native language. Thus, the important questions that we tried to answer in

this paper were: can humans consciously control precisely the movement of the body of their tongue when asked to imitate or reproduce a tongue gesture? Are humans aware of their tongue gestures? Is it easy for humans to perform tongue movements mechanically? Does training with visual feedback of articulatory movements improve tongue control awareness? Answering these questions should give insights on how to improve the use of ECAs as tutor in language learning and make the technique more effective.

To answer these questions, we designed an experimental study based on two groups: a control group (10 participants) and an experimental group (14 participants). The general scheme of this study was a pre-test/post-test design. The control group did not receive any feedback, and the experimental group had a short training session (about 15 to 20 minutes) where participants observed their tongue movements in real time using an ultrasound machine. Each group had pre-test and post-test sessions. Their realizations were recorded using an ultrasound machine and evaluated offline by observing how well they succeeded in achieving the different gestures. This experimental design can be seen as two experiments. In the first one, we investigate how well the participants succeeded in achieving the different tongue gestures, without any *a priori* knowledge. In this experiment we examine only the pre-test sessions. In the second experiment, the goal is to investigate whether a short training session improves their awareness of their own tongue gestures. In this experiment, we examine the pre-test and post-test sessions and we compare the performance of the experimental group against the control group.

Tongue Control Study

Tongue Gestures

The tongue is a complex organ controlled by a set of intrinsic and extrinsic muscles (Abd-el Malek, 1939; Bole & Lessler, 1966; Carpentier & Pajoni, 1989). Intrinsic muscles control the shape of the tongue and the movement of its tip. Extrinsic muscles move the tongue back and upward (styloglossus muscle), forward (genioglossus muscle), downward and back (hyoglossus muscle) and raise the tongue (mylohyoid muscle). Other muscles exist that also participate in controlling the tongue (for instance, palatoglossus, palatopharyngeus and geniohyoid). The tongue muscles are finely controlled during speech and can be smoothly adjusted to produce rapid articulation if needed. 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. 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 Figure 1 and can be organized in two sets. The purpose of the first set of gestures was to observe how humans could 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. Therefore, this study addresses 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 are 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.

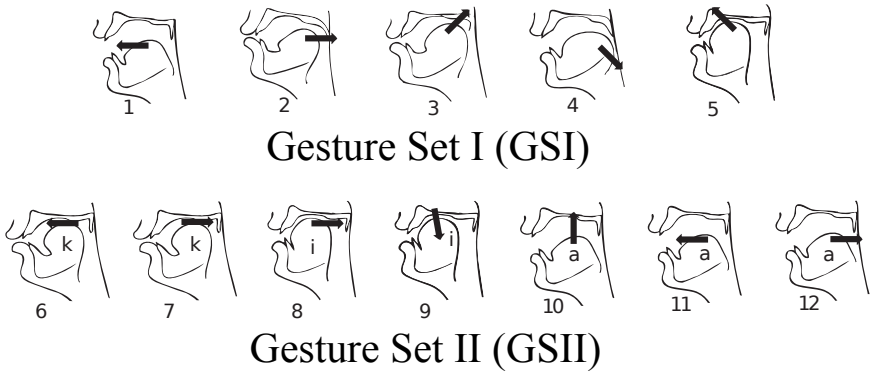

Figure 1. The different directions used for the two sets of gestures. The vocal tract is just for illustration to show the tongue movement. The arrows show the direction of each gesture.

Gesture Set I (GSI)

This set of gestures represents five directions arbitrarily selected in a way that they are 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 this selection 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 may correspond to reaching an articulatory target of an existing phoneme (as in gesture 5, where the final position 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, the task is to ask participants 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, as in (Massaro et al., 2008). 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.

Tongue movement observation

In this study, we observed the motion of the tongue using ultrasound imaging. An ultrasound transducer (probe) 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 per second. A static ultrasound image is difficult to read (see Figure 2 for some examples). We can barely see the body of the tongue, where the contour is sometimes difficult to detect. In addition, it is not possible to see the palate, as the ultrasound beam does not go through the air between the surface of the tongue and the palate. However, while the tongue is moving, it is possible to interpret the movement and make sense of the tongue gestures. However, the interpretation requires some effort and some reference gestures are needed to help with interpretation. Therefore, some experience is needed to be able to correctly interpret the ultrasound images. Despite these difficulties in interpreting ultrasound

images, the technique has the advantage of providing the images of the tongue in real time, not altering the articulation of the participants, and it is not invasive. This is not the case of other more accurate techniques.

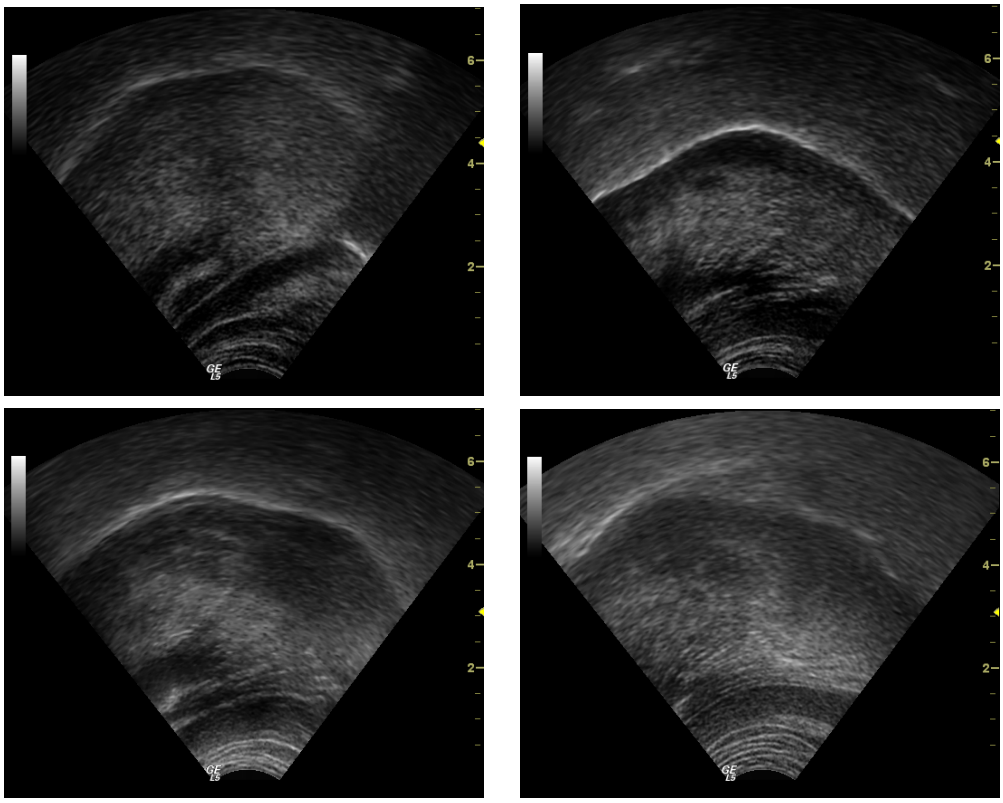

Figure 2. Examples of ultrasound images of the tongue of 4 different speakers. They show midsagittal views of speakers looking to the left.

Experimental Design

Pre-test and post-test sessions

The realizations of the participants were recorded using an ultrasound machine to be analyzed offline. Participants did not receive any visual feedback, i.e., they did not see the ultrasound images during the pre-test and post-test sessions. They handled the ultrasound probe themselves. They were instructed how to handle it correctly. Furthermore, the experimenter checked the orientation of the probe on the screen before asking the participant to perform a given gesture. The recording was analyzed offline. 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 some phonemes was recorded to serve as a reference in the data analyses. However, they were not given any information about the aim of the study before and during the experiment.

Training session

A group of participants was 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 closely related experiments. In the first experiment, we investigate whether humans are aware of their tongue gestures, and whether there is a performance difference in reproducing set GSI compared to 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 follow: (1) a pre-test session; (2) a training session for one group (test group) and nothing for the second group (control 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. The control group has only a paper containing the list of gestures, but they were not given any instructions or asked to do any practice. During the two-test session, no feedback was provided. The only difference between pre-test and post-test is that the sequences of the presented gestures were randomized.

Participants

We considered two groups of participants. The first is the Control group. The participants of this group had pre-test and post-test sessions, separated by a pause of about 15 minutes. This group did not get any visual feedback. They were 10 native speakers of French, all between 24 and 38 years old. They reported no history of a speech or language disorder, and did not have any particular training in phonetics. We call the second group Ultrasound group. The participants had a pre-test and a post-test session separated by a training session of about 15-20 minutes. During this training session, they received visual feedback as explained above. They were 14 native speakers of French, all between 24 and 36 years old. They reported no history of a speech or language disorder, and did not have any particular training in phonetics.

Results

The ultrasound data were examined and the different gestures were evaluated. The evaluation was performed by two judges, who have experience in interpreting ultrasound images. They verified the adequacy of the different gestures. Each participant repeated twice the realization of each gesture. 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 global evaluation. Thus, each realization was scored over 5: (5) completely correct and (0) completely incorrect. A

gesture was rated as completely correct if a displacement of the tongue toward the correct direction and final target was observed. 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). Table 1 shows the main guidelines used by the judges to evaluate the participant realizations. The experimenter elaborated these guidelines after the observation of several recordings, and testing on several scenarios, with a view to providing the most interpretable final score. The judges, obviously, did not participate in this task or see the recordings before starting the task of the evaluation.

Table 1. The main guidelines used for the evaluation

Realization cases	**Score**
During each realization	
a) The whole gesture is correct	5
b) Only the beginning of the gesture is correct, but not reaching the target.	2
c) The gesture toward target is correct but not fully reaching the target.	3 – 4
d) The global gesture is decomposed in two gestures but reaching the final target (one of the gesture should not be contradictory movement toward the final target	3
e) The whole gesture is wrong	0
Penalty There is an extra gesture unrelated to reaching the final target (but within the cases b, c, or d)	-1 or -2 depending on the importance of this extra gesture.
Across two realizations	
Penalty If one realization is completely wrong, but not the other, apply a global penalty	-3
Note The lowest score is zero (no negative score)	

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 who 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. Figure 3 presents the mean score (± standard deviation) for every gesture. The gestures 2 and 8 obtained the highest score (gesture 2: M = 6.78, SD = 2.81; gesture 8: M = 6.7, SD = 2.36). However, they are barely 1 point above the average of all the gestures. These two gestures have in common the backing of the tongue. The two gestures that were the most difficult to reproduce were gestures 1 and 6 (gesture 1: M = 4.65, SD = 2.22; gesture 6: M = 4.74, SD = 2.28). These two gestures have in 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 in many cases not successful. 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 executed a “pre-recorded” movement.

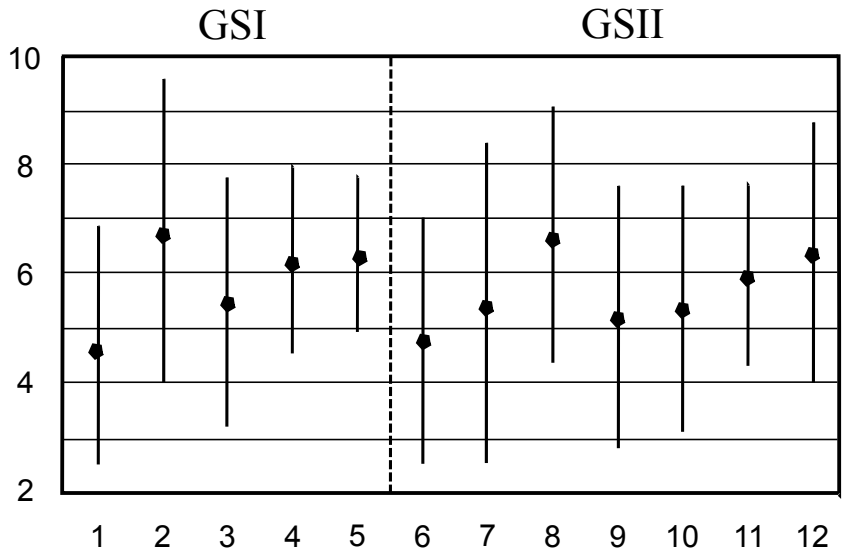

Figure 3. First experiment result. Mean score (± standard deviation) for every gesture. The twelve gestures are presented in two sets: GSI and GSII.

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.

Figure 4 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 deteriorated. One can speculate that the Control group does not have any hint on these gestures, and they are not even able to reproduce them. For the set GSII, as the Control group starts from a known position, this may reduce the possibility of making many wrong gestures, and thus the difference between pre- and post-test is reduced.

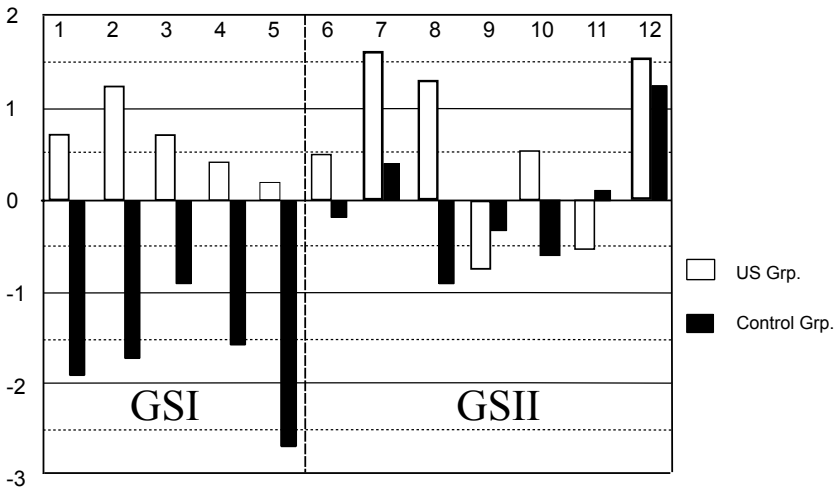

Figure 4. Mean score differences of the 12 gestures across Pre-test Post-test. Results are presented for Ultrasound group and Control group.

Figure 5 shows the average ratings for the different gestures for the Ultrasound group and the Control group, before and after training. For the Ultrasound group, production was overall better after (M = 6.34, SD = 1.99) than before training (M = 5.72, SD = 2.22). The Control group gestures did not present any improvement during the pre-test (M = 5.86, SD = 2.23) and the post-test (M = 5.86, SD = 2.72). This result was significant (F (1, 42) = 3.82, p = 0.005) and there was an interaction between group and test (F (1, 42) = 5.69, p = 0.02).

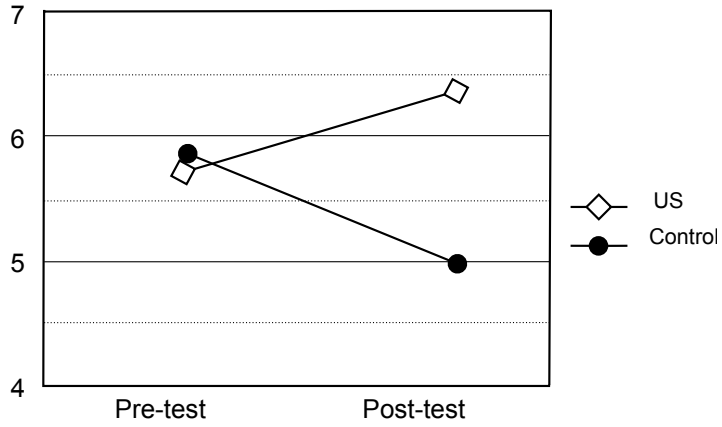

Figure 5. Average ratings for Ultrasound group and Control group before (Pre-test) and after (Post-test) training.

Figure 6 presents more details. We added another level: 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 was no interaction between group and test, and no interaction between group and gesture type.

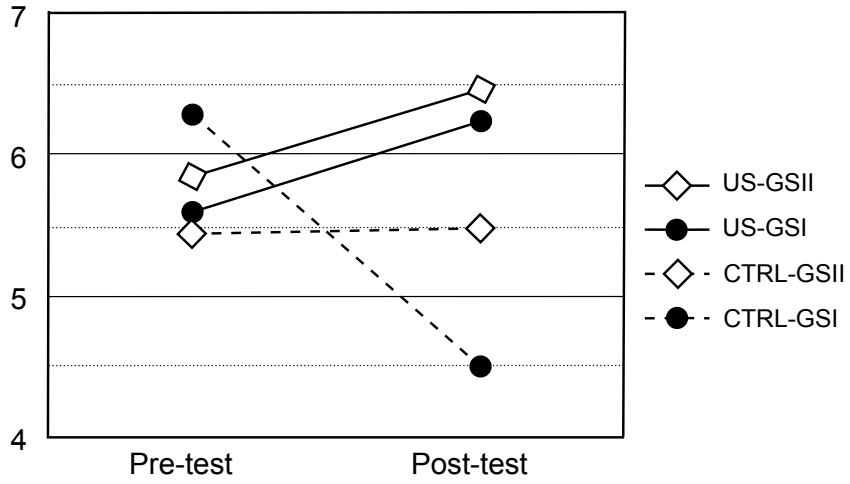

Figure 6. Average ratings for the two sets of gestures (GSI and GSII) across Ultrasound group (US) and Control group (CTRL) before (Pre-test) and after (Post-test) training.

Discussion

The main finding of this 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 of language acquisition, after several repetitions and the

retention became permanent. In our study, 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 learned articulatory path, and that it is not very easy to split up 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. Visual feedback seems to be helpful for these training methods. In fact, this study showed that learners benefit from having visual feedback available, even if only during an extremely short session of practice. We should highlight, however, that we could not confirm that the improvement is totally related to the visual feedback. In fact, the participants of the Control group were not asked to perform any particular task, as effectively practicing the two sets of gestures, and we cannot tell if some of them did do so or not.

During the training session, which is a trial and error 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. During this practice session, participants were consciously trying to control the different gestures starting with awkward movements and they made many errors before starting to produce some correct gestures. We should notice that the effect of this session of practice lasts even when the feedback was removed, i.e., during the post-test session. Another finding is that visual feedback helped to increase the awareness of participants' articulations of phonemes of their first language used in this experiment. In fact, during the tongue observation sessions, participants stated that they did not know that the tested phonemes are produced in

such way. This implies that pronunciation training based on contrasts can be efficient if it is preceded or combined with visual feedback of the learner's articulation.

As final remark, pronunciation training based on illustrating speech articulation should take into account the awareness of learners of the used gestures. More generally, we highly recommend the use of some training based on visual feedback of the learner's articulation preceding the use of ECAs as tutor in language learning lessons. In fact, we believe that the use of ECA in language learning is very effective when used in the right conditions. We recommend that pronunciation training should be based on some explicit 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. In addition, the persistence of the training using some visual feedback technique is not known and should be evaluated in dedicated studies.

Acknowledgments

The author wishes to thank Marie-Odile Berger and Yves Laprie for providing the ultrasound machine facility, and Alexandra Jesse for her help in the statistical analysis.

References

Abd-el Malek, S. (1939). Observations on the morphology of the human tongue. *Journal of Anatomy,* 73 (pt. 2), 201–210.

Badin, P., Elisei, F., Bailly, G., & Tarabalka, Y. (2008). An audiovisual talking head for augmented speech generation: Models and animations based on a real speakers articulatory data. Lecture Notes in Computer Science 5098/2008, 132–143.

Badin, P., Tarabalka, Y., Elisei, F., Bailly, G. (2010). Can you 'read' tongue movements? evaluation of the contribution of tongue display to speech understanding. *Speech Communication* 52, 493–503.

Beskow, J. (2004). Trainable articulatory control models for visual speech synthesis. *International Journal of Speech Technology* 7 (4), 335–349.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's na- tive phonological system. *The Journal of the Acoustical Society of America* 109 (2), 775–794.

Bole, C. T., & Lessler, M. A. (1966). Electromyography of the genioglossus muscles in man. *Journal of Applied Physiology*, 21 (6), 1695–1698.

Bosseler, A., & Massaro, D. W. (2003). Development and Evaluation of a Computer-Animated Tutor for Vocabulary and Language Learning for Children with Autism. *Journal of Autism and Developmental Disorders*, 33, 653-672.

Carpentier, P., & Pajoni, D. (1989). La langue: un ensemble musculaire complexe. *Revue d'Orthopédie Dento-Faciale* 23, 19–28, (*in French*).

Cohen, M., & Massaro, D. W. (1993). Models and techniques in computer animation. Springer, Berlin, Ch. Modelling coarticulation in synthetic visual speech, 139 –156.

Cohen, M. M., Massaro, D. W., & Clark, R. (2002). Training a talking head. *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI'02)*. IEEE Computer Society, Washington, DC, USA, (pp. 499– 505).

Colotte, V., Laprie, Y., & Bonneau, A. (2001). Perceptual experiments on enhanced and slowed down speech sentences for second language acquisition. *Proceedings of Eurospeech*. Aalborg, Denmark, (pp. 469–473).

Cosi, P., Caldognetto, E., Perin, G., Zmarich, C. (2002). Labial coarticulation modeling for realistic facial animation. *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI'02),* IEEE Computer Society, Washington, DC, USA, (pp. 505–510).

Crawford, R. (1995). Teaching voiced velar stops to profoundly deaf children using EPG - two case studies. *Clinical Linguistics & Phonetics* 9 (3), 255–269.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America* 107 (2), 989–999.

Dent, H., Gibbon, F., & Hardcastle, W. (1995). The application of electropalatography (epg) to the remediation of speech disorders in school-aged children and young adults. *European Journal of Disorders of Communication* 30, 264–277.

Engwall, O. (2003). Combining mri, ema and epg measurements in a three-dimensional tongue model. Speech Communication 41 (2-3), 303 – 329.

Engwall, O. (2012). Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. *Computer Assisted Language Learning*, 25 (1), 37 – 64.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication.* 51, 832–844.

Fadiga, L., Craighero, L., Buccino, G., Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a tms study. *European Journal of Neuroscience* 15 (2), 399–402.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of english vowels. *Journal of Phonetics* 25 (4), 437 – 470.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America* 97 (5), 3125–3134.

Grauwinkel, K., Dewitt, B., & Fagel, S. (2007). Visualization of internal articulator dynamics and its intelligibility in synthetic audiovisual speech. *Proceedings of 16th International Congress of Phonetics Sciences (ICPhS)*. Saarbrucken, Germany, 2173 – 2176.

Grauwinkel, K. & Fagel, S. (2007). Visualization of internal articulator dynamics for use in speech therapy for children with Sigmatismus Interdentalis, *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP 2007)*. Hilvarenbeek, The Netherlands, (pp. 142 – 145)

Hardcastle, W., & Gibbon, F. (1997). Electropalatography and its clinical applications. (pp. 149–193). In *Instrumental Clinical Phonetics*. London, UK, Whurr.

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America* 119 (3), 1740– 1751.

Hazan, V., & Simpson, A. (2000). The effect of cue-enhancement on consonant intelligibility in noise: speaker and listener effects. *Language and Speech* 43 (3), 273–294.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87 (1), B47 – B57.

Katz, W., Garst, D., Carter, G., McNeil, M., Fossett, T., Doyle, P., & Szuminsky, N. (2007). Treatment of an individual with aphasia and apraxia of speech using ema visually-augmented feedback. *Brain and Language* 103, 213–214.

Kröger, B.J., Graf-Bortscheller, V., Lowit, A. (2008). Two- and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders. *Proceedings of Interspeech 2008*. Brisbane, Queensland, Australia, (pp. 2639 – 2642).

Kuhl, P. (1991). Human adults and human infants show a ”perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Attention, Perception, and Psychophysics* 50, 93–107.

Lambacher, S. (1999). A CALL Tool for Improving Second Language Acquisition of English Consonants by Japanese Learners. *Computer Assisted Language Learning*. 12 (2), 137-156.

Li, M., Kambhamettu, C., & Stone, M. (2005). Automatic contour tracking in ultrasound images. *Clinical Linguistics & Phonetics* 19, 545–554.

Massaro, D., Bigler, S., Chen, T., Perlman, M., & Ouni, S. (2008). Pronunciation Training: The Role of Eye and Ear. *Proceedings of Interspeech 2008*. Brisbane, Queensland, Australia, (pp. 2623 – 2626).

Massaro, D. W., & Light, J. (2003). Read my tongue movements: Bimodal learning to perceive and produce non-native speech /r/ and /l/. *Proceedings of Interspeech*. Geneva, Switzerland, (pp 2249–2252).

Massaro, D. W. (2003). A computer-animated tutor for spoken and written language learning. *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI'03)*. ACM Press, Vancouver, British Columbia,Canada, (pp. 172 – 175).

Massaro, D. W., Liu, Y., Chen, T. H., & Perfetti, C. A. (2006). A Multilingual Embodied Conversational Agent for Tutoring Speech and Language Learning. *Proceedings Interspeech 2006, Pittsburgh, PA*, (pp. 825 – 828).

Neri, A., Mich., O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, 21(5), 393 – 408.

Pantelemidou, V., Herman, R., & Thomas, J. (2003). Efficacy of speech intervention using electropalatography with a cochlear implant user. *Clinical Linguisitcs & Phonetics* 17 (4-5), 383–392.

Pisoni, D., Lively, S., & Logan, J. (1994). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception (pp. 121 – 166). In *Development of speech perception: The transition from recognizing speech sounds to spoken words*. Cambridge, MA, MIT Press.

Probst, K., Ke, Y., & Eskenazi, M. (2002). Enhancing foreign language tutors - in search of the golden speaker. *Speech Communication* 37 (3-4), 161 – 173.

Strik, H., Neri, A., & Cucchiarini, C. (2008). Speech technology for language tutoring. *LangTech*, Rome, Italy.

Wang, L., Qian, Y., Scott, M., Chen, G., Soong, F., (2012). Computer-Assisted Audiovisual Language Learning, *Computer*, 45 (6), 38 – 47.

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia* 41 (8), 989 – 994.

Wik, P. (2009). Embodied conversational agents in computer assisted language learning. *Speech Communication*, 51 (10), 1024 – 1037.