Paper to be presented at Interspeech 2008, Brisbane, Australia, September 22-26, 2008.

# Pronunciation Training: The Role of Eye and Ear

*Dominic W. Massaro*[1], *Stephanie Bigler*[1], *Trevor Chen*[1], *Marcus Perlman*[1], *Slim Ouni* [2]

[1]Perceptual Science Lab, Department of Psychology, University of California, Santa Cruz
[2]LORIA, Campus Scientifique, BP 239, 54506 Vandoeuvre lès Nancy Cedex, France

Massaro@ucsc.edu, stephbig@gmail.com, t8chen@ucsc.edu, mperlman@ucsc.edu, Slim.Ouni@loria.fr

## Abstract

For speech perception and production of a new language, we examined whether 1) they would be more easily learned by ear and eye relative to by ear alone, and 2) whether viewing the tongue, palate, and velum during production is more beneficial for learning than a standard frontal view of the speaker. In addition, we determine whether differences in learning under these conditions are due to enhanced receptive learning from additional visual information, or to more active learning motivated by the visual presentations. Test stimuli were two similar vowels in Mandarin and two similar stop consonants in Arabic, presented in different word contexts. Participants were tested with auditory speech and were either trained 1) unimodally with just auditory speech or bimodally with both auditory and visual speech, and 2) a standard frontal view versus an inside view of the vocal tract. The visual speech was generated by the appropriate multilingual versions of Baldi [1]. The results test the effectiveness of visible speech for learning a new language. Preliminary results indicate that visible speech can contribute positively to acquiring new speech distinctions and promoting active learning.

**Index Terms**: visible speech synthesis, pronunciation training

## 1. Introduction

One of the magical realizations about speech is how seamlessly the ear instructs articulation. Non-sighted individuals acquire spoken language almost as well as sighted individuals except for a few contrasts that are auditorily difficult and visibly more prominent [2,3]. At first glance, imitating speech from sound alone seems much more remarkable than the ability to imitate our visual observation of someone's movement. We see a moving image of the action, and we then duplicate this action by tracking the changing image. Putatively, imitating action is now better understood with the discovery of mirror neurons. A mirror neuron fires both when an animal performs an action and when the animal observes the same action performed by another animal [4]. Mirror neurons could serve as basis for the imitation of movement and therefore learning specific bodily actions.

Imitation appears to be a natural act for our movements in the visual world because we have an image of the movements and therefore simply have to move our effectors to mimic the gestures in the image. Imitation in our auditory world, however, strikes us as somewhat less natural in that we don't see an obvious isomorphism between what we hear and what we speak. However, this difference between the two modalities disappears if we understand that visual perception is not really more directly linked to action any more than auditory perception of a spoken segment is. In both cases, we have to substitute our own actions for sensory, perceptual, and cognitive impressions. There is evidence suggesting that some mirror neurons are also sensitive to the sounds associated with action, responding when a particular action is performed and when the action-related sound is heard [5]. Especially germane to audiovisual pronunciation training is the observation that some of these mirror neurons fire in response only to *both* sight and sound of the action. Related to this issue is whether speech could be trained by eye as we know it can be by ear [6,7].

### 1.1. Previous Literature

Using visible speech in language learning is a relatively new enterprise, and little is known about its effectiveness. One question is how meaningful is visible speech to the naïve language learner? Certainly, instructors often illustrate various articulations to the student but little systematic tests of its effectiveness have been done. There is even less known about the role of observing tongue movements. Recent work has found that observing tongue movements might benefit perception, particularly after some experience and training [8,9]. However, it remains an open question whether this information can facilitate perception and production of speech segments in a new language.

In an early study, the effectiveness of Baldi was investigated for teaching non-native phonetic contrasts, by comparing instruction illustrating the internal articulatory processes of the oral cavity versus instruction providing just the normal view of the tutor's face [10]. Eleven Japanese speakers of English as a second language were bimodally trained under both instruction methods to identify and produce American English /r/ and /l/ in a within-subject design. Speech identification and production improved under both training methods and this learning transferred to the new words, although training with a view of the internal articulators did not show an additional benefit. There were several reasons why a difference between the two training methods might not have been observed. Participants were trained in both conditions, two of the three training stimuli had a ceiling effect, and the amount of training was relatively short. Given these methodological limitations, it would be premature to conclude that views of the internal articulatory movement do not benefit language learning, and this observation serves as another motivation for the current study.

### 1.2. Current Study

The specific question in the current study involved native English speakers learning a pair of similar speech segments in Arabic and in Mandarin. One member of the pair is basically identical to a segment in English and the other member is similar but does not occur in English. In Mandarin, the pair of segments is /i/ and /y/. In Arabic, the pair is /k/ and /q/.

In Mandarin, /i/ and /y/ are fairly similar based on their psychoacoustic properties, and their visual mouth movements are relatively more distinctive. The hypothesis to be tested is whether watching a close up of the lip and face movements of the segments during the training period will facilitate learning to perceive and produce these segments. We made close-up movies of our animated talker, Bao [11], whose articulation was specifically modeled on a real speaker.

Normal views of the segments /k/ and /q/ in Arabic look identical even though their tongue movements are significantly different. To provide informative visible speech, we therefore provided a cutaway view of the inside of the vocal track of Badr, our Arabic animated talker [12], we created an animated sagittal view of the articulation at the back of the throat, including the tongue, palate, and velum.

We also wondered whether participants engaged in the pronunciation training differently as a function of the practice conditions. For example, when able to watch a frontal view of the animated talker, participants might be more likely to physically practice pronunciation in coordination with the model. Given the sagittal view, they might even practice the gesture of a particular articulator, of which they might not otherwise be aware or only passively aware. To address this question, participants were videotaped during pronunciation training.

| Exercise | Description |
|---|---|
| Pre-Test | The student is instructed to “click on the (word)”. |
| Presentation | One image is highlighted and the student is told “this is the (word)”. Then the student is told to “show me the (word). |
| Identification | The student is instructed to “click on the (word)”. |
| Imitation | One of images is highlighted and the item is named. The student is instructed to repeat the name just said. |
| Elicitation | One of images is highlighted and the student is asked to name it. |
| Post-Test | The student is instructed to “click on the (word)”. |

Table 1. Description of the 6 sections in the training and testing task taken from Timo Vocabulary [12].

# 2. Method

This study involved the use of a recently-released application, Timo Vocabulary, which provides 8 optional exercises to test and train language skills, such as vocabulary and pronunciation [13]. The same testing and training regimen was used in both the Mandarin and Arabic studies. Each trained and tested 6 words consisting of 3 minimal pairs that differed only in the target segment (/k/ vs. /q/ in Arabic and /i/ vs/ /y/ in Mandarin). The six words in Arabic were kalb, qalb, kayd, qayd, kalla, and qalla. The six words in Mandarin were nuu, ni, yuu, yi, luu, and li.

## 2.1. Procedure

Sixteen students recruited from the University of California, Santa Cruz were tested in the design shown in Table 1. The two contrasting speech segments were tested and trained in different word contexts, with either all or a subset of the regimen of exercises shown in Table 2.

The Pre-test section involved the presentation of the images of a given lesson on the screen with Timo’s request to click on one of the items, e.g., “Please click on the word /li/.” No feedback was given, and in all conditions presentation was either auditory only in Mandarin or a simple frontal view in Arabic.

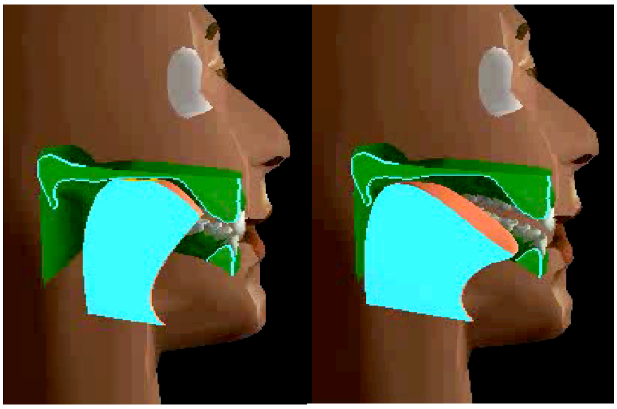

Figure 1. A frame from the movie illustrating the internal articulatory processes of /k/ in the left panel and /q/ in the right panel.

In contrast to the Pre-test, the Presentation and Identification sections involved training. In the Presentation section, one of the six items was highlighted and Timo gave the participants the pronunciation of the highlighted word, and then asked them to click on the corresponding highlighted written form. If the word was responded to correctly (initial responses were usually correct in this particular section), Timo repeated it once more before moving to the next item. If the word was responded to incorrectly, the response word was pronounced, and then the correct word was indicated and pronounced. After all six of the words were presented in this manner, the program moved on to the Identification section. In Identification, Timo gave the participants the pronunciation of the highlighted word, and then asked them to click on the written form of the target word that was pronounced but without the aid of the highlighting of the target word. As before, the student selected the appropriate word by clicking on it, and feedback was given as in the Presentation section. For both sections, in the visual conditions, pronunciation feedback was illustrated with a visual presentation of the animated talker pronouncing the word in natural citation speech, either in a close-up frontal view in Mandarin or a sagittal view in Arabic (see Figures 1 and 2).

Next, explicit pronunciation training and testing followed in the Imitation and Elicitation sections, respectively. In Imitation, the participant was asked to repeat the word when it was highlighted and said by Timo with only auditory speech. In the subsequent Elicitation exercise, the participant was asked to say the item indicated by highlighting the written word, without a spoken cue. The spoken responses from these two exercises were recorded.

The whole session ended with a Post-Test, which was identical to the Pre-Test.

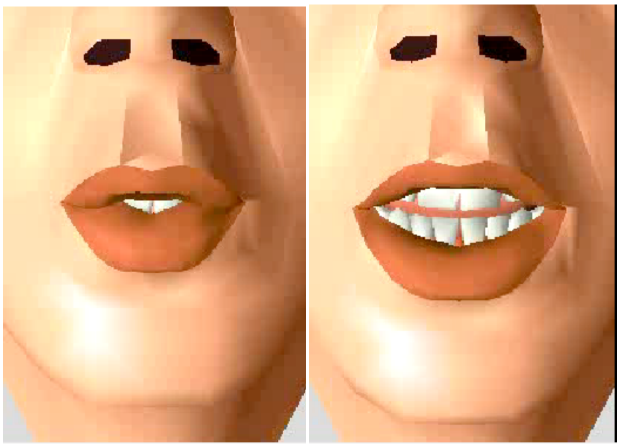

Figure 2. Frontal view of /ly/ and /li/ at the point of roughly maximum articulation of the vowel.

Table 2. The experimental design.

| | First Session | Second Session |
|---|---|---|
| Group 1 | Arabic Inside | Mandarin Frontal |
| Group 2 | Arabic Inside | Mandarin Audio |
| Group 3 | Arabic Frontal | Mandarin Frontal |
| Group 4 | Arabic Frontal | Mandarin Audio |

The experimental design involved the testing and training of both of the languages, as illustrated in Table 2. The design was a between-subjects comparison of two types of feedback: a sagittal view of Badr (the Arabic speaking incarnation of Baldi) with the tongue, palate, and velum visible as shown in Figure 1, or a frontal view of Bao (the Mandarin speaking version) pronouncing the test item in Mandarin as shown in Figure 2. The speech for the Arabic inside view was slowed down by a factor of 1/2 its normal rate. The Mandarin design was identical except that the feedback was either a close-up to the lips or was presented only in auditory speech (with a blank screen where the face would normally occur).

Each subject each participated in two sessions, one in Arabic and one in Mandarin. The order of the sessions was counterbalanced across participants as illustrated in Table 2. Subjects were video recorded as they participated in the experiment, with the camera focused on their face to record pronunciation practice during the course of the experiment.

## 3. Results

The Production results from only 8 of the 16 participants were analyzed, and the presented results must remain tentative. The perception results and the complete production results will be presented at the conference, and archived in another publication.

We evaluated to what extent the training conditions influenced production learning to determine if the visible speech condition shows a significant benefit. Production learning of the 8 participants was evaluated by analyzing the pronunciations in the Imitation and Elicitation exercises. In the initial analysis, a native speaker of Arabic (SO) and a native speaker of Mandarin (TC) rated the accuracy of pronunciation without knowledge of the training and testing conditions. The rating was made on a seven-point Likert scale indicating the quality of the pronunciation from completely wrong (1) to completely accurate (7).

The ratings of 8 participants (half of the participants tested) were completed for the initial analysis. The results were pooled across the three blocks of training to increase the reliability of the ratings, and because an initial analysis showed no effect of this variable.

For the Mandarin results, an analysis of variance was carried out with Experimental versus Control (4 subjects per group), Imitation vs. Elicitation (2 levels, within), Consonant Context (3 levels, within), and Vowel (2 levels, within). The Experimental group (M=5.46, SD=.53) had an average rating score that was 0.87 higher than the Control group (M=4.59, SD=.53), but this result was not significant, F(1,6) = 1.35, p = .29. However, there were only 4 subjects in each group, and the power was low (power = .17).

Imitation (M=5.42, SD=.38) yielded significantly higher rating scores than Elicitation (M=4.625, SD=.37), F(1,6) = 74.38, p < .001. The advantage of Imitation did not interact with Consonant [F(2,12) = 2.10, p = .17], but this Imitation advantage had a significant interaction with Vowel [F(1,6) = 11.62, p < .05]. Specifically, the Imitation advantage over Elicitation occurred for both vowels, but this advantage was greater in magnitude for vowel /i/ (a 1.10 point increase) than for vowel /y/ (only an .49 increase). This outcome is reasonable because learning should improve pronunciation of /y/ (which is a Mandarin-unique vowel) more than the non-unique vowel /i/ (which occurs in English and Chinese).

There was a significant difference between the Vowel types, F(1,6) = 9.15, p < .05. Overall, vowel /i/ (which occurs in both English and Mandarin) yielded significantly higher scores than vowel /y/ (a Mandarin unique vowel). There was no significant difference among the Consonant environments, F(2,12) = 1.52, p = .26 and no significant interaction between Consonant and Vowel, F(2,12) = .01, p = .99.

The analysis carried out for the Arabic results was exactly analogous to the Mandarin analysis. There was no significant difference between the Experimental group (M = 4.701) and the Control group (M = 4.660), F(1,6) = .02, p = .89, Elicitation and Imitation [F(1,6) = .31, p = .60], or Word-Ending (codas), F(2,12) = .19, p = .83. The only significant effect was the main effect for Consonant, F(1,6) = 512.53, p < .001. Overall, /k/ (M = 6.681, SE = .078) ratings were significantly higher than /q/ ratings (M = 2.681, SE = .230). None of the possible 2-way and 3-way interactions were significant.

In addition to performance data, we also analyzed the video recordings to observe how learners interacted with each type of training. In particular, this analysis focused on how frequently subjects engaged in imitative practice with respect to the presentation and feedback conditions. We wondered whether they were more likely to initiate imitative practice in response to the audiovisual versus audio-only conditions in the Mandarin experiment, as well as in the frontal versus sagittal views compared in Arabic.

Subjects were videotaped as they participated in the experiment, and afterwards the videotapes were coded for imitative behaviors. These included any articulatory movements of the mouth that were unambiguously an attempt to practice the target distinctions; sometimes these movements were vocalized, sometimes they were not. Imitative behaviors were added up for each participant for each language session, and considered in terms of both the number of trials during which participants engaged in imitative behavior and the total number of imitative behaviors throughout the session. One subject in the Arabic control condition was discarded because she misunderstood the instructions (her misunderstanding affected only the imitative practice analysis and not her performance in the task).

With small power, there was no significant difference in imitative behaviors between learning conditions in either language. However, for Mandarin, there appeared to be a trend in which the experimental audiovisual condition was more likely to lead to imitative behavior than the control audio only. This was true for both the number of trials with imitative behavior (control M = 6.3, SE = 10.6; experimental M = 19.5, SE = 7.7) and total number of imitative behaviors (control M = 13.0, SE = 22.8; experimental M = 37.0, SE = 29.1). To investigate this further, we looked at Arabic imitative behavior for just the imitation and elicitation activities, allowing a comparison between a more general audiovisual condition (now n = 11) and an audio only condition (still n = 4). The addition of the Arabic data amplified the original trend, again for both number of trials with imitative behavior (audio M = 6.3, SE = 10.6; audiovisual M = 20.1, SE = 6.3) and total number of imitative behaviors (audio M = 13.0, SE = 22.8; audiovisual M = 49.6, SE = 36.6). Although it cannot be completely ruled out that it is the Arabic language and not the audiovisual presentation that is driving the difference, it nevertheless appears that audiovisual presentation leads to more imitative practice by subjects.

## 4. Discussion

The initial analyses offer some support for the value of visible speech in learning a new language. In agreement with recent work, the outside of the face appears to more easily processed than a sagittal viewing illustrating the tongue, palate, and velum. These findings remain tentative until a complete analysis can be carried out.

Designing pronunciation training lessons present a few challenges. First, it is important to determine what the best views to present are and whether or not these views should be the same for all the phonemes to learn or they should be adapted to the particularity of each phoneme [14].

Learning pronunciation based on contrasts is an effective technique. Articulatory phonetic knowledge in both the learner's first language and the language being learned is necessary to choose the vocabulary to design an efficient pronunciation lesson. For phonemes that do not exist in the learner's native language, it is very helpful to start the training by a phoneme from the learner's first language that is very close from an articulatory point of view (i.e. place of articulation is very close to the target phoneme). Learners can see the differences in the articulation of the two sounds, and try to adjust their articulation based on what is seen. It is also possible to provide a more precise indication on how to go from that position to the target position.

The effectiveness of showing internal articulatory movements for pronunciation training is hard to prove and evaluate. Although we might demonstrate that showing internal articulatory movements improves learning the contrast between two phonemes, this does not necessarily mean that the learner was *consciously* imitating the tongue gesture presented by a talking head. For example, when a student watches Baldi's tongue moving back, it may not be that easy to deliberately imitate the movement with his own tongue. Yet, although difficult, it also seems plausible that such imitative skill might be acquired with sufficient practice. Issues along these lines raise questions concerning the long-term gains of this manner of pronunciation training. Future research is important if we are to evaluate the effectiveness of pronunciation training by talking heads.

## 5. Conclusions

The results of this study must remain tentative until a complete analysis is completed. It is challenging to determine the effect of such training in the long run. For example, research should address the permanency of learning effects, whether training is most effective when repeated at various intervals, and whether the temporal nature of training should take into account the modality of pronunciation practice. The present experiment is therefore only the first of many studies that are necessary to inform the field of second language acquisition.

## 6. Acknowledgements

The technology used in this research was supported in part by NSF and PHS. The authors thank Michael M. Cohen for technical help, and Jared Bernstein and Bill Rowe for productive discussions.

## 7. References

[1] Massaro, D. W. (1998). Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, Massachusetts: MIT Press.

[2] Massaro, D. W. (1987). Speech Perception by Ear and Eye: A paradigm for Psychological Inquiry. Hillsdale, N.J.: Lawrence Erlbaum Associates.

[3] Mills, A. E. (1987). Acquisition of speech sounds in the visually-handicapped child. In B. Dodd and R. Campbell (Eds). Hearing by Eye: The psychology of lip-reading. Hillsdale, N.J.: Erlbaum, pp. xx.

[4] Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. Annual Review of Neuroscience, 27, 169-192.

[5] Kohler, E., Keysers, C., Umilta, A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Representation in mirror neurons. Science, 297, 846-848.

[6] Hardison, D. (2002). Sources of variability in the perceptual training of /r/ and /l/: Interaction of adjacent vowel, word position, talkers' visual and acoustic cues. Proceedings of the 7th International Conference of Spoken Language Processing, ICSLP - 2002.

[7] Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y. I. and Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/: III. Long-term retention of new phonetic categories. Journal of the Acoustical Society of America 96, 2076–2087

[8] Grauwinkel, K., B., Dewitt, B., & Fagel, S. (2007). Visualization of Internal Articulator Dynamics and its Intelligibility in Synthetic Audiovisual Speech. in ICPhS. 2007. Saarbrucken, Germany.

[9] Tarabalka, Y., Badin, P., Elisei, F., & Bailly, G. (2007). Can you read tongue movements? Evaluation of the contribution of tongue display to speech understanding. In Conference Internationale sur 'Accessibility et les systemes de suppleance aux personnes en situation de Handicaps (ASSISTH). 2007. Toulouse - France. p. 187-193.

[10] Massaro, D. W. & Light, J (2003). Read My Tongue Movements: Bimodal Learning To Perceive And Produce Non-Native Speech /r/ and /l/. In Proceedings of Eurospeech (Interspeech), 8th European Conference on Speech Communication and Technology. Geneva, Switzerland.

[11] Liu, Y., Massaro, D. W., Chen, T. H., Chen, D., & Perfetti, C. A. (2007). Using Visual Speech for Training Chinese Pronunciation: An In-vivo Experiment. SLaTE Workshop on Speech and Language Technology in Education. ISCA Tutorial and Research Workshop. The Summit Inn, Farmington, Pennsylvania USA. October 1-3, 2007.

[12] Ouni, S., Cohen, M. M., & Massaro, D. W. (2005). Training Baldi to be multilingual: A case study for an Arabic Badr. Speech Communication, 45(2), 115-137.

[13] Timo Vocabulary (http://animatedspeech.com).

[14] Massaro, D. W., Cohen, M. M., Tabain, M., Beskow, J.,& Clark, R. (in press). Animated speech: Research progress and applications. In E. Vatiokis-Bateson, G. Bailly,& P. Perrier (Eds.), Audiovisual Speech Processing (pp.xx-xx). Cambridge, MA: MIT Press.