# Articulatory Training on Facial Movements Using the Webcam Pronunciation Mirror: A Pilot Study

*Steven W. Carruthers*

## Abstract

In this paper, I review the role of articulatory gestures involving facial movements in sound production and intelligibility, focusing on the more extreme examples, /w/, /I/, and /iy/, and report on a pilot research experiment to determine whether incorporating video and auditory feedback of the learners' own productions into pronunciation lessons would serve as a viable mode of instruction that effectively enhances pronunciation of these sounds. Three types of articulatory training were compared: training with no visual feedback, feedback from a hand mirror, and feedback through learners examining their own audio and video recorded via a webcam (the Webcam Pronunciation Mirror). Results of the pilot experiment are inconclusive, as there were only four participants. However, results support previous findings that articulatory training improves overall production and suggest that both the hand mirror and WPM are effective methods for self-monitoring, but using the WPM may be more effective for English /w/ and /iy/.

## Introduction

As a pronunciation tutor and ESL instructor, I have become interested in the role of articulatory gestures involving facial movements in sound production and intelligibility. I believe such knowledge would be helpful to English language learners (ELL), from beginning level learners working on intelligibility to advanced learners focused on accent reduction; however, I feel that I have made a few assumptions in arriving at this view. Why should learners of English be concerned with facial movements associated with pronunciation? Does training really help? What is the better mode of instruction? In this paper, I shall review the role of articulatory gestures involving facial movements in production and intelligibility, focusing on the more extreme examples, /w/, /I/, and /iy/, and report on a pilot experiment to determine whether incorporating video and auditory feedback of the learners' own productions into lessons would serve as a viable mode of instruction.

## Facial Movement in Production and Perception of English Sounds

*Facial Movements in Producing /w/ and /iy/*

*Lip rounding.* To articulate the English /w/, the lips are rounded. Leshen (1975) described lip rounding as making "a circle small enough to impede the breath and cause friction" (p. 67). But lip rounding actually involves two movements. Ladefoged and Maddieson (1996) described these as "vertical lip compression" (i.e., decreasing aperture) and "protrusion" (p. 295). MacKay (1987) stated that these movements are driven by two muscles. The *orbicularius oris* muscle is mainly responsible for the sphincter action of lip compression; for lip protrusion, the *mentalis* muscle curls the lower lip outward, also termed *eversion* (MacKay, p. 236). Figure 1 contrasts shows lip positions for English /w/. In English and other languages, these are coordinated movements, but some languages, such as Japanese and Korean, utilize the former but not the latter. Similarly to increasing protrusion, decreasing lip aperture "tends to lower all formant [peak] frequencies" (Ladefoged & Maddieson, 1996, p. 295). That is, changes in lip position change the length of the vocal tract (Mackay, 1987, p. 268), which alters the acoustics of a sound and can reduce intelligibility. For example, learners may produce /Yd/ for *wood* (Avery & Ehrlich, 2002, p. 136).

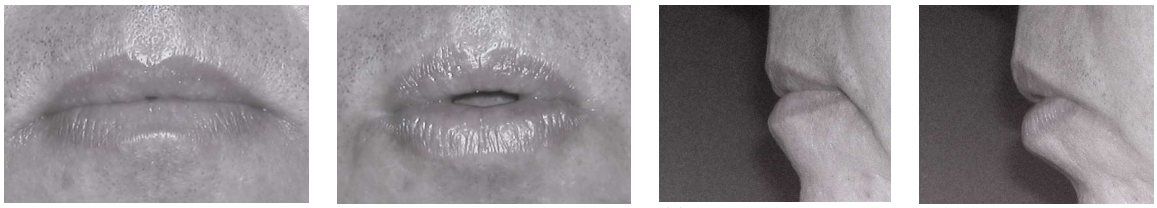


*Figure 1*. Demonstration of lip positions: (a) neutral, (b) rounded for /w/, (c) neutral, (d) and protrusion for /w/. Photo by S. Carruthers, ©2007.

*Lip spread.* Lip spread refers to the degree that the corners of the mouth are pulled back laterally. Celce-Murcia, Brinton, and Goodwin (1996) described the lips as being positioned with "extreme spreading" (p. 96), "widely spread, smiling" (p. 103), for the tense vowel /iy/. Leshen (1975) described the lips being "retracted at the corners causing them to spread in a wide narrow formation" (p. 120). In contrast, for the lax /I/, the lips are "relaxed, slightly parted and spread" (Celce-Murcia, Brinton, & Goodwin p. 103). MacKay (1987) explained that the *buccinator* muscle, with assistance from the *risorius* muscle, pulls the corners outward and upward (p. 235). The *orbicularius oris* muscle controls the height of the opening (p. 235), contributing to the amount of teeth exposure. Celce-Murcia, Brinton, and Goodwin (1996) claimed that lip spread is a "determinant of vowel quality" (p. 95); it changes the sound. Figure 2 shows a waveform for /iy/ with natural lip spread (top) and the slightly smoother form created when the sound is performed with the same tongue and jaw position but without lip spread (bottom).

*Jaw movement.* Jaw movement includes motion and position of the jaw. For /w/ in syllable-initial position, the jaw is almost closed at the onset and opens into the following vowel sound (Leshen, 1975, p. 67). In diagramming the movements of the jaw, Vatikiotis-Bateson and Ostry (1995) found that jaw movement is a significant component of pronunciation, position and direction differing "according to the consonant-vowel composition of the utterance" (p. 115). Thus, it is important that the learner be aware of the jaw movements that are coordinated with manner and place of articulation.

## *The Importance of Facial Cues for Comprehension*

Facial gestures are critical to sound perception. Indeed, these are not only what make lip reading by the hearing impaired possible, but those with normal hearing rely on facial movement as well. Some languages, such as English, are rich in visemes, "the number of 'visual categories' that are identifiable using

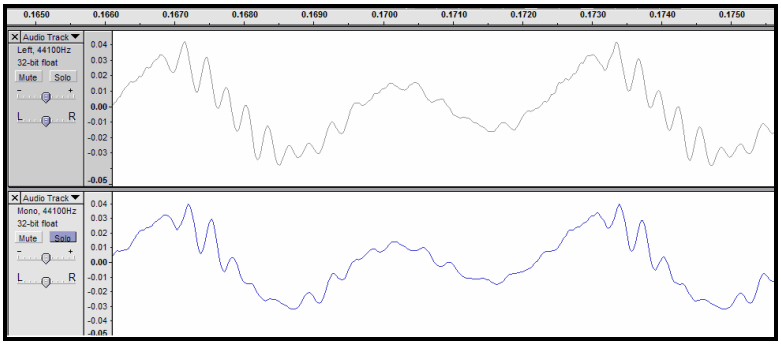


*Figure 2*. Waveform of the tense vowel /iy/ with (top) and without (bottom) natural lip spread. Image produced using Audacity Digital Audio Editor.

lipreading alone" (Hazan, Sennema, Iba, & Faulkner, 2005, p. 362). Sekiiyama et al. (2003, as cited in Hazan et al.) found that adult native speakers of English relied increasingly on visual cues, more so than Japanese speaking adults, and suggested that lesser reliance on visual cues may be due to a low degree of these in the learner's first language (p. 362). By not utilizing visual cues, English language learners are missing out on "a significant source of segmental information for speech perception" (Hazan et al., p. 362).

Poorly executed facial movements can affect intelligibility. Honda, Kurita, Kakita, and Maeda (1995) warned, "Deformations of the tongue and lips are reflected by the sound" (p. 244). Moreover, these deformations affect visual perception of the sound (p. 244). The "extreme vowels /iy/, /a/, and /uw/…have a corresponding extreme articulation" (Honda et al., p. 252) and these "appear to be exploited in signaling the vowel identity" (p. 252). Consequently, listeners use lip shape to anticipate phones. McGurk and MacDonald (1975, as cited in Traunmüller & Öhrström, 2005) described speech perception as a "bimodal process in which information in the auditory and visual modality is integrated" (p. 3). In a study by Lisker & Rossi (1992, as cited in Traunmüller & Öhrström, 2005) in which subjects identified French vowels, the results showed that visible lip rounding increased *perception* of rounding by 30% (p. 4). In a study on the perception of Swedish vowels, Traunmüller & Öhrström (2005) found that listeners benefited from visual cues "even under ideal auditory conditions" (p. 10). That is, even in a quiet environment, listeners use facial movement to help discriminate sounds. Greenberg and Arai (2004) agreed, reporting, "Manner of articulation can be used to deduce the number (and often the identity) of the 'underlying' segments, even when they are acoustically absent or exceedingly reduced" (p. 1068).

In sum, English language learners (ELL) do need to be concerned with facial movements. These are coordinated with articulation mechanisms within the mouth, and distortions in formation result in distortions of sound. Moreover, listeners rely on visual cues to predict place of articulation and discern phones. Thus, the learner is more likely to be understood when appropriate facial movements are incorporated into the production of English sounds.

## Prior Research on Articulatory Training and Its Effectiveness

### *The Effect of Explicit Perception and Articulatory Phonetic Training on Pronunciation*

In considering the worth of my own experiment, I questioned my assumption that perceptual and articulatory training improves production. Was this conclusion founded on experimental results or is it merely accepted practice? Catford and Pisoni (1970) designed an experiment in which participants were trained and measured in their ability to discern and produce foreign sounds. Participants were divided into two groups, those who received only perception training and those who primarily received explicit articulatory instructions (p. 479). They found that subjects who received "systematic training" (p. 477) on production performed better on discrimination and production tasks than those who only received discrimination training; the group receiving articulatory instruction performed "twice as well" (p. 481) as those who received only perception training. Based on their findings, Catford and Pisoni argued that although all types of training improved production, systematic articulatory training improved production and perception better than auditory training alone (p. 481). Still, perception training does have an effect on production. In an experiment to determine whether perceptual training transferred to improvement in /r/ and /l/ pronunciation in Japanese speakers of English, Akahane-Yamada, Tohkura, Bradlow, and Pisoni (1996) found that participants who received perceptual training improved significantly in perception and production in both known and novel tokens, even three and six months after training.

Beyond improved pronunciation and perception, phonological awareness, a by-product of articulatory training, may have other positive effects on the learner. Rey-

nolds (1998) generally defined phonological awareness as "metalinguistic knowledge that languages are composed of…smaller units of sound" (p. 152). Studies by Montgomery (1981, as cited in Yamada, 2004) and Griffiths and Frith (2002, as cited in Yamada) suggested a correlation between "articulatory aware-ness" (p. 96) and the reading ability of dyslexics. Yamada replicated these studies with non-dyslexic English language learners and too found a "significant correlation" (p. 100). Yamada proposed that improved articulatory awareness may result in improved reading acquisition (p. 103). Reynolds suggested that for Japanese learners of English, low phonological awareness, specifically phonemic awareness, may be linked to differences in Japanese and English writing systems, (i.e., logographic versus phonetic) (p. 153). In sum, for ELL, articulatory training does have a positive effect on pronunciation and perception, and may also contribute to the improvement of other skills such as reading.

*Traditional Methods and Strategies for Improving Pronunciation*

Traditional instruction includes having learners imitate native speech, recite tongue twisters and minimal pair drills, read passages aloud, and implement visual aids such as "charts, rods, pictures, mirrors, [and] props," (Celce-Murcia, Brinton, & Goodwin, 1996, pp. 8-10). Along the lines of mimicry, Celce-Murcia et al. recom-mended that learners mirror a speaker, live or on video, "repeating" utterances and "imitating all the speakers' gestures" (p. 310) simultaneously or immediately after the speaker. Aside from that, students are directed to dictionaries to improve word stress (Avery & Erlich, 1996, p. 216).

Learners are too often left to their own devices in improving pronunciation, receiving minimal explicit instruction. Some are advised to use a mirror to monitor their articulatory gestures (Avery & Ehrlich, 1996, p. 316; Leshen, 1975, p. 69). Using videotaping has also been recommended for its "mirror function" (Berman, 1974. p, 20). In looking for data on these strategies, I have not yet found empirical studies. Apparently, these are simply common practices, but generally, authors agree that there are potential benefits to viewing ones own production.

Although teachers and students may recognize that increased articulatory knowledge has a role in improving pronunciation, ELL may not implement such knowledge as a pronunciation strategy. To find out what strategies learners selected to improve their pronunciation, Osburne asked them to describe their mental processing through "oral protocols" (Osburne, 2003, p. 132), verbal description of the strategy implemented. Some learners commented that they used knowledge of articulatory phonetics, such as focusing on the place of articulation attending to an individual phoneme or syllable or prosodic feature, or just using imitation, and a few recalled "articulatory descriptions," (Celce-Murcia, Brinton, & Goodwin, 1996, p. 9), detailed explanation of place and manner. Osburne reported, however, that learners relied most heavily on three other strategies: mimicking the native speaker, speaking slowly, or focusing on an individual sound (p. 132).

*Computer-Based Visual Feedback for Improving Pronunciation*

*Electronic visual feedback.* Many types of computer-based feedback produce electronic visual feedback (EVF), often in the form of a waveform, spectrogram, or other visual format displaying volume, pitch, and duration. For example, Hirata (2004) studied the acquisition of pitch and durational contrasts in English speakers learning Japanese using Kay Elemetrics' CSL-Pitch Program (CSLP), which produces a waveform illustrating pitch and duration (p. 357). Optical Logo-Therapy (OLT), a more visually complex training program developed by Hatzis (1999), integrates training from an articulation instructor and feedback from sophisticated EVF software. OLT was developed to aid both native speakers with impaired hearing and ELL focusing on accent reduction. The actual software, Optical Logo-Therapy Toolkit (OLTK), is similar to other programs mentioned above in that it receives the speakers' output through a mi-

crophone and produces a two-dimensional visual display (p. 57), but it does so from a nine-point metric, detecting aspects such as lip rounding, nasalization, and front-to-back retraction (p. 84). In the sessions, the patient receives articulatory training from a therapist and several modes of feedback, including audio recording and a variety of colored displays. The therapist can also adjust the sensitivity of the program (at a more generous setting, the beginners' utterances are rated at 70%, which may lessen instances of demotivation) (Hatzis, p. 117) and award the patient with stars on the screen for improved performance (p. 118). One of the challenges encountered by Hatzis was developing a measurement that accepts variation in pronunciation and identifies those deviating from an acceptable range, while still agreeing with the "judgment of the instructor" (p. 68). Despite the motivation from working with the program and therapist, "It was difficult for the child to concentrate on the computer environment and instructions given by the therapist at the same time" (p. 118). Moreover, the program did not give corrective instruction for improving the sound production.

*Automatic speech recognition.* Programs such as *Pronto* rely on speech recognition (Dalby & Kewley-Port, 1999, p. 426). With this program, the trainer works with the learner to perfect and record individual templates, utterances that meet an articulatory standard; once the template is established, the learner can work with the program independently of the trainer. The program compares the quality of the utterance with the recorded segment and responds graphically when the learner produces it well (p. 428), but the program provides no articulatory feedback. Kawai and Hirose (1998) developed a method for measuring native and nonnative phones. Their software divides an utterance into recognized phones of English and Japanese (the broad metric produced by several speakers of the respective native languages), and "detects errors in choice of phones" (Kawai & Hirose, p. 1). This program likewise provides no articulatory instruction, but it does maintain an "accent loss gage" (p. 2) to show the learner's progress.

*Visual articulation training programs.* Some programs provide visual articulation feedback and instruction. One example is the Center for Spoken Language Understanding (CSLU) Speech Toolkit, referred to as *Baldi*, after its famous animated talking head. *Baldi* is a 3-D language tutor that models the precise movement of the lip, tongue, and jaw. (This research platform is available free for private, non-commercial use from http://speech.bme.ogi.edu/toolkit.) Massaro and Light (2004, as cited in Hazan et al., 2005) conducted training with *Baldi*, which models articulatory gestures, including facial cues, but does not assess learner output.

### *Effect of Computer-Based Visual Feedback on Pronunciation*

A number of studies using EVF and other computer-based instructional enhancements have been shown to be successful. OLTK participants showed improvement in overcoming speech impairments and reducing accent (Hatzis, 1999, p. 144), but the program was weak in specific areas. The motivating effects of the stars attracted more of the participants' attention than the meaning of the graphic displays. Hatzis reported, "[the participants] tended to get very engrossed in winning reward points and to enthusiastically pursue the same placement" (p. 147). In Hirata's (2004) study, using CSLP, participants received training on interpreting the graphic acoustic displays. They were permitted to listen to samples at any time during training and advanced to the next task when they felt they had matched the pitch contrast of the model. Those receiving training through CSLP improved in the "ability to perceive the pitch and durational patterns of the words, and the increased intelligibility of the subjects' production" (pp. 371-372).

Commercially available speech recognition software, such as *Baldi* (described earlier), has also been used in studies on perception. Massaro and Light's (2004, as cited in Hazan et al., 2005) recent study that used *Baldi* found that learners had "significantly improved the identification and production

of /l/ and r/ but the visible articulation condition did not lead to a greater improvement" (p. 363). In another study using *Baldi* as the virtual teacher for distinguishing /b/, /p/ and /v/, Hazan et al. (2005) reported that the audiovisual (AV) trained learners "showed improvement both in their use of acoustic and visual cues" (p. 368). The test group receiving only auditory training improved more on the auditory test, but "showed little evidence of improved sensitivity to visual cues" (p. 368), and the AV group improved more overall. In a second experiment on /l/ and /r/ discrimination, Hazan et al. exposed one group of participants to the *Baldi* synthetic face, a second group to a human face, and the third received auditory training only. They found that auditory and AV training were both effective, but participants exposed to natural faces did better on discrimination tasks using visual cues from *Baldi*'s synthetic face (p. 372). Overall, Hazan et al. found that the effectiveness of audiovisual training depended on the "visual distinctiveness of the contrasts" (p. 375); that is, the training was more effective with sounds that had clearly visible differences in movement or position of articulators.

### *Further Considerations Influencing Experimental Design*

In addition to general effectiveness, what the above software programs have in common is the complexity of the visual feedback. In my mind, all of them require too much expertise on the part of the user, even the instructor. As Hirata (2004) stated, "An acoustic display of speech on a computer is not easily interpretable by non-phoneticians" (p, 363). This triggers the question whether some form of visual and audio feedback that does not require advanced interpretation on the part of the learner (e.g., Hazan et. al.'s use of human faces) may better enhance acquisition of the target sounds. In determining how to design and implement a study, I have considered these issues, and I have come across three additional effects that deserve attention.

*The McGurk effect.* In studying infant speech perception, McGurk and McDonald (1976, as cited in Brancazio, 2004) discovered that auditory speech perception is affected by visual perception of the sound produced (p. 445). When an audio track playing /ba/ was synchronized with a video of someone saying /ga/, the listener perceived /da/ (p. 445). "This effect clearly indicates that information from the acoustic signal is perceptually integrated with information from the optical signal to arrive at a unified phonetic interpretation" (p. 445). Brancazio (2004), citing MacDonald and McGurk's 1978 study, explained, "The acoustic signal determines the perceived manner of articulation…and voicing, whereas the optical signal affects the perceived place of articulation" (p. 445). Thus, the listener may be influenced by facial movement and miscues (Summerfield & McGrath, 1984, as cited in Traunmüller & Öhrström, 2005, p. 4). Traunmüller and Öhrström (2005) believed, "Perception is dominated by the modality that provides the more reliable information" (p. 2). That is, in a situation in which the auditory signal is compromised, whether by background noise or unclear ELL production, the perceiver will rely more heavily on visual cues.

In an experiment in which auditory and visual signals were de-synchronized, Greenberg and Arai (2004) found that listeners were more "tolerant" (p. 1067) (i.e., indicated better comprehension) of the visual signal *preceding* the auditory than the reverse. In fact, they found that "having knowledge of the visual signals in advance of the audio stream actually improves intelligibility slightly" (Greenberg & Arai, 2004, p. 1068), supporting the idea that visual cues help the listener predict phones (Honda et al., 1995; Greenberg & Arai, 2004).

*Lexical effect.* As potentially influential as the McGurk effect is the "lexical effect" (Brancazio, 2004). Brancazio's (2004) experiments showed that "when perceivers are presented with audiovisual speech stimuli, they automatically integrate the information across modalities" (p. 251). That is, at times of "perceptual uncertainty" (p. 451), listeners are more likely to interpret a set of sounds as a word from the lexicon or identify a particular phoneme as belonging to a

sound from their first language sound inventory (p. 446). More simply, when the response to visual input would result in a word rather than a non-word or familiar phone, listeners were more likely to perceive a word or sound in their language.

*Microphone effect.* One final concern is the effect that the act of recording itself has on the sample. Wilcox (1998) observed the effect of speaking into a microphone on "pronunciation clarity" (p. 2) of Japanese learners of English in several settings such as student debates and karaoke. She believed that using a microphone may tie into a learner's kinesthetic learning style, sense of power, and desire for clarity (pp. 3-4). Amplification and recording also provide much-needed auditory feedback and help the speaker focus (p. 5). In consideration of these tendencies, it is important to note that the act of recording may encourage a better performance by the speaker.

## Research Questions

In consideration of the importance of articulation instruction and feedback and the effects described above, I designed an experiment to determine whether viewing and analyzing a video image (with audio) from a webcam, the Webcam Pronunciation Mirror (WPM), would aid the learner in improving production of sounds that have a distinct facial movement component. Unlike EVF, as described above, interpreting visual images of the face require minimal specialized training, mainly knowledge of the specific articulatory gestures that are involved in producing a sound. Because of the extreme facial movement for /w/ and /iy/, and its lax counterpart /I/ (Hazan, 2005), I have selected these as target sounds. I hypothesized that learners who received articulatory training combined with the opportunity to view and analyze their own articulatory performance would improve in their ability to perform such gestures and in the accuracy of their pronunciation. Furthermore, because the video can be reviewed, learners would be better able to attend to facial movements and resulting acoustic change than when viewing their reflection in a hand mirror. Hence, the research questions to be answered are as follows.

1. Does articulatory training result in improved overall production of English /w/ and production and differentiation of /I/ and /iy/?
2. Does using a hand mirror to monitor ones own productions result in improved overall production of English /w/ and production and differentiation of /I/ and /iy/?
3. Does viewing and analyzing ones own video-recorded productions result in improved overall production of English /w/ and production and differentiation of /I/ and /iy/?
4. Is visual feedback in the form of self-monitoring using the WPM more effective than using a hand mirror?

## Methodology

*Participants*

Volunteers were drawn from the population of English language learners at Hawai'i Pacific University, Honolulu, HI. The participants, two native speakers of Japanese and two native speakers of Korean, were randomly assigned to one of four groups, a control group and three experimental groups. For the purpose of piloting the study, Group A, Group B, Group C, and Group D each consisted of one participant, signified as A1, B1, C1, and D1, respectively. Participants in the experimental groups received articulatory instructions and feedback from an articulatory training assistant (ATA), an experienced pronunciation tutor. For this pilot experiment, I acted as both ATA and rater.

*Group A: Control (no training).* For this study, participants in Group A, the control group, received no articulatory training or feedback. Their only task was to listen to the basic stimuli (pre-recorded model utterances) and perform the speech samples at the specified time points. Given the microphone effect cited above, Group A participants were presented the basic stimuli and their performance recorded in identical circumstances to other participants. They were video-recorded with a webcam but did not view their own images or hear playback of

their sounds. This group was included as a control group to determine how much improvement might occur naturally over the time period with minimal exposure to the basic stimuli.

*Group B: With training (no visual feedback).* Participants in experimental Group B received articulatory instruction and verbal feedback from the ATA but did not receive any form of visual feedback (with the exception of natural tracking ELL may perform voluntarily when repeating model utterances). They were video-recorded with a webcam but do not see their own webcam images or hear the audio track. This group was included to determine a baseline effect of articulatory instruction alone.

*Group C: Training with mirror.* Participants in experimental Group C receive articulatory training and feedback from the ATA. They received visual feedback from a hand mirror that they use to monitor their own productions. They are video recorded but did not see their own webcam images or hear the audio track.

*Group D: Training with WPM.* In addition to articulatory instruction and feedback from the ATA, participants in experimental Group D were provided visual feedback via the WPM (i.e., playback of video/audio recordings of their own recorded utterances). They did not use a mirror to self-monitor.

*Stimuli*

All participants were presented with the basic stimuli: 20 video clips with sound, each with a one-syllable word containing the target sounds. Each word was performed three times by a native speaker model. Word Bank 1 (WB1) contained six tokens beginning with /w/: /wik/, /wIk/, /weyk/, /whΘk/ /wak/, and /wowk/. Word Bank 2 (WB2) had seven sets of tokens containing minimal pairs with the lax vowel /I/ and the tense vowel /iy/: /It/, /iyt/; /sIk/, /siyk/; /tIk/, /tiyk/; /bId/, /biyd/; /dId/, /diyd/; /wIk/, /wiyk/; and /pIt/, /piyt/. To eliminate any lexical effect advantage, words in the lexicon were chosen exclusively over nonsense words. Additionally, nasal consonants and liquids were avoided to eliminate the coloring effect of retroflex /r/ and [ᘐ] on vowels as a factor. Participants in Groups B, C, and D were verbally provided "tactile reinforcements" (Celce-Murcia, Brinton, & Goodwin, 1996, p. 310), clear articulatory descriptions of the facial cues involved in producing /w/ and producing and differentiating /I/ and /iy/. No other information, such as voicing or tongue placement, was provided.

*Equipment*

Video and audio tracks were recorded and played using a Toshiba Satellite 1905-S301 laptop computer with Pentium 4 processor and 1.0 speed USB port. Video images of model utterances and participant productions were recorded through a Logitech QuickCam Pro 3000 webcam (capable of recording 30 frames per second) connected to the USB port. The audio tracks were recorded through a Sony ECM-MS907 stereo condenser microphone connected to the computer's microphone jack. Audio and video tracks were recorded using PY Software Active Webcam version 6.8 (http://pysoft.com). Participants and evaluators listened to the model utterances and playback of participant productions through Sony MDR-7506 dynamic stereo headphones.

*Procedure*

As the ATA, I recorded the participants' productions at four intervals. For the pretest, at the beginning of the first session, Time 1 (T1), all participants listened to and watched the model utterances (the basic stimuli), and recorded both audio and video tracks via the webcam and external microphone. Each time, each participant records the token three times with a 1-second (approximately) pause between tokens. At each time point, the participant recorded a total of 18 /w/ tokens, 21 /I/ tokens, and 21 /iy/ tokens. Groups B, C, and D receive 20 minutes of instruction, practice, and feedback. When needed for training or requested by the participant, the model utterance for a specific target sound was played. After 20 minutes, at T2, Groups B, C, and D were recorded again. About one week later, at T3, all participants were played the

basic stimuli and they were recorded. Groups B, C, and D received 20 minutes of instruction, practice, and feedback. When needed for training or requested by the participant, the model utterance for a specific target sound was played. After 20 minutes, at T4, participants B1, C1, and D1 were played the model utterance one last time and recorded. (See basic timeline in Figure 3.)

| | | |
|---|---|---|
| First Session | Time 1 (T1) | All Groups (A, B, C, and D) hear/see models and are recorded |
| | Training Session (20 min.) | B, C, and D receive training |
| | Time 2 (T2) | After receiving training, B, C, and D are recorded again |
| Second Session (one week later) | Time 3 (T3) | B, C, and D hear/see models and are recorded |
| | Training Session (20 min.) | B, C, and D receive training |
| | Time 4 (T4) | All Groups hear/see models and are recorded |

*Figure 3*. Basic timeline of data collection.

The participant productions were rated by the ATA according to four specific categories: two aspects of movement, oral production, and overall production. Data were entered on a rating sheet (see Figure 4). In light of the McGurk effect, video and audio tracks were isolated and evaluated separately for assessment of individual facial movements and oral production, respectively, and viewed in concert for assessment of overall production.

Facial Movement and Oral Rating Sheet for English /w/

| Token | Lip rounding (1-5) | Exaggerated? (1=yes; 0=no) | Lip protrusion (1-5) | Exaggerated? | Oral Performance | Overall Performance |
|---|---|---|---|---|---|---|
| week | 1 | | 2 | | 3 | 2 |
| week | 3 | | 2 | | 3 | 3 |
| week | 3 | | 2 | | 3 | 3 |
| wick | 1 | | 1 | | 3 | 2 |
| wick | 3 | | 4 | 1 | 3 | 3 |
| wick | 3 | | 4 | 1 | 3 | 3 |
| wake | 4 | | 1 | | 2 | 2 |

*Figure 4*. Sample rating sheet for /w/. Similar sheets were used for other sounds.

For tokens from WB1, the initial /w/ sounds were rated on two aspects of lip rounding, compression and protrusion, with only the visual field available. Likewise, for tokens from WB2, vowel productions were rated on both lip spread and teeth exposure. The rating scale for all types of movement was 1 (no movement) to 5 (target movement). Exaggerated movements (e.g., extreme protrusion or teeth exposure) are also noted on the score sheet by the rater (see Appendix A). The productions were rated aurally, based solely on the audio portion of the track on a scale of 1 (undefined or undifferentiated) to 5 (target oral pronunciation). The utterances were also rated for overall production on a scale of 1 (undefined or undifferentiated) to 5 (target overall production), the rater listening to the audio and video tracks played simultaneously. At each time point, based on this rating system and the number of utterances, the range of possible scores was as follows: /w/ 18-90, /I/ 21-105, and /iy/ 21-105. That is, a participant receiving a rating of 1 for all utterances in a given category would achieve the lower score in the range; a participant achieving a perfect score for all utterances in a given category would achieve the higher score in the range.

## Pilot Experiment Results

Initially, a pre-recorded model production of all target sounds was played for the participants, and immediately following, at T1, all participants performed a pre-test. Subsequently, participants B1, C1, and D1 received training. All groups were tested at the conclusion of the experiment, at T4.

### *Pre-Test*

At T1, all participants were recorded producing all tokens. Each participant displayed predicted pronunciation difficulties, such as non-target lip protrusion or teeth exposure, but to varying degrees. For tokens with /w/ from WB1, the participants were rated for lip compression, lip protrusion, oral production, and overall production (see Table 1 for average scores for each category). For tokens with /I/ or /iy/ from WB2, all participants were rated for lip spread, teeth exposure, oral production, and overall production of the vowel (see Tables

2 and 3). Although at the commencement of the experiment all participants were considered to perform at the low- to high-intermediate level, there was marked variation among the groups. Specifically, C1 had better initial performance of the facial movements for /w/, lip rounding and lip protrusion, and, accordingly, better oral production and overall production. B1 and D1 were notably weaker on these. For /I/, the participants *appeared* more equal, but there was still a greater than 1.1-point difference in the average rating for overall production between lowest and highest rated participants. Additionally, B1 rated low on lip spread. For /iy/, participants scored within a range of 1.1 points for overall production. However, participants differed as much as 2-point in the average ratings for lip spread. Although the participants differed in initial rating for these features and overall production, all had room for significant improvement (see *Issues for Future Study* toward the end of this paper for further discussion).

Table 1

*Pre-Test (T1) and Post-Test (T4) Rating of Production of /w/*

| | Average rating for each category | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Lip rounding | | Lip protrusion | | Oral production | | Overall production | |
| Participant | T1 | T4 | T1 | T4 | T1 | T4 | T1 | T4 |
| A1 | 2.5 | 3.1 | 2.3 | 2.1 | 2.3 | 2.6 | 2.8 | 2.6 |
| B1 | 1.9 | 4.1 | 1.9 | 3.7 | 1.9 | 3.8 | 2.2 | 3.8 |
| C1 | 3.5 | 4.1 | 3.2 | 3.7 | 3.2 | 3.9 | 3.6 | 3.5 |
| D1 | 1.9 | 4.8 | 1.6 | 4.1 | 1.6 | 4.2 | 1.7 | 4.1 |

*Note.* The range of possible scores is 1 to 5 (target movement or production)

Table 2

*Pre-Test (T1) and Post-Test (T4) Rating of Production of /I/*

| | Average rating for each category | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Lip spread | | Teeth exposure | | Oral production | | Overall production | |
| Participant | T1 | T4 | T1 | T4 | T1 | T4 | T1 | T4 |
| A1 | 2.8 | 1.5 | 2.8 | 1.9 | 3.0 | 2.8 | 2.6 | 2.7 |
| B1 | 1.7 | 3.9 | 2.9 | 3.6 | 2.9 | 3.3 | 3.0 | 3.5 |
| C1 | 2.2 | 4.1 | 2.1 | 4.6 | 3.5 | 4.0 | 3.6 | 4.2 |
| D1 | 2.2 | 3.6 | 2.8 | 4.0 | 3.6 | 4.1 | 3.7 | 4.1 |

*Note.* The range of possible scores is 1 to 5 (target movement or production)

Table 3

*Pre-Test (T1) and Post-Test (T4) Rating of Production of /iy/*

| | Average rating for each category | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Lip spread | | Teeth exposure | | Oral production | | Overall production | |
| Participant | T1 | T4 | T1 | T4 | T1 | T4 | T1 | T4 |
| A1 | 2.2 | 1.7 | 2.1 | 1.6 | 2.6 | 2.7 | 2.9 | 2.6 |
| B1 | 1.2 | 3.6 | 1.4 | 2.8 | 2.2 | 3.2 | 2.0 | 3.1 |
| C1 | 3.2 | 4.5 | 2.5 | 4.7 | 3.1 | 3.6 | 3.1 | 3.7 |
| D1 | 2.7 | 3.6 | 2.0 | 3.4 | 2.3 | 3.1 | 2.0 | 3.5 |

*Note*. The range of possible scores is 1 to 5 (target movement or production)

*Post-Test Results*

Participants B1, C1, and D1 were tested again at the end of the first training session (T2) and one week later, at the beginning of the second training session (T3). These tests were intended to provide additional opportunity for practice and attention to the target facial features. At T4, the post-test, participant A1, who received no training or feedback, was tested for a second time. Participants B1, C1, and D1 were also tested at the end of the second training session (T4).

For tokens with /w/ from WB1, all participants were rated for lip compression, lip protrusion, oral production, and overall production. Ratings for participant A1, the control who received no training, declined slightly in overall production of /w/, but increased in the average rating for lip rounding (see Figure 1 for details). At T4, participant B1 had improved measurably, from 2.2 to 3.8, in average rating for overall production of /w/. The average rating for overall production of participant C1 declined slightly, although C1 exhibited modest gains in other categories. In overall production of /w/, D1 more than doubled his average score in all aspects rated.

For tokens with /I/ from WB2, all participants were rated for lip spread, teeth exposure, oral production, and overall production of the vowel (see Table 2). Ratings for participant A1 declined in all categories with the exception of an insignificant increase in overall production. Participant B1's rating nearly doubled for lip spread at T4 and increased slightly in all other areas, although B1 had only received articulatory training. Similarly, C1 was rated much higher in average rating of facial movements and showed a 0.6-point improvement in overall production. For D1, ratings for all categories measurably increased yet reflected only a 0.4-point increase in the average rating for overall production.

For tokens with /iy/ in WB2, all participants were rated for lip spread, teeth exposure, oral production, and overall production of the vowel (see Table 3). Participant A1's average ratings declined for all categories with the exception of an insignificant increase for oral production. Participant B1 was rated higher in all categories, nearly double in the rating of facial movements. Similarly, C1 improved 1.3 points in the rating of lip spread and 2.2 for teeth exposure, although she achieved only a 0.6-point average rating overall production. Participant D1's average rating for overall production increased 2.5 points; this change was echoed by significantly increased ratings for all other categories.

## Discussion

As there were only four participants in the pilot experiment and a single rater, who was also the researcher, any findings are merely a suggestion of what the results of a larger sample might conclude. Nevertheless, the results do support the findings of previous studies and mirror some predicted outcomes. To review the findings, I will address each of the research questions posited earlier in this paper.

1. Does articulatory training result in improved overall production of English /w/ and production and differentiation of /I/ and /iy/?

Results for the three participants, B1, C1, and D1, showed an increase in the accuracy rating for the target facial movements and an improvement in overall production. As expected, results for A1, the control, did not show a measurable difference in improvement in overall production; in fact, for some attributes of production, namely lip spread and teeth exposure for /I/, results for A1 showed a decline. The participant's initial performance might have been a better performance that could be attributed to the microphone effect or similar effect caused by the novelty of the experience. Results for participants B1, C1, and D1 showed increases in average ratings. With the exception of the accuracy ratings of overall production of C1 on tokens with /w/, generally, ratings for overall production increased. I suggest that the increase in accuracy rating of overall production for both /I/ and /iy/ results in greater differentiation of the sounds. Thus, with few exceptions, articulatory training (with or without visual feedback) did result in improved

overall production of English /w/ and production and differentiation of /I/ and /iy/.

2. Does using a hand mirror to monitor ones own productions result in improved overall production of English /w/ and production and differentiation of /I/ and /iy/?

Results for the participant in Group C, who used a hand mirror, showed an increase in the accuracy rating for the target facial movements and an improvement in overall production. Interestingly, participant C1 offered that she had never used a mirror to self-monitor production for any sound. Moreover, in Korean, her first language (L1), exposing ones tongue and teeth in speech is considered rude. Thus, in addition to the potential for L1 influence (resulting from an absence of extreme facial movement in L1 production), cultural norms may interfere with an ELL's acquisition of /w/, /I/, and /iy/, and other sounds. Nevertheless, for participant C1, using a hand mirror to self-monitor productions did result in improved overall production of English /w/ and production and differentiation of /I/ and /iy/. However, the improvement was not notably better than articulatory training alone.

3. Does viewing and analyzing ones own video-recorded productions result in improved overall production of English /w/ and production and differentiation of /I/ and /iy/?

Results for the participant in Group D, the only participant using WPM, showed an increase in the accuracy rating for the target facial movements and an improvement in overall production. The improvement was markedly better than that of other participants. Additionally, participant D1 offered unsolicited positive feedback on the new articulatory information he learned and on the experience of viewing his own productions. He stated that ordinarily he does not like to view his image, but found analyzing his own facial moments interesting and enjoyable. For participant D1, using the WPM to self-monitor productions did result in improved overall production of English /w/ and production and differentiation of /I/ and /iy/.

4. Is visual feedback in the form of self-monitoring using the WPM more effective than using a hand mirror?

Based on the small number of participants and the evaluations of a single rater, it cannot be said with any certainty that the WPM is more effective than using a hand mirror to self-monitor production. For participant D1, who received training with the WPM, results indicate that the average rating for overall production of /w/ increased from 1.7 to 4.1, or 141%. For participant C1, who received training with a hand mirror, average rating remained virtually unchanged from T1 to T4, but at T3, results showed an increase from 3.6 (at T1) to 4.6 (at T3), or a 27% increase in accuracy. It is unclear exactly why the accuracy rating for overall production of /w/ declined from T3 to T4. Participant C1 volunteered that she had practiced with the mirror before T3 and may have given a better performance at that time. Also, as noted above, there was only one rater, so any tendency indicated by the results should consider rater error as a potential factor. For production of /I/, results for D1 showed that the average rating for overall production of /w/ increased from 3.7 to 4.1, or 11%. For C1, results showed that the average rating increased from 3.6 to 4.2, or about 16%. For overall production of /iy/, results for D1 showed an increase in the accuracy rating from 2.0 to 3.5, or 60%. For C1, results showed that the accuracy rating increased from 3.1 to 3.7, or 19%. In sum, results suggest that both methods of self-monitoring are effective, but using the WPM may be more effective for /w/ and /iy/, which have more extreme movement. However, rater reliability, individual variation, or the short duration of the experiment could also account for the differences.

*Issues for Future Study*

Before embarking on additional study, I must address a few concerns. First, two technological issues must be dealt with. Although the camera and software selected are

capable of recording up to 30 frames per minute, on average, only 17 frames per minute were captured, probably due to the processing speed of the laptop. An upgraded machine with a 2.0 high-speed USB port may be needed in order to record at the highest quality available. Also, the microphone picked up a high-pitch buzz possibly caused by the CPU fan, which may affect rating. Using a longer cable and isolating the computer from the recording environment may reduce the noise.

Second, the pilot study highlighted some shortcomings in the implementation of the experiment. Obviously, the number of participants needs to be increased to about at least 20 per group. The pre-test or other standardized test should be used to determine initial ability and participants assigned to groups to ensure some level of relatively equal heterogeneity among groups. The number of raters should be at least two, and these raters would undergo norming. Moreover, the raters would not be involved in articulatory training or other aspects of the session. Furthermore, to give the learners more exposure to the stimulus and an opportunity to receive more training and feedback, the number of sessions and overall duration of the experiment need to be extended. A follow-up test might also be scheduled a month or more after the final training session to determine whether the training had a lasting effect on participant production.

Finally, in a full implementation of the study, the scope should be refined. The study would focus on either /w/ or the minimal pairs /I/ and /iy/. The number of tokens for the sound could be increased and include novel and/or contextualized items. If through subsequent research the WPM were to be found an effective method for improving productions of sounds involving significant facial movements, a variety of phones could be studied. Eventually, an experiment designed to compare the results of WPM to a form of EVF, such as a spectrogram, is needed.

## Conclusion

I earlier hypothesized that learners who receive articulatory training combined with the opportunity to view and analyze their own articulatory performance will improve in their ability to perform such gestures and in the accuracy of their pronunciation. Although there is not sufficient data at this time to support that hypothesis, the results of the study offer some evidence of the potential effectiveness of training with the WPM to that end. As described in the literature review, many modes of EVF involve viewing spectrograms or other shapes in order to interpret the quality of the learner's production. My work with participant D1 involved no complicated explanation, merely aurally describing the facial movements involved and directing of his attention with the mouse pointer toward the particular feature of focus, yet the initial results indicated significant improvement. For him, attending to facial movements and resulting acoustic change—despite an initial shyness, the presence of technical equipment, and strict procedures—appeared to be a relatively non-stressing experience. In looking for techniques and strategies to enhance acquisition of pronunciation, it is critical to look for modes that are effective while not making undue demands on the learner for complex interpretation of data or feedback. Sadly, all experimental participants commented that they had not experienced such focused and individualized feedback, even without the aid of a webcam. Thus, fundamentally, it may be important to provide learners with quality feedback on pronunciation so that they can respond to it and improve.

## References

Akahane-Yamada, R., Tohkura, Y., Bradlow, A. R., & Pisoni, D. B. (1996, October 30). Does training in speech perception modify speech production? *ICSLP 96*. Presented at the Fourth International Conference on Spoken Language Processing, Philadelphia, PA. Retrieved February 12, 2006, from http://www.asel.udel.edu/icslp/cdrom/vol2/277/a277.pdf

Anderson-Hsieh, J. (1993, Winter). Interpreting visual feedback on suprasegmentals in computer assisted pronunciation instruction. *CALICO Journal*, *11*(4), 1-22. Retrieved March 18, 2006, from http://calico.org/journalarticles/Volume11/vol11-4/Anderson-Hsieh.pdf

Avery, P., & Ehrlich, S. (1992). *Teaching American English pronunciation.* Oxford, UK: Oxford University Press.

Berman, J. P. (1974). The role played by closed-circuit television in the teaching of modern languages. *Educational Media International*, *3*, 20-22.

Brancazio, L. (2004, June). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology*, *30*(3), 445-444. Retrieved February 12, 2006, from EBSCO Business Source Premier database.

Catford, J. C., & Pisoni, D. B. (1970, November). Auditory vs. articulatory training in exotic sounds. *Modern Language Journal*, *54*(7), 477-481. Retrieved February 12, 2006, from EBSCO Communication & Mass Media Complete database.

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages.* Cambridge, UK: Cambridge University Press.

Dalby, J., & Kewley-Port, D. (1999, Fall). Explicit pronunciation training using automatic speech recognition technology. *CALICO Journal*, *16*(3), 425-445. Retrieved March 18, 2006, from http://calico.org/journalarticles/ Volume16/vol16-3/Dalby_Kewley-Port.pdf

Greenberg, S., & Arai, T. (2004, May). What are the essential cues for understanding spoken languages? *IEICE Transactions on Information and Systems*, *87*(5), 1059-1070. Retrieved Feb. 12, 2006, from http://labrosa.ee.columbia.edu/Montreal2004/papers/greenberg-arai.pdf

Hatzis, A. (1999, October). *Optical logo therapy (OLT): Computer-based audio-visual feedback using interactive visual displays for speech training* [Electronic version]. Unpublished doctoral thesis, University of Sheffield, UK.

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English [Electronic version]. *Speech Communication*, *47*, 360-378.

Hirata, Y. (2004). Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts [Electronic version]. *Computer Assisted Language Learning*, *17*(3-4), 357-376.

Honda, K., Kurita, T., Kakita, Y., & Maeda, S. (1995). Physiology of the lips and modeling of lip gestures. *Journal of Phonetics*, *23*, 243-254.

Kawai, G., & Hirose, K. (1998). A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training. Retrieved March 30, 2006, from http://www.gavo.t.u-tokyo.ac.jp/~kawai/goh/981130a.pdf

Kurita, T., & Kakita, Y. (1995). Physiology of the lips and modeling of lip gestures. *Journal of Phonetics*, *23*, 243-254.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages.* Oxford, UK: Blackwell.

Leshen, G. (1975). *Speech for the hearing-impaired child.* Tucson: University of Arizona Press.

MacKay, I. R. H. (1987). *Phonetics: The science of speech production.* Boston: Little, Brown.

Osburne, A. G. (2003). Pronunciation strategies of advanced ESOL learners. *International Review of Applied Linguistics in Language Teaching*, *41*(2), 131-143. Retrieved February 12, 2006, from EBSCO Academic Search Premier database.

Reynolds, B. (1998). Phonological awareness in EFL reading awareness. *JALT 98 Proceedings: The Proceedings of the JALT 24th Annual Conference on Language Teaching/Learning and Educational Materials Expo.* Retrieved March 31, 2006 from EBSCO ERIC database. (ED436094)

Traunmüller, H., & Öhrström, N. (2005). *Audiovisual perception of openness and lip rounding in front vowels*. Manuscript submitted for publication. Retrieved February 12, 2006, from http://www.ling.su.se/staff/hartmut/audiovis_V_perc.pdf

Vatikiotis-Bateson, E., & Ostry, D. J. (1995). An analysis of the dimensionality of jaw motion in speech. *Journal of Phonetics*, *22*, 101-117.

Yamada, J. (2004). Implications of articulatory awareness in learning literacy in English as a second language [Electronic version]. *Dyslexia*, *10*, 95-104.