The Effects of EMA-based Augmented Visual Feedback on the English Speakers' Acquisition of the Japanese Flap: a Perceptual Study

June S. Levitt[1], *William F. Katz*[2]

[1] Department of Communication Sciences & Disorders, Texas Woman's University, U.S.A.
[2] Department of Communication Sciences & Disorders, The University of Texas at Dallas, U.S.A.

jlevitt@twu.edu, wkatz@utdallas.edu

Abstract

Electromagnetic Articulography (EMA) was used to provide augmented visual feedback in the learning of non-native speech sounds. Eight adult native speakers of English were randomly assigned to one of the two training conditions: (1) conventional L2 speech production training or (2) conventional L2 speech production training with EMA-based kinematic feedback. The participants' speech was perceptually judged by six native speakers of Japanese. The results indicate that kinematic feedback with EMA facilitates the acquisition and maintenance of the Japanese flap consonant, providing superior acquisition and maintenance. The findings suggest augmented visual feedback may play an important role in adults' L2 learning.

Index Terms: L2, non-native, Japanese, flap, training.

1. Introduction

The present study investigated the role of visual information during L2 speech production learning. Clinical applications of visual information during speech production were developed in the 1960s (e.g., [1,2]). With these early studies, two approaches were introduced, including (1) displays of speech acoustics [1] and (2) displays of speech articulators [2]. The methods and instrumentation used in these early studies required learners to first build skills to interpret the visual displays before they could use them to improve their speech.

Recent technological advancements have yielded instrumentation that directly displays speech articulator movement visually. These technologies include X-ray microbeam, electropalatography (EPG), glossometry, ultrasound, Magnetic Resonance Imaging (MRI), and electromagnetic articulography (EMA) systems.

EMA, the technology used in the present investigation, tracks speech articulator movements in the mid-sagittal plane. EMA employs low-field strength, alternating electromagnetic fields to track small sensors attached to the articulators [3]. The sensor (magnetic) is a small insulated coil attached to articulatory structures at midline, using dental adhesive. As the alternating magnetic fields pass through the sensor, they induce an alternating signal that is tracked by a computer. EMA has been used in clinical speech remediation studies with individuals with apraxia of speech (AOS) following stroke [4,5].

Studies of the role of kinematic feedback in speech have been largely driven by the needs of individuals with communication disorders. In contrast, a relatively small number of studies have addressed basic questions concerning the speech motor learning mechanisms of healthy individuals. The present research seeks to establish a baseline for the effect of visual feedback on speech by investigating non-native sounds as stimuli. That is, visual feedback was used to improve foreign accents in L2 learning. Specifically, eight adult monolingual speakers of American English were trained to produce the Japanese flap /ɽ/, either in (1) conventional L2 speech production training or (2) conventional L2 speech production training with EMA-based visual feedback. The participants' probe data were audio-recorded through three baseline sessions, eight training sessions, and two 4-week post training sessions.

To date, an acoustic analysis of the participants' improvement was conducted in terms of flap duration [6]. The results showed a noticeably greater training effect in the EMA condition than in the non-EMA condition, measured by Cohen's *d*-prime figures of 13.91 and 4.78, respectively.

The present study conducted a perceptual analysis to determine whether these acoustic measures correspond to the perceptual judgments of accentedness. For this purpose, six native speakers of Japanese judged the participants' speech as being either (1) "Japanese flap" (on target) or (2) "not Japanese flap" (off target).

2. Method

2.1. Participants

The participants were eight female monolingual speakers of American English with a minimum of 12 years of formal education (mean age=28, $\underline{SD}$=9.35, range=22-48) and no reported history of neurological, language, learning, speech, or hearing deficits. To rule out individuals who have difficulties perceiving non-native speech sounds, all participants took a brief screening test prior to participation in the present study. This was a same-different (AX) task with CV-syllables, in which two of the three pre-recorded syllables (American English /dɑ/, /lɑ/ and Japanese flap /ɽɑ/) were presented as pairs, via the Direct RT Research software program. The pre-recorded sounds were produced by a single talker (JL), amplitude balanced, and acoustically matched in intonation contour and formant frequencies over their vowel portions. Participants adjusted the intensity of the stimuli to a comfortable listening level. All six possible combinations of the three sounds were repeated five times, for a total of 30 pairs. Participants were required to score 80% or higher on this task to qualify to participate in the training study. The participants' mean score was 96% correct discrimination ($\underline{SD}$=5.8%, range=83-100%), and all participants passed this screening procedure.

Participants were randomly assigned to one of two training conditions: (1) non-EMA condition: conventional L2 training and (2) EMA condition: conventional L2 training with EMA-based kinematic feedback on tongue tip position.

2.2. Stimuli

Participants were trained to produce 16 disyllable words that include Japanese flap /ɽ/ in the word-initial or word-medial position. All words have primary stress on the first syllable. Although some of these words are meaningful in Japanese,

they were effectively used as non-words for this non-native speech sound training.

Table 1. *The non-words used for training*

	/ɾ/:first syllable		/ɾ/:second syllable	
Set 1	/ˈɾakɯ/	/ˈɾekɯ/	/ˈhaɾa/	/ˈhɛɾɛ/
	/ˈɾata/	/ˈɾeta/	/ˈtoɾa/	/ˈtoɾɛ/
Set 2	/ˈɾikɯ/	/ˈɾokɯ/	/ˈhiɾi/	/ˈhoɾo/
	/ˈɾita/	/ˈɾota/	/ˈtoɾi/	/ˈtoɾo/

2.3. Procedure

Participants completed eight training sessions preceded by three baseline recording sessions and followed by two post-training recording sessions. The post-training sessions were conducted four weeks after the last training session. Training sessions were conducted two times per week in the Speech Production laboratory at the University of Texas at Dallas. Participants were trained individually. The present experiment used a single-subject ABA design. A single-subject design was selected because each subject's baseline data served as her control to assess the changes that occurred during the training and retention (post-training) phases.

Participants practiced eight words (Set 1) during the first four training sessions. Next, they practiced the other eight words (Set 2) during the second series of four training sessions. The words were practiced in randomized order within each training session.

Three repetitions of the stimuli were recorded as probes during the three baseline sessions, eight training sessions, and two post-training sessions. Audio stimuli used to elicit these productions were pre-recorded by a female native speaker of Japanese (JL). Participants in both non-EMA and EMA conditions recorded the probe data in the same manner. Each participant was seated in front of a computer monitor and fitted with a headphone microphone. The stimuli were randomized for each recording session and presented one at a time using Microsoft PowerPoint software (v. 2003). Prior to the recording, participants adjusted the volume of the stimuli to a comfortable listening level. A subject-controlled stimulus procedure [7] was used. Participants produced each Japanese stimulus immediately following a pre-recorded audio model, as prompted by the program at their own pace. The participant's speech was digitally audio-recorded to the disk through a headphone microphone (Labtec 342), using WaveSurfer speech analysis software (v. 1.8.5) and Audacity Digital Audio Editor (v. 1.2.6) at a sample rate of 16,000 Hz.

The training procedure was similar for both non-EMA and EMA conditions. Participants practiced each of the speech sounds 20 times in a blocked fashion, resulting in a total of 160 productions (8 words x 20 repetitions) during a training session. The instructor (JL) modeled each word and asked participants to practice the word, five repetitions at a time. The instructor provided summarized verbal feedback regarding the articulatory positions and the timing of tongue movements (KP: knowledge of performance) after every five repetitions. Occasional general comments were made to keep participants motivated.

In the EMA condition, a Carstens AG100 electromagnetic articulography system was used. Participants wore an EMA helmet (during training phases) with a small sensor (2 mm high x 2.5 mm wide x 3.7 mm long) attached approximately one cm posterior to the tongue tip, using a biocompatible adhesive. The sensor was connected to the analog unit by fine wires. The sensor moved as the participant produced speech, and the tongue tip position was displayed on the main computer screen in real time and on a secondary monitor placed in front of the participant.

Prior to each training session, a general area in the mouth was marked on the computer screen with a circle. The participant next produced the word "daddy," to identify the alveolar ridge region corresponding to correct production of the flap /ɾ/. The target zone for the Japanese flap was then set at a region marked slightly posterior to that area, using a mouse-controlled drawing tool. These two circles and the tongue trace were shown to the participant.

3. Data Analyses

Perceptual analyses were conducted using the probe data of each participant recorded throughout sessions. Six individuals unfamiliar with the data served as listeners. Listeners were required to be phonetically-trained native speakers of Japanese with no reported history of neurological, language, speech, or hearing deficits. Six students of Sophia University served as listeners (mean age=27.67, SD=8.48, range=21-44; three male and three female listeners). All six listeners speak standard Japanese without any distinctive regional accent.

The speech materials consisted of 2,496 tokens of the American participants' speech recorded during the baseline, training, and post-training sessions (8 words x 3 repetitions x 13 sessions x 8 participants). Eight of 16 words used for the training were selected for the perceptual analyses. To randomly assign these stimuli across the six listeners for this perceptual experiment, a *Latin Square* technique was used. Each Japanese listener judged one of the three repetitions of all American participants' speech throughout baseline, training, and post-training phases. The stimuli were blocked into two files by syllable position: one each for word-initial flap and word-medial flap tokens. In all, each listener judged a total of 832 stimuli (2 blocks x 1 repetition x 4 words x 13 sessions x 8 talkers). The stimuli were randomized across talker, training phase, training condition, and vowel/syllable contexts.

The perceptual experiment was conducted in a sound-proof booth at Sophia University in Tokyo, Japan. Listeners participated in this perceptual experiment one at a time. Stimuli were presented using DirectRT Research software (v. 2004), one at a time. The procedure was self-paced, with listeners proceeding to the next slide by pressing the space bar. Each slide displayed the target word in Japanese and played its associated sound clip. Listeners indicated whether the presented sound was a "Japanese flap (on target)" or "not Japanese flap (off target)" by pressing one of two keys marked on the PC keyboard. Following a procedure described in previous studies [8,9], each stimulus was presented twice. After completion of the practice session, the main battery of 832 stimuli was completed in two blocks. Listeners were given a short break between blocks. Critically, listeners were blinded with respect to the training phases and conditions of training (EMA vs. non-EMA).

To examine intra-rater (test-retest) reliability, each listener judged approximately 5% of randomly selected probes twice (*n*=40 tokens). This procedure was performed to ensure listeners' consistency in judging the stimuli at different times. The raw agreement score (% agreement) and Cohen's *kappa* were computed to assess the reliability. The average raw agreement score was 94.6% (range=90-97.5%), and the average *kappa* value was 0.82 (range=0.61-0.94), indicating an "almost perfect" level of agreement [10]. Each listener also judged approximately 5% of the tokens randomly extracted from the files assigned to other listeners (*n*=40 tokens). To examine inter-rater reliability, Cohen's *kappa* values for token-by-token agreement between two listeners (i.e., L1 vs. L2, L1 vs.

L3 and so on) were computed. The average *kappa* value was 0.68 (range=0.48-0.93), indicating a "substantial" level of between-listener agreement [10]. The average raw agreement score was 88. 3% (range=80-97.5%).

4. Results

The percentages of the words judged as "Japanese flap" are plotted for baseline, training, and post-training phases for EMA and non-EMA conditions in Figures 1 and 2. Because the listeners judged word-initial and -medial flaps in separate blocks, these data are plotted separately.
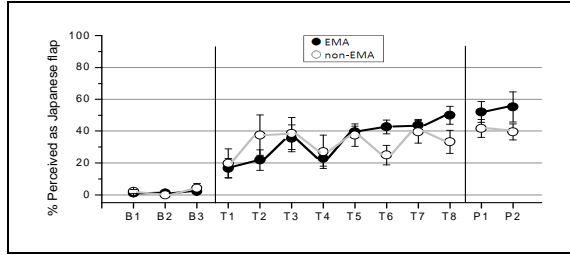

Figure 1. Perceptual judgment for word-initial flap stimuli (/ˈɾɑtɑ, ˈɾɛtɑ, ˈɾitɑ, ˈɾotɑ/).
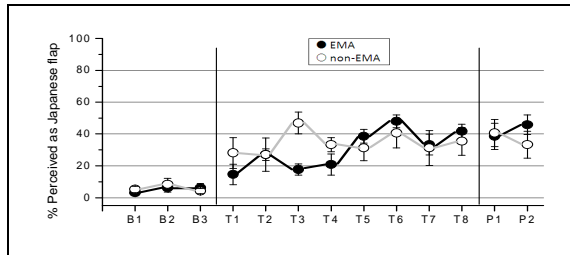

Figure 2. Perceptual judgment for word- medial flap stimuli (/ˈhɑɾɑ, ˈhɛɾɛ, ˈhiɾi,' hoɾo/).

In Figures 1 and 2, sessions (baseline/training/post-training) are indicated on the *x* axis, and the tokens perceived as Japanese flap (%) are shown on the *y* axis. The error bars show standard error. As shown in Figures 1 and 2, the results suggest that the American participants in both non-EMA and EMA training conditions improved their production of the Japanese flap consonant. Participants in the EMA condition showed smaller variability within the subjects in the training phase than the participants in the non-EMA condition. Figure 1 shows that subjects in the EMA condition also reached noticeably higher retention levels than subjects in the non-EMA condition (for words with word-initial flaps). As both conditions started at similar levels of baseline, a greater training effect is suggested for the EMA condition. This was not the case, however, for words with a flap in word-medial position (Figure 2). For these words, the Japanese flap consonant productions were perceived by native speakers approximately 60% and 40% by post-training, for the participants in the EMA condition and in the non-EMA condition, respectively.

4.1. Effect size

For these perceptual data, Cohen's *d* could not be used to determine effect sizes due to a lack of variance in both baseline and post-training phases. As an alternative, Percentage of Non-overlapping Data (PND) [11] analyses were conducted.

PND is a non-parametric method used to examine a training effect or treatment efficacy by computing the percent of training data points that do not overlap with the baseline data points [12]. PND is computed by counting the number of data points that are higher than any one of the baseline figures, and dividing it by the number of the training or retention phases. PND scores range from 0% to 100%, and a criteria for interpretation was outlined by [11]. According to these authors, a score below 50% indicates "unreliable treatments," a score between 50-70% is "questionable effectiveness," between 70-90% is "fairly effective," and a score greater than 90% is "highly effective."

Visual inspection of Figures 1 and 2 suggests that perceptual scores show different trends for the former and the latter halves of the training sessions (i.e., T1-T4 and T5-T8). Therefore, three separate sets of PND scores were computed: (1) between baseline and the first half of the training phase (T1-T4), (2) between the baseline and the second half of the training phase (T5-T8), and (3) between the baseline and the post-training phase (P1 and P2).

Table 2. *Effect Size: PND: Baseline vs. T1-4*

Condition	Participant	PND (%)
EMA	1	25
	2	59
	3	44
	4	56
$\bar{x}$		46
non-EMA	5	56
	6	72
	7	69
	8	63
$\bar{x}$		65

Table 3. *Effect Size: PND: Baseline vs. T5-8*

Condition	Participant	PND (%)
EMA	1	91
	2	72
	3	81
	4	75
$\bar{x}$		80
non-EMA	5	56
	6	91
	7	59
	8	59
$\bar{x}$		66

Table 4. *Effect Size: PND: Baseline vs. Post-training*

Condition	Participant	PND (%)
EMA	1	100
	2	63
	3	81
	4	75
$\bar{x}$		80
non-EMA	5	56
	6	100
	7	56
	8	56
$\bar{x}$		67

Table 2 lists the first set of PND scores (i.e., between the three baseline and the first four training sessions). The mean score for the EMA condition was 46% (range=25-59%), and

that of the non-EMA condition was 65% (range=56-72%). These data suggest that a more immediate effect of training was evident for the non-EMA condition than the EMA condition. By the above-mentioned criteria, however, training by this point had not yet reliably occurred (the effects indicate "unreliable" and "questionable effectiveness," respectively).

Table 3 lists the PND scores for the latter half of the training sessions (i.e., between the three baseline and the last four training sessions). The mean score for the EMA condition was 80% (range=72-91%: "fairly effective" to "highly effective"), and that for the non-EMA condition was 66% (range=56-91%: "questionable effectiveness" to "highly effective"). These data suggest that, with the exception of non-EMA participant #6, the improvement of the non-EMA condition leveled off during the latter half of the training sessions, while that of the EMA condition increased.

Table 4 lists the PND scores between the baseline and the post-training recording sessions. The mean PND score for EMA was 80% (range=63-100%), and that for non-EMA was 67% (range=56-100%), both in the "questionable effectiveness" to "highly effective" range. These perceptual data suggest a greater retention for the EMA condition than the non-EMA condition, with the EMA mean effect size (.80) being "fairly effective" by PND standards.

5. Discussion

The results suggest that American participants in both the EMA and non-EMA conditions improved their Japanese flap consonant production over the course of training, as judged by native Japanese listeners. Participants in the non-EMA condition showed more immediate improvement during the earlier training sessions, although this improvement levelled off after the first half of the training. In contrast, the participants in the EMA condition monotonically improved their Japanese flap productions over the course of training. A possible explanation for these different patterns between the two training conditions is that the auditory/visual integration during EMA training (i.e., eye to tongue coordination) may have required some time to establish before triggering "superadditive" benefits [13]. Also, the benefits of the EMA condition were maintained through the retention phase, as found in the post-training scores.

The results of the perceptual analyses suggest that the EMA-based visual feedback can facilitate the normal adult speakers' learning of non-native speech sounds, as indexed by judgment scores of native Japanese listeners. This result is broadly consistent with the results of the acoustical analyses of duration of the present training data [6].

However, this conclusion must be weighed with caution for several reasons. First, the experimental design was not double-blinded. Although the instructor (JL) followed a written procedure designed to guide both EMA and non-EMA conditions in the same manner, instructor bias could have affected the results. This type of methodological shortcoming is unfortunately rather common with most of the speech training/remediation studies to date [9,13,14,15,16]. Nevertheless, the current experiment should be replicated to ensure validity.

Second, because this study used an ABA design, it did not explore skill transfer and generalization, as could be examined with a multiple baseline design. Also, because the training was concluded for a fixed number of sessions (i.e. eight sessions), participants were not trained to mastery criteria. Without further training, it is not known whether the speech production of the participants in EMA condition could have reached the native Japanese speakers' norm.

Despite these limitations, the present data have advanced our understanding of the effects of augmented visual kinematic feedback on healthy adults' non-native speech production learning. If future studies examine varied feedback presentation schedules, a wider variety of motor target types, and series of other factors known to affect motor learning, they will provide important information for a deeper understanding of non-native speech training from a motor learning perspective.

7. References

[1] Boone, D. (1966). Modification of the voices of deaf children. *Volta Revie*, 68, 686-692.

[2] Bridges, C. C. (1964). An appratus for the visual display of speech sounds. *American Journal of Psychology*, 77, 301-303.

[3] Schönle, P., Grabe, K., Wenig, P., Höhne, J., Schrader, J., & Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language,* 31, 26-35.

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

[5] McNeil, M. R., Fossett, T. R. D., Katz, W. F., Garst, D., Carter, G., Szuminsky, N., & Doyle, P. J. (2007). Effects of on-line kinematic feedback treatment for apraxia of speech. *Brain and Language*, 103, 223-225.

[6] Levitt, J. S. & Katz, W. F. (2008). Augmented visual feedback in second language learning: Training Japanese post-alveolar flaps to American English speakers. *Proceedings of Meetings on Acoustics, Acoustical Society of America*, Vol. 2, 060002, 1-13.

[7] Logan, J. S., & Pruitt, J. S. (1995). Methodological issues in training listeners to perceive non-native phonemes. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 351-377). Baltimore, MD: York Press.

[8] Pruitt, J. S., Jenkins, J. J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *Journal of Acoustical Society of America*, Vol. 119 (3), 1684-1696.

[9] Ertmer, D. J., & Maki, J. E. (2000). A comparison of speech training methods with deaf adolescents: spectrographic versus noninstrumental instruction. *Journal of Speech,* Language, and Hearing Research, Vol. 43, 1509-1523.

[10] Landis, J. R. and Koch G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159-174.

[11] Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single subject research methodology: Methodology and validation. *Remedial and Special Education,* 8, 24-33.

[12] Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification,* Vol. 28, No. 2, 234-246.

[13] Massaro, D. W., & Light J. (2004). Using visible speech to train perception and production of speech for individuals with hearing loss. *Journal of Speech, Language, and Hearing Research*, Vol. 47, 304-320.

[14] Ballard, K. J., Maas, E., & Robin, D. A. (2007) Treating control of voicing in apraxia of speech with variable practice. *Aphasiology*, 21(12), 1195-1217.

[15] Fletcher, S. G., Dagenais, P. A., & Critz-Crosby, P. (1991). Teaching consonants to profoundly hearing-impaired speakers using palatometry. *Journal of Speech, Language, and Hearing Research,* Vol. 34, 929-942.

[16] Grillo, E. U. (1997). *The role of sensory feedback on the coordination dynamics of a limb and a voice task.* Unpublished doctoral dissertation, University of Pittsburgh.