Investigation of the effect of articulatory-based second language production learning on speech perception

Atsuo Suemitsu, Takayuki Ito, Jianwu Dang, Mark Tiede

Buenos Aires – 5 to 9 September, 2016
Acoustics for the 21st Century…

PROCEEDINGS of the 22nd International Congress on Acoustics

Speech Communication: Paper ICA2016-526

Investigation of the effect of articulatory-based second language production learning on speech perception

Atsuo Suemitsu[a], Takayuki Ito[b], Jianwu Dang[c], Mark Tiede[d]

[a] Sapporo University of Health Sciences, Japan, sue@sapporo-hokeniryou-u.ac.jp
[b] CNRS, GIPSA-Lab, France, takayuki.ito@gipsa-lab.grenoble-inp.fr
[c] JAIST, Japan, jdang@jaist.ac.jp
[d] Haskins Laboratories, United States, tiede@haskins.yale.edu

Abstract

The effect of second language production training on perception has been previously explored, but it remains unclear whether such training by itself influences the perception of speech sounds. In previous work participants heard the correct pronunciation of the target while simultaneously undergoing production training, making it unclear what component of improvement was due to the production training alone. In the current study we have therefore modified our electromagnetic articulometer-based training system, which provides estimates of learner-specific head-corrected tongue positions for a target utterance in real time, to eliminate simultaneous presentation of audio stimuli. Japanese learners of the American English vowel /æ/ performed ABX perceptual testing on this vowel before and after the visually presented articulatory-based pronunciation training. We examined whether or not the production-driven pronunciation improvement also induces a change in the perception of the second language sounds.

Keywords: Electromagnetic Articulometer, EMA, Pronunciation, L2 Training

Investigation of the effect of articulatory-based second language production learning on speech perception

1 Introduction

The links between the production and perception of speech has been investigated such as the Motor Theory of Speech Perception [1]. The speech motor system may be involved in both generation and decode of the gestures of speech. In this work we probe this hypothesis by examining whether appropriate production training transfers to perception in an L2 learning task. While many researchers have previously explored the effects of production training on perception, the results are clouded by participants hearing the correct pronunciation of the target while simultaneously undergoing production training, making it unclear what component of improvement was due to the production training alone.

Recently we have developed an electromagnetic articulometer (EMA)-based training system, which presents articulatory positions in real time together with the target articulatory positions estimated from speaker acoustics and articulatory data [2]. Using this system, short-term training with no audio cue for the American English (AE) vowel /æ/ was shown to improve Japanese learners' pronunciation of this non-native sound.

In this study, we examined whether the production-driven pronunciation improvement can also improve the perception of the second language (L2) sounds. For this purpose, three Japanese learners of the American English vowel /æ/ were tested using an ABX perceptual paradigm on this vowel before and after the visually presented articulatory-based pronunciation training. Preliminary results suggest that acquisition of a motor command of the target sound does transfer to improved perception.

2 EMA-based, real-time, visual feedback system

Figure 1 shows an overview of the developed system. The 3D EMA system (AG500, Carstens Medizinelektronik) tracks positions of sensors glued to the speech articulators and reference points. The sensors were placed as shown in Figure 2: the tongue tip (TT), blade (TB), and dorsum (TD), lower incisors (LI), upper lip (UL), and lower lip (LL), together with reference sensors on the upper incisors, nasion, and mastoid processes tracked to compensate for head movement (all midsagitally placed apart from the mastoid references). Articulatory movement and speech acoustics are digitized at sampling rates of 200 Hz and 16 kHz, respectively. For visualization, sensor position data were transformed to a coordinate system based on each participant's occlusal plane and corrected for head movement. Acquisition and transformation of the data was repeated every 50 ms, that is, the articulatory presentation was updated at a 20 Hz rate. No perceptually apparent latency between sensor motion and its visualization was observed. Figure 3 shows an example of the real-time visual feedback display. The tongue

surface contour was obtained using cubic spline interpolation through the three tongue sensor positions.

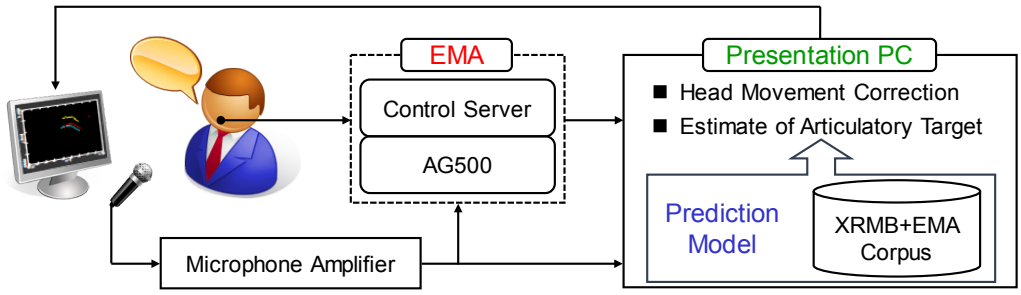


Figure 1: Overview of the developed system

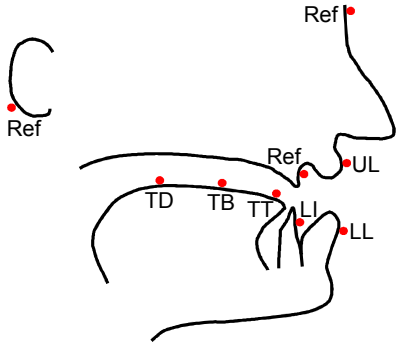


Figure 2: Sensor coil locations

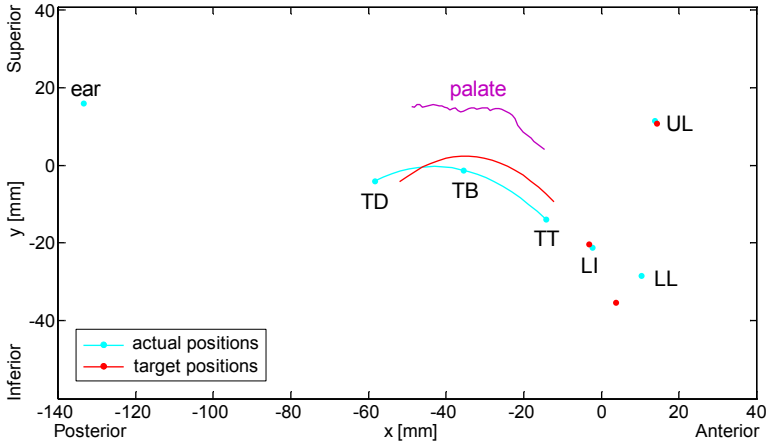


Figure 3: Example of the real-time visual feedback display

The speaker-specific target articulatory positions for /æ/ were estimated using a multiple linear regression model trained on native AE speakers, using as predictors each learner's acoustic and kinematic productions of AE vowels /a/, /i/, and /u/, which overlap reasonably well with Japanese /a/, /i/, and /ɯ/. This approach enabled us to provide an estimate of /æ/ target position compatible with each learner's particular vocal tract shape.

The prediction model was constructed using the acoustic and kinematic data of 49 native AE speakers from the University of Wisconsin X-ray microbeam (XRMB) speech production corpus [3] (19 males and 21 females), augmented by EMA data (5 males and 4 females) collected at Haskins Laboratories. Specifically, twelve models (6 articulatory attachment points (TT, TB, TD, LI, UL, and LL) x 2 coordinates (x,y)) were built using stepwise selection of these predictors: the first (F1) and second formant (F2) frequencies; posterior/anterior (X) and inferior/superior (Y) coordinate values of TT, TB, TD, LI, UL, and LL for AE vowels /a/, /i/, and /u/; the area of the triangle defined by these vowels in the F1xF2 plane; and the area of the triangle defined by the TDxy positions associated with these vowels. Here, TB was estimated by calculating a midpoint between T2 (mid-ventral) and T3 (mid-dorsal) pellets from the Wisconsin XRMB corpus, and TT and TD corresponded to T1 (ventral) and T4 (dorsal) pellets, respectively.

3 Methods

3.1 Participants

The participants were 3 male monolingual native speakers of Japanese in their 20s, with no self-reported hearing deficits or speech disorders. All participants had received some English instruction in school, but had no overseas living experience. The Ethical Committee of the Japan Advanced Institute of Science and Technology (JAIST) approved the experimental procedures, and all participants provided written informed consent.

3.2 Production session

The experimental sequence for the production session consisted of four phases: preparation, pre-test, training, and post-test.

In the preparation phase, sensors were attached to the speech articulators as shown in Figure 2, palatal shape was measured for real-time visualization, and the occlusal plane was sampled to provide a consistent coordinate system during real-time display. In addition, the articulatory positions and speech data for the participant's production of the sustained Japanese vowels /a/, /i/, and /ɯ/ were collected for estimation of that speaker's articulatory /æ/ position.

In the pre-test and post-test phase, the participants were asked to produce the target words (‘back’, ‘sad’, ‘had’, or the vowel /æ/). Each word was presented 5 times in randomized order, on a computer screen with no audio stimuli. Each participant, fitted with 10 EMA sensors attached as shown in Figure 2, was seated in front of the computer screen in a quiet room at JAIST. Articulatory recordings were collected using an EMA (AG500) at 200 Hz synchronized with concurrent audio recorded with a directional microphone (NTG-3, RODE) at 16 kHz.

In the training phase, the participant was first asked to fit his tongue contour and the displayed UL, LL, and LI positions to the estimated target positions without producing speech sounds for about 5 minutes, in order to facilitate learning motor control of this novel articulation without phonation. Then, the participant practiced the production of the vowel /æ/ after matching his/her articulation to the target, as elicited by visual stimulus presentation. This task was repeated 20 times.

To assess the effect of pronunciation training, the F1 and F2 of the vowel /æ/ were obtained from the acoustic recordings from the pre- and post-test phases. Formant estimates were obtained using a 14th order LPC analysis using a 0.97 pre-emphasis factor, and a Hanning window length of 25 ms with 15 ms overlap between frames. Using the corresponding spectrogram to verify spectral stability, five frames taken from the approximate center of each utterance were analyzed and the resulting formants averaged. Additionally, taking the human auditory system and the anisotropy of the F1xF2 frequency plane into account, all formant values were converted to the Equivalent Rectangular Bandwidth (ERB) units [4]. Then the Euclidean distances between the produced sounds and a reference sound, established as the median of the native AE speaker productions, in the F1xF2 ERB space, was calculated.

3.3 Perception session

In the perception session, the perception test was conducted before and after the production session in the same room as the production session. The perception test was based on the ABX design. The stimuli for the perception test consisted of 75 minimal pairs contrasted /æ/-/ɑ/, /æ/-/ʌ/, or /ɑ/-/ʌ/ in CVC words. These utterances were collected from a male native AE speaker. Each stimulus was presented 4 times in counterbalanced order. In a trial, the participants were presented with a sequence of three sounds (e.g., ABA or ABB) through headphones (HDA 200, Sennheiser) and then asked to indicate whether the third sound was the same as the first (A) or second (B) sounds.

To investigate the effect of pronunciation training on perception, correction rates were obtained for three conditions: '/æ/-/ɑ/', '/æ/-/ʌ/', and '/ɑ/-/ʌ/' in the perception test before and after the production session.

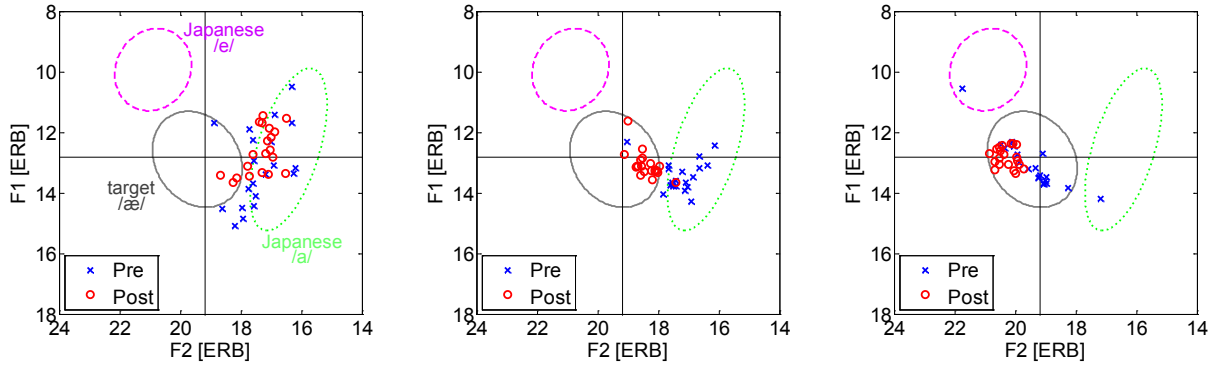
4 Results

Figure 4 shows the distribution of the produced /æ/ sounds in the F1 x F2 space at pre- (cross) and post-test (circle) phases for all participants. Central gray ellipses represent the 95% confidence limits for the /æ/ distribution for the male AE speakers obtained from the XRMB and EMA corpus, and vertical and horizontal lines indicate the medians of the F1 and F2 values of the native /æ/ distribution, respectively, and dashed and dotted ellipses are the 95% confidence limits for the Japanese /e/ and /a/ distributions from acoustic data (21 male Japanese speakers) collected at JAIST, respectively. Figure 5 compares the average acoustic distance from target at the pre- and post-test for each participant, where error bars represent standard error of the mean (n=20).

From these figures, it can be seen for participants A and B, that the produced /æ/ sounds after training were distributed closer to the center of the native /æ/ distribution. For participant C, the acoustic distance slightly increased, but almost all of his utterances were within 95% confidence limits for the native /æ/ distribution at pre-test as well as post-test phases. Therefore, C's pronunciation did not worsen. There may be little room for improvement.

Figure 6 shows correction rates for all participants for three conditions '/æ/-/ɑ/', '/æ/-/ʌ/', and '/ɑ/-/ʌ/' before and after pronunciation training. For participants A and B, the correction rates for

conditions '/æ/-/ɑ/' and '/æ/-/ʌ/' are higher after than before training, but those for condition '/ɑ/-/ʌ/' are lower after than before training. These results show that the ability to identify the vowel /æ/ increased. Thus, pronunciation improvement by the visually presented articulatory-based pronunciation training can improve the perception of the second language sounds. For participant C, the correction rate for condition '/æ/-/ɑ/' is higher after than before training, but that for condition '/æ/-/ʌ/' is lower after than before training. These results show that for this participant the ability to identify the vowel /æ/ was not improved. This suggests that pronunciation improvement of the L2 sounds is needed to facilitate the ability to identify them. In other words, acquisition of appropriate motor commands for producing L2 sounds might be important for correctly perceiving them.



(a) Participant A (b) Participant B (c) Participant C

Figure 4: Scatterplots of utterances in the F1 x F2 space

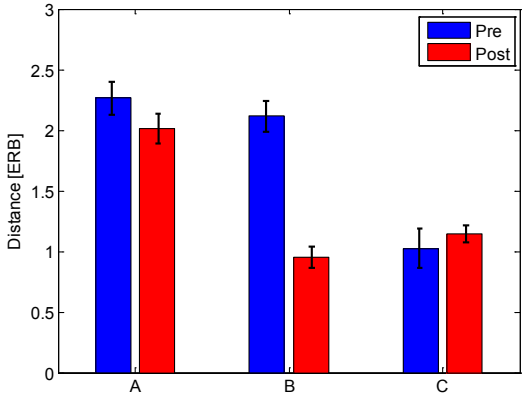


Figure 5: Euclidean distance between the produced and reference sounds at the pre- and post-test phases.

5 Conclusions

We have investigated the effect of articulatory-based second language production learning on speech perception. Japanese learners of the American English vowel /æ/ were tested on perception of the vowel before and after the visually presented articulatory-based pronunciation

training. Experimental results showed that pronunciation improvement transferred to a perception improvement. This provides support for the link between speech production and perception, suggesting that the motor commands for producing L2 sounds can be utilized in perceiving them.

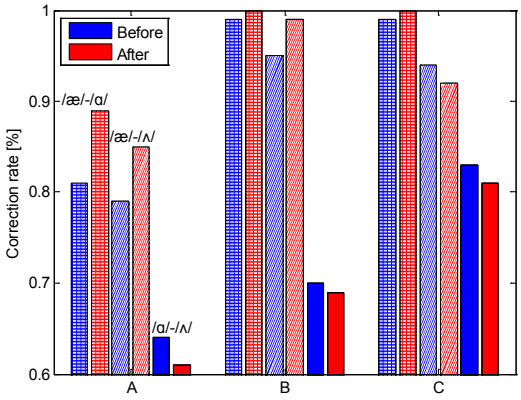


Figure 6: Correction rates for three conditions before and after pronunciation training

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number JP25240026. A special thanks to Jeff Moore for your support.

References

[1] Liberman, A. M.; Cooper, F. S.; Shankweiler, D. P.; Studdert-Kennedy, M. Perception of the speech code, *Psychological Review*, Vol 74, 1967, pp 431-461.

[2] Suemitsu, A.; Dang, J.; Ito, T.; Tiede, M. A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning, *JASA Express Letters*, Vol 138(4), 2015, pp EL382-387.

[3] Westbury, J. *X-Ray Microbeam Speech Production Database User's Handbook*. University of Wisconsin, Madison (USA), 1994.

[4] Glasberg, B.; Moore, B. Derivation of auditory filter shapes from notched-noise data. Hearing Research, Vol 47, 1990, 103–138.