A cross-language study of compensatory response to formant-shifted feedback

Takashi Mitsuya [1], *Ewen N. MacDonald* [1], *David W. Purcell* [2], *and Kevin G. Munhall* [1]

[1] Queen's University, Kingston, ON, Canada
[2] University of Western Ontario, London, ON, Canada
takashi.mitsuya@queensu.ca

Abstract

Learning new sounds in a second language requires the acquisition of new motor routines and new sensorimotor planning systems needed to ensure coordination. Auditory feedback is an important part of the planning and control system required for fluent speech production. ESL vowel production was studied using a real-time formant perturbation technique to modify auditory feedback. Three groups of subjects (Native English, Japanese ESL, and Korean ESL) produced tokens of the English word "Head" with the first formant (F1) shifted either up or down in frequency. When F1 was shifted up, compensations by Native English speakers were larger than either ESL group. The F1 lowering perturbations produced more similar compensations by all three groups. This direction asymmetry in magnitude of compensation is discussed in relation to differences in native vowel inventories and the nature of auditory feedback processing.

Index Terms: Speech production, Vowel learning

1. Introduction

Learning new vowels is thought to involve new speech motor skills as well as the perceptual learning of new category boundaries. In this study we examine the role of auditory feedback in shaping ESL articulation patterns for vowels. Recent studies have shown that talkers use the sound of their own voice as part of a regulatory system that coordinates vocal pitch [1, 2, 3], loudness [4] timing [5] and spectral details of speech sounds [6, 7, 8]

In previous studies using real time perturbation of formant frequencies, talkers compensated by producing formants shifted in frequency opposite to the direction of perturbation [6, 7, 8]. This rapid compensation demonstrates that the speech motor control system is attempting to correct a perceived error in the auditory feedback talkers receive. In the present study, this role of perception in producing speech is explored by comparing compensation results from three groups of talkers speaking in English.

Previous work has demonstrated that while native Japanese speakers can discriminate between the English vowels /ɛ/ and /æ/ [9], native Korean speakers have difficulty perceiving the contrast [10, 11]. The consequences of this perceptual difference in categorization were explored by comparing how native English, native Japanese, and native Korean speakers compensated in response to real time formant perturbation of the English vowel /ɛ/.

2. Method

2.1. Subjects

Forty-four students at Queen's University participated in the current study. Eighteen of them were female undergraduate students whose first language was Canadian English. Another 18 were female Japanese ESL students. For the Korean speakers, data collection is ongoing but data from 8 Korean ESL students (5 males and 3 females) are presented here. The majority of the ESL students of both language groups had just arrived in Canada with little exposure to an English-speaking culture prior to the experiment. All of the subjects had normal threshold with a range of 500 – 4000 Hz (< 20dB HL), and none reported a history of language or speech impairments.

2.2. Equipment

The equipment used was the same as that previously reported in Purcell and Munhall [7]. The talkers were recorded using a headset microphone (Shure WH20), amplified using a Tucker-Davis Technologies MA3 microphone amplifier and low-pass filtered at a cutoff frequency of 4500 Hz (Frequency Devices 901 filter). This signal was digitized at 10 kHz sampling rate. When altered auditory feedback was desired, the signal was filtered in real time to produce formant shifts using a National Instruments PXI-8176 controller. For both normal and altered auditory feedback, noise was added using a Madsen Midimate 622 audiometer and the voice signal and noise were presented to the subject using headphones (Sennheiser HD 265) at 85 and 50 dB SPL respectively. The manipulation of auditory feedback was achieved by filtering the voice in real-time. Voicing was detected using a statistical amplitude threshold technique. Formants in the speech were determined using an iterative Burg algorithm [12]. The formant estimates were used to calculate the filter coefficients so that a pair of spectral zeroes was positioned at the location of the existing formant frequency and a pair of spectral poles was positioned at the desired frequency of the new formant. The formant frequency estimate and new filter coefficients were computed every 900 μs.

2.3. Screening procedure and model order estimation

Subjects were tested individually in a sound attenuated room (Industrial Acoustics Company). They sat in front of a computer monitor where a target word was presented. Subjects were instructed to say the visually prompted word with consistent loudness and pitch. Prior to the experimental conditions, a screening procedure was also run in order to determine the best model order to estimate their formant structure for the perturbation in the experimental conditions. During the screening procedure, the subjects produced seven English monophthongs /i, ɪ, e, ɛ, æ, ɑ, u/ five times in an /hVd/ contexts. These vowels were randomly presented. For each talker, a model order (the number of coefficients in the autoregression analysis), which ranged in value from 8 to 12, was selected to achieve the most stable and smooth tracking of formants near the trained vowel (/ɛ/ in "head"). The heuristic used was based on minimum variance in formant frequency over a 25 ms segment midway through the vowel.

Along with the English vowels, the ESL subjects also produced their native language's vowels in an /hV/ context in order to examine their native vowel space. For Japanese subjects, 5 Japanese monophthongs /i, ɐ, a, o, ɯ/ were produced, and for Korean subjects, 8 Korean monophthongs /i, e, ɛ, a, o, ʌ, u, ɨ/ were produced (/ø/ was not included because this vowel is often pronounced as /we/). To ensure that the ESL subjects produced the vowels of their native language, the visual prompts were presented in their native orthography (i.e., Hiragana for Japanese, and Hangul for Korean).

2.4. Procedure and experimental conditions

The experiment consisted of two sessions. In one session, the F1 perturbation was positive (F1-Up) while in the other the perturbation was negative (F1-Down). Each session consisted of four phases over the course of which talkers were prompted to produce the English word "head" (/hɛd/) 140 times. In the first phase, Baseline (first 20 utterances), subjects received normal feedback. In the second phase, Ramp (utterances 21-70), F1 was perturbed with the magnitude of the perturbation increasing by 4 Hz with each utterances, resulting in a 200 Hz shift by utterance 70. In the third phase, Hold (utterances 71-90), the F1 perturbation of 200 Hz was held constant. Finally, in the Return phase (utterances 91-140), the feedback was returned to normal. The order in which talkers received the F1-Up and F1-Down sessions was counterbalanced. A schematic of the experiment is shown in Figure 1. Thus, in the Hold phase, if talkers did not compensate, the altered feedback in the F1-Up condition sounded somewhat like "hid" (/hɪd/), whereas in the F1-Down condition, it sounded like "had" (/hæd/).

With our Korean subjects, the F2 was also perturbed. In the F1-up condition, the F2 was shifted downward with a 5 Hz increment during the Ramp phase, resulting a maximum of 250 Hz perturbation, whereas in the F1-down condition, the F2 was shifted upward by the same increment. With the F1/F2 perturbation, the categorical shift of the vowels between the actual production and the altered feedback was accentuated[1].

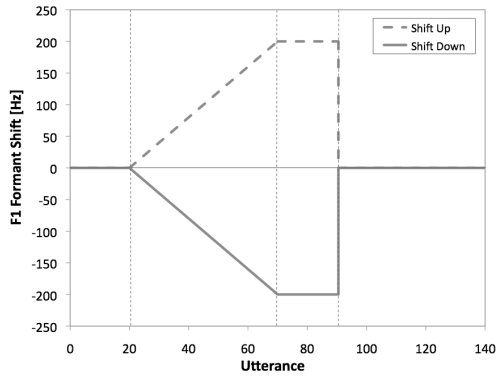

Figure 1: *Schematic of procedure for experiment.*

3. Results

3.1. Vowel space comparisons

Since the current study was designed to compare the compensatory pattern across the language groups while producing the English vowel /ɛ/, it was important to compare the acoustics of this vowel as produced by each group. ESL speakers' acoustic properties of the vowel were compared with those of native English speakers. Moreover, we examined whether the vowel was assimilated toward one of their native vowels.

Vowel spaces were estimated based on utterances collected from the screener procedure. In normal production, there is a large gender difference in the average F1 and F2 of each vowel. To account for this when estimating the vowel space, the F1 and F2 data from the Korean male speakers were normalized using the process of Nordström & Lindblom [13]. These normalized data were then pooled with data collected from the female Korean talkers. The English vowels as produced by an average individual native English speaker are plotted in Figure 2. Similarly, the Japanese and Korean vowels by average individual native speakers of each language are plotted in Figure 3. Finally, the English vowels, as produced by average individual native Japanese and Korean speakers are plotted in Figure 4. In Figures 2-4, the center of each ellipse represents the mean F1/F2 frequency for that vowel, while the solid and dashed ellipses represent one and two standard deviations respectively.

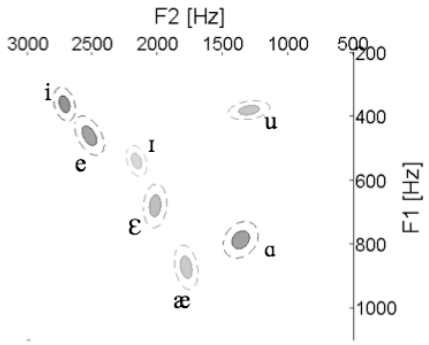

Figure 2: *English vowel space of native English speakers in an /hVd/ context*

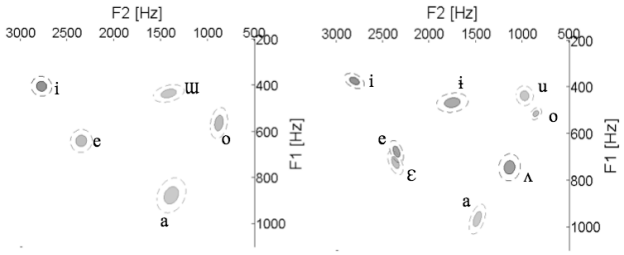

Figure 3: *Vowel spaces of native Japanese (left) and Korean (right) speakers produced in an /hV/ context.*

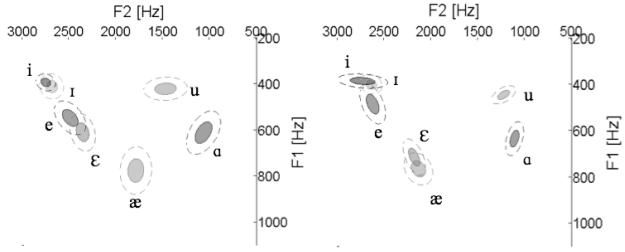

Figure 4: *English vowel spaces by Japanese (left) and Korean (right) speakers produced in an /hVd/ context.*

Both Japanese and Korean vowels were less densely distributed compared to English vowels, especially in the front

area where there are fewer vowels (our Korean speakers did not differentiate /e/ and /ɛ/). Moreover, the acoustical distance between the high front vowel /i/ to open mid vowel /a/ in both languages was larger because their /a/ had slightly higher F1 and lower F2 values, compared to the English corner vowel /æ/. This makes the sparse vowel distribution even sparser.

For English vowels produced by our ESL speakers, the most notable difference is how /ɛ, æ/ were produced. Among Japanese, there was a clear productive distinction between the vowels, however, our Korean speakers did not differentiate the two vowels. This finding is consistent with what has been reported in the literature [14] and suggests the two vowels were perceptually categorized into a single category by our Korean speakers.

Over all, the English /ɛ/ produced by our ESL speakers were very similar to the mid open front vowel of their native language, such that, among Japanese, /ɛ/ was assimilated toward /ę/, and among Korean, it was assimilated toward their version of /ɛ/. It is arguable that they were simply producing their native vowel instead of trying to produce /ɛ/, however, the larger variability with the English /ɛ/ suggests these vowels are represented as two distinct speech gestures.

3.2. Magnitude of compensation

In order to examine the difference of compensatory response for the perturbed vowel, we examined each individual's average change in production during the Hold phase. This is the phase in which the maximum perturbation was applied. For each individual, the change in production was quantified by subtracting the average F1 of the last 15 utterances of the Baseline phase from the average F1 of the 20 utterances from the Hold phase.

As the perturbations in the F1-Up and F1-Down condition were opposite in sign, the average results of each condition were analyzed separately.

The results for the F1-Up condition are plotted in Figure 5. As can be seen in the figure, the native English talkers altered their production of F1 more than the ESL talkers. This observation was confirmed by an ANOVA ($F[1, 2] = 11.095$, $p < 0.01$, $\eta^2 = .35$). Post hoc analyses with Bonferroni correction revealed that the Japanese and Korean speakers compensated significantly less than the native English speakers ($p < 0.05$), but the ESL groups did not differ from each other ($p > 0.05$).

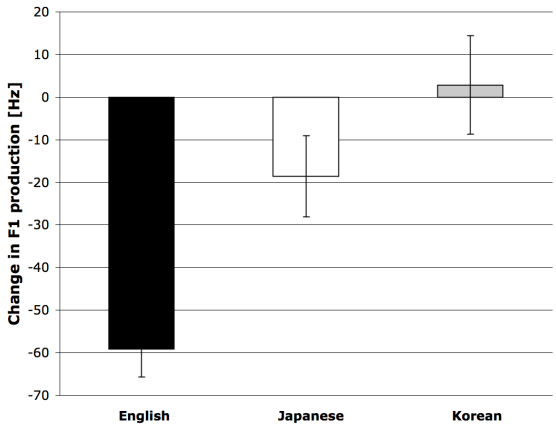

Figure 5: *The change in production during the Hold phase in the F1-Up condition.*

The results for the F1-Down condition are plotted in Figure 6. Overall, the three language groups exhibited more similarity than the F1-Up condition. While the English group had the largest change in F1, all three groups showed significant levels of compensation. An ANOVA showed no significant difference between the groups ($F[1,2] = 0.882$, $p = 0.42$, $\eta^2 = .04$).

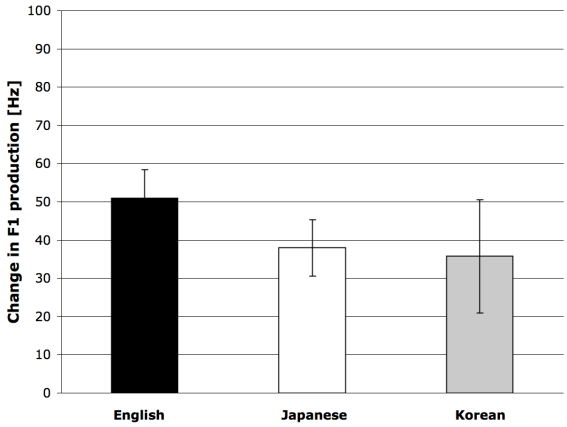

Figure 6: *The change in production during the Hold phase in the F1-Down condition.*

4. Discussion

In the current study, cross language differences in the compensatory response to altered auditory feedback were examined. Talkers of three different native languages were recruited (English, Japanese, and Korean) and received perturbed feedback in which the formant frequency of F1 was either increased or decreased while talkers produced English utterances. When the F1 frequency of the acoustic feedback was decreased (F1-Down), all three groups exhibited a similar compensatory response. However, when the F1 frequency of the acoustic feedback was increased (F1-Up), the English talkers exhibited a larger response than that of the other two groups. Given the experimental paradigm, there are three possible explanations for the observed between group differences: differences in the perceptual categorization of adjacent English vowels, differences in the familiarity with producing English utterances, and differences in the distribution of vowels in talker's native language vowel space.

Previous work has demonstrated that while native Japanese speakers can discriminate between the English vowels /ɛ/ and /æ/, native Korean speakers have difficulty perceiving the contrast [10, 11, 14]. Thus, in the F1-Up condition of the experiment where the altered feedback of the English vowel /ɛ/ would be similar to /æ/, native Japanese talkers would categorize the feedback as different from the intended vowel but native Korean talkers would not. The difference in compensatory response observed between the native Japanese and Korean speakers in the F1-Up condition showed this trend but the difference was not statistically significant. However, this small effect, even if reliable with more subjects, does not explain the differences between the two ESL groups and English speakers.

It is possible that the larger compensatory response of native English talkers in the F1-Up condition is a result of these talkers having more experience in producing English utterances compared to the talkers in the ESL groups. As Flege [15] and Best [16] postulate, when talkers gain experience with a new language, they are thought to establish precise

perceptual or gestural targets, which would lead to smaller variance in normal production (i.e., the precision of control increases). Based on experience differences in producing English utterances, the native English talkers should exhibit more precise control compared to the talkers in the two ESL groups. As auditory feedback is involved with maintaining the precision of control, one would expect the native English talkers to exhibit a larger compensatory response compared to talkers in the two ESL groups in both the F1-Up and F1-Down conditions. In the F1-Down condition, a trend of smaller ESL compensatory response was observed but was not statistically significant. However, even if reliable with more subjects, the modest difference in compensatory response between English and the ESL groups in F1-Down is much smaller than the differences observed in the F1-Up condition. Thus, the between group differences in familiarity with producing English vowels does not explain the asymmetry of the between group differences in the compensatory response.

When compared to English, the vowel spaces of both Korean and Japanese are less densely packed, particularly in the open frontal region. Because their production of the English vowel /ɛ/ was assimilated to their native counterpart, it is possible that the altered feedback was perceived and operated in their native vowel space. If so, the modified feedback of the F1-Up condition would have been located in the L1 acoustic space where there is no vowel representation, and the feedback might have been perceived as a tolerable instance of the vowel the speakers were producing. Hence, no compensation was needed. The F1-Down feedback, however, would come close to the vowel categorical boundary of the L1 high front vowel. If this is the case, then one would expect a talker to exhibit a similar compensatory response to altered feedback when producing utterances in L1 and L2. Further investigation is needed to fully disentangle these accounts.

5. Acknowledgements

This research was supported by the National Institute of Deafness and Communicative Disorders Grant No. DC-08092 and the Natural Sciences and Engineering Research Council of Canada.

6. Note

[1]Korean speakers' data were collected under a different experimental protocol. We are including their data here to increase the number of native vowel spaces considered.

7. References

[1] Kawahara, H. (1995). "Hearing voice: transformed auditory feedback effects on voice pitch control," Proceedings of the international joint conference on artificial intelligence: workshop on computational auditory scene analysis, pp. 143–148. Montreal, Canada.

[2] Burnett, T. A., Senner, J. E. & Larson, C. R. (1997). "Voice F0 responses to pitch-shifted auditory feedback: a preliminary study," Journal of Voice, 11, 202–211.

[3] Jones, J. A. & Munhall, K. G. (2000) "Perceptual calibration of F0 production: evidence from feedback perturbation," J. Acoust. Soc. Am. 108, 1246–1251.

[4] Bauer, J. J., Mittal, J., Larson, C.R., & Hain, T.C. (2006). "Vocal responses to unanticipated perturbations in voice loudness feedback: an automatic mechanism for stabilizing voice amplitude," J. Acoust. Soc. Am. 119, 2363–2371.

[5] Mitsuya, T., MacDonald, E. N, & Munhall, K. G. (2009). "Auditory feedback and articulatory timing," J. Acoust. Soc. Am. 264, 2223.

[6] Houde, J. F., & Jordan, M. I. (1998). "Sensorimotor adaptation in speech production," Science 279, 1213–1216.

[7] Purcell, D. W., & Munhall, K. G. (2006). "Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation," J. Acoust. Soc. Am. 120, 966–977.

[8] Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception," J. Acoust. Soc. Am. 122, 2306–2319.

[9] Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S. A., & Nishi, K. (2001). "Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners," J. Acoust. Soc. Am. 109, 1691–1704.

[10] Nishi, K., & Kewley-Port, D. (2008). "Nonnative speech perception training using vowel subsets: Effects of vowels in sets and order of training," Journal of Speech, Language, and Hearing Research, 51, 1480- 1493.

[11] Ingram, J. C. L., & Park, S.-G. (1996). "Inter-language vowel perception and production by Korean and Japanese listeners," Proceedings of the Fourth International Conference on Spoken Language Processing, Philadelphia, PA.

[12] Orfandidis, S. J. (1988). "Optimum Signal Processing, An introduction," New York, NY: MacMillan.

[13] Nordström, P. E. & Lindblom, B. (1975). "A normalization procedure for vowel formant data," Paper 212 at the international congress of phonetic sciences in Leeds, UK, August.

[14] Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels," J. Phonetics, 25, 437–470.

[15] Flege, J. E. (1995). "Second-language speech learning: Theory, findings, and problems," In W. Strange (Ed.), Speech perception and linguistic experience: Theoretical and methodological issues (pp. 229 –273). Timonium, MD: York.

[16] Best, C. T. (1994). "The emergence of native-language phonological influence in infants: A perceptual assimilation model," In J. Goodman & H. Nusbaum (Eds.), The development of speech perception: The transition from speech sounds to spoken words (p. 167-224). Cambridge, MA: MIT.