THE LINK BETWEEN PERCEPTION AND PRODUCTION OF ENGLISH VOWEL CONTRASTS IN NON-NATIVE SPEAKERS

Chih-Chao Chang[1], Chun-Ting Chien[2], Yu-An Lu[1]

National Yang Ming Chiao Tung University[1], National Taipei University of Nursing and Health Sciences[2]
c.c.willchang@gmail.com; alixchien.ct@gmail.com; yuanlu@nycu.edu.tw

ABSTRACT

This study investigates the link between L1-Mandarin speakers' perception and production of English /i, ɪ/ (high) and /ɛ, æ/ (mid) vowels. Through perceptual categorization and phonetic imitation experiments, we found that L2 speakers have difficulty integrating spectral and durational cues for English vowel categorization, as indicated by either a trade-off or a null correlation between the perceptual weights of the two cues at the individual level. For the perception-production link, the imitation of English high vowels was constrained by the reliance on durational cues in perception and mediated by L1 categories at the phonological level. In contrast, the perception-production link for English mid vowels was only established at the acoustic-phonetic level for vowel quality but not for vowel duration. Our findings provide a look into the level of abstraction of L2 sound categories in non-native speakers across different phonetic aspects and vowel contrasts.

Keywords: L2 perception-production link, English tense-lax vowels, imitation, vowel categorization

1. INTRODUCTION

While some second-language (L2) learning models propose a strong connection between speech perception and production [4], other models do not make a clear association between the two [1, 2]. Most L2 studies explored the two modalities separately in different tasks, yielding discordant claims about the perception-production link. To rectify this, the current study uses perceptual categorization and phonetic imitation experiments to investigate the perception and production of English /i, ɪ/ and /ɛ, æ/ vowel contrasts by L1-Mandarin speakers and the perception-production link at the individual level through engaging the two modalities simultaneously.

In English, tense vowels have more peripheral F1/F2 values and longer durations than lax vowels [6]. Perceptually, however, English native speakers rely more on spectral than durational cues for categorizing high (/i, ɪ/) and mid (/ɛ, æ/) vowels [10]. Furthermore, the perceptual weights of spectral and durational cues are positively correlated at the individual level in mid vowels [11], whereas the correlation is underexplored in high vowels. In addition, [11] used a phonetic imitation task to examine the perception-production link for mid vowels in English native speaker. They found less faithful spectral imitation in ambiguous tokens than in unambiguous /ɛ/ and /æ/ tokens, suggesting that native phonological categories constrain phonetic imitation. In contrast, duration, the secondary cue in L1-English speakers' vowel categorization, was imitated gradiently. At the individual level, the speakers with a stronger reliance on spectral cues in perception exhibited a greater degree of duration imitation in production, implying that some English native speakers were more sensitive to acoustic-phonetic information in general, which was reflected in their duration imitation and showed the perception-production link at the acoustic-phonetic level.

In contrast, previous studies showed that L1-Mandarin speakers heavily rely on duration to distinguish English high vowels /i, ɪ/, while the role of duration is less clear for mid vowels /ɛ, æ/ [7, 8]. This highlights the fact that L1-Mandarin speakers use acoustic-phonetic cues to categorize English vowels differently from native speakers. The correlation between spectral and duration cue weights in L1-Mandarin speakers, however, remains unclear. Regarding the L2 perception-production link, most studies correlate perception and production measures through separate tasks and thus lead to mixed conclusions [5, 8, 17]. In addition, in terms of phonology, Mandarin lacks tense-lax vowel contrasts but has /i/ and [ɛ] (an allophone of /a/) in its inventory; also, Mandarin contrasts vowels in spectral but not durational aspects [13]. These indicate a possible modulation of L1 linguistic knowledge in the L2 perception-production link [5, 8, 17].

Given the differences in phonetic cue weightings and native phonological categories compared to English native speakers, we aim to investigate the perception-production link in L1-Mandarin speakers for different L2 English vowel contrasts.

2. METHODS

2.1. Participants

Twenty-five L1-Mandarin speakers were recruited. However, 9 and 6 participants were excluded from the

data of high and mid vowels, respectively, due to their mislabeling responses in the perception experiment (reversed pattern of identification), resulting in a sample of 16 participants for high vowels (8F, 8M; *M* = 24.3) and 19 participants for mid vowels (11F, 8M; *M* = 24.5). All participants were native Mandarin speakers and had learned English as a second language. Their self-reported English proficiency was 4.38 and 4.33 on a scale of 1 to 7 (1 being poor and 7 being proficient), and the onset of L2 exposure was 6.25 and 5.94 years of age, for the high vowel group and mid vowel group, respectively. None of the participants reported speech or hearing impairments.

2.2. Stimuli

2.2.1. Perceptual categorization

Bit-beat [bɪt-bit] and *bet-bat* [bɛt-bæt] continua were created from natural tokens produced by a phonetically trained male native English speaker in a sound-attenuated booth. To resynthesize the tokens, we first used Tandem-Straight [9] to manipulate spectral F1/F2 values of the vowels and created a seven-step continuum from /bɪt/ to /bit/ and from /bɛt/ to /bæt/. Next, the PSOLA technique was implemented in Praat [3] on each spectral step to manipulate the vowel durations in seven equal steps (40 ms/step) from 100 ms to 340 ms. In total, 49 stimuli (7 spectral steps x 7 duration steps) were created for each of the high and mid vowel pairs for the vowel categorization task (Fig.1, left panel).

2.2.2. Phonetic imitation

We included a subset of stimuli for the vowel categorization (steps 1, 4 and 7 at 140ms, 220ms and 300ms) along with additional perceptually salient tokens (steps 1, 4, 6 at 60ms and 380ms). In total, 15 stimuli (3 spectral steps [steps 1, 4, 7] x 5 duration steps [60ms, 140ms, 220ms, 300ms, 380ms]) were created for each of the high and mid vowel pairs for the phonetic imitation task (Fig.1, right panel).

2.3. Procedure

The participants completed the two-alternative forced choice perceptual vowel identification task before the phonetic imitation task. For the perception task, they were asked to categorize the stimuli they heard as *bit/beat* or *bet/bat* upon hearing stimuli randomly sampled from the *bit-beat* and *bet-bat* continua via E-Prime [15] (49 tokens x 5 repetition = 245 trials). For the imitation task, they were asked to imitate the stimuli presented to them twice in each trial. To establish the degree of imitation for each participant, baseline productions of the target words *bit/beat* and *bet/bat* were collected prior to the experiment for high and mid vowel groups, respectively (6 tokens from the 8 repetitions for each word).

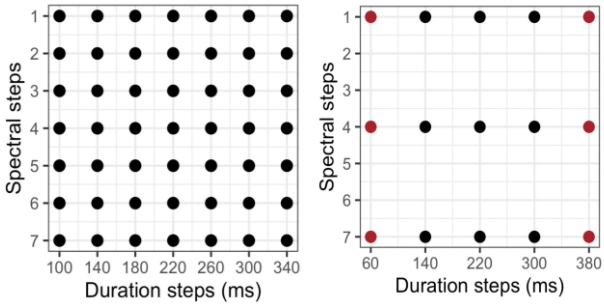

Figure 1: The stimuli for the perceptual categorization (left) and the phonetic imitation (right) task

3. RESULTS

3.1. Perceptual categorization

Participants' responses of *bit/bet* (lax vowels) were coded as 0 and *beat/bat* (tense vowels) as 1, as shown in Fig. 2. A number closer to 1 thus indicates a higher proportion of *beat/bat* responses. To examine the participants' responses as a function of spectral and durational steps at the group level, the data were submitted to two mixed-effect logistic regression models, one for high vowels and the other for mid vowels, using the *glmer* function in the *lme4* package in R [14]. The results showed that L1-Mandarin speakers' perceptual categorization of high vowels was significantly predicted by both spectral (β = 1.85, p <.01) and durational steps (β = 1.48, p <.001), with sharper discontinuities along the durational continuum; in contrast, mid vowel categorization was only significantly predicted by spectral steps (β = 2.46, p <.001; durational steps: β = -0.16, p =.47).

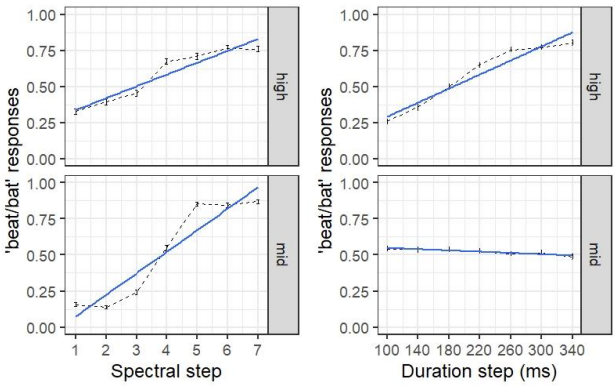

Figure 2: Proportion (dashed black lines) and estimation (solid blue lines) of the participants' *beat* responses (upper panel, high vowel) and *bat* responses (lower panel, mid vowel) along the spectral (left panel, step 1 = *bit/bet*; step 7 = *beat/bat*) and duration continua (right panel)

Next, we extracted by-participant spectral and durational coefficients from the omnibus model to

serve as the cue weights in perception at the individual level (Fig. 3). Further inspection revealed a *trade-off* relationship between the two cues for the high vowels in which the *more* the speakers relied on durational cues to categorize /i/ and /ɪ/, the *less* they relied on spectral cues (r = -.61, p < .05). No direct link was found between the two cues in /ɛ, æ/ categorization (r = .09, p = .69). These findings suggest the L1-Mandarin speakers' failure to integrate the two cues in English vowel perception.

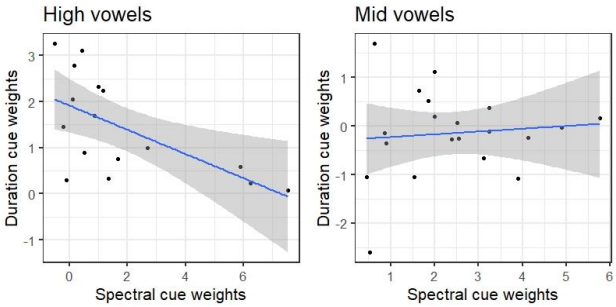

Figure 3: Correlation between individual spectral and duration cue weights in perception of high (left) and mid (right) vowels

3.1. Phonetic imitation

Considering the gender effect on formant values and the distribution of duration data, we standardized the raw F1/F2 values by participant and converted the values into Bark scales, and performed log-transformation for millisecond values of duration data for the subsequent analyses. In the speakers' imitated speech, ambiguous high vowels were produced categorically as /i/ indicated by the overlapping *beat* and *ambiguous* vowel ellipses, while ambiguous mid vowels were imitated gradiently along the spectral steps (Fig. 4). This can be interpreted as an L1 phonological effect: Mandarin /i/ warped the imitation of the ambiguous high vowels, but no such native magnet was available for the mid vowels. Duration, on the other hand, was imitated gradiently for both high and mid vowels, suggesting a faithful imitation of the durational cues in the stimuli (Fig. 5).

We further calculated the degrees of imitation for vowel quality (F1-F2 Euclidean distance) and duration, using the formula $|X_{target}-X_{baseline}| - |X_{target}-X_{imitation}|$ [16]. At the individual level, the speakers who weighted duration cues for the English high vowels more heavily in perception were able to imitate spectral (Fig. 6B, t = 2.64, p < .01) and duration (Fig. 6D, t = 3.91, p < .001) information more faithfully; in contrast, the spectral cue weights did not correlate with either spectral (Fig. 6A, t=0.82, p=.41) or duration (Fig. 6C, t=-1.45, p=.15) imitation. This suggests that both spectral and duration imitation of English high vowels seems to be tied to the perceptual sensitivity in duration, rather than spectral cues, in L1-Mandarin speakers, indicating that the formation of L2 /i/ and /ɪ/ categories by L1-Mandarin speakers is based on duration to a larger extent.

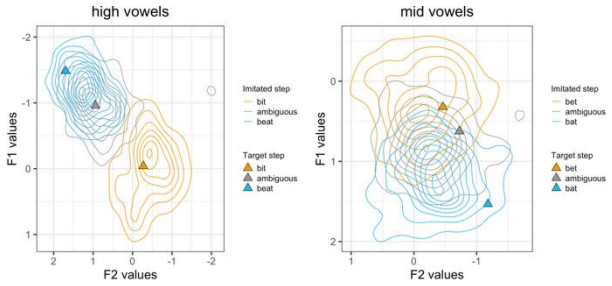

Figure 4: Imitated (colored ellipses) and target F1/F2 values (triangles) of high (left) and mid (right) vowels

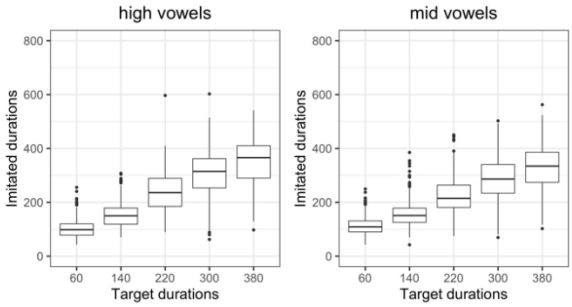

Figure 5: Imitated and target duration values of high (left) and mid (right) vowels

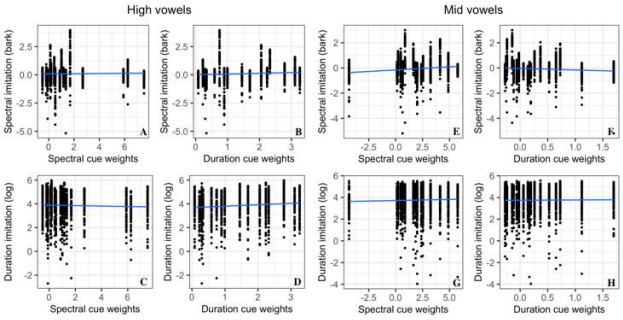

Figure 6: Correlation between individual perceptual cue weights (spectral/duration) and imitation (spectral/ duration) for high (left, A-D) and mid vowels (right, E-G)

For the mid vowels, however, the speakers who weighted spectral cues more heavily in perception displayed a higher degree of spectral imitation (Fig. 6E, t = 5.58, p <.001), while those who relied more on duration in perception showed lower degree of spectral imitation (Fig. 6F, t = -3.28, p < .01); on the other hand, the duration imitation of the mid vowels was not predicted by either spectral (Fig. 6G, t = 1.47, p = .14) or duration (Fig. 6H, t = 0.39, p = .70) cue weights. These results indicate that, for English mid vowels /ɛ/ and /æ/, the two acoustic-phonetic cues in perception constrained the production of vowel quality in a trade-off manner.

6. DISCUSSION

The current study adopted perceptual categorization and phonetic imitation tasks to examine the link between perception and production of English high (/i, ɪ/) and mid (/ɛ, æ/) vowel pairs in L1-Mandarin speakers.

In perception, L1-English speakers rely primarily on spectral cues across vowel contrasts, and less on durational cues. In contrast, we found that (1) L1-Mandarin speakers rely more on durational cues for the /i, ɪ/ contrast, with spectral cues being secondary, and (2) they rely more exclusively on spectral cues to contrast /ɛ, æ/ vowels. It appears that the L2 speakers developed contrast-specific phonetic cue weighting for vowel perception in a non-native way. Furthermore, at the individual level, we observed a trade-off relationship between the perceptual weights of spectral and durational cues for the high vowels, while a null relationship between the two was revealed for the mid vowels. Together, these results imply that L2 speakers may have difficulty simultaneously integrating various acoustic-phonetic cues to map sounds onto L2 categories.

With regard to the perception-production link, our findings suggest that the L1-Mandarin speakers' spectral imitation of L2 English high vowels was mediated by their L1 phonemic categories. During phonetic imitation, L1-Mandarin speakers may perceptually map the L2 English high vowel /i/ onto the native sound category /i/ in Mandarin [5, 8, 17] and thus condition the phonological contrast in their production, leading to the assimilated imitation of the ambiguous and /i/ vowel tokens. On the other hand, the speakers were able to imitate the spectral information of the mid vowels in a gradient manner. This is presumably due to the lack of influence from L1 phonology given no corresponding phonemic categories for both /ɛ/ and /æ/ in Mandarin, though [ɛ] sound does appear as an allophone under certain contexts. The vowel durations, on the other hand, were faithfully imitated for both high and mid vowels, which is in line with previous findings suggesting that suprasegmental properties of speech can be picked up more easily for imitation [11].

At the individual level, for high vowels, the speakers who relied more on durational cues in perception were better able to imitate *both* durational and spectral information of targets; in contrast, spectral cue weights did not correlate with either duration or spectral imitation. This suggests that the imitation of English high vowels, in terms of both vowel quality and duration, is linked to the sensitivity of the primary duration cue, but not that of the secondary spectral cues, in perception. Partly supporting phonological imitation hypothesis claiming all phonologically-relevant phonetic cues to be enhanced in production, our finding exhibits asymmetries between perception and imitation (cf. [12]). Despite the fact that spectral information serves as a secondary cue in L1-Mandarin speakers' perception and is thus encoded to some extent in their English /i, ɪ/ categories, duration, the primary cue, is better represented to contrast the /i, ɪ/ categories. As such, a greater perceptual sensitivity to duration, but not spectral cues, may help activate the L1-Mandarin speakers' L2 representations more robustly and lead to more accurate L2 imitation across different phonetic aspects.

For the mid vowels, however, the speakers who relied more on spectral cues in perception displayed a higher degree of spectral imitation, while those with larger reliance on durational cues imitated the spectral information more poorly. This finding suggests a direct perception-production link at the acoustic-phonetic level for the vowel quality of English mid vowels (cf. [11]); that is, greater sensitivity to spectral cues in perception leads to better imitation of vowel quality in the same phonetic aspect. On the other hand, if the speakers attended more to duration, they were less capable of imitating the spectral information closely. These results provide further evidence that L1-Mandarin speakers utilize primary and non-primary cues in a trade-off manner. In addition, the speakers' duration imitation was not predicted by either of the perceptual cue weights, which may reflect the minor role of vowel duration for the perceptual categorization of English mid vowels and hence result in an absence of perception-production link.

Taken together, L2 English vowel production in L1-Mandarin speaker is linked to perception only when the corresponding phonetic aspect(s) is critical for their perceptual categorization. For English high vowels, both the vowel quality and duration in production were related to the efficiency in using the primary (duration) cue in perception that is more robustly represented to distinguish L2 sounds; in addition, the imitation of spectral information, which is contrastive between Mandarin vowels, was further mediated by the L1 categories at the phonological level. On the contrary, for English mid vowels, the perception-production link is only established in vowel quality, directly constrained by a trade-off in perceptual sensitivity to the primary (spectral) and non-primary (duration) cue at the acoustic-phonetic level. Our findings highlight the roles of phonetic cue weightings and native phonological categories in shaping the abstraction of L2 sound categories during second language acquisition, leading to disparate perception-production links across various phonetic aspects in different L2 vowel contrasts.

7. ACKNOWLEDGEMENTS

We would like to thank Meng-Hsuan Lin and Yin-Ching Chang for their help on collecting data and data annotation. This work was supported by Ministry of Science and Technology of Taiwan (MOST110-2628-H-A49-001-MY2) to Yu-An Lu.

8. REFERENCES

[1] Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 171-204). Baltimore: York Press.

[2] Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege* (pp. 13-34). Amsterdam: John Benjamins.

[3] Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer (Version 6.0.26). Retrieved from www.praat.org.

[4] Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233-276). Baltimore: York Press.

[5] Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. Journal of Phonetics, 25(4), 437–470.

[6] Hillenbrand, J., Getty, L., Clark, M., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of Acoustic Society of America*, 97(5), 3099-3111.

[7] Hsieh, B.-r., & Pan, H.-h. (2010). L2 experience and non-native vowel categorization of L1-mandarin speakers. Paper presented at the Eleventh Annual Conference of the International Speech Communication Association.

[8] Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *The Journal of the Acoustical Society of America*, 119(2), 1118-1130.

[9] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. Paper presented at the Acoustics, Speech and Signal Processing, Las Vegas, NV.

[10] Kim, D., Clayards, M. & Goad, H. (2017). Individual differences in second language speech perception across tasks and contrasts: The case of English vowel contrasts by Korean learners. Linguistics Vanguard, 3(1), 20160025.

[11] Kim, D., & Clayards, M. (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*, 34(6), 769-786.

[12] Kwon, H. (2019). The role of native phonology in spontaneous imitation: Evidence from Seoul Korean. Laboratory Phonology: *Journal of the Association for Laboratory Phonology*, 10(1).

[13] Lin, Y.-H. (2007). *The Sounds of Chinese*. Cambridge, UK: Cambridge University Press.

[14] R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria. URL http://www.R-project.org/: R Foundation for Statistical Computing.

[15] Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.

[16] Walker, A., & Campbell-Kibler, K. (2015). Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Frontiers in Psychology*, 6, 546.

[17] Wang, X. (1997). The Acquisition of English Vowels by Mandarin ESL Learners: A Study of Production and Perception. Master's thesis, Simon Fraser University, Burnaby, BC, Canada.