PRECEDING VOWEL DURATION AND SYLLABLE-FINAL STOP VOICING: AN EXAMINATION OF L2 ENGLISH PRODUCTION AND PERCEPTION BY CHINESE LEARNERS OF ENGLISH

Hongwei Ding[1], Yuqing Zhan[1], Sishi Liao[1], Jiahong Yuan[2]

[1]Institute of Cross-Linguistic Processing and Cognition, Shanghai Jiao Tong University, China
[2]Linguistic Data Consortium, University of Pennsylvania, USA
{hwding; lancaster_zhan; liao_ssh}@sjtu.edu.cn, jiahong@ldc.upenn.edu

ABSTRACT

It is widely accepted that vowels are longer before voiced consonants than before voiceless ones in many languages (such as English). However, neither voiced-voiceless stop contrasts nor long-short vowel contrasts exist in Mandarin Chinese. This study investigates whether the Chinese learners exhibit a difference in vowel duration as a function of the stop voicing in their L2 English production and perception. The production measurements in the Chinese L2 English TIMIT database demonstrated a similar effect as that of the native speakers. In the perceptual experiment with 56 Chinese participants, the results showed that there was a general tendency for the subjects to choose voiced stops when the preceding vowels were lengthened, but there was no significant difference between the intermediate- and the advanced-level learners. However, the Chinese learners displayed different perceptual patterns in different vowels and stops. The results have some implications for L2 English speech learning.

Keywords: voiced/voiceless stops, preceding vowel duration, voiced/voiceless perception, L2 English, L1 Mandarin Chinese

1. INTRODUCTION

It is generally agreed that adults' production and perception of speech are patterned to some extent by the phonological system in their native language (L1). While learning a second language (L2), the learners usually have difficulty in producing and perceiving different sounds that only have distinguishing features in L2 but not in L1. Several studies have shown that foreign language learners differ from native speakers in the production and perception of L2 phonetic contrasts [7, 5]. For instance, unlike English in which stops have voicing contrasts and can occur in syllable final positions, Mandarin has no voicing contrast for stops and allows stops only at syllable onset. Therefore, most Mandarin learners have difficulties in producing and perceiving voicing contrast of word-final stops in English. Previous studies have shown that Chinese speakers tend to devoice English voiced stops, delete voiced and voiceless stops, and insert vowels after syllable-final stops [5, 18] in their production.

Mandarin Chinese learners may come across more difficulties in perceiving word-final stops because the stops may not be released by the native English speakers in connected speech. It was reported that about 40% of around 1,130 sentence-final stops occurring in the TIMIT database were produced without audible releases [10]. Previous studies reported that Mandarin Chinese speakers are able to perceive released syllable-final voicing very well, but they show difficulties in perceiving unreleased syllable-final stop voicing [6, 20]. For example, Mandarin Chinese learners can distinguish *mad* and *mat* if the stops are released, but they cannot distinguish these two words if the stops are not released. However, in this case the native English speakers usually rely on the secondary acoustic cue, that is the preceding vowel length, in their production and perception of unreleased word-final stops.

It was reported that it is presumably a language-universal phenomenon that vowel duration varies as a function of the voicing of the following consonant, such a tendency has been confirmed for English in various studies [9, 15, 8, 12]. It was noted that in many minimal pairs of CVC type in English, the vowel followed by a voiced consonant is longer than the same vowel followed by a voiceless consonant by a ratio of approximately 3:2 [15, 4], and it was found that differences in vowel length are sufficient to cue the perceptual distinction between voiced and voiceless consonants. It was further reported that the obstruent following any vowel of less than about 200 ms was identified as voiceless, while that following a vowel of more than 300 ms tended to be recognized as voiced [16]. However, the extent to which an adjacent voiced or voiceless consonant affects its preceding vowel duration is determined by the language-specific phonological structure [4]. Only few studies have been devoted to examine the performance of Mandarin Chinese learners on this secondary acoustic cue [6, 4, 20]. In this study, we aimed to focus on this issue to answer the following questions related to Chinese learners:

- Do they exhibit a difference in the preceding vowel duration as a function of the following word-final stop voicing in their L2 English production and perception?
- Is there any difference in the perception between intermediate and advanced learners?
- Do vowel categories and place of articulation of stops influence their perception results?

2. METHOD

Production materials were taken from a database and perception stimuli were specially designed. For the convenience of display, the ARPAbet transcription [11], which represents phonemes of General American English with distinct sequences of ASCII characters, are employed in the figures in this paper.

2.1. Production

For the production investigation, we employed sentences from the English L2 TIMIT [19, 3].

2.1.1. Materials

Although numerous words ending with stops exist in the English L2 TIMIT, but only three monosyllabic minimal word pairs could be found in the database [19], which are listed in Table 1.

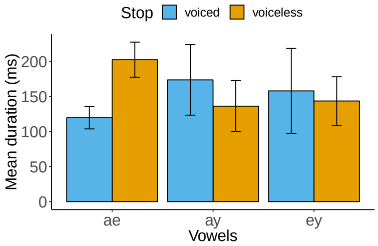
Table 1: Monosyllabic minimal word pairs selected from L2 English TIMIT.

Vowel	Word	Occurrences
[æ]	had	6
	hat	5
[aɪ]	ride	14
	right	18
[eɪ]	made	4
	mate	5

2.1.2. Results

The speakers were recruited from a key university in Shanghai, so their English levels were equivalent to the intermediate level and above. Though they were slightly different in oral performance, all of them produced the final stops with release. We only measured the absolute duration of the vowel, and found that they did produce the same vowel ([aɪ] and [eɪ]) longer before voiced stop [d] (in *ride* and *made*) than before the corresponding voiceless stop [t] (in *right* and *mate*) except for the vowel [ae]. The comparison can be observed in Fig 1. The transcription “ae, ay, ey” correspond to IPA “[æ], [aɪ], [eɪ] ” respectively.

Figure 1: Mean vowel duration of L2 production



2.2. Perception

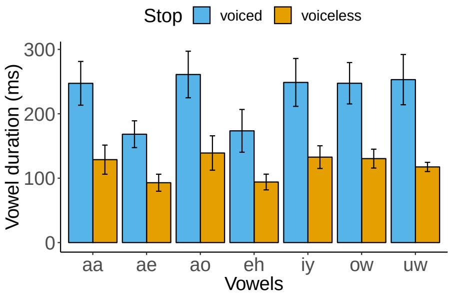
We took great care in the preparation and conduct of the perception experiment.

2.2.1. Stimuli

In designing our test data, we used words of the canonical form of hVC. We chose /h/ as the initial consonant, for /h/ exerts little influence on the following vowels [13]. Seven vowels in American English [i, ɛ, æ, ɑ, ɔ, u, oʊ] were selected as stimuli to cover high-mid-low, front-back, and monophthong-semidiphthong-diphthong vowels [14]. They were combined with three pairs of voiced/voiceless (labial [b, p], dental [d, t], and velar [g, k]) stops at word-final positions; therefore, 21 minimal pairs were obtained. An adult native speaker of standard American English was selected as the speaker. The recording was carried out in 16 bit and 44.1 kHz in a soundproof room. After a short training, the speaker was able to read these words accurately in IPA transcription. Finally, the speaker managed to produce each 21 pairs for five times with normal and consistent prosody. The absolute duration values of the target vowels from the 210 stimuli are presented in Fig. 2.

Similar to the results reported in previous investigations, vowels were significantly longer in voiced contexts than in voiceless ones, with a voiced-to-voiceless mean ratio of 1:1.91. In order to avoid unnaturalness, the maximum and minimum vowel duration values in each group produced by the speaker were selected as the two ends for stimuli, which are listed in Table 2. The transcription “aa, ae, ao, eh, iy, ow, uw” in ARPAbet correspond to “[ɑ], [æ], [ɔ],

Figure 2: Mean vowel duration (and standard deviation) of the target vowels in stimuli recording



[ɛ], [i], [oʊ], [u]" in IPA, and C, M, R stand for consonant, mean, and ratio, respectively.

Table 2: Maximum and minimum vowel duration and their ratio in stimuli words hVC.

C	Vowels							M
	[i]	[ɛ]	[æ]	[ɑ]	[ɔ]	[u]	[oʊ]	
[*b*]	272.7	184.5	177.6	250.1	321.6	260.8	272.3	248.5
[*p*]	99.7	81.6	78.4	105.0	119.7	107.6	101.5	99.1
R	2.7	2.3	2.3	2.4	2.7	2.4	2.7	2.5
[*d*]	315.9	230.3	186.3	314.3	320.4	346.8	269.4	283.3
[*t*]	122.2	71.4	74.1	121.3	110.3	104.6	113.7	102.5
R	2.6	3.2	2.5	2.6	2.9	3.3	2.4	2.8
[*g*]	301.5	241.5	217.6	262.7	254.8	278.7	299.5	265.2
[*k*]	113.5	85.8	86.0	101.3	94.3	118.7	120.9	102.9
R	2.7	2.8	2.5	2.6	2.7	2.3	2.5	2.6

The naturally produced words with the maximum duration were selected for manipulation. The audible voiced parts of word-final stops were removed, and the intensity peak scales were normalized to ensure the same loudness. Along each maximum-minimum continuum of the vowel duration, six equal steps were divided and seven points were obtained with point 1 for the shortest and 7 for the longest. All the stimuli were created by resynthesis with PSOLA in Praat [2], and 147 stimuli were developed (7 vowels x 7 vowel durations x 3 stop pairs). A pilot test was conducted to ensure that every stimulus was fairly natural without any signal distortion.

2.2.2. Subjects

We recruited students at a key university in Shanghai as subjects with two degrees of English proficiency. Subjects who had passed TEM8 (Test for English Majors-Band 8) or IELTS with scores higher than 7.0 were classified as advanced level; the others were regarded as intermediate level. Thus we obtained 56 participants, including 33 advanced and 23 intermediate learners. All of them were born and grew up in China, and used Mandarin as their native language. They claimed to have normal hearing and participated in the experiment voluntarily, and received a small amount of payment for participation.

2.2.3. Procedures

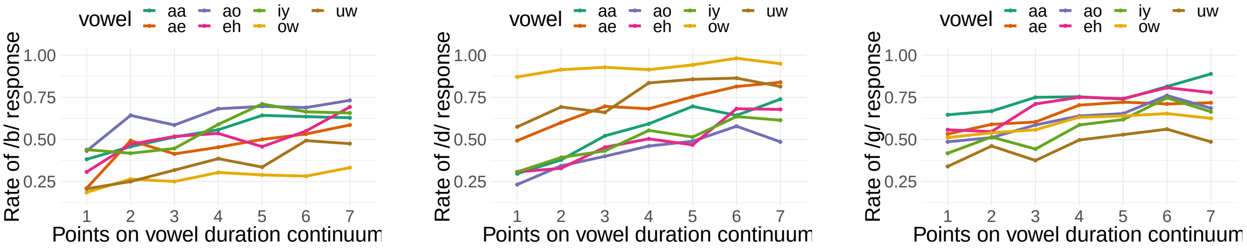
The participants completed the voiced-voiceless 2AFC (two-alternative forced choice) identification task with E-prime 2.0 in a quiet room individually. Voiced-voiceless stimuli were presented binaurally to listeners via Sennheiser headphones. The participants heard a sound stimulus hV and simultaneously saw voiced-voiceless stop pair on the computer screen, e.g. "1 = b" on the left side and "2 = p"on the right side of the screen for the pair of [b]-[p]. They were asked which stop was unreleased by pressing "1" or "2" on the keyboard to indicate their choice. Listeners should respond within 10 seconds for each sound stimulus. Participants were trained and practiced for about 5 minutes by the instructor to ensure that they understood and could follow the task instructions precisely before the test started. Identification tasks were divided into three blocks, namely to distinguish between [b]-[p], [d]-[t], and [g]-[k]. Each block contained 245 stimuli (7 vowels x 7 steps x 5 repetitions), which were presented in a random order. Participants were given a short break between every two blocks, and it took about 30 minutes to complete the whole experiment. The choice and response time for every stimulus were recorded.

2.2.4. Results

Fig. 3 illustrates the overall picture of listeners' choices in every step on the vowel duration continuum. Despite several ups and downs, all the lines show an upward trend. The x-axis represents the vowel duration with point 1 as shortest, and 7 as longest, while the y-axis indicates the mean percentage of choosing voiced stops as the unreleased consonants. The results showed that there was a general tendency for the Chinese subjects to choose voiced stops when the preceding vowels were lengthened.

The percentage values in raw data were first transformed into rationalized arsine unit (RAU) to make them more suitable for statistical analysis [17]. Then a series of linear mixed-effects analyses were conducted on participants' responses using the lme4 package in R [1]. Subjects were set as random factor as usual, subjects' response of RAU as the dependent variable. Likelihood ratio tests were performed to evaluate four fixed factors. The main effects analyses showed that there was no significant difference for fixed variable of speaker groups (x^2 = 0.587, df =

Figure 3: Judgements in percentage of corresponding voiced stops as a function of steps along the vowel duration continuum for all participants



(a) hVb(p)

(b) hVd(t)

(c) hVg(k)

1, p = 0.443). However, significant differences were observed for other three variables (place of articulation of stops: x^2 = 421.94, df = 2, p < 0.001; vowel category: x^2 = 82.194, df = 6, p < 0.001; continuum point: x^2 = 614.42, df = 6, p < 0.001).

Moreover, we adopted lmerTest to examine whether RAU values differ between different duration steps for every vowel and each place of articulation, and it was found there were no significant differences if the preceding vowel is a diphthong [oʊ] or the syllable-final is a velar stop. No significant differences in RAU could be found between any two duration steps for /hæ(kg)/, /hoʊ(kg)/, /hu(kg)/, which can be easily observed in Fig. 3.

3. DISCUSSION

Based on the results, we can now answer the aforementioned questions concerning Chinese learners:

- They tend to display different vowel lengths as to the following stop voicing in their production and perception, but not with certainty.
- There is no difference between learners of intermediate and advanced levels.
- Vowel categories and places of articulation of stop can influence L2 perceptual performance.

Most Mandarin Chinese learners are not aware of the English phonological rule that vowels are longer before voiced stops, but they do use this rule unconsciously in their production and perception.

In the production data, Chinese learners usually use unaspirated/aspirated to replace voiced/voiceless stops. They produce an extremely long aspirated stop for a voiceless one, making the preceding vowel relatively shorter. This provides more evidence to support the idea that longer vowels before voiced stops may be universal. The reason that minimal pair *hat* and *had* did not support this result is that *had* was unstressed as an auxiliary verb.

In the perception experiment, Chinese learners also showed similar behavior as the native speakers that diphthong vowels and velar stops can interfere with their perceptual performance [14] though their performance is not so good as the native speakers reported in the literature [6, 16], because there is no 100% choice for a voiced stop if the vowel duration exceeds 300 ms. However, in the current study we can hardly determine whether the interference of diphthong vowels and velar stops in the secondary perceptual cue is universal or language-specific, which demands more investigation. Moreover, the fact that no significant difference exists between intermediate and advanced English levels may indicate that these subtle phonetic features can hardly be improved for adult learners. The advanced learners may rely on other prosodic, syntactic, or semantic cues in English listening comprehension, but this secondary acoustic cue may be effective in distinguishing non-native from native speakers.

4. CONCLUSIONS

This study provided more evidence to support the fact that Chinese learners employ the English phonological rules that vowels preceding voiced stops are longer than voiceless ones unconsciously in their production and perception. Their perceptual performance are degraded if the vowel is a diphthong or the stop is a velar, which is also the same as the native speakers. However, experiments should be extended to native speakers and English learners from other countries to discover more linguistic facts.

5. ACKNOWLEDGEMENTS

We sincerely thank the three anonymous reviewers for their meticulous and insightful comments. The work is sponsored by Shanghai Social Science Project (2018BYY003), major Program of National Social Science Foundation of China (No.15ZDB103), and the Interdisciplinary Program of Shanghai Jiao Tong University (14JCZ03).

6. REFERENCES

[1] Bates, D., Mächler, M., Bolker, B. M., Walker, S. C. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.

[2] Boersma, P., Weenink, D. 2017. Praat: doing phonetics by computer [computer program]. Version 6.0.24, retrieved 31 Januray, 2017.

[3] Chanchaochai, N., Cieri, C., Debrah, J., Ding, H., Jiang, Y., Liao, S., Liberman, M., Wright, J., Yuan, J., Zhan, J., Zhan, Y. 2018. Global-TIMIT: Acoustic-phonetic datasets for the world's languages. *Proceedings of Interspeech* 192–196.

[4] Chen, M. 1970. Vowel length variation as a function of voicing of the consonant environment. *Phonetica* 22, 129–159.

[5] Flege, J. 1987. The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics* 15, 47–56.

[6] Flege, J. 1989. Chinese subjects' perception of the word-final english /t/-/d/ contrast: Performance before and after training. *Journal of the Acoustical Society of America* 86, 1684–1697.

[7] Flege, J. 1993. Production and perception of a novel, second-language phonetic contrast. *The Journal of the Acoustical Society of America* 93(3), 1589–1608.

[8] House, A. S. 1961. On vowel duration in English. *The Journal of the Acoustical Society of America* 33, 1174–1178.

[9] House, A. S., Fairbanks, G. 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America* 25(1), 105–113.

[10] Keating, P. A., Byrd, D., Flemming, E., Todaka, Y. 1994. Phonetic analyses of word and segment variation using the TIMIT corpus of American English. *Speech Communication* 14, 131–142.

[11] Klautau, A. 2001. ARPABET and the TIMIT alphabet.

[12] Kluender, K., Diehl, R., Wright, B. 1988. Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics* 16, 153–169.

[13] Ladefoged, P., Maddieson, I. 1996. *The Sounds of the World's Languages*. Blackwell Publishers.

[14] Lisker, L. 1999. Perceiving final voiceless stops without release: Effects of preceding monophthongs versus nonmonophthongs. *Phonetica* 56, 44–55.

[15] Peterson, G. E., Lehiste, I. 1960. Duration of syllable nuclei in English. *Journal of the Acoustical Society of America* 32(6), 693–703.

[16] Raphael, L. 1972. Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *The Journal of the Acoustical Society of America* 54(4), 1296–1303.

[17] Studebaker, G. 1985. A rationalized arcsine transform. *Journal of Speech, Language, and Hearing Research* 28(3).

[18] Weinberger, S. 1987. The influence of linguistic context on syllable simplification. In: Ioup, G., Weinberger, S., (eds), *Interlanguage Phonology: The Acquisition of a Second Language Sound System*. Cambrige, MA: Newbury House 401–417.

[19] Yuan, J., Ding, H., Liao, S., Zhan, Y., Liberman, M. 2017. Chinese TIMIT: A TIMIT-like corpus of standard Chinese. *O-COCOSDA*.

[20] Zhang, Y. 2017. *Distributional learning of extrinsic vowel duration differences by Mandarin native speakers*. PhD thesis University of Amsterdam.