Durational correlates of Japanese phonemic quantity contrasts by Cantonese-speaking L2 learners

Albert Lee[1], *Peggy Mok*[2]

[1] Department of Linguistics, University of Hong Kong.
[2] Department of Linguistics and Modern Languages, The Chinese University of Hong Kong.
albertlee@hku.hk, peggymok@cuhk.edu.hk

Abstract

This paper reports a production study of Japanese phonemic quantity contrasts by native speakers of Japanese, beginner learners, and advanced learners speaking Cantonese as L1. The three groups were compared using various standard durational measures. It was found that both learner groups successfully distinguished all the quantity conditions, although they did so differently from their Japanese peers. Specifically, whereas the short vs. long contrasts were enhanced in slow speech by native speakers, such enhancement was absent in both learner groups. The pedagogical and typological implications of these data are discussed.

Index Terms: Japanese, geminate, L2 production

1. Introduction

The phonemic quantity contrasts in Japanese have been extensively studied in the research literature (e.g. [1], [2]). Both consonants (*kita* 'came' vs. *kitta* 'cut') and vowels (*kita* vs. *kiita* 'heard') contrast in terms of length, and present a challenge to many learners of the language (e.g. [3]–[5]). Numerous durational correlates of these phonemic quantity contrasts have been identified in the last two decades [1], [6]–[8]. For short vs. long vowels, the ratio of duration is approximately 1:2.4~2.7 [1] and is greater at slow speech rate. For short vs. long consonants, the ratio of closure duration is 1:2.8 [9], with a 11% lengthening in the vowel preceding, and 9% shortening in the vowel following the geminate [10], see also [11]. The lengthening of the vowel preceding geminate appears to violate Maddieson's [12] typology where vowels are shorter before geminates across languages.

Little is known about the acquisition of these quantity contrasts by Cantonese speaking learners. In Cantonese, there are vowel pairs that contrast in length (e.g. */ka:i/街* 'street' vs. */kai/雞* 'chicken'), as well as the 'cat tail' type geminates (e.g. */tsi:.tso:/ 知咗* 'knew' vs. */tsit.tso:/ 唧咗* 'squeezed') given Cantonese allows an unreleased stop coda in its syllable structure. These partial uses of quantity contrasts beg the question of whether Cantonese speaking learners of Japanese could distinguish *kita* vs. *kitta* vs. *kiita* successfully.

In [13], the production and the perception of Swedish quantity by American English, Latin American Spanish, and Estonian learners were investigated. Though not as good as the Estonian learners, the English speakers performed better than their Spanish-speaking counterparts, presumably due to the partial use of durational cues to vowel length contrasts in their L1. By implication, the partial use of durational cues in Cantonese might mean that Hong Kong learners would also be able to distinguish Japanese quantity contrasts, but only to some extent. This paper thus tests the hypothesis that Cantonese learners can distinguish long vs. short consonants and vowels in Japanese, but will also manifest evidence of incomplete acquisition in certain contexts. Our results will shed new light on the role of L1 on the acquisition of L2 speech sounds.

2. Methodology

2.1. Speakers and materials

We conducted a production study with four native speakers of Japanese as controls, 10 advanced learners in their final year of BA Japanese Studies programme, and 10 beginners who were in their first year of the same programme. The advanced learners had all stayed in Japan for one year as exchange students. Both learner groups are native speakers of Hong Kong Cantonese. All participants reported no speech and hearing impairment. Table 1 shows the 27 (quasi-)real Japanese words and 18 non-words used as stimuli. They contrasted in vowel and consonant quantity (CV.CV, CVV.CV, CVC.CV), and were each repeated three times at three speech rates. All words were displayed in *kana* syllabary (*hiragana* or *katakana*) as well as *kanji* where applicable. To obtain true minimal triplets, infrequent words, place names and personal names had to be used.

	CV.CV	CVV.CV	CVC.CV
Real words	*kita* 'came'	*kiita* 'heard'	*kitta* 'cut'
	shite 'do'	*shiite* 'lay'	*shitte* 'know'
	seto (place name)	*seito* 'pupil'	*setto* 'set'
	ato 'after'	*aato* 'art'	*atto* 'at'
	nita 'resembled'	*niita* (place name)	*nitta* (personal name)
	seki 'seat'	*seiki* 'century'	*sekki* 'solar term'
	jaku 'weak'	*jaaku* 'jerk'	*jakku* 'Jack'
	mito (place name)	*miito* 'meat'	*mitto* 'mitt'
	kato 'transition'	*kaato* 'cart'	*katto* 'cut'
Non-words	*sasa*	*saasa*	*sassa*
	sese	*seese*	*sesse*
	soso	*sooso*	*sosso*
	tata	*taata*	*tatta*
	tete	*teete*	*tette*
	toto	*tooto*	*totto*

Table 1. *Stimuli used in the present study*

2.2. Procedures

Recording took place in a quiet room in the Chinese University of Hong Kong, using a ZOOM H2n voice recorder. Stimuli were presented on a computer screen using a Javascript-based sentence randomiser. Speakers were briefed about the experimental task and granted their written consent before recording. For the non-word blocks, speakers were instructed to

use the HL accent pattern. Utterances were collected over six randomised blocks, namely Real Word normal⇒Slow⇒Fast⇒ Non-word normal⇒Slow⇒Fast. Within each block, each token were repeated three times. Altogether, 15 tokens (9 real and 6 non-words) × 3 speech rates × 3 quantity × 3 repetitions × 24 speakers (4 native+10 beginners+10 advanced) = 9,720 utterances were collected.

Speech data were manually labeled by the segment FormantPro (described in [14], [15]). It is a Praat [16] script for extracting formant trajectories, as well as intensity and duration values. Since all target words were disyllabic, four segments (henceforth $C_1V_1C_2V_2$) were labelled. Vowel boundaries were located at the onset and offset of voicing. Subsequently, for each labelled interval FormantPro extracted the duration and mean intensity values as well as time-normalised formant values.

3. Results

3.1. Average syllable duration

The mean syllable duration of all target words was checked to assure that the three speech rates were produced correctly. Table 2 shows that in all speaker groups, average syllable duration was the shortest in fast speech and the longest in slow speech. One-way ANOVA confirms that the main effect of speech rate was significant $F(2,3372) = 2225$, $p<0.001$. All speech rate conditions had a significantly different mean syllable duration from one another, according to post-hoc Bonferroni tests ($p<0.001$). Thus it is safe to conclude that any significant effects of speech rate observed in subsequent analyses are robust.

Group		Mean	STD	N
Advanced	Fast	180.00	34.685	450
	Normal	223.99	41.743	450
	Slow	319.30	56.324	450
Beginners	Fast	201.05	28.628	450
	Normal	240.61	38.383	450
	Slow	334.10	54.722	450
Native	Fast	160.11	22.097	225
	Normal	208.54	38.629	225
	Slow	310.52	106.530	225

Table 2. *Mean syllable duration (ms) of target words across speech rate and speaker group conditions*

3.2. Short vs. long vowels

First the absolute duration of V1 is considered. **Table 3** shows the duration of V1 in different speech rate, vowel length, word type, and speaker group conditions. As expected, in all speaker groups and word types, V1 is longer in CVV than in CV, and the longest in slow speech and the shortest in fast speech.

To verify that the learner groups produced the long vs. short distinction consistently, averaged data (tokens and repetitions collapsed) were submitted to a mixed ANOVA with Speech Rate (Fast/Normal/Slow) and Quantity (CV/CVV) as fixed factors, and Speaker Group (Advanced/Beginners/Native) as between-subject factor. The main effects of Speech Rate ($F(1.13,24.78) = 96.2$) and Quantity ($F(1,22) = 213.3$), as well as their interaction ($F(1.22,26.86) = 58.4$) reached statistical significance (all $p<0.001$), whereas Speaker Group and its interaction with other factors did not. This shows that both learner groups made a clear distinction between long and short vowels in terms of the absolution duration of V1 across different speech rates, but the two groups were not significantly different from each other, nor did they differ from the native speakers.

		Real words		Non-words	
		CV	CVV	CV	CVV
Native	Fast	65	123	66	129
	Normal	77	160	80	188
	Slow	112	276	122	300
Advanced	Fast	73	143	84	155
	Normal	82	167	111	201
	Slow	128	259	159	285
Beginner	Fast	83	137	88	172
	Normal	96	174	106	198
	Slow	129	250	161	294

Table 3. *Duration (ms) of V1 in different speech rate, vowel length, word type, and speaker group conditions*

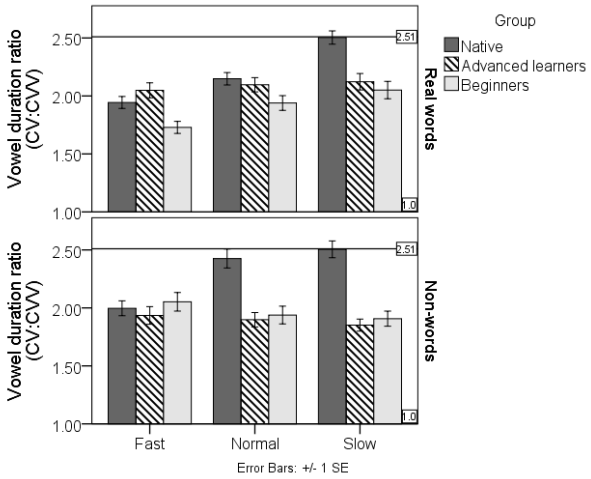

Figure 1. *Duration ratio (CV:CVV) of V1 in different speech rate, word type, and speaker group conditions*

We further examined V1 duration ratio (CV:CVV), following Hirata [1]. The mean ratio of the native, advanced, and beginner groups were respectively 1:2.24, 1:2.01, and 1:1.93. In **Figure 1**, if the ratio exceeds the 1:1 reference, long vowels are longer than short vowels. There is also the 1:2.51 line for reference, which is the ratio reported by Hirata [1] for accented vowels (or 1:2.22 for unaccented vowels in her study). As is clear from the diagram, for all speaker groups the vowel duration ratio exceeded by far the 1:1 threshold (grand mean 1:2.03, SD 0.58). There was also an effect of speech rate on V1 duration ratio in native speakers but not in the learner groups. Paired T-tests revealed that for native speakers, there was a significant difference in V1 duration ratio between fast and normal speech $t(74) = -7.1$, $p<0.001$ and between normal and slow speech $t(74) = -6.7$, $p<0.001$, whereas no significant difference was observed between speech rates in either of the learner groups ($p>0.1$), with the exception of fast vs. normal speech of beginners, which were significantly different $t(142) = -2.1$, $p<0.042$, though the difference was small.

The same holds true for word duration ratio (**Figure 2**). Here if the ratio exceeds 1:1, a CVVCV word is longer than a CVCV word. The 1:1.4 reference is adapted from Hirata (2004), where the word duration ratio of CVCV:CVVCV was 2:2.7~2.95 (i.e. ~2:2.8) . For all speaker groups, words with a long vowel were longer than otherwise. For native speakers,

slow speech had the effect of enhancing the long vs. short contrast, but the same effect was not observed in the learner groups. In other words, although the learners consistently distinguished long vs. short vowels, they used a strategy different from the native speakers across different speech rates.

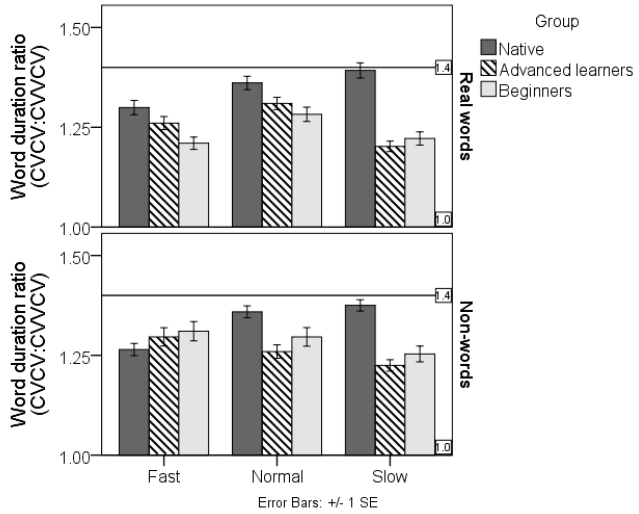

Figure 2. *Duration ratio (CVCV:CVVCV) of target words in different speech rates, word types, and speaker groups*

3.3. Singleton vs. geminate consonants

The production of singleton vs. geminate consonants is analyzed in terms of the duration ratio of C2 as well as the duration of surrounding vowels. In **Figure 3**, the 1:1 threshold means that singleton and geminate consonants are equal in C2 duration. The 1:2.8 reference was taken from [9][1]. The native speakers were much closer to the 1:2.8 reference (mean = 2.37, SD = 0.55) than the learners (advanced learners mean = 1.73, SD = 0.62; beginners mean = 1.71, SD = 0.62). This time, the contrast-enhancing effect of slow speech was observed in all three groups. **Table 4** shows that for fast vs. normal speech, C2 duration ratio was not significantly different in the learner groups; elsewhere, it was consistently greater in slower speech.

	A	B	t	df	*p*	A-B
Advanced	Fast	Normal	1.54	149	0.125	0.071
	Normal	Slow	- 2.76	149	***0.007***	-0.122
Beginners	Fast	Normal	-0.11	149	0.913	-0.004
	Slow	Normal	4.91	149	***<.001***	0.19
Native	Fast	Normal	-4.72	74	***<.001***	-0.217
	Normal	Slow	-6.56	74	***<.001***	-0.386

Table 4. *Paired T-tests comparing C2 duration ratio among Speaker Group ×Speech Rate conditions*

Place of articulation appears to affect C2 duration ratio too. **Table 5** shows that, in the present study, for all speaker groups /t/ had a greater C2 duration than /k/, like the native speaker group in [9]. The same was true for our learner groups, unlike the American English speakers in [9] who manifest no such tendency.

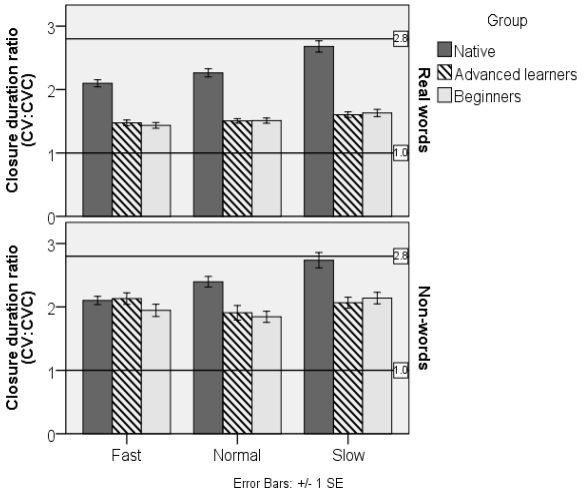

Figure 3. *Duration ratio (CV:CVC) of C2 in different speech rate, word type, and speaker group conditions*

Consonant	Han (1992)		The present study		
	Native[2]	American	Native	Advanced	Beginners
/k/	2.4	1.8	2.2	1.4	1.3
/s/	N/A		2.1	1.9	1.9
/t/	3.0	1.7	2.5	1.7	1.7

Table 5. *Effect on consonant on C2 ratio (CV:CVC)*

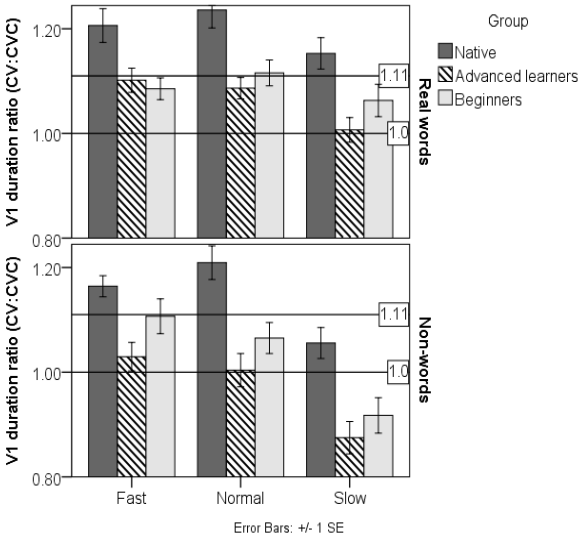

Figure 4. *Duration ratio (CV:CVC) of V1 in different speech rate, word type, and speaker group conditions*

Next, the effect of consonant quantity on V1 was examined, following [10] and [11]. Han [10] reported that V1 was 11% longer (see the 1:1 threshold and the 1.11 reference in **Figure 4**) before and V2 9% shorter after a geminate. Like in previous studies, our native speakers lengthened V1 before a geminate, whereas the learner groups did not always do so. For example,

[1] Note that VOT was not measured in the present study.

[2] Mean value from Table 5 in [9], where VOT duration is included as part of the consonant, like in the present study.

in non-words spoken at slow speed, V1 was even shorter before a geminate. Moreover, unlike other measurements reported so far, slow speech does not seem to enhance the quantity contrast in terms of V1 duration ratio.

For V2 duration ratio, **Table 6** shows that our native speakers always shortened V2 after a geminate, as did the beginners; whereas there was no discernable pattern in the advanced learners' production.

	Real words			Non-words		
	Fast	Normal	Slow	Fast	Normal	Slow
Native	0.94	0.96	0.94	0.95	0.99	0.98
Advanced	1.03	1.02	0.93	1.00	1.00	0.96
Beginner	0.96	0.97	0.91	0.91	0.97	0.97

Table 6. *Duration ratio (CV:CVC) of V2 in different speech rate, word type, and speaker group conditions*

4. Discussion and conclusion

The present study has yielded a range of evidence to show that Cantonese-speaking learners of Japanese were able to distinguish phonemic quantity contrasts, albeit using a different strategy from that of their native speaker counterparts. Recall that in Cantonese there are vowel pairs (e.g. /ai/ vs. /a:i/) that contrast in length, as well as the 'cat tail' geminates (i.e. unreleased stop coda+initial stop/fricative/afficate sequences), our learner groups' ability to distinguish Japanese quantities may thus be attributed to these partial uses of duration in their L1, much like the American English participants in [13].

However, the acquisition of Japanese quantities was not complete for both learner groups. While both groups were obviously capable of using duration to mark quantity contrasts in both consonants and vowels, in most cases the enhancing effect of slow speech was only observed in the native speakers. This, together with all the smaller durational ratios in their production reported above, shows that they had not mastered the control of duration in different speech rate conditions. Considering also the lack of significant differences between the learner groups in several of the measures reported above, it seems that these Cantonese-speaking learners started with some advantage from their L1, but their production never became native-like even after years of exposure and time spent in Japan. The challenge they faced as beginners persisted through their proficiency curve and remained after they had become much better speakers. These observations support our hypothesis that Cantonese learners can easily acquire Japanese quantity distinctions, but their acquisition would be incomplete, as a result of partial use of duration in their L1.

For short vs. long vowels, the learner groups showed a smaller V1 duration ratio (1:2.01 for advanced learners and 1:1.93 for beginners) than the native speakers (1:2.24), but the two learner groups did not differ from each other significantly. We also replicated the contrast-enhancing effect of slow speech on vowel duration and word duration ratios in the native speakers [1], but it was absent in both learner groups.

With regards to singleton vs. geminate consonants, again there was clear evidence that the learners were capable of making the quantity distinction, but they also differed from the native speakers in terms of several duration ratios and of the lesser enhancement effect in their slower speech. Note that this effect is not to be confused with the speech rate-independent duration ratios reported in [2]. In [2], the ratios of **C2:V1** and **C2:V2** within the same word were found to be stable across speaking rates, and thus served well to distinguish CV.CV and CVC.CV. In the present study, using duration ratios of **CV.CV:CVC.CV**, the distinction between the two phonemic quantities was found to be speech-rate dependent, and greater at slower speaking rates.

Taken together, our results lead to two theoretical implications. Firstly, L1 transfer benefit (e.g. [17]–[20]) appears to be based on phonetic dimensions (e.g. [13]) rather than on actual phonemes [21]. That is, the use of duration as a cue to only a subset of vowels in Cantonese seems to help learners distinguish quantity conditions in different L2 vowels. Our results also point to the fact that learners can benefit from their L1 even if the phonetic dimension in question is not used phonemically. That is, Cantonese has no phonemic geminates but the derived geminates seem to have helped our learners acquire Japanese geminates. Secondly, our data suggest that for production quantity distinction is harder to master in slower speech, while the opposite is true for perception [22]. The reason underlying this discrepancy is surely an interesting question to explore.

The effect of place of articulation on C2 ratio is believed to be due to longer duration of /k/. One reason is that, in our corpus all the cases where /k/ occupies C2 have a high vowel /i/ or /u/ in V2, which is prone to V2 devoicing. As C2 in these cases are longer, the resulting CVCV:CVCCV ratio naturally becomes smaller. Another source of a longer /k/ is that velar stops are known to have longer VOT [23], which in our data is included as part of C2. Although this effect was observed in all our speaker groups, it was not observed in the American speakers in [9]; the source of the discrepancy is unclear.

With regards to Maddieson's typology, **Figure 4** suggests that the learners were only lengthening their V1 in some conditions, unlike their native peers who consistently did so across all speech rates and word type conditions. In some cases, the advanced learners were lengthening V1 less than the beginners as if their pronunciation had deteriorated. It appears that the learners performed V1 lengthening better at normal speech rate than slow speech rate, better in real words than non-words. Then in the most challenging condition, namely non-words at slow speech, the learners shortened V1 instead, somehow conforming to Maddieson's typology. Although with the present data we are unable to conclude whether this is idiosyncratic, or a residue of typological influence that only resurfaced when learners had the most difficulty, V1 lengthening shows us again how a hard-to-acquire feature can persistently continue to be challenging to rather advanced speakers who have had extensive natural exposure to their L2.

From a pedagogical point of view, our data show that distinguishing long and short per se is not difficult even for the beginners. The real challenge of acquiring a native-like pronunciation lies in adjusting the long vs. short duration ratio according to speech rate. Teachers of Japanese should consider exposing L2 students to input at various speech rates. Besides, teachers of Japanese should also be aware of non-local cues to quantity contrasts (e.g. V1 lengthening before geminate) to help students acquire the most native-sounding pronunciation possible. More work needs to be done to fully understand the acquisition of Japanese phonemic quantity contrasts by Hong Kong L2 learners. Analyses of the effect of phonemic quantity on the vowel space are underway. In the future, we also aim to conduct perception studies to examine what cues these listeners rely on the most.

5. References

[1] Y. Hirata, "Effects of speaking rate on the vowel length distinction in Japanese," *J. Phon.*, vol. 32, no. 4, pp. 565–589, 2004.

[2] K. Idemaru and S. Guion-Anderson, "Relational timing in the production and perception of Japanese singleton and geminate stops," *Phonetica*, vol. 67, no. 1–2, pp. 25–46, 2010.

[3] M. Kurihara, "中国語北方方言話者の日本語長音の知覚特徴 [Perception of Japanese long vowels by speakers of Mandarin]," *Tohoku Univ. J. Linguist. Sci. [言語科学論集]*, vol. 8, pp. 1–12, 2004.

[4] X. Ren, "中国人日本語学習者の促音の知覚について," *Meikai Japanese Lang. J. [明海日本語]*, vol. 16, pp. 93–95, 2011.

[5] M. Kurihara, "中国語北方方言話者の日本語長音と短音の産出について [On the production of Japanese long vs. short vowels by speakers of Mandarin]," *Tohoku Univ. J. Linguist. Sci. [言語科学論集]*, vol. 9, pp. 107–118, 2005.

[6] Y. Hirata, "Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2384–2394, 2004.

[7] Y. Hirata and J. Whiton, "Effects of speaking rate on the single/geminate stop distinction in Japanese," *J. Acoust. Soc. Am.*, vol. 118, no. 3, pp. 1647–1660, 2005.

[8] Y. Hirata and K. Tsukada, "Effects of speaking rate and vowel length on formant frequency displacement in Japanese," *Phonetica*, vol. 66, no. 3, pp. 129–149, 2009.

[9] M. S. Han, "The timing control of geminate and single stop consonant in Japanese: A challenge for nonnative speakers," *Phonetica*, vol. 49, pp. 102–127, 1992.

[10] M. S. Han, "Acoustic manifestations of mora timing in Japanese," *J. Acoust. Soc. Am.*, vol. 96, no. 1, pp. 73–82, 1994.

[11] K. Idemaru and S. G. Guion, "Acoustic covariants of length contrast in Japanese stops," *J. Int. Phon. Assoc.*, vol. 38, no. 2, pp. 167–186, 2008.

[12] I. Maddieson, "Phonetic cues to syllabification," in *Phonetic linguistics: Essays in honor of Peter Ladefoged*, V. A. Fromkin, Ed. Orlando, FL: Academic Press, 1985, pp. 203–221.

[13] R. McAllister, J. E. Flege, and T. Piske, "The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian," *J. Phon.*, vol. 30, no. 2, pp. 229–258, 2002.

[14] C. Cheng and Y. Xu, "Articulatory limit and extreme segmental reduction in Taiwan Mandarin," *J. Acoust. Soc. Am.*, vol. 134, no. 6, pp. 4481–4495, 2013.

[15] F. Chiu, L. Fromont, A. Lee, and Y. Xu, "Long-distance anticipatory vowel-to-vowel assimilatory effects in French and Japanese," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*, 2015, no. 1008.

[16] P. P. G. Boersma and V. J. J. P. van Heuven, "Speak and unSpeak with PRAAT," *Glot Int.*, vol. 5, no. 9/10, pp. 341–347, 2001.

[17] P. K. Kuhl, "A new view of language acquisition," *Proc. Natl. Acad. Sci. United States Am.*, vol. 97, no. 22, pp. 11850–11857, 2000.

[18] J. E. Flege, "Second language speech learning: Theory, findings, and problems," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 233–277.

[19] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementaries," in *Language experience in second language speech learning: In honor of James Emil Flege*, M. J. Munro and O.-S. Bohn, Eds. Amsterdam: John Benjamins, 2007, pp. 13–34.

[20] C. T. Best, "A direct realist view of cross-language speech perception," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 171–204.

[21] J. E. Flege and R. F. Port, "Cross-language phonetic interference: Arabic to English," *Lang. Speech*, vol. 24, no. 2, pp. 125–146, 1981.

[22] Y. Hirata, E. Whitehurst, and E. Cullings, "Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates.," *J. Acoust. Soc. Am.*, vol. 121, no. 6, pp. 3837–45, 2007.

[23] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.