PRODUCTION OF CANTONESE TONES BY MANDARIN-SPEAKING IMMIGRANTS: ACOUSTIC AND PERCEPTUAL MEASUREMENTS

Yike Yang[1], Dong Han[1], Sze Man Wong[1], Chak Sum Chan[1], Chun Yeung Leung[1], Xiaocong Chen[2]

[1]Hong Kong Shue Yan University, [2]The Hong Kong Polytechnic University
yyang@hksyu.edu, dhan@hksyu.edu, 201042@hksyu.edu.hk,
221043@hksyu.edu.hk, 221039@hksyu.edu.hk, xiaocong.chen@polyu.edu.hk

ABSTRACT

Although Cantonese has a complex tonal system, there is a lack of research on adult learners' acquisition of second language (L2) Cantonese tones, particularly studies of learners with a tone language background. The present study attempted to explore whether Mandarin-speaking immigrants could acquire the Cantonese tonal system and whether there would be category assimilation or dissimilation of lexical tones in their L2 Cantonese. A tone production experiment involving 41 participants was conducted, and both acoustic and perceptual measurements were employed to analyse the speech samples. The immigrants showed a smaller tonal space in comparison with the native speakers; they also had very low accuracy rates in their tone production, indicating that they had not fully acquired the Cantonese tonal system. Explanations for the confusion patterns are provided, and the effects of the first language on L2 tone acquisition are discussed.

Keywords: speech production, speech prosody, lexical tone, second language, Cantonese

1. INTRODUCTION

1.1. Lexical tones in Cantonese and Mandarin

In Cantonese and Mandarin, lexical tones contribute to differences in word meanings. As shown in Table 1, Cantonese has six lexical tones, with three level tones (T1, T3 and T6), two rising tones (T2 and T5) and one falling tone (T4) [1]. Unlike Cantonese, the four lexical tones in Mandarin are distinguished by pitch contours [2]. The differences between the two tonal systems may bring about interactions of bilingual speakers' Cantonese and Mandarin.

Table 1: The tonal systems of Cantonese and Mandarin.

	Cantonese tonal system		Mandarin tonal system	
Name	**Category**	**Letter**	**Category**	**Letter**
Tone 1 (T1)	High Level	55	High Level	55
Tone 2 (T2)	High Rising	25	Rising	35
Tone 3 (T3)	Mid Level	33	Dip-rising	214
Tone 4 (T4)	Mid-low Falling	21	Falling	51
Tone 5 (T5)	Mid-low Rising	23	N/A	N/A
Tone 6 (T6)	Mid-low Level	22	N/A	N/A

According to recent studies, several Cantonese tone pairs (such as T3 and T6) are merging in either the production or the perception [3], which is probably part of a sound change in contemporary Cantonese [4]. If immigrants in Hong Kong are exposed to different types of Cantonese input (some merged and some not), it is interesting to explore whether they can maintain the contrasts amongst the six tones in their second language (L2) Cantonese.

1.2. Modelling L2 speech production

To account for the differences in the learnability of L2 speech, the Speech Learning Model (SLM) [5] and its revised version, the Revised Speech Learning Model (SLM-r) [6], have proposed that the processes and mechanisms that guide first language (L1) speech acquisition remain intact and accessible for L2 speech learning across the lifespan. According to the SLM and the SLM-r, there is a common phonetic space in a bilingual speaker's mind that stores the phonetic categories of both the L1 and the L2, and in which the L1 and L2 categories exert mutual influence.

One hypothesis in the SLM is the category assimilation hypothesis (CAH), which posits that an L2 sound that is perceived as being similar to an L1 sound does not form a new category in the common space and is understood as a variant of the L1 sound at an allophonic level. In this case, only one single phonetic category is used to process the two linked diaphones, and this mapping will eventually give rise to a new merged category, a phenomenon that has been documented in several studies [7]–[9]. Another claim in the SLM is the category dissimilation hypothesis (CDH). A new category will be established if an L2 sound is absent in the L1 system, which will make the combined phonetic space more crowded, resulting in the phonemes tending to disperse to ensure that the phonetic contrast is maintained. Recent support for the CDH comes from [10], which showed that Spanish-Catalan bilinguals had developed two categories to accommodate the mid-back vowels in the two languages.

1.3. The present study

Studies of Cantonese tone acquisition have mainly focused on younger populations, such as monolingual children [11], bilingual children [12] and bilingual youths [13], but adult L2 learners' acquisition of Cantonese tones has not been investigated. Adults are fundamentally different from younger populations because adult learners have fully acquired their L1 when they begin to acquire their L2. In addition, the acquisition of L2 tones by learners with a tonal language background has received little attention. Thus, the present study attempts to examine whether adult learners with a tonal language background (Mandarin-speaking immigrants) can acquire the complex Cantonese tonal system, and whether there is category assimilation or dissimilation of lexical tones in their L2 Cantonese due to the influences of their L1 Mandarin, with the aim of determining whether the hypotheses of the SLM and the SLM-r hold for lexical tones. The purpose of this study is to answer the following research questions:

1) Are there acoustic and perceptual differences in the production of Cantonese tones by native speakers and immigrants?

2) Is there any category assimilation or dissimilation in the Cantonese tone production of the immigrants?

2. METHODS

2.1. Participants

Two groups of participants attended a tone production experiment in a soundproofed booth at a local university. The participants were 32 Mandarin-speaking immigrants who were born and raised in Northern China, and who had spoken Mandarin as their only Chinese dialect prior to their arrival in Hong Kong after puberty. To assess their language profile, the immigrants completed a language background questionnaire [14] prior to the recording session. The results of the questionnaire indicated that the immigrants were fluent speakers of Cantonese, although they were more dominant in Mandarin. Nine native Cantonese speakers were included as the control group. The Cantonese speakers was born and raised in Hong Kong, where Cantonese is the dominant language. None of the participants reported any history of speech, language or hearing disorders.

2.2. Materials and procedures

The target stimuli were 12 monosyllabic words contrasting the six lexical tones in two base syllables (/si/ and /fu/), as listed in Table 2. Another ten monosyllabic words with different consonant and vowel combinations were included as the filler trials. Both the target and filler syllables are of high frequency in Cantonese. The syllables were presented in two contexts; that is, in isolation and in a carrier phrase '我讀__呢個字 *ngo5 duk 6_ nei1 go3 zi6* (I read the character _)'. Each stimulus appeared twice. In total, there were 1,968 target trials (2 target syllables * 6 tones * 2 context * 2 times * 41 speakers).

Table 2: The target syllables.

Tone	Syllable
T1	詩 /si1/ 'poem', 夫 /fu1/ 'husband'
T2	史 /si2/ 'history', 苦 /fu2/ 'bitter'
T3	試 /si3/ 'to try', 富 /fu3/ 'rich'
T4	時 /si4/ 'time', 扶 /fu4/ 'to hold'
T5	市 /si5/ 'market', 婦 /fu5/ 'married woman'
T6	事 /si6/ 'thing', 父 /fu6/ 'father'

The participants were briefed about the task requirements and were allowed to read the stimulus list prior to the experiment. During the experiment, the stimuli were presented randomly in E-Prime 2.0 [15] on a computer screen, and the production sessions were recorded in Audacity [16] at a sampling rate of 44,100 Hz.

2.3. Data processing and analysis

The vowel portions of the target syllables, which bear lexical tones, were segmented manually in Praat [17]. Following the segmentation, 20 time-normalised F0 values were extracted for each vowel using a Praat script. To eliminate individual differences in the F0 range and make direct cross-group/speaker comparisons possible, the F0 values, which were originally measured in Hz, were converted to a five-point scale from 1 to 5 with Equation 1:

$$c_m = ((5-1) * \frac{o_m - min_n}{max_n - min_n}) + 1 \qquad (1)$$

where c_m and o_m represent the converted and original F0 values of the m^{th} point, respectively, and max_n and min_n stand for the maximal and minimal values of all the original F0 values of the n^{th} speaker.

The F0 values were analysed using generalised additive mixed models (GAMMs) via the 'mgcv' package [18] in R [19], [20], in which *the converted F0 values* was the dependent variable, and *time*, *group*, *tone*, *context* and *syllable* were included as the predictors. The GAMM was adopted because it does not assume a linear relationship between the dependent variable and the predictors, which makes it appropriate for modelling time-dependent datasets such as the data in this study. The figures were plotted using the 'ggplot2' package in R [21].

Perceptual evaluations were also included in this study. Two native speakers of Hong Kong Cantonese who did not exhibit tone merging were invited to listen to all the target trials. The trials were presented randomly to the listeners; thus, the listeners did not know whether the speaker was a native speaker or an immigrant before they began the judgement task. The listeners were instructed to judge which character (corresponding to a tone) each trial represented and were allowed to listen to the trials several times if they deemed it necessary. The listeners completed the task independently, and the agreement amongst them was 77.69%. We adopted the more stringent criterion from [22]: Only the trials that both listeners considered to be the intended tones were counted as having been pronounced correctly.

3. RESULTS

3.1. F0 contours of the tone production

Figure 1 presents the F0 contours of the six tones produced by Cantonese speakers and immigrants. According to Figure 1(a), the Cantonese speakers clearly distinguished the six tones in their production. For the immigrants, as shown in Figure 1(b), T1 was extremely high and far from the remaining five tones, and the five tones were crowded in a narrower space in comparison to the native speakers. In addition, the native speakers showed a greater F0 range than did the immigrants.

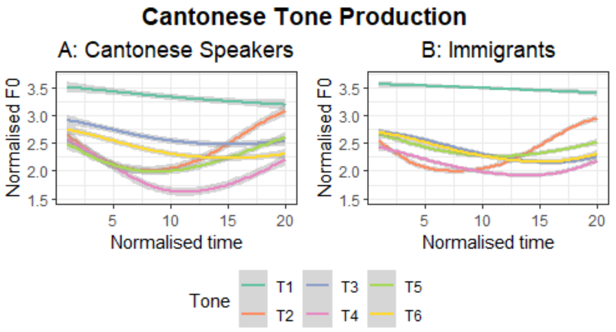

Figure 1: Production of Cantonese tones by Cantonese speakers and immigrants.

3.2. Acoustic analyses

A GAMM was first fitted with the entire dataset. There were main effects of *time*, *group*, *tone*, *context* and *syllable* (*ps* < .05). The predictors *context* (in isolation and in sentences) and *syllable* (/si/ and /fu/) are beyond the scope of this study. In the following of this section, the predictors *group* and *tone* will be investigated in more detail, with *time* always being included as a one-dimensional smooth.

A GAMM was constructed for native speakers' F0 data, with *tone* as the predictor and T1 as the reference. The results suggested that T1 was statistically different from the remaining tones, and it always had higher F0 values (*ps* < .001). More GAMMs were fitted to test whether the proposed merging-in-progress tone pairs were distinguishable. The models indicated that none of the tone pairs (T2 and T5, T3 and T6, T4 and T6, and T4 and T5) had merged in our data (*ps* < .001).

The same procedures were applied to the immigrants' data. The immigrants also showed very high F0 values for T1 compared to the remaining five tones (*ps* < .001). Although the immigrants exhibited a narrower tonal space for the remaining five tones, many of the tone pairs were separable by GAMMs: T2 and T5 (*p* < .001), T3 and T4 (*p* < .001), T3 and T5 (*p* = .027), and T4 and T6 (*p* < .001). However, T3 and T6 produced by the immigrants were merged as one category (*p* = .467).

Next, separate GAMMs were fitted to compare each tone produced by the native speakers and the immigrants, with data from the native speakers as the reference. Table 3 lists the statistics from the GAMMs, which suggests divergence between the two speaker groups for all the tones except for Tone 6, which was also marginally significant. Specifically, while the immigrants' T2, T3 and T6 were not as high as the same tonal categories that were produced by the native speakers, their T1, T4 and T5 were higher than were those of the native speakers. Given that the F0 values had been normalised to the same scale, it can be inferred that the immigrants failed to pronounce the Cantonese tones in a native-like way.

Table 3: Statistics from the GAMMs (with native speakers as the reference).

Tone	Estimate ± SE	T value	P value
T1	0.152 ± 0.028	5.429	< .001
T2	-0.053 ± 0.022	-2.402	< .001
T3	-0.252 ± 0.028	-11.09	< .001
T4	0.144 ± 0.023	6.644	< .001
T5	0.174 ± 0.023	7.696	< .001
T6	-0.042 ± 0.023	-1.835	0.067

3.3. Perceptual evaluations

The overall accuracy rates of the two groups were calculated first. As introduced in Section 2.3, only the tokens that were judged as being the intended ones by both listeners were counted as accurate. The tones produced by the native Cantonese speakers were generally perceived as being correct, with the accuracy rate of each tone ranging from 66.67% to 91.67% (average rate: 79.40%). However, the immigrants made many tone production errors, with an average accuracy rate of 44.99% (ranging from 20.83% to 68.75%); in other words, the two native listeners considered more than half of the trials to have been pronounced incorrectly.

Figure 2 displays the production accuracy rates for each tone. As can be seen in Figure 2(a), the native speakers' T2 and T4 were always perceived as the intended tones (100% accuracy rate), and T1 was also almost always perfect (97.22% accuracy rate). The accuracy rate for T3 was 72.22%, and the majority of the incorrectly identified T3 syllables were perceived as being T6. For T5 and T6, the accuracy rates were slightly above 50%. The T5 syllables tended to be perceived as T2, and the T6 syllables were likely to be perceived as T3.

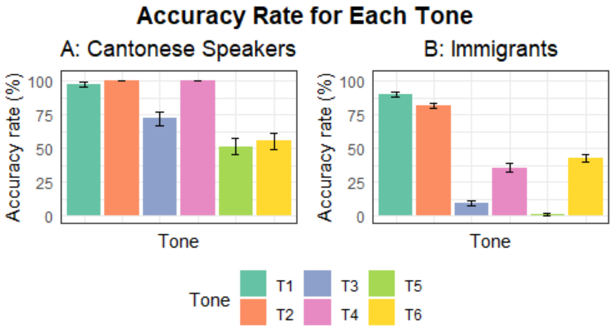

Figure 2: Accuracy rate of Cantonese tone production by Cantonese speakers and immigrants.

The accuracy rates for each tone produced by the immigrants are presented in Figure 2(b), according to which T1 and T2 had the highest accuracy rates, although they were not as accurate as were the native speakers' T1 and T2. The accuracy rates for the remaining four tones were very low, and the tones that the four tones were most likely to be perceived as are presented in Table 4. Both T3 and T4 were very likely to be perceived as T6, and T6 tended to be perceived as T4 or T3, thus suggesting that the immigrants had difficulty in distinguishing the T3-T6 and T4-T6 pairs. Another observation was that T5 was perceived more frequently as T6 than it was as T2.

Table 4: Listeners' judgements of immigrants' tones.

Tone	Perceived as
T3	T6 (68.18%), T4 (24.12%)
T4	T6 (54.48%), T2 (29.60%)
T5	T6 (45.40%), T2 (31.43%)
T6	T4 (48.94%), T3 (31.34%)

4. DISCUSSION AND CONCLUSION

This study investigated Mandarin-speaking immigrants' production of Cantonese tones using acoustic analyses and perceptual evaluations. The acoustic results suggested that the native speakers had a larger tonal space than did the immigrants, which is in line with previous findings that Cantonese speakers exhibit larger F0 range than Mandarin speakers [23]. Consequently, the native speakers clearly distinguished the six tones, and the immigrants' T2 to T6 were extremely crowded and even revealed the phenomenon of tone merging [13]. The perceptual evaluations confirmed the acoustic data. Although the native listeners generally considered the immigrants' T1 and T2 to have been pronounced correctly, the overall accuracy rates for the immigrants' tone production were very low, which suggested that they had not acquired the Cantonese tonal system completely.

A closer examination of the data showed common confusion patterns in the immigrants' tone production: T3-T6, T4-T6 and T5-T6. The first two pairs have been reported as being tones that are undergoing the process of merging [4], which could partially account for the immigrants' pattern from their linguistic input. However, more explanations are needed as native speakers did not always exhibit T3-T6 or T4-T6 confusion patterns. It is possible that T3, T4 and T6 are challenging for Mandarin speakers because these tonal categories are missing in their L1 system, and the acoustic and perceptual differences between these tones and their L1 tones are huge. It is therefore difficult for Mandarin speakers to establish the new categories when they are learning Cantonese, given the similarities between T3 and T6 (in pitch direction) and between T4 and T6 (in pitch height) within the tonal system of Cantonese.

The confusion between T5 and T6 was unexpected because the two tones differ in pitch direction, and native speakers do not confuse these two tones. A possible source of the confusion might be the learners' difficulty in correctly establishing the tonal categories because T5 and T6 share similar pitch height (23 vs 22). This indicates that the acoustic or perceptual similarity within a tonal system might surpass linguistic input and cause confusion in L2 category formation. Further studies should be conducted to compare the weight of linguistic input and phonetic similarity in L2 speech acquisition.

The data provide support for CAH and CDH from the SLM and the SLM-r. The identification accuracy rate for T5 was below 2%, suggesting that the T5 category was missing in the immigrants' Cantonese. As there is only one rising tone in Mandarin, it is likely that the learners had merged T5 with T2 to form one rising category, which is an example of category assimilation. Note that the immigrants also merged T5 with T6, which may be explained by the more similar pitch height between T5 and T6. Moreover, the immigrants attempted to raise their T1 to make it farther away from the remaining tones, and their T1 category was apparently different from that of the native speakers based on the acoustic analysis, which lends support to the CDH, although whether the immigrants' Cantonese T1 and Mandarin T1 are separate still needs to be confirmed when their L1 and L2 T1 categories are compared directly. Such investigation will further our understanding of L1 and L2 speech interactions (e.g. [24]).

5. ACKNOWLEDGEMENTS

The authors would like to thank all the informants for their participation and the ICPhS committee and reviewers for their constructive suggestions. This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No.: UGC/FDS15/H15/22) and an ASA International Student Grant from the Acoustical Society of America to the first author.

6. REFERENCES

[1] R. S. Bauer and P. K. Benedict, *Modern Cantonese Phonology*. Berlin: Walter de Gruyter, 1997.

[2] S. Duanmu, *The Phonology of Standard Chinese*. Oxford University Press, 2007.

[3] R. S. Y. Fung and C. K. C. Lee, “Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception,” *J. Acoust. Soc. Am.*, vol. 146, no. 5, pp. EL424–EL430, Nov. 2019, doi: 10.1121/1.5133661.

[4] P. P. K. Mok, D. Zuo, and P. W. Y. Wong, “Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese,” *Lang. Var. Change*, vol. 25, no. 3, pp. 341–370, Oct. 2013, doi: 10.1017/S0954394513000161.

[5] J. E. Flege, “Interactions between the native and second-Ianguage phonetic systems,” in *An integrated view of language development: Papers in honor of Henning Wode*, T. Piske, A. Rohde, and P. Burmeister, Eds. Trier: Wissenschaftlicher Verlag, 2002, pp. 217–244.

[6] J. E. Flege and O.-S. Bohn, “The Revised Speech Learning Model (SLM-r),” in *Second Language Speech Learning: Theoretical and Empirical Progress*, R. Wayland, Ed. Cambridge: Cambridge University Press, 2021, pp. 3–83.

[7] R. C. Major, “Losing English as a First Language,” *Mod. Lang. J.*, vol. 76, no. 2, pp. 190–208, 1992, doi: 10.2307/329772.

[8] C. B. Chang, “Rapid and multifaceted effects of second-language learning on first-language speech production,” *J. Phon.*, vol. 40, no. 2, pp. 249–268, 2012, doi: 10.1016/j.wocn.2011.10.007.

[9] M. L. Sancier and C. a. Fowler, “Gestural drift in a bilingual speaker of Brazilian Portuguese and English,” *J. Phon.*, vol. 25, no. 4, pp. 421–436, 1997, doi: 10.1006/jpho.1997.0051.

[10] M. Simonet, “Production of a catalan-specific vowel contrast by early Spanish-Catalan bilinguals,” *Phonetica*, vol. 68, no. 1–2, pp. 88–110, 2011, doi: 10.1159/000328847.

[11] P. P. K. Mok, H. S. H. Fung, and V. G. Li, “Assessing the Link Between Perception and Production in Cantonese Tone Acquisition,” *J. Speech, Lang. Hear. Res.*, vol. 62, no. 5, pp. 1243–1257, May 2019, doi: 10.1044/2018_JSLHR-S-17-0430.

[12] Y. Yao *et al.*, “Cantonese tone production in pre-school Urdu–Cantonese bilingual minority children,” *Int. J. Biling.*, vol. 24, no. 4, pp. 767–782, Aug. 2020, doi: 10.1177/1367006919884659.

[13] A. C. L. Yu, C. W. T. Lee, C. Lan, and P. P. K. Mok, “A New System of Cantonese Tones? Tone Perception and Production in Hong Kong South Asian Cantonese,” *Lang. Speech*, vol. 65, no. 3, pp. 625–649, Sep. 2022, doi: 10.1177/00238309211046030.

[14] D. Birdsong, L. M. Gertken, and M. Amengual, “Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism,” *COERLL, University of Texas at Austin*, 2012. https://sites.la.utexas.edu/bilingual/.

[15] W. Schneider, A. Eschman, and A. Zuccolotto, *E-Prime User’s Guide*. Pittsburgh: Psychological Software Tools Inc, 2012.

[16] Audacity Team, “Audacity(R): Free Audio Editor and Recorder.” 2019, [Online]. Available: https://audacityteam.org/.

[17] P. Boersma and D. Weenink, “Praat: doing phonetics by computer.” 2015, [Online]. Available: http://www.praat.org/.

[18] S. N. Wood, *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2017.

[19] R Core Team, “R: A Language and Environment for Statistical Computing.” R Foundation for Statistical Computing, Vienna, Austria, 2018, [Online]. Available: https://www.r-project.org.

[20] RStudio Team, “RStudio: Integrated Development for R.” RStudio, Inc., Boston, MA, 2016, [Online]. Available: http://www.rstudio.com/.

[21] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Cham: Springer, 2016.

[22] P. P. K. Mok and A. Lee, “The acquisition of lexical tones by Cantonese–English bilingual children,” *J. Child Lang.*, vol. 45, no. 6, pp. 1357–1376, Nov. 2018, doi: 10.1017/S0305000918000260.

[23] Y. Yang, S. Chen, and X. Chen, “F0 Patterns in Mandarin Statements of Mandarin and Cantonese Speakers,” in *Proc. Interspeech 2020*, Oct. 2020, pp. 4163–4167, doi: 10.21437/Interspeech.2020-2549.

[24] Y. Yang, “First Language Attrition and Second Language Attainment of Mandarin-speaking Immigrants in Hong Kong: Evidence from Prosodic Focus,” Ph.D dissertation, The Hong Kong Polytechnic Univ., Hong Kong, 2022.