Effect of exposure on production and perception of ongoing level tone mergers in Hong Kong Cantonese

Yuhan Lin[1], Yao Yao[2], Jin Luo[3]

[1,2]The Hong Kong Polytechnic University, [3]University of Groningen
yuhan.cn.lin@polyu.edu.hk; y.yao@polyu.edu.hk; j.luo6@student.rug.nl

ABSTRACT

This paper examines the plasticity of speech production and perception in sound change. It focuses on the merger between the mid-level (T3) and low-level tones (T6) in Hong Kong Cantonese and investigates 1) whether exposure to an unmerged talker affects the production and perception of this tonal contrast and 2) how speakers' baseline performance interacts with the exposure effect. Fourteen young speakers (F=10) participated in four production blocks (baseline, two shadowing blocks, post-task) in which they read T3 and T6 monosyllables as well as AX discrimination tasks on T3/T6 minimal pairs.

Significant exposure effect was only found in production among speakers who were more merged in the baseline production: T3/T6 difference was significantly greater in the shadowing and post-task blocks compared to the baseline. No exposure effect was found for perception. This finding indicates that the plasticity of speech production in sound change is likely subject to phonological constraints.

Keywords: Cantonese, tones, imitation, perception, mergers

1. INTRODUCTION

A critical issue in the sound change literature lies in the plasticity of speech production and perception. To investigate this issue, this study adopts the auditory naming paradigm to examine the ongoing merger between mid-level (T3) and low-level (T6) tones in Hong Kong Cantonese.

Hong Kong Cantonese has six lexical tones, as illustrated in Table 1. Following the tradition in Chinese linguistics literature, we use the five-point system [4] to represent the tonal contours in the language: five refers to the ceiling of one's F0 and one represents the floor. Recent research has documented several ongoing tonal mergers in Hong Kong Cantonese, especially those involving the mid and low tones. The T3/T6 merger, the focus of the current study, has been documented by Mok and colleagues [10]. They categorized the participants as merging and non-merging based on an auditory screening test and compared their production and perception of three Cantonese tonal mergers. Of the 169 speakers screened, only 28 were identified by the authors as potentially merging participants. Acoustic analysis showed that the merging participants had reduced "tone space" while retaining six tonal categories. They also observed much inter-speaker variability: for example, the misclassification rate of T3 as T6 based on predictive discriminant analysis ranged from 2.9% to 52.8%. In the AX discrimination task, the merging participants had nearly perfect accuracy rates, but were significantly slower than their non-merging counterparts in reaction time. These findings indicate that this merger is still incipient, which provides an excellent opportunity to examine the dynamics of the perception-production link in the process of sound change.

Table 1: Cantonese tones with examples.

Tone	Contour	Example	Gloss
T1	high-level	si^{55}	'poetry'
T2	high-rising	si^{25}	'history'
T3	mid-level	si^{33}	'try'
T4	low-falling	si^{21}	'time'
T5	low-rising	si^{23}	'market'
T6	low-level	si^{22}	'be'

Babel et al. [1] investigated the flexibility of New Zealand speakers' production of the ongoing NEAR/SQUARE merger using the auditory naming paradigm, in which participants imitated a model talker who distinguishes the two vowel classes. Acoustic analysis revealed that participants only became more unmerged in the post-task block, but not during shadowing.

Luo and Yao [8] adopted this paradigm to examine whether young Hong Kong Cantonese speakers (18-25 y/o) could reverse the T3/T6 merger via the imitation of an unmerged old male speaker (age=60). They found a significant increase in T3/T6 distinction in the two shadowing blocks, and the post-task reading had the greatest mean T3/T6 difference. However, since the study did not report the inter-speaker variability in baseline production, it remained

unclear whether all participants were merging T3 and T6 prior to the experiment.

Expanding on previous research, this study examines how the exposure to an unmerged model talker affects the production and perception of T3 and T6 among young Hong Kong Cantonese speakers. Additionally, it investigates how their baseline production and perception interact with this effect.

2. METHODOLOGY

2.1. Subjects

This paper reports an analysis based on data from fourteen Hong Kong Cantonese speakers (F=10) aged 18 to 25 years old.

2.2. Procedures

The experiment was carried out in the following order: baseline production, baseline AX discrimination, shadowing block 1, shadowing block 2, post-task production, post-task AX discrimination, and post-task questionnaire. The order of trials was randomized by block. For each block, practice trials (two for production, eight for perception) were provided.

In the baseline and post-task production blocks, the participants read out the characters in isolation. For each trial, a character was displayed in the middle of the screen for 2500ms after a fixation point was shown for 500ms. In the two shadowing blocks, the stimulus was played 200ms before the character was displayed. The participants were instructed to follow the talker in reading out the characters. In the AX discrimination trials, the question "are the two Cantonese pronunciations the same?" and the responses "same" (left) and "different" (right) were presented on the screen. The participants pressed the 'f' or 'j' key to indicate their responses. The correspondence between keys and responses was balanced across subjects. Reaction time was calculated from the onset of the second syllable.

The subjects participated in the experiment individually in a sound-attenuated booth. OpenSesame 3.2.5 was used for stimuli presentation and data collection.

2.3. Stimuli

Monosyllables were used in both production and perception parts of the study, and all reading materials were presented to the participants in traditional Chinese characters. All characters had more than 3,500 occurrences in the *Chinese Character Database* [12], a corpus for Cantonese pronunciations. Three T3/T6 minimal pairs, 12 T3 and 12 T6 syllables with no minimal pairs were included in the production tasks. In the baseline and post-task blocks, all 30 syllables were produced. In the two shadowing blocks, two of the minimal pairs, 12 of the non-minimal-pair syllables (six for each tone) were used.

In the pre- and post-exposure AX discrimination tasks, the same 13 T3/T6 minimal pairs were used, including the three pairs from the production tasks. These monosyllables were used to construct 13 AX trials and 12 AA trials for each block. Of the 13 AX trials, seven were in the order of T3/T6, and six were presented as T6/T3. Six T3 and six T6 syllables were used in the AA trials.

Fillers were included in all blocks. All filler materials were of either T1, the high-level tone, or T2, the high-rising tone. Specifically, 13 minimal pairs and 26 non-minimal-pair syllables (six for each tone) were included. The minimal pairs differed in their segmental features. Similar to the critical trials, three minimal pairs and 26 non-minimal-pair syllables were used in the baseline and post-task production blocks, and two minimal pairs and 12 non-minimal-pair syllables were included in the shadowing blocks. For each perception block, 13 AX and 12 AA filler trials were included, matching the critical trials.

Stimuli in the shadowing blocks and the AX discrimination tasks were produced by the same phonetically-trained 24-year-old male Cantonese speaker who was born and raised in Hong Kong. The recordings were made in a sound-attenuated booth at a sampling rate of 44.1 kHz. The speaker was aware of the experimental design and self-reported that he regularly makes the distinction. The mean T3/T6 difference of the stimuli is 23.02Hz or 1.85(T) after normalization (see Section 2.4). The T3/T6 difference tested significant in a paired t-test for the 13 T3/T6 minimal pairs ($t(13)=46.982$, $p < 0.001$) and in a t-test for the 12 T3 and 12 T6 syllables ($t(20.423)=21.524$, $p < 0.001$).

2.4. Analysis

In total, 1288 critical syllables (92 syllables * 14 participants) were collected for the production study, however, 5 tokens were excluded from the analysis due to speech errors (e.g. producing T3 as T2). Syllable boundaries were automatically marked in Praat and hand-corrected for alignment errors. The F0 for each syllable was extracted at 12 equidistant points using a script. In order to reduce the variation due to physiological differences across subjects, F0 values in Hz were normalized using formula (1). This widely adopted formula [6] [13] transforms the F0 values into T, a scale that is comparable to the traditional five-tone tone system. The mean of the middle two-thirds of the normalized F0 values was

taken as the dependent variable in the statistical analysis.

(1) $T = 5 * ((\log(F0_x) - \log(F0_{min})) / (\log(F0_{max}) - \log(F0_{min}))$

where $F0_x$ is the pitch value at a given time point, $F0_{max}$ and $F0_{min}$ represent the maximum and minimum pitch value of a given speaker respectively

For perception, accuracy and reaction time data were collected from 364 (26 pairs * 14 participants) AX trials and 336 (24 pairs * 14 participants) AA trials respectively. In order to eliminate inattentive responses, those with reaction time (ms) two standard deviations away from the mean were excluded, resulting in the removal of 4% of the dataset. In order to satisfy the normal distribution model assumption, log-transformed reaction time was used in the hypothesis testing.

The statistical analysis of both production and perception data was conducted with mixed-effects modeling in R [11] using the lme4 package [3]. P-values for factors were determined using log-likelihood comparison, and p-values for levels were generated using the lmerTest package [7].

3. RESULTS

3.1. Production

The production tasks examine whether participants converge to the unmerged talker, and how their baseline production and perception interacts with the convergence effect. Therefore, two mixed-effects models with similar structure were run with normalized F0 (T) as the dependent variable, one for the production interaction, and the other for the perception interaction. In the production interaction model, the independent variables included tone (T3, T6), block (baseline, shadow 1, shadow 2, post-task), by-speaker mean T3/T6 difference in baseline (normalized F0), the three-way and two-way interactions between these factors, as well as minimal pair presence (yes, no). In the perception interaction model, the by-speaker T3/T6 difference was replaced with the by-speaker mean log-transformed reaction time for the AX pairs. The presence of a significant three-way interaction would indicate the effect of baseline performance on imitation. Design-driven maximal random effect structure [2] was included: by subject intercept, tone by subject slope, block by subject slope, and by syllable intercept.

Log-likelihood model comparisons revealed significant three-way interaction for the production interaction ($\chi^2(3) = 9.258$, $p=0.026$), but not the perception interaction model. In order to further explore the effect of baseline production, participants were divided into two groups based on their normalized F0 (T). Given that T3 and T6 are traditionally represented as 33 and 22 respectively in the five-point system as shown in Table 1, we decided to use 0.5 as the cut-off line, resulting in eight merging and six non-merging participants.

For the merging participants, the interaction between tone and block was significant ($\chi^2(3) = 30.283$, $p<0.001$). As shown in Table 2, which reports the output for the merging model, the T3/T6 differences in the shadowing blocks and post-task are significantly greater than in the baseline. Figure 1 illustrates the mean normalized F0 (T) by block: T3/T6 difference increases from the first to the second shadowing block, but decreases in the post-task block. The F0 (T) for both T3 and T6 are raised during the shadow and post-task blocks, but the shift in T3 was much greater. This pattern is likely a result of the greater acoustic space available for the mid- than low-level tone in Hong Kong Cantonese. For the non-merging group, as displayed in Figure 2, the mean T3/T6 difference increased during the shadowing blocks, but became even smaller than the baseline in the post-task condition. Such an interaction between tone and block was only trending in the statistical analysis ($\chi^2(3) = 6.557$, $p=0.087$).

Table 2: Model estimates and standard errors

	Estimate (std. error)
(Intercept)	1.583 (0.137)***
Tone = T6	-0.265 (0.084) **
Block = post-task	0.393 (0.155) *
Block = shadow 1	0.466 (0.09) ***
Block = shadow 2	0.579 (0.104) ***
Minimal pair = yes	-0.07 (0.055)
T6 : post-task	-0.184 (0.064) **
T6 : shadow 1	-0.314 (0.079) ***
T6 : shadow 2	-0.410 (0.079) ***

* p < 0.05, ** p < 0.01, *** p < 0.001

for normalized F0 (T) for merging participants

3.2. Perception

The accuracy rates for the AX discrimination tasks were extremely high: only four errors were found across all participants, which corroborates Mok and colleagues' [10] findings.

For log-transformed reaction time, the AX and AA pairs were tested separately. As is with the case for normalized F0, for each pair type (AA, AX), two mixed-effects models were built to examine how the baseline production and perception performance interact with the exposure effect on log-transformed reaction time. The independent variable included

block (baseline, post-task), mean baseline normalized F0/mean baseline log-transformed reaction time, and their interaction, and the random effects included by subject intercept, block by subject slope, and by syllable intercept. For both AX and AA pairs, the mean log-transformed reaction time was shorter in the post-task block, but the difference was not significant. The two-way interaction did not test significant in any model, suggesting that baseline performance did not interact with the effect of exposure. This pattern contradicts Mok et al. [10], which found that the merging participants were significantly slower than their non-merging counterparts.

Figure 1: Normalized F0 (T) for T3 and T6 by block for merging participants. Error bars represent 95% confidence intervals.

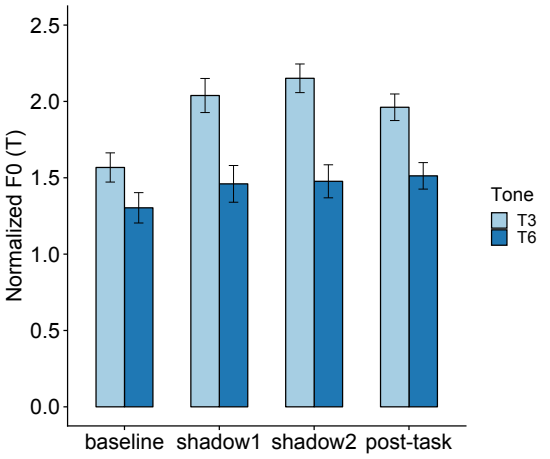
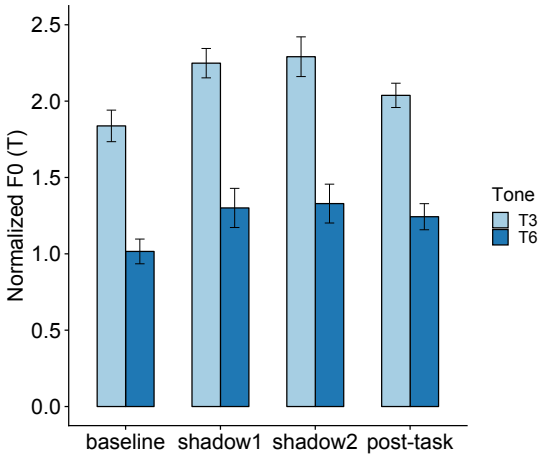


Figure 2: Normalized F0 (T) for T3 and T6 by block for non-merging participants. Error bars represent 95% confidence intervals.



4. DISCUSSION

The current study examines the plasticity of speech production and perception in an ongoing sound change. Specifically, it investigates whether the exposure to an unmerged talker affects the production and perception of a tonal pair (T3/T6) undergoing merger in Hong Kong Cantonese, and how speaker's baseline production and perception interact with this process. The significant effect of exposure was only found in the production of speakers who were more merged in the baseline production: T3/T6 difference was significantly greater in the shadowing and post-task blocks compared to the baseline.

This finding also has some implications for the link between perception and production, which constitutes another key question in sound change [14]. If there were a direct and unmediated perception-production link, we would expect all participants to show similar patterns of imitation regardless of their baseline production. Nonetheless, only the merging participants, namely, those who showed less T3/T6 distinction in the baseline exhibited imitation and maintained greater distinction in the post-task block. For the non-merging speakers, the trend was that the T3/T6 distinction only increased during shadowing but did not persist afterwards. The absence of significant imitation effect among the non-merging speakers suggests that the plasticity of speech production in a sound change is likely subject to phonological constraints. This finding corroborates Mitterer and Ernestus' [9] work which shows that phonologically irrelevant information is less likely to be imitated.

The effect of exposure for perception was absent in either accuracy or log-transformed reaction time. Given that Cantonese speakers consistently performed at ceiling in the standard AX discrimination task with monosyllables in Mok et al. [10] and the current study, it could be the case that significant perceptual confusability of tonal pairs would only arise in more challenging listening conditions.

With regard to perception, with contrasting the results from Mok and colleagues [10], I found no significant difference in reaction time between the merging and non-merging participants. This discrepancy may result from the different methods for distinguishing between merging and non-merging participants in the two studies. In Mok et al. [10], the merging participants were identified through an auditory screening by the authors, whereas in the current study, the distinction of the two groups was determined by acoustic measures.

5. REFERENCES

[1] Babel, M., McAuliffe, M., Haber, G. 2013. Can mergers-in-progress be unmerged in speech accommodation? *Frontiers in psychology*. 4, 653.

[2] Barr. D., Levy R., Scheepers. C., Tily H. J. 2013. Random effects structure for confirmatory hypothesis testing. *J. Mem Lang.*, 68(3), 255-278.
[3] Bates, D., Maechler, M., Bolker, B. Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Stats. Software.* 67(1), 1-48.
[4] Chao, Y. -R. 1930. A system of tone-letters. *Le Maître Phonétique*. 45, 24–27.
[5] Evans, B. G., Iverson, P. 2007. Plasticity in vowel perception and production: A study of accent change in young adults. *J. Acoust. Soc. Am.* 121(6), 3813-3826.
[6] Fung, R. S. Y., Wong, C.S.P. 2011. Acoustic analysis of the new rising tone in Hong Kong Cantonese. *ICPhS XVII*, 715-718.
[7] Kuznetsova, A., Brockhoff. P. B., Christensen, R. H. B. 2017. LmerTest package: Tests in linear mixed effects models. *J. Stats. Software*. 82(13), 1-26.
[8] Luo, J., Yao, Y. 2017. Undoing the "lazy accent": Gender, age and language attitude in reversing Hong Kong Cantonese tone merger via phonetic imitation. *4th Workshop on Sound Change*, Edinburgh.
[9] Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, 109(1), 168-173.
[10] Mok, P. P. K., Zuo, D., Wong, P. W. Y. 2013. Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese, *Language Variation and Change*. 25, 341-370.
[11] R Development Core Team. 2018. R: A language and environment for statistical computing [Computer software]. http://www.r-project.org.
[12] Research Centre for the Humanities Computing, Chinese University of Hong Kong. 2003. *Chinese character database: With word-formations phonologically disambiguated according to the Cantonese dialect. https://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/.*
[13] Shi, F., Wang, P. 2006. Beijinghua danziyin shengdiao de tongji fenxi [A statistical analysis of the tones in Beijing Mandarin]. *Zhongguo Yuwen.* 1, 33-40.
[14] Stevens, M., & Harrington, J. (2014). The individual and the actuation of sound change. *Loquens*. 1(1), 003.