Language Variation and Change, **25** (2013), 341–370.
© Cambridge University Press, 2013 0954-3945/13 $16.00
doi:10.1017/S0954394513000161

Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese

PEGGY P. K. MOK, DONGHUI ZUO AND PEGGY W. Y. WONG
The Chinese University of Hong Kong

ABSTRACT

Cantonese has six lexical tones (T), but some tone pairs appear to be merging: T2 [25] vs. T5 [23], T3 [33] vs. T6 [22], and T4 [21] vs. T6 [22]. Twenty-eight merging participants and thirty control participants in Hong Kong were recruited for a perception experiment. Both accuracy rate and reaction time data were collected. Seventeen merging participants also participated in a production experiment. Predictive discriminant analysis of the fundamental frequency data and judgments by native transcribers were used to assess production accuracy. Results show that the merging participants still had six tone categories in production, although their "tone space" was more reduced. Tones with lower type frequency were more prone to change. The merging group was significantly slower in tone perception than the control group was. In illustrating the patterns of the ongoing tone merging process in Cantonese, this study contributes to a better understanding of the forces of sound change in general.

INTRODUCTION

In recent years, the tone system in Hong Kong Cantonese appears to be changing in that some speakers no longer distinguish all six lexical tones. However, only a few studies have reported evidence for this. The present study aims to address this gap by examining both the production and perception of the potential tone mergers, and also to investigate possible causes of such changes.

Cantonese tone system

Hong Kong Cantonese, a variety of Cantonese, is a Chinese language well known for its complex tone system, which has attracted much research interest (e.g., Fok-Chan, 1974; Gandour, 1981; Khouw & Ciocca, 2007). There are six lexical tones (T) and three allotones in Hong Kong Cantonese (Bauer & Benedict, 1997). Each syllable (usually corresponding to a morpheme) carries a tone. The six lexical tones are: T1 (high-level [55]), T2 (high-rising [25]), T3 (mid-level [33]), T4 (low-falling

The authors would like to thank the Department of Linguistics and Modern Languages for supporting this project. Thanks also go to Prof. Shui-duen Chan for allowing us to use the written Cantonese corpus, and to Angel Chan for providing the perception stimuli and the reference production materials. The authors also thank the anonymous reviewers for their very helpful comments and suggestions.

TABLE 1. *Cantonese tones with examples*

Register	Lexical tones			Allotones
High	T1 high-level [ji55] ‘clothes’	T2 high-rising [ji25] ‘chair’	T3 mid-level [ji33] ‘idea’	T7 high-stopped [jɪk5] ‘benefit’ T8 mid-stopped [jak3] ‘eat’
Low	T4 low-falling [ji21] ‘suspicious’	T5 low-rising [ji23] ‘ear’	T6 low-level [ji22] ‘two’	T9 low-stopped [jɪk2] ‘wing’

[21]), T5 (low-rising [23]), and T6 (low-level [22]). The numbers in [ ] represent the relative starting and ending pitch levels of each tone, with 5 being the highest and 1 being the lowest pitch level of a speaker’s normal pitch range (Chao, 1930, 1947). The six lexical tones appear in open syllables or syllables with nasal endings [-m, -n, -ŋ]. The three allotones are traditionally called the “entering tones” in Chinese phonology. They only appear in syllables ending with unreleased stops [-p, -t, -k]: T7 (high-stopped [5]), T8 (mid-stopped [3]), and T9 (low-stopped [2]). They are much shorter in duration and are considered allotones of the three corresponding unstopped level tones T1, T3, and T6, respectively (Bauer & Benedict, 1997; Chao, 1947). Table 1 shows all the tones with examples.

Previous studies have shown that fundamental frequency (F0) is the primary and sufficient cue for tonal distinction in Cantonese (Fok-Chan, 1974; Khouw & Ciocca, 2007; Vance, 1977). Creakiness in T4 [21] can also contribute to the identification of that tone (Yu & Lam, 2011), but it is an optional supplementary feature. Figure 1 shows for reference the F0 traces of the six lexical tones produced by the first author who distinguishes all six tones clearly in production. It can be seen that pitch height, pitch direction, and magnitude of change are important features for tonal distinction in Cantonese (Gandour, 1981; Khouw & Ciocca, 2007). The three level tones (T1, T3, T6) are distinguished by tone height. T1 [55] is well separated from all other tones by being at the top of the speaker’s normal pitch range. There is a larger difference in F0 between T1 [55] and T3 [33] than between T3 [33] and T6 [22] (only around 30 Hz for the tokens in Figure 1; see also data in Khouw & Ciocca, 2007). T2 [25] and T5 [23] have a rising contour, but differ in the magnitude of change. T2 rises to the highest pitch level, whereas T5 only has a minimal rise. T4 [21] and T6 [22] differ slightly in contour with T4 having a small fall throughout the syllable. T5 [23] and T6 [22] differ in the opposite direction, with T5 having a small rise. It can be seen from Figure 1 that the Cantonese “tone space” is very crowded in the lower pitch range, with four tones sharing the same low starting point [2], and four tones ending in the low to mid pitch range.

Given such subtle differences in a narrow pitch range, it is not surprising that some tone pairs are easily confusable, especially in isolation. Tone pairs with similar contours and/or little contrast in height—that is, T2 [25] vs. T5 [23], T3 [33] vs. T6 [22], T4 [21] vs. T6 [22]—can be confusable even for adult native

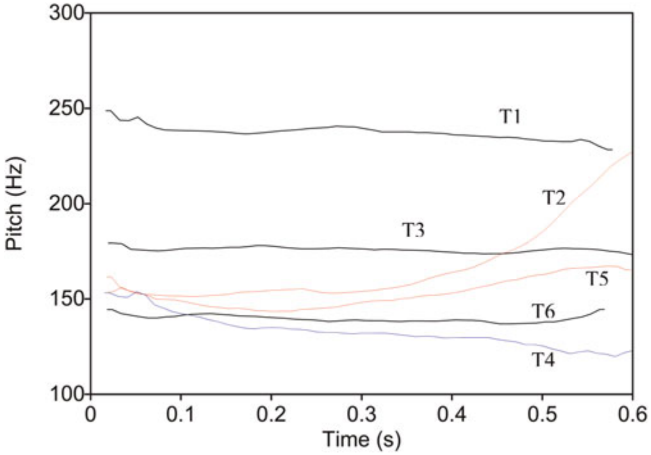

FIGURE 1. F0 traces of the six lexical tones on the syllable [ji] spoken by a female speaker.

speakers (Fok-Chan, 1974; Varley & So, 1995). Typically developing children acquiring Cantonese also find T2 [25] vs. T5 [23], T3 [33] vs. T6 [22] particularly hard to distinguish and acquire them last (e.g., Ciocca & Lui, 2003; Wong, Ciocca, & Yung, 2009). Wong et al. (2009) also found that tone pairs with large differences in F0 height or contour—such as, T1 [55] vs. T4 [21] and T1 [55] vs. T2 [25]—are the easiest for children to distinguish.

Tone merging in Hong Kong Cantonese

In the past decade, researchers in Hong Kong have noticed that some Cantonese speakers no longer distinguish all six tones in their production, most notably T2 [25] vs. T5 [23]. Kei, Smyth, So, Lau, and Capell (2002) reported several cases. They compared the tone production of 15 participants against a control group of 56 speakers who were screened for accuracy of their tone production. They found that 6 of their 15 participants did not clearly distinguish the two rising tones (T2 [25] and T5 [23]) in their production. Kei et al. (2002) considered that these speakers made "tone production errors," as the purpose of their study was to establish a normative profile of Cantonese tones for speech therapists. Bauer, Cheung, and Cheung (2003) followed up on Kei et al.'s (2002) findings from a tone merging perspective. They investigated the tone production of eight male speakers and found that two of their speakers produced the rising tones unconventionally. They concluded that there are tonal subvarieties coexisting in the Hong Kong speech community.

These two studies only investigated the production of the T2/T5 merger. A more recent study investigated both the production and perception of this tone pair. Yiu (2009) found that 5 of her 15 participants confused some of the T2/T5 pairs in perception. In production, three participants appeared to be merging the two rising tones. Yiu (2009) found that the participants' production performance could not be used to predict their perception performance, and vice versa. Taken together, these three studies suggest that some Hong Kong Cantonese speakers are merging the two rising tones, with different possible patterns: low-rising merging into high-rising (T5 [23] → T2 [25]); high-rising merging into low-rising (T2 [25] → T5 [23]); and having a novel intermediate realization.

Impressionistically, some young speakers also confuse the mid-level (T3 [33]) and the low-level tones (T6 [22]), as well as the low-level (T6 [22]) and the low-falling tones (T4 [21]) in their production. However, thus far, very few studies have investigated the production or perception of other potential tone mergers in Hong Kong Cantonese (see Fung, Kung, Law, Su, & Wong, 2012, for T4/T6; and Yu, 2007, for a morphological tone merge). In addition, Killingley (1985) argued that the mid-level T3 [33] and the low-rising T5 [23] have merged in Malayan Cantonese and proposed that the same is also true in Hong Kong Cantonese. However, there is no concrete evidence in that study to support this claim, and the data in Fok (1974) and Kei et al. (2002) show that these two tones are clearly distinct. Therefore, more studies are needed to investigate the merging tone patterns in Cantonese and the factors contributing to such patterns.

Additionally, it is interesting to note that in Kei et al.'s (2002), Bauer et al.'s (2003), and Yiu's (2009) studies, the speakers who confused the two rising tones did not merge all words with the two target tones. They were able to differentiate some T2/T5 words. Their results indicate that these speakers probably still have two separate rising tone categories, although one category is weaker. A logical question to ask is: What contributes to this pattern? However, these three studies gave no clear answer.

A possible factor affecting tone merging is word frequency. The effect of word frequency is an important factor in language change. Hooper (1976) and Bybee (2007) distinguished two types of word frequency effects on sound change. Reductive sound change (i.e., changes involving deletion/weakening of speech sounds) tends to affect high-frequency words first, whereas analogical sound change (i.e., changes happening based on an analogy taken from related patterns) usually starts from low-frequency words first. Bybee (2007; cf. Hooper, 1976) argued that reductive sound change originates from the automation of speech production, whereas analogical sound change stems from imperfect learning. Similarly, Philips (1984) suggested that reductive sound change is motivated by physiological factors, whereas analogical sound change is motivated by conceptual factors. It is unclear to which type of sound change tone merging belongs, and what its effects on tone merging are.

High-frequency words are shown to be more susceptible to surface variations and are often more reduced in speech (Bybee, 2007; Jurafsky, Bell, Gregory, & Raymond, 2000; Wright, 2004). Zhao and Jurafsky (2007, 2009) showed that word frequency can affect tone production in Cantonese. They found that low-frequency words are hyperarticulated and are produced with a higher pitch. Low-frequency words also have a more expanded pitch range than that of high-frequency words, that is, the tones of low-frequency words are more dispersed in the tonal space. Although they did not investigate tone mergers, it is conceivable that more tone mergers may be found in high-frequency words because they are more susceptible to surface variations and sound change, and because they have a more compressed pitch range, which renders the tones less distinct. In addition, high-frequency words are often articulated more quickly and less distinctly. Nokes and Hay (2012) showed that an increase in speech rate can decrease pitch variation. Their data support the preceding assumption about word frequency (high) and tone production (reduced) quite well.

Yiu's (2009) data also show some support for the effects of word frequency. In Cantonese, there are more T2 [25] words than T5 [23] words (Fok-Chan, 1974; Leung, Law, & Fung, 2004). She found that her participants appeared to confuse T2 words more often than T5 words. However, the results only suggest that type frequency (i.e., how many distinct items are represented by a particular pattern) may affect tone merging, but it is unclear whether and how token frequency (i.e., the number of times a unit appears in a defined discourse) can affect the merging of tones in Cantonese, as none of the previous studies on Cantonese tone merging has manipulated token frequency systematically.

In fact, very few studies have investigated the effects of word frequency on tone merging. Myers and Li (2009) investigated the effects of lexical (token) frequency on syllable contraction in Southern Min. The tonal contours of two syllables are merged into one if the syllables are contracted. They found that the degree of this tone merge correlated positively with word frequency. Although the reductive tone merge in Southern Min syllable contraction is qualitatively different from the phonological tone merge in Cantonese being considered here, their results do indicate that token frequency can indeed affect tone production. In addition, Hildebrandt (2003) found that word frequency could affect the retention/merging of tones in Manange (a Tibeto-Burman language of Nepal), but the effects were also dependent on the degree of access to Manange and Nepali. Therefore, in view of Yiu's (2009), Myers and Li's (2009), and Hildebrandt's (2003) results, it would be interesting to see whether and how tone merging in Cantonese is affected by word frequency.

The present study

Our study investigated both the production and perception of the potential tone mergers in Hong Kong Cantonese. In addition to the T2/T5 pairs that previous studies focused on, we included other potentially merging pairs as well. Previous studies compared fundamental frequency data for tone production, but they had very limited data from only a few merging speakers (e.g., only two merging speakers in Bauer et al., 2003). We systematically screened a large number of speakers and collected production data from 17 merging speakers (see details in the participants section), in order to reveal a more comprehensive picture of the tone merging phenomenon.

As tone contour is a time-varying feature, it can be characterized using multiple point measurements of fundamental frequency, each of which can be treated as a variable for modeling purposes. The present study made use of predictive discriminant analysis to examine whether the merging tones were distinct from each other. Predictive discriminant analysis is a multivariate test that shows how different cases group themselves based on a set of variables. Some of the previous studies on tone production have also used this statistical test to quantify how distinctive two tone categories are. For example, Kratochvil (1986) used predictive discriminant analysis to predict whether the sandhi-ed T3 in Mandarin (when two T3s are put side by side, the first T3 becomes a T2; see Lin, 2007:97, for examples) was distinct from the canonical T2. One of the main goals of the current study is to see whether the merging participants distinguish the merging tone pairs in their production. We believe that multivariate predictive discriminant analysis is a proper test for this purpose.

In addition to comparing acoustic data, we also include data that were auditorily transcribed by native speakers who clearly distinguish all six tones in their production and perception. This is designed to ensure that both intermediate and merged tokens can be accounted for, which can also show us perceptual judgment of the participants' production by native listeners, in addition to the results by machine recognition (discriminant analysis).

Moreover, we investigate the effects of word frequency on tone merger by manipulating token frequency systematically in the production experiment. We would like to see whether high-frequency and low-frequency words show different patterns in the ongoing tone merger.

Although Yiu (2009) included perception data in her study of the T2/T5 merge, using both accuracy and reaction time data of many more nonmerging and potentially merging participants in the present study can give us a more comprehensive picture of tone perception. Finally, previous studies recruited the participants randomly, so they only had a small number of participants who appeared to be merging tones. In order to target the merging speakers, we screened a large number of young participants for our study.

METHODS

Participants

Previous studies suggest that the merging process is a relatively recent phenomenon. We therefore targeted young speakers for our study. To recruit merging speakers, a simple screening test for production was conducted. Each potential participant was recorded reading a list of 18 words (3 different syllables × 6 tones) embedded in a short carrier phrase. Their recordings were auditorily checked by the first and the third authors (both native speakers of Cantonese who clearly distinguish all six tones) to determine which speakers were likely to merge the tones. The screening was based on impressionistic judgments. Only those who were judged to have noncanonical tone production (of any tone) by both judges were selected as merging participants. In total, 169 people were screened, and 28 potentially merging participants were identified by the first and the third authors based on their tone production. Due to time limitations and various logistic constraints, 17 of the 28 identified merging participants participated in both the production and perception experiments, and the other 11 merging participants participated in the perception experiment only. An additional 30 participants who clearly distinguish all six tones in their production were used as a control group for the perception experiment. Thus, 58 participants participated in the perception experiment, of which 28 were merging (21 female, 7 male) and 30 were control (23 female, 7 male), and 17 merging participants (14 female, 3 male) participated in the production experiment. They were all local undergraduate students at the Chinese University of Hong Kong between 18 and 22 years of age. There was no age difference between the two groups. The participants reported having no history of hearing problems or language disorders. They were paid to participate in the experiments.

Materials

Word frequency was included as a factor in the production experiment. Because no corpus of spoken Cantonese with information of word frequency was available to us when we conducted our study, we used a corpus that is based on written text in order to manipulate word (token) frequency systematically. High- and low-frequency monosyllabic and disyllabic words were selected from an electronic database of around 33,000 Cantonese words extracted from a 1.7 million character corpus of Hong Kong newspapers (see details in Chan & Tang, 1990). Only the monosyllabic data are presented in this paper. Six high-frequency and six low-frequency monosyllabic words were chosen for each of the six tones. Word frequency was based on the log frequency (base 10) of the token frequency in the corpus. The chosen words were selected from the highest (logged frequency ranging from 3.14 to 4.10) and lowest frequency (logged frequency ranging from 0 to 1.59) brackets in the corpus in order to maximize their difference in word frequency. This resulted in 72 target monosyllabic words (6 tones × 6 words × 2 frequencies). Monosyllables ending with a stop coda (i.e., the allotones mentioned in the Cantonese tone system section) were not selected. All the monosyllabic words were randomized and embedded in a short carrier phrase: [ŋɔ^{23} tʊk^{2} ___ tsi^{22}] 'I read the word ___' when presented to the speakers in the production experiment. All materials were presented in Chinese characters, which show no information about tones.

The perception experiment using Cantonese monosyllables was an AX discrimination task with two types of materials: 120 AA pairs and 120 AB pairs. The AA pairs had the same stimuli (e.g., T1/T1 of the same syllable [ji 55]), and the AB pairs had different stimuli that differed only in tones (e.g., T2/T5 of the same segmental syllable [ji 25] and [ji 23]). Two criteria were used in selecting the stimuli: (i) 10 different syllables of each tone were included; (ii) 5 of the 10 syllables had all six tones attested in Cantonese, for example, a full set such as [ji] shown in Table 1, while the other five syllables do not appear in all six tones, for example, a deficient set such as [wan] with T3 [33] being unattested. Altogether, 60 target monosyllables were chosen; 56 of the target

TABLE 2. *Mean durations (ms) of the perception monosyllables produced by a female speaker in isolation*

Tone	Mean	Minimum	Maximum
T1	513	445	605
T2	514	389	666
T3	568	450	735
T4	515	434	646
T5	533	417	661
T6	577	469	727

monosyllables were also used in the production experiment. Word frequency was not manipulated systematically in the perception experiment, but the numbers of high- and low-frequency items for each tone were balanced. The 60 target monosyllables all appeared in the AA pairs (same stimuli). An additional 60 dummy AA pairs with completely different materials were used to balance the number (120) of the AB pairs (different stimuli). The dummy items were excluded from analysis. Two syllables were used to construct the AB pairs (2 syllables × 6 tones = 12 tokens). Each token was paired up with other tokens having the same segments, but with different tones (e.g., T1 vs. T2, T1 vs. T3). The order of the AB pairs was counterbalanced, that is, both A/B and B/A sequences were included. This resulted in 120 AB pairs (2 syllables × 6 tones × 5 matching tones × 2 orders). A female researcher who is also a qualified speech therapist in Hong Kong produced all the monosyllables in isolation for the perception experiment. The average duration of these monosyllables can be found in Table 2. Unattested matching tones for the AB pairs were presented to her in transcription. She had no difficulty in producing these tokens. The 120 AA and 120 AB pairs were randomized in the perception experiment.

Procedures

The 17 potentially merging participants participated in the production experiment first, lest they should become aware of the purpose of the experiments at the outset and thus be extra careful in their performance. Informal inquiry after the production experiment confirmed that the participants were oblivious to the aim of the experiment. The participants were recorded reading the monosyllabic materials in Chinese characters. Two randomized lists of the same materials were used for counterbalancing. Nine participants read the first list and eight participants read the second list. Before the actual recording, the participants practiced by reading the materials for as long as they liked to ensure natural production. Short breaks were given during the recording. The recording took place in a sound-treated booth at the Chinese University of Hong Kong. Their speech was recorded directly onto hard disk with a sampling rate of 22,050 Hz using Praat (Boersma, 2001) via a condenser microphone placed approximately 20 cm away from the participants. They were recorded reading the materials with a normal speech rate. Three repetitions of the materials were taken, so altogether there were 216 tokens per speaker (72 target monosyllables × 3 repetitions).

The 17 merging participants were invited back for the perception experiment at least two weeks after the production experiment. An additional 11 merging participants and 30 control participants were also included. The perception experiment was an AX discrimination task using the 120 AA and 120 AB pairs previously mentioned. The participants were asked to judge whether the two syllables they heard were of the "same" or "different" tones. There was a short practice session to familiarize the participants. The experiment was divided into four blocks. The stimuli within each block were randomized for each subject. Short breaks were scheduled between blocks.

The participants participated in the perception experiment individually in a quiet room at the Chinese University of Hong Kong. The stimuli were presented to them at a comfortable level via a Logitech Clearchat Premium PC Headset through a sound card using E-Prime 2.0 Professional (Psychology Software Tools, Pittsburgh, PA) with a desktop computer. Both accuracy and reaction time data were collected using the PST Serial Response Box. No feedback, either visual or auditory, was given. The interstimulus interval was 500 ms. Reaction time was calculated from the onset of the second monosyllable. All reaction time values longer than 3000 ms were counted as missing responses and were excluded from analysis. This resulted in the loss of 0.6% of the data.

Acoustic measurements

The beginning and end of periodicity of the target syllable (if the whole syllable consisted of sonorants) or the vowel (if the syllable had a nonsonorous initial consonant) in the production data were marked manually based on the spectrogram (using the start of F1 as the beginning and the end of F2 as the end of periodicity). Ten equidistant measurement points were taken between these two marks, so the data were time-normalized for each tone. The fundamental frequency (F0) values of each point were tracked automatically using the autocorrection algorithm and then checked manually using Praat (Boersma, 2001). Due to the perturbation from initial consonants or creakiness, data from the first (i.e., the beginning of periodicity) and the tenth (i.e., the end of periodicity) measurement points were excluded from analysis (Rose, 1987). Only data from the second to the ninth measurement points were used for comparison. Very creaky tokens (e.g., T4 [21]) were excluded from acoustic analysis because F0 values could not be measured, which resulted in the loss of 9% of the total dataset (about 60% of which were T4 [21] tokens).

In addition to acoustic measurements, all tokens of tone production by the 17 merging speakers were identified by two independent native Cantonese transcribers with phonetic training. The transcribers did not participate in either the production or the perception experiments. They were prescreened using a tone identification task to ensure that they were not merging any of the tones. The transcribers heard all the monosyllables in carrier phrases so they could determine the tones with reference to the context.

RESULTS

Production

The main questions in the production experiment were whether the 17 merging participants distinguished the merging tone pairs in their production, and whether word frequency had any effect on their tone production. Because previous studies showed that there would be much variation in the participants' merging patterns, it is important to consider their production data individually. To this end, we analyzed the acoustic data using predictive discriminant analysis for three reasons. First, discriminant analysis is a multivariate test that can evaluate between-group differences on the basis of a combination of several variables (Tabachnick & Fidell, 2007:375), e.g., F0 measured at several time points along the tone contour. Second, discriminant analysis makes use of functions that maximize between-group differences relative to within-group differences. As F0 data always involve some within-speaker differences, discriminant analysis is an ideal method to deal with this issue. Third, predictive discriminant analysis generates classification accuracy rates ("classification rate" hereafter) for the data, which is a straightforward way to quantify the distinctiveness of different groups.

Predictive discriminant analysis predicts group membership from a set of predictors. In the current study, the dependent variable Tone was taken as the grouping variable, and four independent variables were used as the predictors: the F0 values at the second, fifth, sixth, and ninth measurement points of the pitch contour, representing the onset, middle, and offset pitch values of the tone. Because T1 [55] is well separated from the other tones and is not involved in any merging tone pair, we did not include T1 in the statistical test. We subsequently conducted discriminant analysis again including T1 to check for the robustness of the classification rates, and there was no significant difference between the results based on only five tones and those including T1. Therefore, we only reported the results based on five tones here for a more focused investigation of the merging tone pairs.

We first conducted discriminant analysis using SPSS (SPSS Inc., Chicago, IL) for high- and low-frequency words separately for each subject, but there was no significant difference between the high- and low-frequency words in their classification rates within each subject: high-frequency words (mean = 87.206, SD = 7.866) vs. low-frequency words (mean = 87.694, SD = 9.990) [$t(16) = -.291$, $p = .775$]. To have an overall picture of the merging tone pairs with more data points, high-frequency tokens and low-frequency tokens were then analyzed together. Therefore, for each speaker, the dataset contained five groups (T2, T3, T4, T5, T6) with 36 members each (6 syllables × 2 word frequencies × 3 repetitions), which adds up to at most 180 tokens per speaker, as creaky tokens and outliers were not included in the analysis. Both univariate and multivariate outliers were excluded before classification. Cases with a *z*-score greater than 3.29 were regarded as univariate outliers, and cases with a

TABLE 3. *Misclassification rates (%) of the merging tone pairs and the overall misclassification rates (%) of two reference speakers (R) and 17 merging subjects (M)*

	Misclassification rates						
	T2 → T5	T5 → T2	T3 → T6	T6 → T3	T4 → T6	T6 → T4	Overall
R1 (F)	5.7	0	5.6	8.6	0	2.9	5.8
R2 (F)	8.6	0	5.6	5.6	2.9	0	5.2
R mean	7.2	0	5.6	7.1	1.5	1.5	5.5
M1 (F)	13.9	8.3	13.9	2.8	2.8	0	9.4
M2 (M)	40.6	20.0	22.9	8.6	60.0	5.7	29.3
M3 (F)	6.7	0	24.1	14.3	0	0	13.9
M4 (F)	15.6	11.4	35.5	11.1	8.6	5.6	21.3
M5 (F)	11.1	2.8	19.4	8.3	0	0	10.1
M6 (M)	19.4	11.1	19.4	25.0	11.1	2.8	15.6
M7 (M)	14.7	8.6	11.1	17.6	27.3	2.9	20.5
M8 (F)	20.6	5.6	52.8	41.7	0	0	27.7
M9 (F)	11.5	4.5	11.1	14.3	100.0	3.6	17.6
M10 (F)	2.9	0	11.1	5.7	0	2.9	6.8
M11 (F)	2.9	0	8.8	9.1	46.2	6.1	12.5
M12 (F)	3.8	3.0	25.7	28.6	5.6	2.9	17.7
M13 (F)	26.5	17.6	5.7	21.2	37.9	27.3	41.2
M14 (F)	26.5	14.7	25.0	8.6	11.1	5.7	25.5
M15 (F)	5.6	0	17.1	5.6	5.9	0	8.5
M16 (F)	16.7	31.6	30.3	21.2	0	0	22.9
M17 (F)	22.2	17.1	2.9	2.9	11.1	0	14.0
M mean	15.4	9.2	19.8	14.5	19.3	3.9	18.5

Note: Misclassification rates higher than 10% are shaded for reference. (M) = male; (F) = female.

Mahalanobis distance greater than the critical chi-square value were regarded as multivariate outliers (Tabachnick & Fidell, 2007:73–75).

A high classification rate in discriminant analysis indicates clear group membership, that is, clear separation of tones. If the tone pairs produced by the participants were merging, these pairs would have similar distributions so that some tokens were likely to be misclassified into a wrong category (e.g., T2 tokens misclassified as T5). This would result in a low classification rate.

We obtained reference data from two native speakers—the female researcher who produced the perception materials and the first author—who clearly distinguished all six tones using the same experimental materials for comparison. Both speakers had a very high overall classification rate (94.2% and 94.4%), whereas the 17 merging participants had a much lower overall classification rate (mean = 80.5%, SD = 9%). Among the 17 merging participants, the highest classification rate was 93.2% and the lowest classification rate was 58.5%. Table 3 shows the misclassification rates of the merging tone pairs of the two reference speakers and the 17 merging participants. A larger number indicates more misclassification, that is, more overlap between the tones. We only report the misclassification rates for the merging tone pairs here for two reasons. First, this is the focus of the present study. Second, there are 25 tone pairs total, and the misclassification rates of the nonmerging pairs were very low (most of them were 0%). Other details about the results of the predictive discriminant analysis can be found in Appendix A.

We can see from Table 3 that even for the two reference speakers, there was slight overlap in the merging tone pairs given their acoustic similarity. Nevertheless, the merging participants had at least one tone pair being misclassified for more than 10% of time, and some pairs were misclassified for a substantial degree (over 30%). Some participants also had very substantial overlap for more than one merging tone pairs (e.g., M2 and M16).

In addition to the T2/T5 pair identified by previous studies, there is clear evidence in the misclassification rates that the T3/T6 and T4/T6 pairs are also merging at different paces. The data also suggest that the merging of tones is not complete, as most speakers show only partial overlap (albeit to a substantial degree for some) in the merging tone pairs, in other words, the tones were classified above chance level (50%). Only one speaker, M8, seemed to merge T3 and T6 completely; 52.8% of her T3 tokens were misclassified as T6 (binomial test, $p = .868$) and 41.7% of her T6 tokens were misclassified as T3 (binomial test, $p = .608$), indicating that her T3/T6 discrimination was approaching chance level. Moreover, the individual misidentification rates reveal that the merging of tones is not symmetrical. For most speakers, T2 tokens were more often misidentified as T5 by discriminant analysis than T5 tokens were misidentified as T2, and T4 tokens were more often misclassified as T6 than T6 tokens as T4.

Figure 2 shows the distribution of the production data in the space of the first two canonical discriminant functions[1] of four female speakers for visual illustration: R2 (overall classification rate of 94.4%), M10 (highest overall classification rate of 93.2%), M13 (lowest overall classification rate of 58.5%), and M8 (near-complete merge between T3/T6). The tone groups of R2 are well separated, whereas those of M13 are clustered together. Although the overall classification rate of M10 is very similar to R2, her T3 and T6 groups are much closer together and overlap more than those of R2. The T3 and T6 tokens of M8 largely overlap and the group centroids are also very close, suggesting a near-complete merge between T3 and T6 for this speaker.

Discriminant analysis only predicts group membership based on the acoustic data at four static points. Although the results show that the speakers still have distinct tone categories (i.e., tone merging is still in progress), it is unclear whether these categories are canonical or not, and whether these tone categories differ between merging and nonmerging participants. To further explore these issues, we compared the F0 values at the ninth measurement point (i.e., the tone offset) of the three merging tone pairs between merging and nonmerging participants. We divided the participants based on the predictive discriminant analysis results in Table 3. Those with misclassification rates below 10% for a particular tone pair were considered nonmerging participants for that tone pair. Therefore, for T2/T5, R1, R2, M3, M10, M11, M12, and M15 are nonmerging participants; for T3/T6, R1, R2, M11, and M17 are nonmerging participants;

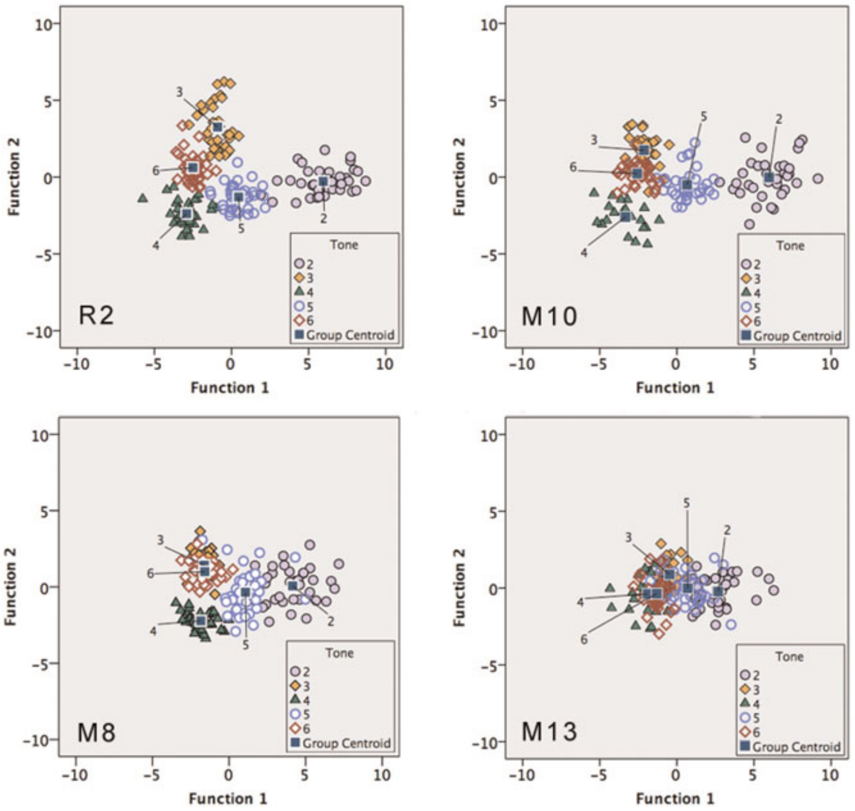

FIGURE 2. Scatterplots of the first two canonical discriminant functions of four female speakers.

and for T4/T6, R1, R2, M1, M3, M4, M5, M8, M10, M12, M15, and M16 are nonmerging participants.

We calculated a quotient for each merging tone pair for each speaker in order to normalize the data for comparison as we had both male and female speakers. The offset of T2 [25] should be higher than that of T5 [23], T3 [33] higher than T6 [22], and T6 [22] higher than T4 [21]. We took the F0 values at the ninth measurement points of the lower tones in these three merging tone pairs as the reference (denominator) and divided the F0 values at the same measurement points of the higher tones (numerator) by this reference. The resultant quotients can represent the distance between the tone pairs. A larger quotient indicates that the tones are more separated and farther apart in the "tone space." We used a very simple concept of tone space: the pitch range defined by the highest pitch [5] and lowest pitch [1] values of the lexical tones produced by individual speakers. Independent *t*-tests confirm that the nonmerging participants had significantly larger quotients than the merging participants for all three tone pairs, respectively: T2/T5 (1.19 vs. 1.01 [$t(17) = 5.845$, $p < .0001$]), T3/T6 (1.10 vs.

1.07 [$t(17) = 2.388$, $p = .029$]), T4/T6 (1.14 vs. 1.08 [$t(17) = 2.398$, $p = .028$]). These results clearly show that although the merging participants still had these tone categories, their tones were acoustically more similar than those of the nonmerging participants.

Furthermore, although discriminant analysis illustrates the grouping of the tokens produced by the merging participants based on their fundamental frequencies at four static time points, human perception of the tones may be different from machine recognition. It is unclear whether listeners can identify the target tones accurately given the fact that the distances between tone categories of the merging participants were smaller than those of the control participants. Therefore, we included a qualitative auditory analysis of the data. The produced tones were transcribed by two independent transcribers (see details in the acoustic measurements discussion). Table 4 shows the inter-rater reliability between the two transcribers. When the two transcribers gave the same judgment on a particular token (e.g., T2), it was counted as “agree.” When one transcriber gave an answer that was different from but was still related to the judgment of the other transcriber (e.g., one is T2, the other is between T2 and T5), it was counted as “partially agree.” When the answers of the two transcribers were not related (e.g., one is T2, the other is T5), it was counted as “disagree.” It can be seen that the inter-rater reliability is quite high, except for T5, which means that the qualitative data based on their judgment are reliable, and that T5 was heard to be produced with much variation.

Figure 3 shows the breakdown of tone judgments by the transcribers. The *x*-axis shows the target tones, and the *y*-axis shows the actual tones perceived by the transcribers (the six perceived tones and the “others” category being arranged from top to bottom). The blocks in each column represent the percentages of that particular target tone being perceived as each of the six tones or as others (i.e., between two tones or could not be clearly specified). Dotted lines indicate 0%. A token is considered to be produced as a particular tone in Figure 3 when both transcribers heard it as that tone (agree). Partially agree or disagree judgments were all grouped under others in Figure 3. We can see that T1 [55] is the most stable tone as all tokens were heard as T1 by both transcribers (except one high-frequency token mispronounced by a speaker as T5). T5 [23] is the most variable tone because many more tokens were heard as the other tones and as others by the transcribers. If we compare the merging tone pairs, we can see that T2 appears to be more stable than T5, T4 more stable than T6, and T3 and T6 appear to be of comparable variability.

Word (token) frequency does not seem to have any consistent effect on these merging tone pairs, as the perception patterns by the native judges were very similar in Figure 3. Separate *t*-tests for each tone comparing the high- and low-frequency words in terms of their percentages of change confirm that there is no significant difference between the high- and low-frequency words (T2: $t(10) = -.199$, $p = .864$; T3: $t(10) = -1.690$, $p = .122$; T4: $t(10) = 1.503$, $p = .164$; T5: $t(10) = .961$, $p = .359$; T6: $t(10) = .863$, $p = .408$).

Finally, an interesting pattern can be observed when we consider the likelihood of having noncanonical tone productions for individual words, as Figure 3 shows

TABLE 4. *Inter-rater reliability (%) between the two native transcribers*

Tone	High-frequency words			Low-frequency words			Overall		
	Agree	Partially agree	Disagree	Agree	Partially agree	Disagree	Agree	Partially agree	Disagree
T1	100	0	0	100	0	0	100	0	0
T2	94.1	3.3	2.6	92.5	3.3	4.2	93.3	3.3	3.4
T3	90.2	5.9	3.9	83.7	7.5	8.8	86.9	6.7	6.4
T4	94.4	2.6	2.9	97.1	0.3	2.6	95.8	1.5	2.8
T5	63.1	25.8	11.1	64.7	20.6	14.7	63.9	23.2	12.9
T6	79.4	8.8	11.8	83	4.2	12.7	81.2	6.5	12.3
Mean	86.9	7.7	5.4	86.8	6.0	7.2	86.9	6.9	6.3

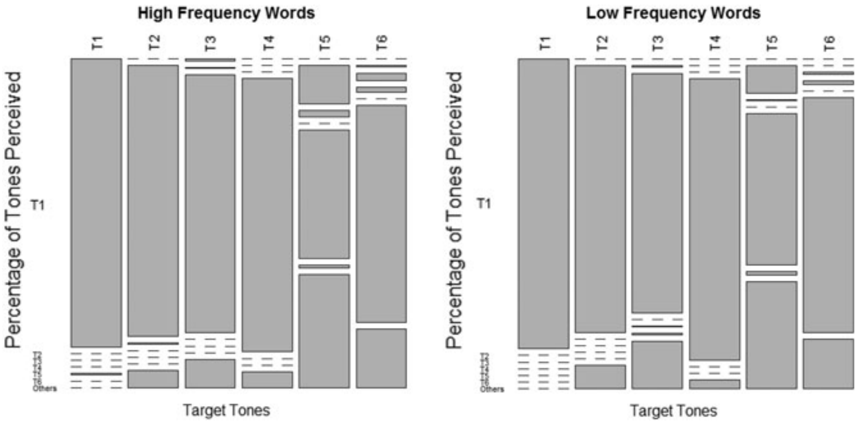

FIGURE 3. Tone judgment by the native transcribers.

that the tones differ in the transcribers' perception of their variability. One may expect that words with the same tone and the same word frequency may have similar susceptibility to change. However, this is not always the case. Six monosyllables were used for each tone in each word frequency group in the production experiment. Some words were much more susceptible to change than others were, even when they shared the same tone and word frequency. Table 5 shows the percentage of change for each monosyllable in different tone pairs. There were 51 tokens for each monosyllabic word (3 repetitions × 17 participants). The percentage shows how many of these 51 tokens were judged to be produced noncanonically. It is obvious that some words were correctly produced by all participants, but others were much more likely to have different pronunciation. High-frequency T2 words represent the most extreme case. Most of the variant pronunciation belongs to the word 者 [tsɛ] ('he who') (33.3%), whereas other words in this tone category remain very stable. Taken together, we can see that T5 is more variable than T2, and T6 more variable than T4, not only in terms of the frequency of change (Figure 3), but also in terms of the scope of change (Table 5). It is interesting to note that although T3 and T6 appear to have similar variability in Figure 3, they do not differ much in terms of the number of words affected.

Perception

Both accuracy and reaction time data were collected in the perception experiment (AX discrimination of Cantonese monosyllables). The production results show that the merging participants still have six distinct tone categories of the merging tone pairs, although with a reduced tone space compared with that of the nonmerging participants. This pattern can also be found in their perception data. Mixed analyses of variance were conducted to see whether merging and control participants show any difference in tone discrimination accuracy and

TABLE 5. *Percentage of change for each monosyllabic word*

High frequency				Low frequency				High frequency						Low frequency					
T2		T5		T2		T5		T3		T6		T4		T3		T6		T4	
點 dim ‘dot’	0	兩 lœŋ ‘two’	60.8	影 jɪŋ ‘shadow’	0	懶 laːn ‘lazy’	25.5	向 hœŋ ‘toward’	0	萬 maːn ‘ten thousand’	9.8	名 mɪŋ ‘name’	2.0	靚 lɛŋ ‘beautiful’	11.8	煉 lin ‘refine’	11.8	皇 wɔŋ ‘king’	0
此 ts^{h}i ‘this’	2.0	有 jau ‘have’	51.0	椅 ji ‘chair’	0	美 mei ‘beautiful’	58.8	次 ts^{h}i ‘times’	7.8	令 lɪŋ ‘order’	18	而 ji ‘and’	3.9	厭 jim ‘hate’	7.8	胃 wɐi ‘stomach’	17.6	圓 jyn ‘round’	0
九 kɐu ‘nine’	0	會 wui ‘can’	35.3	忍 jɐn ‘tolerate’	2.0	養 jœŋ ‘rear’	37.3	對 tɵy ‘right’	8	未 mei ‘not yet’	9.8	由 jɐu ‘from’	3.9	意 ji ‘idea’	21.6	譽 jy ‘fame’	19.6	勞 lou ‘labour’	0
很 hɐn ‘very’	0	每 mui ‘every’	62.7	毀 wɐi ‘destroy’	5.9	旅 lɵy ‘travel’	39.2	至 tsi ‘reach’	19.6	二 ji ‘two’	23.5	來 lɔi ‘come’	3.9	怨 jyn ‘blame’	19.6	潤 jɵn ‘moist’	21.6	忘 mɔŋ ‘forget’	7.8
種 tsʊŋ ‘type’	3.9	與 jy ‘and’	54.9	戀 lyn ‘love’	19.6	語 jy ‘language’	43.1	要 jiu ‘want’	11.8	是 si ‘yes’	39.2	和 wɔ ‘and’	7.8	印 jɐn ‘stamp’	23.5	貌 maːu ‘face’	19.6	聯 lyn ‘join’	3.9
者 tsɛ ‘he-who’	33.3	以 ji ‘by’	86.3	擁 jʊŋ ‘hug’	19.6	議 ji ‘discuss’	84.3	更 kɐŋ ‘more’	17.6	又 jɐu ‘again’	49.0	年 lin ‘year’	11.8	幼 jɐu ‘young’	19.6	護 wu ‘protect’	24.0	遊 jɐu ‘travel’	3.9

Note: Phonemic transcription in IPA symbols and the gloss can be found under each monosyllabic word. Many of these are polysemous words. Only the most common meaning is given here.

log-transformed reaction time. Tone pair (21 levels: T1 vs. T1, T1 vs. T2, T1 vs. T3, and so on) was treated as a within-subject factor, and merging status (two levels: control, merging) was treated as a between-subject factor. The result showed that these two groups of participants did not differ in accuracy ($F(1, 56) = 7.33$, $p = .396$, $\eta_p^2 = .013$). Both groups performed at ceiling (average overall accuracy over 97%). Nevertheless, the reaction time data reveal a very different picture. The merging participants were significantly slower than the control participants ($F(1, 56) = 8.003$, $p = .006$, $\eta_p^2 = .125$). Figure 4 shows their log-transformed reaction time data (LogRT) for both AA and AB pairs. The data were collapsed across presentation order in the experiment. Post hoc independent *t*-tests confirm that the merging participants were significantly slower than the control participants in all AA and AB pairs ($p < .05$), except the 2-4 tone pair (with the same pattern, just not significant $p = .127$). The consistent difference in reaction time was not confined to confusing merging tone pairs alone. It is worth pointing out that the AA pairs were very easy for native speakers to respond to, and yet, the merging participants still performed significantly slower than the control participants did. Their mean actual reaction time data in milliseconds can be found in Appendix B. It can be seen that the merging participants were on average about 150 ms slower than the control participants were, and they had more individual variation as reflected by larger standard deviations. In summary, although the merging participants could still distinguish the six tone categories in perception, they were slower in perceiving the tonal distinction than the control participants were in general, not only for the identified merging tone pairs.

DISCUSSION

Our results illustrate the patterns of the ongoing tone mergers in Hong Kong Cantonese at the beginning stage. Some young speakers are merging T2 with T5, T3 with T6, and T4 with T6 in production, although they still had six tone

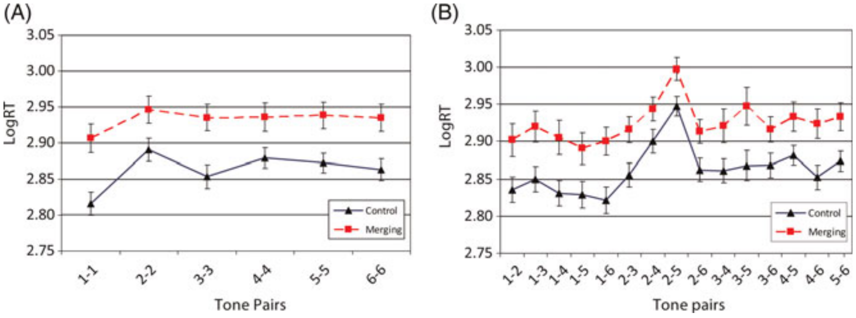

FIGURE 4. Log-transformed reaction time (LogRT) data for the (A) AA tone pairs and (B) AB tone pairs collapsed across presentation order. The error bars show one standard error. Except the 2-4 pair in (B), all comparisons between the merging and control participants are significant ($p < .05$).

categories. Lexical (token) frequency did not have any consistent pattern in the rates of tone change. In terms of perception, although the merging participants could still distinguish the six lexical tones in perception under experimental settings, they were much slower in tone perception than the control participants were in general, which is not confined to the identified merging tone pairs. It is best illustrated by the significant difference in log-transformed reaction time for the T1/T1 pairs between the two subject groups (see Figure 4). The T1/T1 pairs are the easiest for both groups of participants, and T1 is also the most stable and accurate tone for the merging participants. Still, the merging participants performed significantly slower than the control participants in their reaction time. One logical possibility of such a pattern is that the merging participants may be slower doing everything compared with the control participants. However, given our limited reaction time data on only one perceptual task, we cannot simply conclude that the merging participants were systematically slower overall in other activities as well. An alternative explanation is that the merging participants were very cautious in the perception experiment. They had to finish listening to the second syllable and needed a longer processing time in order to make the correct decisions, resulting in ceiling accuracy rates but very long reaction times. Informal enquiry after the perception experiment confirmed this possibility, as some merging participants did report that they were very careful in doing the perception experiment. Some side support for this explanation can be obtained from the control participants as well. They also found the T2/T5 pairs difficult to distinguish, resulting in the longest reaction time among all tone pairs (about 100 to 200 ms longer than other tone pairs, see Appendix B), probably because they were also very careful in distinguishing this tone pair. The extra time needed is consistent with the average difference in reaction time between the control and merging groups (about 150 ms). Thus, we believe that it is safe to say that the merging participants were slower than the control participants in making tonal distinctions in general.

As mentioned, although the merging participants were slower than the control group, their accuracy rate is still quite high (well above 90%), and they did not differ significantly from the control group. In addition to the possibility that they were performing at the ceiling because they still had six tone categories, it is also probable that because the perception data was based on a forced-choice AX discrimination task carried out under time pressure in an experimental setting, the merging participants were focusing their attention to the subtle acoustic cues that they normally would not notice in a naturalistic and noisy environment. Their long reaction time data is good evidence for this behavior. Moreover, the perception task was based on speech materials produced carefully by one trained expert speaker. Distinguishing tones by one careful speaker is much easier for the merging participants than distinguishing tones by multiple speakers, as in normal daily conversation. If we had used a more natural task, the merging participants might not be able to distinguish the tone pairs so well (which may result in an interesting situation similar to the "Bill Peter effect" showing discrepancy between production and perception of near-merger; see Labov, 1994:363–364, for more detail). At any rate, these merging participants should be regarded as incomplete "mergers." Therefore, the tone pairs are just beginning to merge in the language as a whole, and there is much individual variation.

The high degree of individual variation observed in our tone merging data echoes that of segmental sound changes. Production differences both across and within speakers were reported by Gordon and Maclagan (2001) in their 15-year longitudinal study of the NEAR/SQUARE diphthong merger in New Zealand English. They suggested that initial variability is a general characteristic of the process of sound changes. Such individuals provide an informative indication of the amount of variability during the progress of the merger.

The perception results correspond quite well with the idea that perceptual difficulty/confusion and listeners can contribute to sound change (Ohala, 1981, 1983). The merging participants found the acoustically similar tone pairs difficult in perception, which caused confusion in their own production (comparable to "hypo-correction" in Ohala's term). Whereas the merging participants performed the worst for the T2/T5 pairs in perception in terms of reaction time, they also confused more words in these two tones than other tone pairs in their production (Figure 3 and Table 5). Their relatively better perception performance for the T3/T6 pairs also corresponds with relatively better production of these two tones. Both the control and merging groups found the T1/T1 pair easiest to distinguish. This is not surprising given that T1 [55] is well separated from the other five tones by being at the top of the speakers' normal pitch range (Figure 1). It is perceptually more salient than other tones. T1 is also the most stable tone in the production of the merging participants. Our data clearly show that there is a link between perceptual difficulty and production accuracy. A number of reasons can explain why Yiu (2009) found no such link between production and perception in her data: (i) she had very few merging participants (five for production, three for perception); (ii) she only examined the T2/T5 pair in detail, whereas we included other merging tone pairs as well; and (iii) her data (accuracy data for perception and F0 contours for production) and analyses may not be sensitive enough to reveal the link between production and perception. At any rate, the actual mechanisms between production and perception in tone merging need to be further explored.

In addition, the adult data concur well with the acquisition patterns of Cantonese-speaking children. Both monolingual and bilingual Cantonese children easily confuse T2 [25] and T5 [23], and T3 [33] and T6 [22]. They also acquire them last (e.g., Ciocca & Lui, 2003; Wong et al., 2009), and they find T1 [55] the easiest and acquire it well before other tones (So & Dodd, 1995). The parallels between child and adult phonologies (in terms of child acquisition order/difficulty and adult merging pairs) strongly suggest that their patterns stem from the same underlying phonetic cause (Greenlee & Ohala, 1980): subtle differences in a narrow pitch range. The acoustic similarity between the merging tone pairs (see Figure 1) renders them difficult to acquire and particularly susceptible to confusion and sound change. These data suggest that language acquisition, particularly by children, is a likely factor of tone merge.

TABLE 6. *Type frequency of Cantonese tones*

Source	T1	T2	T3	T4	T5	T6	T7	T8	T9
Fok-Chan (1974)	17.1%	19.4%	14.3%	12.2%	11.3%	16.4%	4.6%	2%	2.6%
Leung, Law, & Fung (2004)	17.85%	16.49%	16.79%	11.24%	10.69%	17.87%	3.51%	2%	3.15%

This study systematically manipulated word frequency in the production experiment. However, the results do not seem to support an effect of word frequency, contrary to the tonal merger data that resulted from syllable contraction in Southern Min (Myers & Li, 2009). The F0 patterns of both high- and low-frequency words in the present study are similar, and the transcription data also do not show any consistent difference between high- and low-frequency words. Nevertheless, before we can conclude that word frequency does not have a role in Cantonese tone merging, we should carefully re-examine the effects of word frequency first.

As mentioned in the tone merging in Hong Kong Cantonese section, there are two types of frequency effects on sound change. Reductive sound change tends to affect high-frequency words first, whereas analogical sound change usually starts from low-frequency words first (Bybee, 2007; Hooper, 1976). Bybee (2007) argued that reductive sound change originated from the automation of speech production, whereas analogical sound change stems from imperfect learning. In addition, we also need to distinguish token and type frequency. The frequency manipulated in the present study is token frequency as it is based on how frequent a monosyllable appeared in a corpus of Hong Kong newspapers. However, the six lexical tones themselves also differ in type frequency. Table 6 shows the type frequency of all Cantonese tones according to Fok-Chan (1974) and Leung, Law, and Fung (2004). The type frequency patterns are very similar in both studies. It is obvious that T2 [25] is more frequent than T5 [23] and T6 [22] is more frequent than T4 [21], whereas T3 [33] and T6 [22] are similar in type frequency.

If we consider the differences between reductive and analogical sound change, and between token and type frequency, frequency does seem to play a role in Cantonese tone merging. Tonal merger in Southern Min reported in Myers and Li (2009) is reductive in nature because it is the result of syllable contraction. However, tone merging in Cantonese should not be regarded as reductive sound change, because producing T2 [25] instead of T5 [23] or T3 [33] instead of T6 [22] are not “phonetic shortcuts” due to the automation of speech production. Rather, tone merging in Cantonese should be regarded as analogical sound change resulting from perceptual difficulty and imperfect learning, as we have discussed. If that is the case, we would expect low type frequency words to be affected more than high type frequency words as they are under more pressure to conform. This is indeed what we found. Information given in Figure 3 and Tables 4 and 5 all show that T5 [23] (with the lowest type frequency) is more variable than T2 [25] (with a high type frequency) for words with both high- and low-token frequency. T3 [33] and T6 [22] have comparable type frequency, and they also have comparable variability in tone production. T1 [55] has a high type frequency and it is also very stable. Nevertheless, given the small number of tones in the inventory and the uncertainties about tone stability, it is possible that this observed pattern could be a coincidence. Despite this possibility, we believe that our data do lend some support for the effects of type frequency on tone merging.

Another possible factor contributing to the low type frequency of T5 is the phonetic difficulties in producing tones. Tone type frequency tends to negatively correlate with the phonetic difficulty of producing different tones (e.g., Zhang, 2001). It is possible that the relative difficulty in the finer control of the gentle slope of F0 change may lead to the low type frequency of T5 and thus makes T5 most susceptible to tone merging. If that is the case, articulatory processes might also be involved in the T5 to T2 merge, which could possibly be characterized as an ongoing reductive process. Clearly, more empirical data are needed to substantiate this possibility.

One intriguing question remains, however. T6 [22] has a higher type frequency than T4 [21], but T4 is much more stable than T6. This seems to contradict the patterns of type frequency and analogical sound change that we have discussed. Nevertheless, whereas T2/T5 and T3/T6 differ mainly in F0 patterns, T4 [21] and T6 [22] differ in voice quality as well. T4 [21] ends in the lowest pitch level, which often results in creaky voice. Given that creaky voice is an additional cue to T4 (Yu & Lam, 2011), it is understandable why T4 remains much more stable than T6 despite the difference in type frequency. In fact, Figure 3 and Tables 4 and 5 show that T4 is the second most stable tone after T1. The salient concomitant feature of creaky voice may override the effect of word frequency. Thus, we can conclude that the effects of word frequency (type) and sound change (analogical) can be found in Cantonese tone merging, but language-specific situations must also be taken into consideration.

If we compare the acoustic data with the transcription data, one discrepant pattern can be observed. In the acoustic data, more T2 tokens were misidentified as T5 than T5 tokens were misidentified as T2 by discriminant analysis (Table 3), whereas the transcribers perceived more T5 tokens as T2 than T2 tokens as T5 (Table 5). One possible reason may be because the auditory analysis was based only on the judgments of two independent transcribers; their transcriptions may not be as representative as the objective analyses of the acoustic data of the 17 speakers. Nevertheless, given the high inter-rater reliability between the two transcribers (Table 4), this should not be a major reason for the discrepant pattern. A more probable reason may be that human and machine recognition were based on different sets of data. Machine recognition (discriminant analysis) only relied on the F0 data of four static points, whereas human perception had the benefit of listening to the whole pitch contours in context. In addition, human perception may be more sensitive to differences in the slope of pitch change, which were not represented in the discriminant analysis data, particularly given that the major difference between T2 and T5 lies on the slope toward the end of the tone (see Figure 1). Machine recognition may also suffer from the fact that the endpoint of the pitch contour (the tenth measurement point), which represents the maximal difference between T2 and T5, is omitted from analysis. A further possibility is that statistical distributions may not have direct mappings in the perceptual space. This is best illustrated by the slight misclassification of T2 into T5 (but zero misclassification of T5 into T2) of the two reference speakers who clearly distinguish all six tones (see Table 3). In any case, this discrepant pattern is intriguing and is worthy of further exploration.

Our data can also shed light on how sound change spreads through the lexicon. Wang (1969; see also Wang & Chen, 1977) documented sound changes that occurred gradually over a long period of time and found that not all words were equally affected. He proposed the idea of lexical diffusion to account for such observations and concluded that most types of phonological sound change are phonetically abrupt but lexically gradual. Labov (1981) and Philips (1984, 2001) linked lexical diffusion with different types of frequency effects.

Our synchronic data of sound change in progress support the idea of lexical diffusion as words with the same tone and the same token frequency were affected differently (see Table 5). The extreme case of high-frequency T2 words illustrates lexically gradual change very well. However, our data also show that sound change can be phonetically gradual too. Figure 3 shows that while some tone production was changed categorically (e.g., T5 changed to T2), a significant portion of tone change was classified as others by the transcribers; in other words, they were heard to be intermediate between two tones or could not be clearly specified. These tokens represent small gradual phonetic changes. Therefore, our data not only provide evidence to support the idea of lexical diffusion synchronically at the suprasegmental level, they also demonstrate that sound change by lexical diffusion can be gradual both phonetically and lexically (Bybee, 2007).

A related issue is why some words are more prone to sound change than others are. If we consider high-frequency T2 words, it is possible that the morphological status of a word may be at work as 者 [tsɛ] (‘he-who’) is a suffix. Moreover, polysemy may also be involved, as the high-frequency T5 word 以 [ji] has many meanings. These two words were much more likely to change than other words in the same tone and frequency categories. Although such suggestions are interesting and worthy of further exploration, caution should be taken as many monosyllabic Chinese (Cantonese) words are polysemous, including many of those in Table 5. Moreover, the morphological status of Chinese (Cantonese) words, such as, part of speech, can sometimes be difficult to determine as Chinese lacks morphological marking in association with categorical alternations

(Chao, 1968; Li & Thompson, 1981). Despite these limitations, the relationship between word properties (besides word frequency) and sound change is worth pursuing in future studies.

Our synchronic data on Cantonese tone merging conform well to many important factors in sound change discussed in the literature: phonetic similarity, perceptual difficulty, imperfect acquisition, frequency effects, and lexical diffusion. However, one interesting question remains: Why are the tone pairs merging in recent years? Our phonetic study cannot give any concrete answers to this interesting question, but we can provide some sociolinguistic speculations for future investigations. The dynamic demographic composition and language contact in Hong Kong over the past 60 years may be one of the reasons triggering the merging of tones. Waves of refugees from different parts of China came to Hong Kong in the late 1940s to the early 1970s, bringing with them many different Chinese dialects. Even after 1980 when strict immigration laws were implemented in Hong Kong, different Chinese dialects continue to interact with Hong Kong Cantonese through cross-border marriage (e.g., children of such families may be influenced by the nonstandard Cantonese tones of their mainland mothers). Moreover, the influence of Mandarin is increasing in Hong Kong after the 1997 handover. All this resulted in increased language contact and interaction. Such external forces may provide an explanation for the merging of tones in recent years.

Our study investigated the initial stages of tone merging in progress in Cantonese. Future studies on the same topic are needed to track the merger development longitudinally. Gordon and Maclagan's (2001) study clearly shows how systematic longitudinal investigation of the same sound change can provide a much more comprehensive understanding of the change involved, because unexpected development can occur. In addition, future studies can investigate the effects of gender differences on tone merger more systematically as prior work has shown that women often lead certain types of language change (Labov, 2001), although our data did not show any significant difference between male and female participants. Given the small number of male participants in our study, our results may not reveal the whole picture of gender effects. More detailed background information of the merging participants should also be collected in order to confirm our speculations about the effects of the demographic composition and language contact in Hong Kong, which may have contributed to the merging of tones. Interesting patterns are likely to emerge.

Modern Hong Kong Cantonese is undergoing many changes at both the segmental and suprasegmental levels, but most previous studies only dealt with segmental sound changes. The results of the present study contribute to the overall picture of sound changes happening in Cantonese phonology. In addition to illustrating the patterns of the early ongoing tone merging process in Cantonese, this study also provides synchronic suprasegmental data to elucidate the various important factors in sound change and a better understanding of the forces of sound change in general.

NOTE

[1] Discriminant analysis uses canonical discriminant functions to define the boundaries between different groups. These functions are calculated based on canonical correlation analysis (Hotelling, 1936), which computes the linear combinations of two sets of variables that have maximum correlation with each other. The functions can be regarded as boundaries between different groups in discriminant analysis. If we have two groups, we can use one line (i.e., one function) to discriminate them. The number of possible functions is determined by either the number of independent variables or the number of groups minus one, whichever is smaller. Here we have four independent variables and five groups, so in total we can have four discriminant functions for each subject. The functions are ordered according to how much they can account for the variance between different groups. Appendix A shows the first two canonical discriminant functions for each subject.

REFERENCES

Bauer, Robert S., Cheung, Kwan Hin, & Cheung, Pak Man. (2003). Variation and merger of the rising tones in Hong Kong Cantonese. *Language Variation and Change* 15:211–225.

Bauer, Robert S., & Benedict, Paul. K. (1997). *Modern Cantonese phonology*. Berlin: Mouton de Gruyter.

Boersma, Paul. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:341–345.

Bybee, Joan. (2007). *Frequency of Use and the organization of language*. Oxford: Oxford University Press.

Chan, Shui Duen, & Tang, Zhi Xiang. (1990). Quantitative analysis of lexical distribution in different Chinese communities in the 1990's. *Yuyan Wenzi Yingyong (Applied Linguistics)* 3:10–18.

Chao, Yuen Ren. (1930). A system of tone-letters. *Le Maître Phonétique* 45:24–27.

______. (1947). *Cantonese primer*. New York: Greenwood Press.

______. (1968). *A grammar of spoken Chinese*. Berkeley and Los Angeles: University of California Press.

Ciocca, Valter, & Lui, Jessica. (2003). The development of lexical tone perception in Cantonese. *Journal of Multilingual Communication Disorders* 1:141–147.

Fok-Chan, Yuen Yuen. (1974). *A perceptual study of tones in Cantonese*. Hong Kong: Hong Kong University Press.

Fung, Roxana., Kung, Carmen, Law, Sampo, Su, I Fan., & Wong, Cathy. (2012). Near-merger in Hong Kong Cantonese tones: A behavioural and ERP study. In *Proceedings of the Third International Symposium on Tonal Aspects of Languages (TAL2012)*, Nanjing, China, May 26–29.

Gandour, Jackson. (1981). Perceptual dimensions of tone: Evidence from Cantonese. *Journal of Chinese Linguistics* 9:20–36.

Gordon, Elizabeth, & Maclagan, Margaret A. (2001). "Capturing a sound change": A real time study over 15 years of the NEAT/SQUARE diphthong merger in New Zealand English. *Australian Journal of Linguistics* 21:215–238.

Greenlee, Mel, & Ohala, John. J. (1980). Phonetically motivated parallels between child phonology and historical sound change. *Language Sciences* 2:283–308.

Hildebrandt, Kristine. (2003). *Manange tone: Scenarios of retention and loss in two communities*. Ph.D. thesis, University of California, Santa Barbara.

Hooper, Joan. B. (1976). Word frequency in lexical diffusion and the source of morpho-phonological change. In William Christie (ed.), *Current progress in historical linguistics*. Amsterdam: North Holland. 95–105.

Hotelling, Harold. (1936). Relations between two sets of variates. *Biometrika* 28:312–377.

Jurafsky, Dan, Bell, Alan, Gregory, Michelle, & Raymond, William. (2000). Probabilistic relations between words: Evidence from reduction in lexical production. In Joan. Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins. 229–254.

Kei, Joseph, Smyth, Veronica, So, Lydia K. H., Lau, C. C., & Capell, K. (2002). Assessing the accuracy of production of Cantonese lexical tones: A comparison between perceptual judgement and an instrumental measure. *Asia Pacific Journal of Speech, Language and Hearing* 7:25–38.

Khouw, Edward, & Ciocca, Valter. (2007). Perceptual correlates of Cantonese tones. *Journal of Phonetics* 35:104–117.

Killingley, Siew Yue. (1985). *A new look at Cantonese tones: Five or six?* Newcastle upon Tyne: Grevatt & Grevatt.

Kratochvil, Paul. (1986). The case of the third tone. In The Chinese Language Society of Hong Kong (ed.), *Wang Li memorial volumes*. Hong Kong: Joint Publishing. 253–276.

Labov, William. (1981). Resolving the neogrammarian controversy. *Language* 57:267–308.

______. (1994). *Principles of linguistic change*. Vol. 1. *Internal factors*. Oxford: Blackwell Publishers.

______. (2001). *Principles in linguistic change*. Vol. 2. *Social factors*. Oxford: Blackwell Publishers.

Leung, Man Tak, Law, Sam Po, & Fung, Suk Yee. (2004). Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments & Computers* 36:500–505.

Li, Charles N., & Thompson, Sandra A. (1981). *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.

Lin, Yen-Hwei. (2007). *The sounds of Chinese*. Cambridge: Cambridge University Press.

Myers, James, & Li, Yingshing. (2009). Lexical frequency effects in Taiwan Southern Min syllable contraction. *Journal of Phonetics* 37:212–230.

Nokes, Jacqui, & Hay, Jennifer. (2012). Acoustic correlates of rhythm in New Zealand English: A diachronic study. *Language Variation and Change* 24:1–31.

Ohala, John J. (1981). The listener as a source of sound change. In C. Masek, R. Hendrick, & M. Miller (eds.), *Papers from the parasession on language and behavior*. Chicago: Chicago Linguistic Society, The University of Chicago. 178–203.

______. (1983). The phonetics of sound change. In C. Jones (ed.), *Historical linguistics: Problems and perspectives*. London: Longman. 237–278.

Philips, Betty S. (1984). Word frequency and the actuation of sound change. *Language* 60:320–342.

______. (2001). Lexical diffusion, lexical frequency, and lexical analysis. In J. Bybee & P. Hopper (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins. 123–136.

Rose, Phil. (1987). Considerations in the normalization of the fundamental frequency of linguistic tone. *Speech Communication* 6:343–351.

So, Lydia K. H., & Dodd, Barbara. (1995). The acquisition of phonology by Cantonese-speaking children. *Journal of Child Language* 22:473–495.

Tabachnick, Barbara G., & Fidell, Linda S. (2007). *Using multivariate statistics*. 5th ed. Boston: Allyn and Bacon.

Vance, Timothy J. (1977). Tonal distinctions in Cantonese. *Phonetica* 34:93–107.

Varley, Rosemary, & So, Lydia K. H. (1995). Age effects in tonal comprehension in Cantonese. *Journal of Chinese Linguistics* 23:76–97.

Wang, William S. Y. (1969). Competing changes as a cause of residue. *Language* 45:9–25.

Wang, William S. Y., & Chen, Chin-Chuan. (1977). Implementation of phonological change: The Shaungfeng Chinese case. In William S. Y. Wang (ed.), *The Lexicon in phonological change*. The Hague: Mouton. 86–100.

Wong, Anita M. Y., Ciocca, Valter, & Yung, Sun. (2009). The perception of lexical tone contrasts in Cantonese children with and without specific language impairment (SLI). *Journal of Speech, Language and Hearing Research* 52:1493–1509.

Wright, Richard. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden & R. A. M. Temple (eds.), *Papers in laboratory phonology VI*. Cambridge: Cambridge University Press. 75–87.

Yiu, Carine Y. (2009). A preliminary study on the change of rising tones in Hong Kong Cantonese: An experimental study (in Chinese). *Language and Linguistics* 10:269–291.

Yu, Alan C. L. (2007). Understanding near mergers: The case of morphological tone in Cantonese. *Phonology* 24:187–214.

Yu, Kristine M., & Lam, Hui Wai. (2011). The role of creaky voice in Cantonese tonal perception. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS)*, Hong Kong, August 17–21. 2240–2243.

Zhang, Jie. (2001). *The effects of duration and sonority on contour tone distribution: A typological survey and formal analysis*. Ph.D. thesis. University of California, Los Angeles.

Zhao, Yuan, & Jurafsky, Dan. (2007). The effect of lexical frequency on tone production. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, August 6–10. 477–480.

______. (2009). The effect of lexical frequency and Lombard reflex on tone hyperarticulation. *Journal of Phonetics* 37:231–247.

APPENDIX A

TABLE 7. *Detailed results of the predictive discriminant analysis of two reference speakers (R) and 17 merging subjects (M)*

Speaker	Wilks' lambda	Classification function I	Classification function II
R1 (F)	.023 ($p < .00001$)	$y = -.361P_2 - .821P_5$ $+ .065P_6 + 1.167P_9$ % of variance = 77.3 canonical correlation = .952	$y = -.066P_2 + .462P_5$ $+ .476P_6 + .160P_9$ % of variance = 22.0 canonical correlation = .857
R2 (F)	.018 ($p < .00001$)	$y = -.227P_2 + .075P_5 -$ $.559P_6 + 1.186P_9$ % of variance = 73.1 canonical correlation = .955	$y = -.022P_2 + .619P_5$ $+ .443P_6 + .012P_9$ % of variance = 26.8 canonical correlation = .890
M1 (F)	.023($p < .00001$)	$y = -.337P_2 - .710P_5 + .114P_6 + 1.108P_9$ % of variance = 76.9 canonical correlation = .952	$y = .283P_2 + .201P_5 + .653P_6 + .102P_9$ % of variance = 22.4 canonical correlation = .859
M2 (M)	.093 ($p < .00001$)	$y = -.277P_2 - .288P_5 - .038P_6 + .960P_9$ % of variance = 80.4 canonical correlation = .898	$y = .264P_2 + .470P_5 + .446P_6 + .243P_9$ % of variance = 18.5 canonical correlation = .700
M3 (F)	.036 ($p < .00001$)	$y = -.142P_2 - .101P_5 - .519P_6 + 1.092P_9$ % of variance = 78.7 canonical correlation = .942	$y = -.129P_2 + .218P_5 + .769P_6 + .164P_9$ % of variance = 21.1 canonical correlation = .822
M4 (F)	.061 ($p < .00001$)	$y = -.195P_2 - .624P_5 - .523P_6 + 1.453P_9$ % of variance = 70.8 canonical correlation = .907	$y = .004P_2 - .371P_5 + 1.298P_6 + .060P_9$ % of variance = 29.0 canonical correlation = .809
M5 (F)	.019 ($p < .00001$)	$y = .034P_2 - .617P_5 - .342P_6 + 1.134P_9$ % of variance = 78.2 canonical correlation = .958	$y = -.028P_2 - .571P_5 + 1.359P_6 + .222P_9$ % of variance = 21.3 canonical correlation = .868
M6 (M)	.078 ($p < .00001$)	$y = -.161P_2 - .508P_5 + .525P_6 + .907P_9$ % of variance = 76.5 canonical correlation = .902	$y = .206P_2 + .299P_5 + .762P_6 - .401P_9$ % of variance = 23.0 canonical correlation = .754

Continued

TABLE 7. *Continued*

Speaker	Wilks' lambda	Classification function I	Classification function II
M7 (M)	.095 ($p < .00001$)	$y = -.331P_2 - .285P_5 - .129P_6 + 1.180P_9$ % of variance = 79.0 canonical correlation = .893	$y = .245P_2 + .472P_5 + .628P_6 - .231P_9$ % of variance = 19.2 canonical correlation = .699
M8 (F)	.058 ($p < .00001$)	$y = -.087P_2 - 1.553\ P_5 + 1.027P_6 + .861P_9$ % of variance = 76.4 canonical correlation = .919	$y = .042P_2 + .263P_5 + .691P_6 + .081P_9$ % of variance = 23.5 canonical correlation = .791
M9 (F)	.064 ($p < .00001$)	$y = -.151P_2 - .723P_5 - .351P_6 + 1.254P_9$ % of variance = 91.9 canonical correlation = .942	$y = -.773P_2 + .371P_5 + .479P_6 + .124P_9$ % of variance = 7.1 canonical correlation = .616
M10 (F)	.029 ($p < .00001$)	$y = -.268P_2 - .837P_5 + .043P_6 + 1.239P_9$ % of variance = 87.9 canonical correlation = .961	$y = .114P_2 - .127P_5 + 1.109P_6 - .041P_9$ % of variance = 11.9 canonical correlation = .787
M11 (F)	.037 ($p < .00001$)	$y = -.085P_2 - .551P_5 - .111P_6 + 1.048P_9$ % of variance = 79.3 canonical correlation = .941	$y = .456P_2 + .193P_5 + .608P_6 + .183P_9$ % of variance = 20.5 canonical correlation = .816
M12 (F)	.041 ($p < .00001$)	$y = -.335P_2 - .847P_5 + .083P_6 + 1.234P_9$ % of variance = 82.4 canonical correlation = .941	$y = -.036P_2 + .530P_5 + .512P_6 + .020P_9$ % of variance = 16.6 canonical correlation = .779
M13 (F)	.217 ($p < .00001$)	$y = -.210P_2 - .048P_5 - .417P_6 + 1.150P_9$ % of variance = 91.5 canonical correlation = .853	$y = .422P_2 + .746P_5 - .066P_6 + .122P_9$ % of variance = 8.1 canonical correlation = .438
M14 (F)	.087 ($p < .00001$)	$y = -.361P_2 - .952P_5 - .226P_6 + 1.616P_9$ % of variance = 71.5 canonical correlation = .881	$y = -.473P_2 - .780P_5 + 2.021\ P_6 - .168P_9$ % of variance = 24.7 canonical correlation = .738
M15 (F)	.021 ($p < .00001$)	$y = -.037P_2 - .855P_5 + .536P_6 + .898P_9$ % of variance = 70.8 canonical correlation = .948	$y = .157P_2 + .657P_5 + .271P_6 + .251P_9$ % of variance = 28.9 canonical correlation = .885

Continued

TABLE 7. *Continued*

Speaker	Wilks' lambda	Classification function I	Classification function II
M16 (F)	.050 ($p < .00001$)	$y = .281P_2 + .591P_5 + .125P_6 - .822P_9$ % of variance = 59.5 canonical correlation = .896	$y = -.099P_2 + .145P_5 + .621P_6 + .457P_9$ % of variance = 39.2 canonical correlation = .853
M17 (F)	.034 ($p < .00001$)	$y = -.279P_2 - .315P_5 + .001P_6 + .941P_9$ % of variance = 77.5 canonical correlation = .941	$y = .277P_2 + .343P_5 + .539P_6 + .268P_9$ % of variance = 22.0 canonical correlation = .829

Note: (F) = female speaker; (M) = male speaker. P_2, P_5, P_6, and P_9 stand for the second, fifth, sixth, and ninth measurement points on the pitch contour, respectively.

APPENDIX B

TABLE 8. *Mean reaction time data (ms) of all participants*

	Tone pairs	Control participants		Merging participants	
		Mean RT	SD	Mean RT	SD
AA pairs	1-1	683	146	856	232
	2-2	818	179	935	246
	3-3	752	164	918	224
	4-4	790	147	913	241
	5-5	780	139	915	220
	6-6	758	144	909	220
AB pairs	1-2	726	171	860	244
	1-3	745	170	884	256
	1-4	720	163	865	294
	1-5	722	184	827	231
	1-6	705	176	845	211
	2-3	753	180	863	194
	2-4	843	197	908	195
	2-5	916	158	1032	223
	2-6	764	171	855	198
	3-4	764	184	894	282
	3-5	788	226	969	314
	3-6	780	182	873	186
	4-5	789	148	904	268
	4-6	744	170	885	232
	5-6	778	145	903	225
Average		768	192	894	257

Note: RT = reaction time.