Individual variability of acoustic-perceptual mapping in ongoing merger: Observations from Cantonese Tone 4/Tone 6

Yubin Zhang
The Hong Kong Polytechnic University

Sociophonetic studies reveal several individual manifestations of mergers-in-progress. In the current study, a combined identification and rating task was conducted to investigate how the ongoing Cantonese Tone 4/Tone 6 (T4/T6) (near-)merger manifests itself in individual perceptual space. The results indicated that Cantonese listeners mapped synthesized F_0 trajectories onto perceptual representations differently. Our fine-grained analysis at the individual level enabled us to identity a wider range of manifestations of the ongoing Cantonese T4/T6 merger than previous studies on this tonal merger—clear (complete) distinction, reduced distinction, near-merger with different degree of production overlapping, complete merger and flip-flop. Our data also suggested that the direction of this sound change is from T6 to T4, that is, the canonical level T6 gradually impinges on the space of T4 by exhibiting falling pitch contours. For mergers-in-progress, evidence for individual or sociostylistic variations in phonological representations and the production-perception asymmetry are in favor of exemplar-based models of phonological representations and sound change.

1. Introduction

1.1 The configuration of mergers-in-progress

Individuals in a speech community typically exhibit a wide spectrum of manifestations of mergers-in-progress: complete (clear) distinction, reduced distinction, near-merger / flip flop and complete merger (Labov et al., 1991; Hall-Lew, 2013).

First, at the two ends of the spectrum, lie complete (clear) distinction and complete merger. These two ends are thought to be the beginning and ending points of the life cycle of a sound change. Individuals with complete distinction are considered as having a phonological contrast that is closest to the canonical one, while those with complete merger cease to retain the contrast in both production and perception.

Second, some individuals show a distinction, but their ability to produce and/or perceive the contrast is attenuated. For instance, Labov et al.'s (1991) results on /er/-/ʌr/ merger, e.g., MERRY/MURRAY and FERRY/FURRY, in Philadelphian English, suggest that some Philadelphians, despite the fact that they maintained a distinction between /er/-/ʌr/, were less capable of producing and/or perceiving the distinction than other Philadelphians. Furthermore, compared with most of the non-Philadelphians with an evident distinction, Philadelphians who could still distinguish these two categories typically showed reduced

differences in production and perception. Such cases imply that ongoing sound change can render the phonetic distributions for a phonological category closely approximated without leading to merger.

Third, another well-known indication of merger-in-progress is near-merger (Labov et al., 1991, 1972). Strictly speaking, the term 'near-merger' or 'partial-merger' refers to any transitional stage between complete distinction and complete merger. Thus, there are two kinds of near-merger: (1) merger in production and distinction in perception; (2) merger in perception and distinction in production. The latter one is known as the classic near-merger, where language users display consistent yet somewhat overlapping distinctions in production but have difficulty distinguishing the contrast in perception. The phenomenon of classic near-merger has been reported for vowels in varieties of English (see Labov et al., 1991 for more details), Russian consonants (Diehm & Johnson, 1997) and Cantonese tones (Fung et al., 2012). Take for example Labov et al.'s (1991) data on Philadelphian vowel merger. They found that some speakers maintained a distinction in production, but performed inferiorly in categorizing their own productions or clear tokens produced by other speakers. Another phenomenon or concept similar to near-merger is 'incomplete neutralization', where a phonological contrast is supposed to be neutralized in certain phonological contexts, but small acoustic differences associated with underlying categories are still discernable instrumentally. In German, the voicing opposition has long been thought to be completely neutralized in word-final positions, i.e., {b, d, g} → {p, t, k} / __D (domain-boundary), e.g., for lexical items like Rad /rad/ (traditionally transcribed as [ʁat], 'wheel') and Rat /rat/ ([ʁat] 'advice'). However, a number of acoustic studies have found some small but measurable phonetic differences between these putative neutralized categories, like the duration of the preceding vowel. Perceptually, although listeners might be capable of distinguishing devoiced (voiced) stops from voiceless ones in neutralizing contexts, their performance was inferior compared with non-neutralizing contexts (Röttger et al., 2014). Besides the final devoicing in German, American English t/d flapping (Braver, 2014), where {t, d} → ɾ/ V__V0 (V0: - stressed), and Mandarin tone 3 sandhi (Shu-hui, 2000), where T3 → T2/ __T3, have also been argued to be cases of incomplete neutralization. Taken together, near-merger or incomplete neutralization reveals that in ongoing sound change, supposedly merged categories can be reliably separated in production, but their perceptibility is substantially diminished or even disappear (Braver, 2014; Röttger et al., 2014).

Fourth, flip-flop, a surprising case on the merging spectrum, has also been documented (Labov et al., 1972; Hall-Lew, 2013). Technically speaking, accurately positioning the flip-flop phenomenon within the spectrum is difficult. According to Hall-Lew (2013), flip-flop refers to a production pattern where the distinction is inverted in at least one phonetic dimension and therefore the phonological contrast is maintained but opposite to the canonical one. To take an example, Hall-Lew's (2013) study on the merger of low back vowels in San Francisco English, i.e., /ɑ/ (COT)– /ɔ/ (CAUGHT) with CAUGHT being backer and higher than COT in the vowel space, identified two flip-flopping speakers

who produced either fronter CAUGHT than COT (flip-flop in the F_2 dimension) or lower CAUGHT than COT (flip-flop in the F_1 dimension). The author further claimed that flip-flop is a special case of near-merger as the two speakers seemed to perceive no distinction. Although a clear picture of the flip-flop phenomenon remains to be unveiled, the existing evidence suggests that individuals in a community undergoing merger can develop phonological contrasts that are opposite to the canonical ones in certain phonetic dimensions. In such cases, the production-perceptual asymmetry can also occur.

1.2 Theoretical approaches to phonological representations and sound change

Sociophonetic studies on indications of mergers-in-progress have provided a large amount of empirical evidence on phonological representations and the underlying mechanisms of sound change. Various theoretical approaches have been proposed to elucidate the nature of phonological representations and the patterns of mergers-in-progress.

Among these various approaches, classical modular feedforward models of phonetics and phonology are probably the best-known one (see Pierrehumbert, 2001 and Bermú Dez-Otero, 2007 for a review). These models are 'modular' because they assume the modularity of phonological encoding and phonetic implementation. Phonological representations are discrete and categorical, deprived of continuous phonetic information. The phonetic module is responsible for computing gradient articulatory gestures. These models are also 'feedforward' as information flows unidirectionally from the phonological module to the phonetic one. According to the predictions of these models, sound change commences from some gradient perturbations at the phonetic level. Then, gradient changes accumulate and are stabilized at the phonological level. Thus, all gradient aspects of mergers-in-progress, as displayed at the intermediate stages between complete distinction and complete merger, is dealt with in the phonetic module. For instance, individuals with reduced distinction represent the same set of phonological categories as those with clear distinction, but the phonetic realizations of these categories exhibit a different degree of distinction. However, although postulating a group of symbolic representatations without phonetic gradiency neatly fulfils the requirements of theoretical simplicity and can indeed account for numerous phenomena on the merging spectrum, there are several major drawbacks of the modular feedforward architecture. First, mounting evidence suggests that phonetic detail, such as memory traces of talker-specific voice (Goldinger, 1998, 1996) and word-specific allophonic detail (Pierrehumbert, 2001), may also be part of the long-term representations. It is unclear at which level this kind of knowledge is represented in modular feedforward models. Second, according to Pierrehumbert (2001), these models do not provide treatment of sociostylistic effects, e.g., gender, social class, etc., on mergers-in-progress. Third, they cannot elegantly handle the dissociation between production and perception, as observed in near-merger and incomplete neutralization.

The challenges faced by the traditional theories of phonetics-phonology interface necessitate the modifications to classical theories or the development of alternative

proposals. The exemplar-based models stand out from other approaches by providing new insights into phonological representations and sound change (Yu, 2007). According to this theory, phonological categories are viewed as clusters of similar exemplars, which are the detailed episodic memory traces of phonetic experiences (Pierrehumbert, 2006; Goldinger, 1998). Therefore, phonological representations can be fine-grained in exemplar-based models, compared with classic modular feedforward ones. More specifically for sound variation and change, exemplar theories treat sound change as a consequence of alterations and redistribution of exemplars. Therefore, cases like reduced distinction can be easily explained as approximation of exemplars of distinct categories. Furthermore, as exemplars can be acquired in different sociostylistic contexts, exemplar-based models can readily explain sociostylistic variations in sound change. Finally, because exemplars assembled in perception may not necessarily be employed for production, these models can also explain the mismatch between production and perception, as observed in near-merger and incomplete neutralization.

In light of the aforementioned empirical findings and theoretical innovations, the current study aims at examining how ongoing merger manifests itself in the perceptual space by providing a cross-sectional case study of the perception of the ongoing Cantonese T4/T6 (near-)merger. Previous studies on mergers-in-progress, like some of those reviewed above, mainly relied on acoustic analysis, which can precisely delineate individual acoustic sound space. Although some studies employed perceptual experiments, the analysis was coarse-grained, without revealing individual perceptual space. More specifically for Cantonese T4/T6 (near-)merger, while the pereception of this tonal pair has been examined in previous studies, we are not aware of any fine-grained examination of individual percetual representations of these two tones. To achieve this goal, we seek to model individual perceptual representations of the ongoing Cantonese T4/T6 (near-)merger, using the perceptual data collected from a combined identification and rating task.

2. Backgrounds

2.1 Cantonese and Cantonese tones

Cantonese is one of the major varieties of the Chinese language in Guangdong and Guangxi Provinces of China and overseas Chinese communities. The standard Cantonese, also called Guangfu speech (廣府話), is the prestigious variety spoken in Guangzhou (Canton) and its neighboring areas, like Hong Kong and Macau, and is considered as the lingua franca of the Pearl River Delta region (珠江三角洲地區).

In phonological description, standard Cantonese has a complex tonal system with a total of six lexical tones : high-level T1, high-rising T2, mid-level T3, low-falling T4, low-rising T5, and low-level T6 (Bauer & Benedict, 1997). As illustrated in Figure 1, variations in fundamental frequency (F_0 hereafter) are mainly used to distinguish the six lexical tones of Cantonese (Khouw & Ciocca, 2007; Vance, 1977). Cantonese T4 and T6 occupy the lowest-pitched portion of Cantonese tonal space. T4 is typically transcribed as a low-falling tone (21, ˩), but it can also be realized with an even pitch trajectory (11, ˩). T6 is described

as a low-level tone (22, ˩), but its phonetic realization frequently exhibits a falling F0 contour, which is also common for other Cantonese level tones (Vance, 1977).

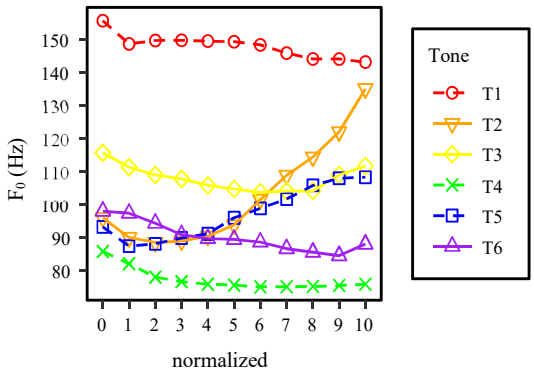

Figure 1. The F0 profile for six Cantonese tones realized on the syllable [si]. The data were obtained from one male Cantonese speaker from Hong Kong, i.e., subject 10 (*s10*) in our dataset.

2.2 Tone merging in Cantonese: T4/T6 as an ongoing (near-)merger

The complex tonal system of Cantonese is in the process of variation and merger (Mok et al., 2013; Ou, 2012). There are mainly four tonal mergers reported in the literature – T3/T6, T3/T5, T2/T5 and T4/T6.

The T4/T6 pair has been identified as a classic near-merger in Hong Kong Cantonese (Fung et al., 2012) and Guangzhou Cantonese (Ou, 2012). For example, Fung et al. (2012) compared the production and perception of the T4/T6 pair for Hong Kong Cantonese speakers with and without near-merger. In production, compared with normal controls, speakers with near-merger produced distinctive but more approximated pitch contours for this tonal pair at the group level. In perception, the near-merger group had difficulty discriminating the T4/T6 pair in an AX discrimination task, and in a later ERP study, no reliable mismatch negativity (MMN) to the T4/T6 contrast was elicited for the near-merger group in a passive oddball paradigm. The T4/T6 merger was also reported in Malaysian Cantonese (Weng, 2014). Although no production and perception studies were conducted in that study, it is likely that the reported T4/T6 merger is also a near-merger in Malaysian Cantonese.

As for the direction of this sound change, there are chiefly two lines of evidence supporting the claim that T6 is encroaching on T4's tonal space, giving rise to (near-)merger. First, when asked to identify natural T4 and T6 tokens, Hong Kong Cantonese listeners tended to misidentify T6 as T4 more frequently than vice versa (Fok-Chan, 1974; Varley & So, 1995). Second, Fung et al.'s (2012) production data revealed that the approximation of the pitch trajectories of T4 and T6 seems to be unidirectionally from T6 to T4. In other words, the T6 tokens produced by the near-merger group exhibited a larger degree of pitch drop than the control group, approaching the tonal space of T4, but the T4 tokens produced by the two groups appeared to be broadly comparable.

3. Methods

3.1 Materials

The speech stimuli for this experiment were generated in PRAAT (Boersma & Weenink, 2017) using *KlattGrid* speech synthesizer (Weenink, 2009). We recorded several repetitions of one male speaker's productions of the Cantonese words wa^6/話 ([wa:ɬ], 'word, utterance') and wa^4/華 ([wa:ɬ], 'China, splendid') and used these tokens as benchmarks for synthesis. The vocal tract parameters were based on the syllable wa^6.

To probe into Cantonese speakers' mental representations of the two tones, we used a series of densely spaced F_0 contours in the low pitch range. We first selected 5 equally spaced F_0 levels in the speaker's comfortable range for T4 and T6 production – 80, 90, 100, 110, 120 Hz. Then, the 5 F_0 values were assigned to the starting and ending points of the syllable respectively. Finally, we aligned the F_0 values of the starting and ending points, resulting in 5 level and 10 falling F_0 trajectories (Figure 2). Note that we did not include rising contours in our study, despite the fact that T4 has a high-rising pitch realization derived from a tonal alternation process (Yu, 2007). The inclusion of this T4 variant would further complicate the experiment.

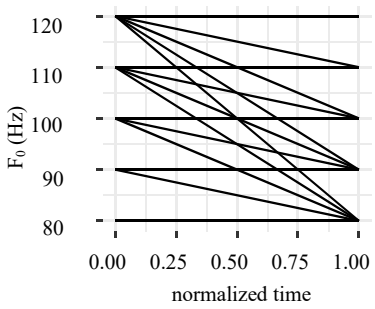

Figure 2. The 15 synthesized F_0 trajectories used in the perceptual experiment. There are 9 levels of F_0 MEAN—80, 85, 90, 95, 100, 105, 110, 115, 120 Hz, and 5 levels of F_0 CHANGE—0, 10, 20, 30, 40 Hz.

3.2 Participants

We recruited a total of 31 Cantonese speakers—11 males (age 24.3 ± 2.3) and 20 females (age 23.9 ± 1.4). Except two overseas Chinese from Malaysia (s1 and s27), almost all participants were born and raised in the Pearl River Delta region, including Hong Kong, Macau and southern parts of Guangdong Province. Nearly all the participants spoke standard Guangfu speech, except *s8* from Zhanjiang (湛江). The participants also had some knowledge of Mandarin Chinese and English. None of them reported any speech, language and hearing impairments. They were paid for their participation in the experiment.

3.3 Procedure

A combined identification and goodness rating task was administered to the participants in a sound-treated room. The speech stimuli were presented in PRAAT over the

headset at a comfortable hearing level. All participants completed 4 practical trials before the actual experiment. A total of 540 tokens[1] were pseudo-randomly presented in three blocks. The participants were required to identify the tone of the speech stimuli—T4 (華) or T6 (話)—by clicking on corresponding buttons. Then, after identifying the tone, they were asked to judge the goodness of the tone on a scale of 1 (very bad) to 5 (very good). Each stimulus was presented only once. After the responses of identification and goodness rating were collected, the next stimulus was presented automatically. The goodness rating task would allow us to identify unnatural tokens due to experiment artifacts, as these artificial tokens would generally receive lower rating scores. More importantly, the rating task could provide us with more information on individual tonal representations. For example, if identification results reveal that a participant's T4 responses increase as F_0 change becomes larger, we can only conclude that F_0 change is a more important cue for his/her T4 identification than T6 identification. It is likely that the participant represent both tones as falling tones, although T4 falls to a greater degree. Moreover, supplementary production data were also collected for some participants (see Appendix, Figure 7).

3.4 Data Analysis

The identification data were analyzed using a linear mixed-effects model implemented in R (R Core Team, 2017). A mixed-effects model, including both fixed and random effects, provides a powerful tool for investigating both the overall group-level effects and individual differences. The fixed intercept and slope(s) are the overall group-level parameters, while individual differences are modelled by the random-effects component, which allows each participant to have their own intercepts and slopes. We fit one generalized linear mixed-effects model (GLMM) with a logistic link function using the R package *lme4* (Bates et al., 2015).

We specified F_0 MEAN, F_0 CHANGE, and their interaction as fixed effects in the GLMM. These predictors were *z*-transformed prior to model fitting. For the random-effects, we first constructed a model with the maximal random-effects structure justified by the design, including by-participant random intercepts, random slopes for F_0 MEAN, F_0 CHANGE, and F_0 MEAN: F_0 CHANGE interaction, and all correlation parameters. Then, we built a series of models with more parsimonious random-effects structures. Model comparison was performed using the log-likelihood test and the best random-effects structure was determined based on the test results. The maximal random-effects structure proved to be the best one. Our final model reported in section 4.1 included three fixed effects and the full random-effects structure. The GLMM is formulated in the following way:

$$logit\left(p_{T4_{ijk}}\right) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})X_{M_{ij}} + (\beta_2 + b_{2i})X_{C_{ik}} + (\beta_3 + b_{3i})X_{M_{ij}}{:}X_{C_{ik}} + \varepsilon_{ijk}$$

[1] We also manipulated phonation and duration for other research purposes. The 540 trials included 15 F_0 trajectories × 4 phonations × 3 durations × 3 repetitions. As the current analysis focused on pitch representations and the effects of phonation and duration were limited, only F_0-related manipulation and results are reported in this paper.

$p_{T4_{ijk}}$ is the probability of a T4 response for the j^{th} level of F_0 MEAN, the k^{th} level of F_0 CHANGE and the i^{th} participant. $X_{M_{ij}}$ and $X_{C_{ik}}$ denote standardized F_0 MEAN and F_0 CHANGE respectively. β_0 represents the fixed intercept while β_1, β_2, and β_3 represent three fixed slopes. b_{0i} is the random intercept and b_{1i} , b_{2i} and b_{3i} are three random slopes.

For the rating data, we implemented an algorithm to select F_0 trajectories that best represent the individual- and group-level T4/T6 space. Taking the exemplar approach, the T4/T6 space is defined as two clusters of exemplars. Three best rated F_0 trajectories were selected as representative exemplars of T4 and T6 clusters respectively.

The algorithm first calculated the identification rate R_{it} of each tone $t = T4\ or\ T6$ for each participant i by the formula $R_{it} = n_{it}/N$, where n_{it} is the total number of T4 or T6 responses given by a participant and $N = 540$ stands for the total number of stimuli. Then, the identification rate of each tone t for each participant i and each F_0 trajectory $l = 1, 2 \dots 15$ (see Figure 2) was also calculated in a similar way—$R_{itl} = n_{itl}/N_l$, where $N_l = 36$ denotes the total number of stimuli for each F_0 trajectory l.

Next, raw rating scores were averaged over each F_0 trajectory l for each participant i and each tone t, resulting in a sequence of 15 rating scores $A_{it} = (a_{itl})_{l=1}^{15};\ 1 \le l \le 15,\ l \in N^*$. For each participant i and tone t, a total number of s elements in sequence A_{it}, with identification rate for trajectory l (R_{itl}) less than the average identification rate (R_{it}), i.e., $R_{itl} < R_{it}$, were pruned away. This process resulted in sequence $B_{it} = (b_{itm})_{m=1}^{15-s};\ 1 \le m \le 15 - s,\ m \in N^*$, which contained rating scores with relatively higher identification rate. This sequence was further sorted in a decreasing order, giving rise to a new sequence $C_{it} = (c_{itr})_{r=1}^{15-s};\ 1 \le r \le 15 - s,\ r \in N^*$, where $c_{itr} = b_{itm},\ c_{itr} \ge c_{it(r+1)}$. The original indices l of the first three elements in sequence C_{it}, that is, the three best rated F_0 trajectories of each tone t, were returned for each participant i.[2]

4. Results

4.1 Identification

Table 1. The fixed-effects structure of the GLMM

	Estimate	Std. Error	Z value	Pr(>|z|)
INTERCEPT	-1.105	0.231	-4.775	< 0.001***
F_0 MEAN	-1.530	0.278	-5.499	< 0.001***
F_0 CHANGE	0.517	0.146	3.542	< 0.001***
F_0 MEAN: F_0 CHANGE	-0.037	0.063	-0.584	0.559

The estimated coefficients for the fixed-effects component of the model are listed in Table 1. The intercept reflects the log-odds of a T4 response when all the predictors are held constant, i.e., at their mean values (β = -1.105, p <0.001). The GLMM revealed a significant main effect of F_0 MEAN (β = -1.530, p <0.001). The negative coefficient for F_0 MEAN

[2] To generate group-level tonal representations, overall identification rate for each tone, identification rate for each tone and each F_0 trajectory, and averaged rating scores for each tone and each F_0 trajectory, were calculated in a similar fashion at the group level.

signifies that the log-odds of T4 responses decreased as F_0 MEAN went upwards. There was also a significant main effect of F_0 CHANGE ($\beta = 0.517$, $p < 0.001$), suggesting that the log-odds of T4 responses increased as F_0 CHANGE became larger. The interaction between F_0 MEAN and F_0 CHANGE failed to reach significance ($\beta = -0.037$, $p = 0.559$).

The random-effects structure is summarized in Table 2. The table lists the standard deviations of by-participant adjustments to group-level coefficients (BLUPs) and the estimated correlation parameters, pertaining to the correlations among random slopes and intercepts. The scatterplot matrix (Figure 3) visualizes the correlation structure of the random effects.

Table 2. The random-effects structure of the GLMM

Groups	Name	Std. Dev.	Correlation		
Participant	INTERCEPT	1.273			
	F_0 MEAN	1.534	0.65		
	F_0 CHANGE	0.799	-0.43	-0.59	
	F_0 MEAN: F_0 CHANGE	0.274	-0.42	-0.38	0.50

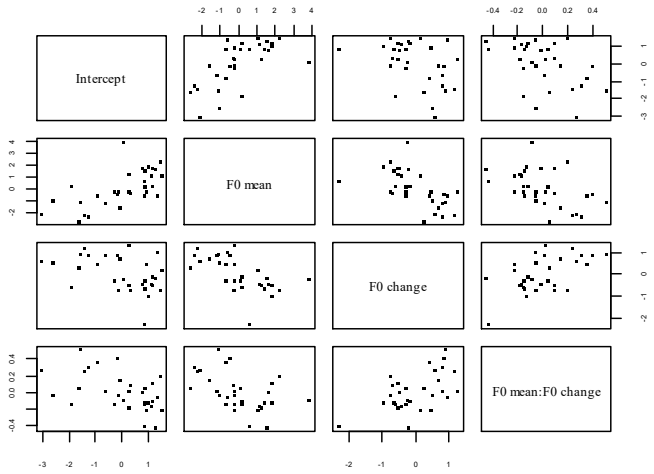

Figure 3. The correlation structure of the by-participant random intercepts and slopes. Each dot represents a participant. For each pairs of covariates, the adjustments to group-level estimates (BLUPs, the best linear unbiased predictors) are shown.

There are three noteworthy adjustments to group-level estimates. First, by-participant adjustments to the fixed intercept, ranging from -3.056 to 1.483, represent individual differences in the baseline log-odds of giving a T4 response. Subjects with large positive BLUPs for the intercept, like *s23* (1.483), tended to give more T4 responses than subjects with large negative BLUPs, like *s30* (-3.056). Second, by-participant adjustments to the group-level slope for F_0 MEAN range from -2.691 (*s7*) to 3.883 (*s15*), indicating substantial individual variability in the F_0 mean dimension. Third, as reflected by adjustments to the group-level slope of F_0 CHANGE, which range from -2.318 (*s28*) to 1.285 (*s21*), variability in the F_0 change dimension is also evident.

The aforementioned three adjustments correlate with each other. The negative correlation parameter pertaining to adjustments for the fixed slopes of F_0 MEAN and F_0 CHANGE reveals that for participants with reduced sensitivity to variations in F_0 mean (upward adjustments resulting in less steep slopes of F_0 MEAN), their sensitivity to variations in F_0 change was also attenuated (downward adjustments resulting in less steep slopes of F_0 CHANGE). In addition, adjustments for the fixed slope of F_0 MEAN also positively correlate with adjustments for the fixed intercept, indicating that for participants with reduced sensitivity to variations in F_0 mean, they tended to give more T4 responses (upward adjustments resulting in higher intercept estimates). Furthermore, adjustments to the fixed slope of F_0 CHANGE and the fixed intercept also show a moderate negative correlation, indicating that for participants with reduced sensitivity to variations in F_0 change, they generally identified more stimuli as T4 than other participants. To summarize, for participants with a larger number of T4 responses than others, they were less sensitive to both F_0 mean and F_0 change dimensions.

By-participant adjustments to the interaction between F_0 MEAN and F_0 CHANGE also seem to show some individual variability. Additionally, they seem to correlate with adjustments for the other fixed effects. However, the interaction is not very interpretable, as F_0 MEAN and F_0 CHANGE are not fully crossed in our design. For instance, there is only one level of F_0 CHANGE at 85 Hz. Furthermore, the interaction might be non-linear, as evidenced by studies on cue-weighting in speech perception (e.g., Kong and Edwards 2016).

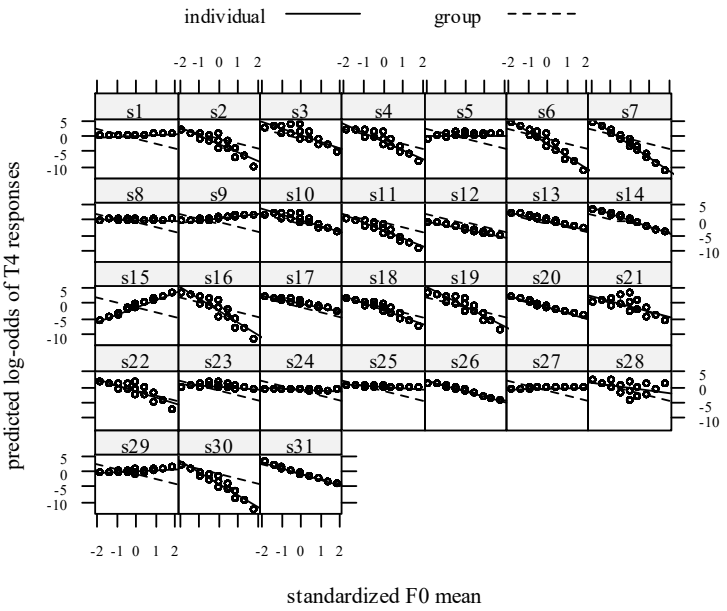

Figure 4. Individual predicted effects of F_0 MEAN on log-odds of T4 responses. The solid line and the dotted line denote individual-level effects and group-level fixed effects respectively. The fitted slopes for *s9*, *s28* and *s29* are not accurate, as these participants have a falling-rising identification curve in the F_0 mean dimension. This indicates that they might use extra labels in identification (see Appendix, Figure 8).

The predicted by-participant effects of F_0 MEAN, i.e., fixed effects plus random effects, on log-odds of T4 responses are displayed in Figure 4. Inspection of Figure 4 indicates that most participants maintained a contrast of T4 and T6, as evinced by their negative slopes, but individuals varied according to the size of F_0 MEAN effect. The fitted regression lines of some participants, like *s7*, have more negative slopes than those of others, like *s12*. The variability indicates that participants exhibited differential perceptual sensitivity to the F_0 mean dimension. The slopes of the fitted lines of four participants (*s1*, *s5*, *s8*, *s27*), who reported no differences between these two tones, approach zero, suggesting that these participants merged these two tones perceptually. Participants *s8* and *s27* were further identified as exhibiting near-merger, as small traces of the underlying tonal distinction were preserved in their production, while participant *s5* with complete merger showed no distinction in both production and perception (see Appendix, Figure 7). The fitted regression line of *s15* has a large positive slope, which greatly diverges from the group estimation. The unexpected pattern suggests that this participant might confuse these two tones and showed a flip-flop pattern in perception. In other words, this participant identified canonical T4 items as T6 and T6 items as T4. However, he seemed to show the canonical contrast in production (see Appendix, Figure 7).

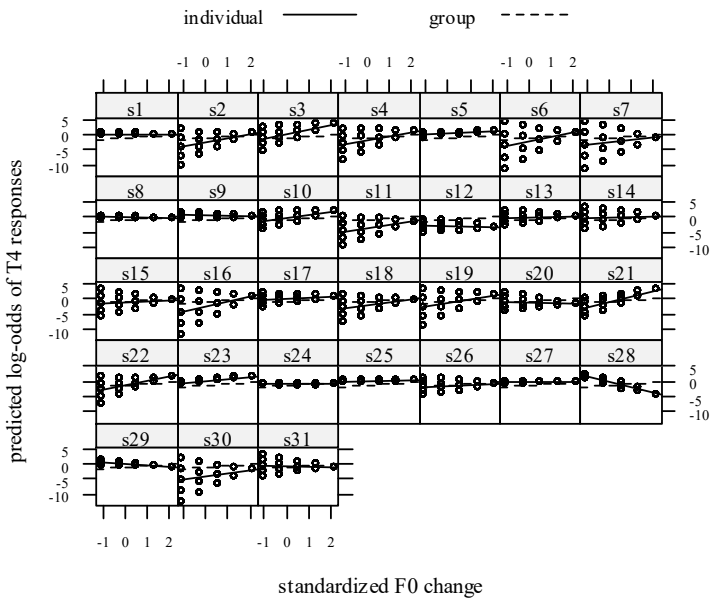

Figure 5. Individual predicted effects of F_0 CHANGE on log-odds of T4 responses. The solid line and the dotted line denote individual-level effects and group-level fixed effects respectively.

The predicted by-participant effects of F_0 CHANGE on log-odds of T4 responses are displayed in Figure 5. In the F_0 change dimension, the slopes of most individual regression lines are positive, indicating increased log-odds of T4 responses with increased F_0 CHANGE. However, the effect of F_0 CHANGE was not constant across participants. The near-horizontal

regression lines of some participants (e.g., *s20*), including the four participants who merged these two tones perceptually (*s1*, *s5*, *s8*, *s27*), reveal that the effect of F_0 CHANGE on their T4 responses was attenuated or even disappeared. The negative coefficients of F_0 CHANGE estimated for some participants, like *s28* and *s29*, indicate that their T4 responses decreased as F_0 CHANGE became larger. This pattern suggests that in their tonal representations, T6 might fall to a greater degree than T4. In other words, these participants could be characterized as flip-flopping in the F_0 change dimension.

To summarize, five types of mergers-in-progress could be identified based on identification results, supplementary production data, and self-report. There were speakers with complete merger (*s5*), near-merger (*s8*, *s27*) and flip-flop (*s15*, *s28*, *s29*). Speakers with reduced distinction could be defined as having less precipitous slopes for both dimensions, i.e., positive adjustments to the fixed slope of F_0 MEAN and negative adjustments to the fixed slope of F_0 CHANGE (*s12*, *s13*, *s17*, *s24*, *s25*). The remaining participants with distinction could be considered as having clear (complete) distinction.

4.2 Goodness rating

The rating results are generally consistent with identification results. The output of the algorithm is given in Figure 6. For the group-level tonal representations (pooled), T4 and T6 were well separated by F_0 mean, with T6 having a higher pitch than T4. Furthermore, both tones exhibited a falling pitch contour, but T4 seemed to fall to a larger degree.

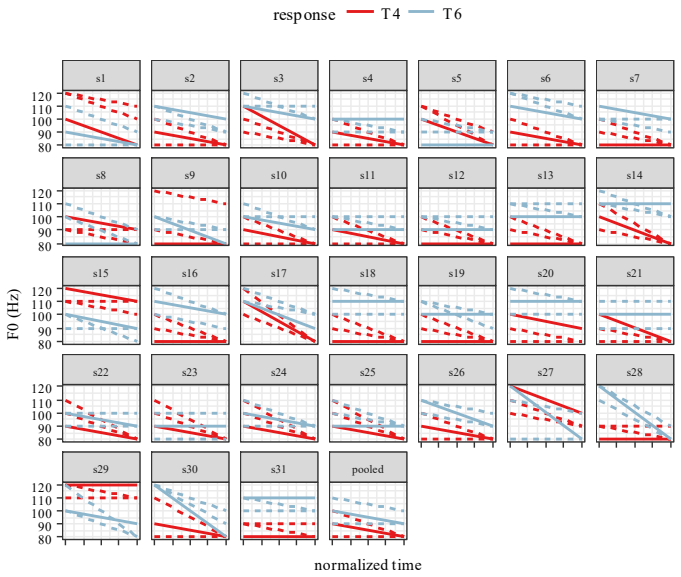

Figure 6. Individual- and group-level (pooled) tonal space of T4 and T6. The three best rated tokens for each participant and the whole group were generated by the algorithm described in section 3.4. Blue and red lines denote T6 and T4 respectively. The solid line represents the best token and the other two dotted lines represent the second and the third best tokens. The generated T4/T6 space for participants with (near-)merger (*s1*, *s5*, *s8*, *s27*) is not reliable. Some T4 exemplars for participants with a falling-rising identification curve in the F0 mean dimension (*s9*, *s28*, *s29*) are also not valid (c.f., Appendix, Figure 8).

For the individual-level representations, there are several notable aspects. Firstly, individual-level representations of T4 and T6 differed in the F0 mean dimension. In the F0 mean dimension, nearly all participants except those with (near-)merger (*s1*, *s5*, *s8*, *s27*) represented higher-pitched T6 than T4, but the perceptual distance between these two tones varied among participants. Secondly, in the F_0 change dimension, the representations with respect to pitch contours were variable across participants. For T4, although most participants considered that both (extra-)low level and low-falling pitch trajectories were its representative exemplars, some participants like *s14* showed preferences for a low-falling pitch representation more than others like *s31*. For T6, while most participants, like *s11*, represented it as a low-level or low-slightly-falling tone, some participants, like *s17*, had a low-falling pitch representation. For flip-flopper *s28* and *s29*, their T6 even showed a larger downward F_0 movement than T4, which was consistent with their identification results. Thirdly, when these two dimensions taken as a whole, some participants, like *s17*, had a more compact T4/T6 space than others, and this pattern might partially explain their reduced sensitivity to both F_0 mean and F_0 change dimensions in identification.

5. Discussion

5.1 Types of mergers-in-progress: Evidence from Cantonese T4/T6

Our results indicated that individual Cantonese listeners mapped F_0 trajectories onto T4 and T6 representations differently. Previous perceptual studies on the T4/T6 merger used natural stimuli and the analysis was performed at the group level, without revealing individual perceptual representations (Fung et al., 2012; Ou, 2012). These studies only identified speakers with distinction and those with (near-)merger. Our fine-grained analysis could capture subtle individual differences in tonal representations and thus unveiled a more comprehensive picture of individual manifestations of the ongoing tonal merger: clear (complete) distinction, reduced distinction, near-merger with different degree of production overlapping, complete merger and flip-flop.

First, while some speakers, like *s19*, showed a clear distinction of this tonal contrast, others exhibited reduced perceptual sensitivity to variations in F_0 mean and/or F_0 change. Speakers with overall reduced sensitivity to this contrast, like *s17*, tended to show less precipitous slope in both dimensions and give relatively more T4 responses than others. In line with previous results (Varley & So, 1995; Fok-Chan, 1974; Fung et al., 2012), our findings suggested that the direction of this tonal merger is from T6 to T4. Put another way, T6 gradually covers the tonal space of T4 by exhibiting falling pitch contours.

Second, some speakers with near-merger, such as *s8*, *s27* and *s28*, were not capable of using F_0 cues to identify T4 and T6, but they could produce distinctive T4 and T6. Fung et al. (2012) showed that the F_0 trajectories of T4 and T6 produced by Hong Kong Cantonese speakers with near-merger exhibited relatively closer approximation than normal controls at the group level. Our analysis further revealed that the degree of approximation in production varied across speakers. For instance, participant *s8* produced

a smaller degree of distinction than *s27* (see Appendix, Figure 7). These results were reminiscent of Labov et al.'s (1991) study on Philadelphian /er/-/ʌr/ merger. They also identified two kinds of near-mergers – near-mergers with and without overlapping distributions of sound classes in production. Moreover, the data from participant *s27*, who came from Malaysia, further indicated that the reported T4/T6 merger in Malaysia Cantonese can also be near-merger. Additionally, as with Labov et al. (1991), we also found one speaker with complete merger (*s5*), who perceived and produced no distinction.

Third, there are two types of flip-flops in the dataset. One speaker (*s15*) flip-flopped with respect to F_0 mean, perceiving lower-pitched T6 than T4. Surprisingly, he produced lower-pitched T4 than T6. Some other speakers, like *s28* and *s29*, flip-flopped in the F_0 change dimension, perceiving or/and producing T6 as having a larger downward pitch movement than T4 (see Figure 5, Figure 6 and Figure 7). Hall-Lew (2013) defined flip-flop as a production pattern co-occurring with near-merger. However, this definition does not seem to fit nicely with our observations. Our data suggested that flip-flop can also be manifest in perception. Moreover, flip-flop may not necessarily be accompanied by near-merger, as T4-T6 flip-flop occurred in both production and perception for *s28*. Therefore, we suggest that flip-flop is probably an advanced production or perception pattern where the merging sounds pass the point of convergence in at least one phonetic dimension. However, in the absence of production data collected from different speaking-style conditions and the speakers' detailed linguistic and social backgrounds, it is currently impossible to draw any firm conclusion on the flip-flop phenomenon. For example, there is an alternative socio-dialectical explanation for the flip-flop pattern exhibited by *s15*, that is, his Cantonese was influenced by a local Cantonese variety, where T6 is generally lower-pitched than T4. This hypothesis seems to be validated by his self-report that he spent his early childhood in Shunde (順德), Foshan, but later migrated to Guangzhou and spoke standard Cantonese. Thus, the flip-flop patterns observed in our perceptual data might reflect the old layer of a speaker's tonal system instead of the advanced features of the ongoing merger. Similarly, for the two flip-flopping speakers reported in Hall-Lew (2013), rather than flip-flopping to negotiate between conflicting local identities, they may had acquired the flip-flop pattern from other varieties of English, which have either fronter CAUGHT than COT or lower CAUGHT than COT, as the author mentioned that these two speakers had spent some years outside their hometown San Francisco.

5.2 An exemplar-based account of the ongoing Cantonese T4/T6 merger

Our perceptual results suggested that phonological representations can be variable across individuals as a result of individual linguistic experiences. Exemplar-based models, which allow representations of individual-particular or sociostylistic phonetic detail and the separation of production and perception (Yu, 2007), can better account for the current data. We do not aim to provide a comprehensive evaluation of various phonological models, but the appropriate model at least needs to deal with individual-particular or sociostylistic phonetic detail and production-perception asymmetry. A modified version of traditional

modular feed-forward models that incorporates phonetic detail, like a hybrid model with both abstractions and exemplars (Pierrehumbert, 2006, 2001) can also deal with our data.

Yu (2007) has already attempted to expound on near-merger from the perspective of exemplar theory. The exemplar-based approach can also shed some light on our data regarding cases of mergers-in-progress. According to the theory, perceptual memory traces, i.e. exemplars, associated with each phonological label, i.e. exemplar cluster or cloud, are constantly updated. As a result, the distributions of exemplars for a phonological category can be shifted as perceptual experiences accrue.

First, when individuals are exposed to sound variations, the exemplar redistribution can give rise to the approximation of phonological categories, leading to reduced phonological distinction without merger. Second, for some individuals, near-merger or incomplete neutralization occurs. when an enormous number of variable and heterogeneous exemplars are accumulating and two exemplar clouds are becoming increasingly approximated, some language users may deem the distinction unreliable and cease to use it for differentiating sound categories in perception (Labov et al., 1991; Braver, 2014). However, the distinction can be preserved in production with a different degree of overlapping, in that the computation of production targets could rely on the guidance of a different set of exemplars. Individual-specific production norms may be established through exposure to perceptual exemplars earlier in life and thus remain relatively independent at the time point when phonological contrast is suspended. This idea agrees with the dual-stream model of cortical organization of speech processing (Hickok & Poeppel, 2007), which predicts that the articulatory-motoric coding in the dorsal stream relies on sensory input initially but becomes more automated and independent with little sensory guidance as learning progresses. Third, probably due to mutual interactions between exemplars for production and perception, the distinction preserved in production might finally disappear for some individuals after the suspension of a phonological contrast. Otherwise, some individuals could simply fail to establish the phonological contrast during language acquisition. Both cases would result in complete merger. Fourth, flip-flopping speakers may over-interpret or over-generalize some extreme exemplars through personal linguistic experiences, developing a unique contrast opposite to the population grammar in certain phonetic dimensions.

5.3 Limitations and future directions

Our study has several limitations. First, our production data are not adequate to capture how the production-perception link changes as a merger progresses. For example, are production and perception closely matched at the stage of clear or reduced distinction? Are there any cases of poor production and good perception occurring at the near-merger stage, especially for the T4/T6 (near-)merger? Future studies may consider administering both production and perception tasks to further examine these questions. Second, the current study cannot pinpoint the ultimate origin of this merger. Why does T6 exhibit falling pitch contours and gradually occupy the space of T4? The variation seems to

originate from Cantonese tonal system per se, as the ongoing tonal merger has been reported in multiple Cantonese dialects (Fung et al., 2012; Ou, 2012; Weng, 2014). Although pitch fall is not used to distinguish meanings, Cantonese level tones are frequently accompanied by pitch fall (Vance, 1977). The falling pitch contours seem to be free variants of level tones, but they may be derived due to prosodic or pragmatic reasons. Impressionistically, Cantonese speakers sometimes employ falling pitch contours to express emphasis or strong emotion. However, listeners are not always capable of compensating for these contextually-induced perturbations (Ohala, 1981). T6 exemplars with context-dependent falling pitch contours may gradually become associated with context-free T6 labels, causing increased overlapping between exemplar clouds of canonical low-falling T4 and low-level T6. Future studies may test this hypothesis by designing experiments to explore the functional load of pitch fall in Cantonese and its relevance to sound change.

6. Conclusion

This paper presents data on individual variability in the perception of the ongoing Cantonese T4/T6 (near-)merger. Our analysis at the individual level demonstrates a great variety of individual manifestations of the ongoing Cantonese T4/T6 merger. We have observed that some individuals represent closely approximated T4/T6 and their T6 shows a large degree of pitch drop. This pattern reveals the direction of this sound change, that is, T6 exemplars with falling pitch contours gradually accumulate in the perceptual space and intrudes into the exemplar space of T4. Compared with traditional modular feedforward models, exemplar-based models provide strong explanations for the observed gradiency and production-perception asymmetry in the ongoing tonal merger.

APPENDIX

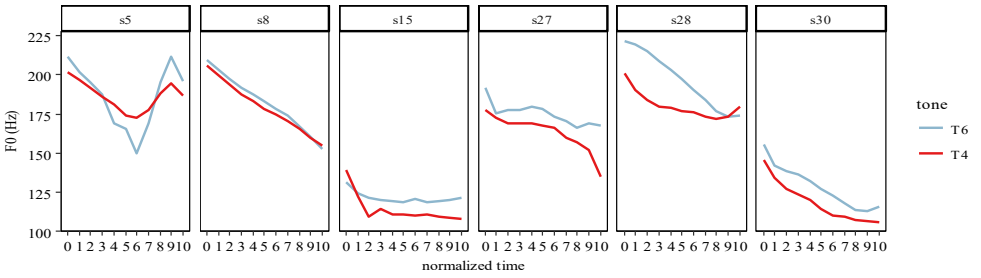

Figure 7. Tone production data from *s5*, *s8*, *s27*, *s28* and *s30*. Grand-averaged F_0 trajectories of T4 and T6, collapsed across test tokens and their repetitions, are displayed for each participant. Participant *s5* frequently produced creaky voice at the end of some tokens of T4 and T6, leading to inaccurate F_0 estimates in the later portions of the vocal segments (time point 6, 7, 8, 9 and 10). The production data from *s10* is displayed in Figure 1.

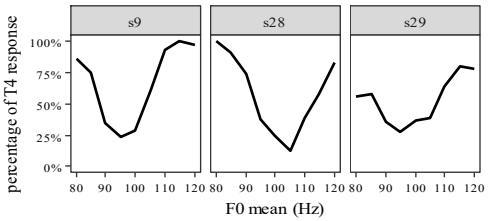

Figure 8. Raw identification curves in the F_0 mean dimension for *s9*, *s28* and *s29*. The U-shaped curve might indicate the use of additional identification labels, like wa^1/嘩 ([wa:˥], 'noise'), which carries the high-level tone (T1) and has orthographic connections with wa^4/華 ([wa:˩], 'China, splendid').

REFERENCES

BATES, DOUGLAS, MARTIN MAECHLER, AND STEVEN WALKER. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67.1: 1–48.

BAUER, ROBERT S., AND PAUL K. BENEDICT. 1997. *Modern Cantonese phonology*. Berlin; New Tork: Mouton de Gruyter.

BERMÚ DEZ-OTERO, RICARDO. 2007. Diachronic phonology. *The Cambridge handbook of phonology*, edited by de Lacy, 497–519. Cambridge University Press.

BOERSMA, PAUL, AND DAVID WEENINK. 2017. PRAAT: Doing phonetics by computer. Available from: http://www.fon.hum.uva.nl/ praat/ [accessed Jan. 2, 2017].

BRAVER, AARON. 2014. Imperceptible incomplete neutralization: Production, non-identifiability, and non-discriminability in American English flapping. *Lingua* 152: 24–44.

DIEHM, ERIN, AND KEITH JOHNSON. 1997. Near-merger in Russian palatalization. *OSU Working Papers in Linguistics 50*, 11–18.

FOK-CHAN, YUEN-YUEN. 1974. *A perceptual study of tones in Cantonese*. Centre of Asian Studies occasional papers and monographs (no. 18).

FUNG, ROXANA, CARMEN KUNG, SAM-PO LAW, I-FAN SU, AND CATHY WONG. 2012. Near-merger in Hong Kong Cantonese tones : a behavioural and ERP study. *The 3rd International Symposium on Tonal Aspects of Languages (TAL 2012)*, 1–6. Nanjing, China.

GOLDINGER, STEPHEN D. 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology* 22.5: 1166–83.

———. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105.2: 251–79.

HALL-LEW, LAUREN. 2013. "Flip-flop" and mergers-in-progress. *English Language and Linguistics* 17.2: 359–90.

HICKOK, GREGORY, AND DAVID POEPPEL. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience* 8.5: 393–402.

KHOUW, EDWARD, AND VALTER CIOCCA. 2007. Perceptual correlates of Cantonese tones. *Journal of Phonetics* 35.1: 104–17.

KONG, EUN JONG, AND JAN EDWARDS. 2016. Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics* 59: 40–57.

LABOV, WILLIAM, MARK KAREN, AND COREY MILLER. 1991. Near-mergers and the suspension of phonemic contrast. *Language Variation and Change* 3.1: 33–74.

LABOV, WILLIAM, MALCAH YAEGER, AND RICHARD STEINER. 1972. *A quantitative study of sound change in progress*. Philadelphia: U.S. Regional Survey.

MATTHEWS, STEPHEN., AND VIRGINIA. YIP. 2011. *Cantonese: A comprehensive grammar*. 2nd ed. London; New York: Routledge.

MOK, PEGGY PIK-KI., DONGHUI ZUO, AND PEGGY WAI-YI. WONG. 2013. Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change* 25: 341–70.

OHALA, JOHN J. 1981. The listener as a source of sound change. *Papers from the parasession on language and behavior*, edited by Carrie S. Masek, Roberta A. Hendrick, and Mary Frances Miller, 2nd ed., 178–203. Chicago: Chicago Linguistic Society.

OU, JINGHUA. 2012. *Tone merger in Guangzhou Cantonese*. Mphil dissertation, The Hong Kong Polytechnic University.

PIERREHUMBERT, JANET BRECKENRIDGE. 2001. Word-specific phonetics. In *Laboratory phonology 7*, edited by Carlos Gussenhoven and Warner Natasha, 2nd ed., 101–39. Berlin & New York: Mouton de Gruyter.

———. 2006. The next toolkit. *Journal of Phonetics* 34: 516–30.

R CORE TEAM. 2017. R: A language and environment for statistical computing.

RÖTTGER, TIMO B., BODO WINTER, SVEN GRAWUNDER, JAMES KIRBY, AND MARTINE GRICE. 2014. Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics* 43.1: 11–25.

SHU-HUI, PENG. 2000. Lexical versus "phonological" representations of Mandarin sandhi tones. *Papers in laboratory phonology V: Acquisition and the lexicon*, edited by Michael B. Broe and Janet Breckenridge. Pierrehumbert, 152–67. Cambridge University Press.

VANCE, TIMOTHY J. 1977. Tonal distinctions in Cantonese. *Phonetica* 34.2: 93–107.

VARLEY, ROSEMARY, AND LYDIA K.H. SO. 1995. Age effects in tonal comprehension in Cantonese. *Journal of Chinese Linguistics* 23: 76–98.

WEENINK, DAVID. 2009. The KlattGrid speech synthesizer. *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2059–62. Brighton: UK.

WENG, HUISHAN. 2014. On the reasons for tone merger in Malaysian Cantonese (试论马来西亚粤语声调简化的原因). *Wen Jiao Zi Liao* (文教资料) 10: 21–22.

YU, ALAN C. L. 2007. Understanding near mergers: the case of morphological tone in Cantonese. *Phonology* 24.1: 187–214.