Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception

Roxana S. Y. Fung, and Chris K. C. Lee

ARTICLES YOU MAY BE INTERESTED IN

Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception

Roxana S. Y. Fung[a)] and Chris K. C. Lee[b)]
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hung Hom, Hong Kong
roxana.fung@polyu.edu.hk, chriskcl@bu.edu

Abstract: This paper examines the tone-merging phenomenon in Hong Kong Cantonese. Both perception and production tasks were administered to 120 participants with ages ranging from 20 to 58 years. After considering the complicated interplay of perception and production confusion, the paper provides statistical evidence that three tonal contrasts have undergone merging in contemporary Hong Kong Cantonese. They are Full-merger T2 and T5, where contrast is collapsed in both perception and production; Partial-merger T3 and T6, where contrast is collapsed in production only; and Near-merger T4 and T6, where contrast is collapsed in perception but maintained in production.
© 2019 Acoustical Society of America
[CCC]

1. Introduction

There has been growing evidence from impressionistic observations and recent experimental data that the tonal system of Hong Kong Cantonese (HKC) is in the process of merging. HKC stands out from other tone languages in the world by having a rich tonal inventory. As shown in Fig. 1, HKC has six contrastive tones with four levels of contrast. Some studies reported the confusion of T2 (high rising) and T5 (low rising), based on a few individuals. Whereas the early works focused mainly on production confusion [Bauer *et al.* (2003) on 8 speakers; Kej *et al.* (2002) on 15 speakers]; later works studied both production and perception confusion [Yiu (2009) on 15 participants and Mok *et al.* (2013) on 17 participants]. Some studies found one more problematic pair: T3 (mid level) and T6 (low level): Peng and Wang (2005) investigated production confusion on 80 speakers; Wong (2011) reported the perception confusion on 137 participants aged 13–33. Some studies found that T6 (low level) may be confused with T4 (extra low level) as well: Law *et al.* (2013) reported the perception confusion on 40 participants using neuroimaging method; Liang (2017) and Zhang (2019) investigated both production and perception confusion on 40 and 50 speakers, respectively. All these studies have shown that perception or production of some tone pairs have exhibited great inter-individual variations. However, no conclusive evidence is found to confirm the mergers, probably due to the limitation of sampling and the resources of appropriate statistical analyses. The tone-merging phenomenon remains anecdotal. The present paper attempts to confirm the tone pairs that exhibit mergers on a larger sample size of HKC speakers using more sophisticated acoustic and statistical methods. For the convenience of discussion, we use the term "tone merger" to refer to the collapse of the phonetic contrast of two tones, and coin the term "tone mergerer" to refer to a person who merged two tone categories.

2. Method

2.1 Participants

A total of 120 native HKC speakers (61 males, 59 females) with ages ranging from 20 to 58 years [$M = 34.43$ years, standard deviation (SD) $= 11.9$] participated in this study. All participants were born and raised in Hong Kong with HKC as their native language and the major communication means at home. None of them has known hearing or speech disorder.

2.2 Speech materials

The speech materials included 48 syllables being derived by eight CV roots. The first set included four roots: /fu/, /sɛ/, /si/, /ji/, which generated 24 real syllables for all six

---

[a)]Author to whom correspondence should be addressed.
[b)]Also at: Department of Linguistics, Boston University, Boston, MA 02215, USA.
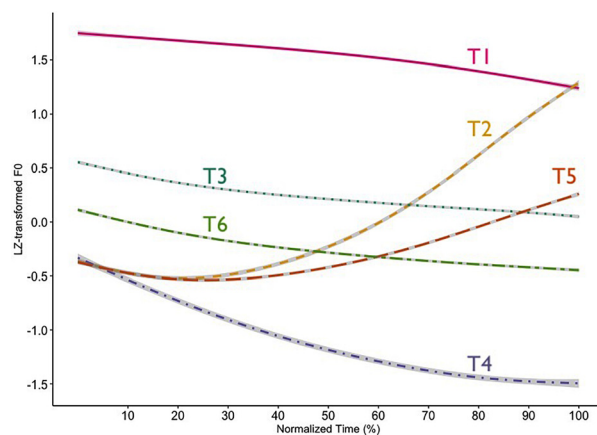
Fig. 1. (Color online) The trajectories of the six contrastive tones based on the production of all participants in our study.

tones. The second set included four roots: /ku/, /pʰɔ/, /ja/ and /jɛ/, which generated 12 real syllables and 12 nonsense syllables for the six tones. The syllables were produced by the first author, a female native HKC speaker who was born and raised in Hong Kong and without any sign of tone merger. The overall intensity of all stimuli was normalized using PRAAT (Boersma, 2001). However, the duration of all stimuli was not normalized to keep the stimuli as natural as possible. The stimuli were acoustically inspected and auditorily checked by three native speakers of HKC to make sure that stimuli are perfectly distinctive. All the speech materials were also analyzed with Pillai score as outlined in Sec. 3.2 and confirmed that they showed no signs of tone merging. All 48 syllables formed the stimuli of the perception task. As for the production task, the stimuli contained the 36 real syllables only. A Chinese character was chosen to represent each of the real syllables. Characters with alternate readings were avoided as far as possible.

2.3 Procedures

Production and perception tasks were administered to the participants using a computer program specifically developed for this study. All tasks were run on a Lenovo notebook computer T400 connected to an external sound card, Creative Sound Blaster Audigy 2NX (Singapore), located in a sound-attenuating recording studio at the Hong Kong Polytechnic University. Instructions of the tasks were displayed in Traditional Chinese on the computer screen and were read aloud in Cantonese to each subject. The production task was performed before the perception task to eliminate the priming effect. Trials were given to the participants at the beginning of each task for practice. There was a break in between two tasks. The whole experiment lasted about 1 h.

2.3.1 Perception task

The AX discrimination test was administered to the participants. Any two syllables differing only in tone from the speech materials were paired by a randomizer of the computer program and were presented to the participants over earphones at a hearing level that could be adjusted by the participant. A total of 168 pairs of syllables [8 CV roots × 21 tonal contrasts (15AB contrast + 6AA contrast)] were presented. The participants were instructed to indicate whether the two aurally presented syllables were identical or not by clicking the "same" or "different" button on the computer screen accordingly. The participants could replay the tone pairs up to 3 times before making their final decision. There was no time limit for participants to submit their responses.

2.3.2 Production task

Each of the 36 real syllables was embedded in the following two sentence carriers of different positions forming 72 stimuli (36 syllables × 2 positions): /ŋɔ ji ka tuk ___ ji/ "I am now reading the ___ character" and /ni kɔ ji hɐi ___/ "This character is ___." The stimuli were randomly drawn by the computer program and presented visually to the participants. After reading the written stimuli on the computer screen, the participants clicked the record button on the screen and read out the stimuli shown. The recording was carried out with a Shure SM48S low-noise unidirectional microphone setting at about 15 cm mouth-to-phone distance. After the recording, each sentence

was low-pass filtered at 22 kHz, digitized at a sampling rate of 44.1 kHz, and stored onto the Lenovo computer as a separate audio file.

3. Statistical analysis and results

3.1 The perception task

The perception accuracy of each tone pair in the discrimination task was measured by the percentage of correct responses out of the total number of trials for each participant. A total of 20 160 trials were analyzed. The perception accuracy of the 21 tonal contrasts is shown by the bar plots in Fig. 2. While the discrimination of most tonal contrasts almost reached the ceiling of 100% correct, some tone pairs seem to be more prone to confusion than others. In particular, T2-T5 displays the greatest individual variation in discriminability [SD = 32.6%, interquartile range (IQR) = 50%]: while the mean accuracy rate is as high as 70.6%, some participants scored as low as 0% ($N = 9$). T4-T6 is the second most varied in terms of discriminability (SD = 20.4%, IQR = 25%). On the other hand, T3-T5 is the only tone pair that is discriminated perfectly by all 120 participants.

To assess whether some tone pairs are more difficult to be discriminated from others, the participants' responses to the AX discrimination task were hand fitted into a binomial mixed effects model in R (R Core Team, 2019) with the *lme4* package (Bates *et al.*, 2015). The dependent variable was the correct response to a trial. The independent variables included the following linguistic factors: tone pair ("TNPAIR"), and whether the trial contained a nonsense syllable ("NSSYL"). The identity of tone pairs was coded with sum coding, so that every level was compared with the grand mean. Since T3-T5 is invariably discriminated 100% correct by all participants, this pair is excluded from comparisons by arranging it as the last level of the independent variable.

The model was hand-fitted using a forward stepwise selection, with pairwise model selection guided by Akaike Information Criterion. The best model contained the separate fixed effects of TNPAIR and NSSYL. For random effects, the model had participant and test trials as random intercepts. The output of the binomial mixed model investigating the relative difficulties of tone discrimination is shown in Table 1. In general, if a trial compares two real syllables, tone discrimination becomes slightly easier ($p = 0.0004$). Recall that the intercept of the model being the grand mean probability of correct responses, T2-T5 ($p = 0.0063$) and T4-T6 ($p = 0.031$) are the only two pairs that show a significant difference from the grand mean. Hence, it is confirmed that T2-T5 and T4-T6 are the two pairs that pose difficulties to native HKC speakers in perception at the community level.

In order to understand the extent of the tonal confusion across the community, we have to set an accuracy score in the discrimination task to determine which participants are tone mergerers. Since 50% correct is the chance level performance of this two response-option forced-choice task, we consider a participant who achieved 62.5% correct (i.e., 5 out of 8 responses correct) or lower in the discrimination task a tone mergerer at the perception side. Table 2 shows the percentage of tone mergerers
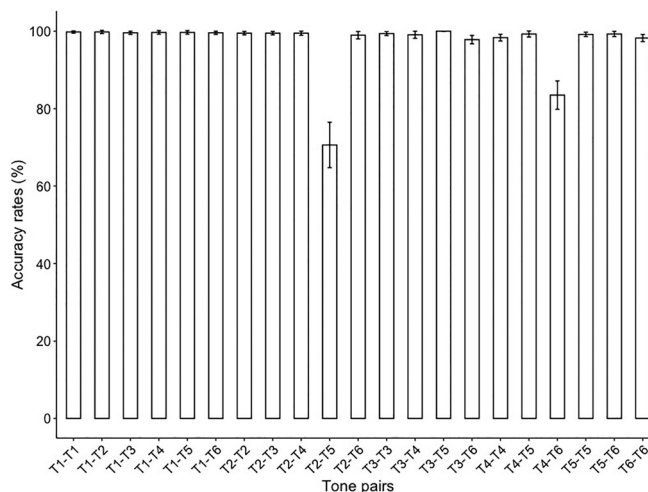

Fig. 2. Boxplots displaying accuracy of discriminating the 21 tonal contrasts in the perception task. Error bars show 95% confidence intervals.

Table 1. Output of the binomial mixed model investigating the relative difficulties of tone discrimination. [#]trials containing no nonsense syllables. Asterisks flag the levels of significance: *, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$.

	Estimate	SE	z	p	
(Intercept)	6.210	1.814	3.423	0.001	***
NONSYL_FALSE[#]	0.345	0.097	3.568	0.0004	***
TNPAIR_T1-T1	0.761	1.942	0.392	0.695	
TNPAIR_T1-T2	0.908	1.949	0.466	0.641	
TNPAIR_T1-T3	0.254	1.888	0.135	0.893	
TNPAIR_T1-T4	0.569	1.903	0.299	0.765	
TNPAIR_T1-T5	0.474	1.907	0.249	0.804	
TNPAIR_T1-T6	0.183	1.890	0.097	0.923	
TNPAIR_T2-T2	−0.188	1.874	−0.100	0.920	
TNPAIR_T2-T3	−0.167	1.874	−0.089	0.929	
TNPAIR_T2-T4	−0.087	1.874	−0.047	0.963	
TNPAIR_T2-T5	−4.984	1.824	−2.732	0.0063	**
TNPAIR_T2-T6	−0.718	1.846	−0.389	0.697	
TNPAIR_T3-T3	−0.337	1.867	−0.180	0.857	
TNPAIR_T3-T4	−0.704	1.851	−0.381	0.704	
TNPAIR_T3-T6	−1.478	1.836	−0.805	0.421	
TNPAIR_T4-T4	−1.429	1.838	−0.777	0.437	
TNPAIR_T4-T5	−0.413	1.863	−0.222	0.825	
TNPAIR_T4-T6	−3.942	1.825	−2.160	0.031	*
TNPAIR_T5-T5	−0.663	1.854	−0.358	0.721	
TNPAIR_T5-T6	−0.488	1.858	−0.263	0.793	
TNPAIR_T6-T6	−1.488	1.838	−0.810	0.418	

in our study. Note that 35% and 18.3% of our participants were unable to discriminate T2-T5 and T4-T6, respectively, in perception.

3.2 The production task

All the audio recordings produced by the participants were manually labeled by the second author, a native speaker of HKC. Only the vocalic part of the target words was labeled. Tokens that contain segmental errors were discarded. Among 8640 tokens, 377 tokens were excluded. The pitch tracks of all labeled target syllables were then estimated by the STRAIGHT (Kawahara *et al.*, 1998) and SHR (Sun, 2000) algorithms included in VoiceSauce (Shue *et al.*, 2011). STRAIGHT was the default algorithm in VoiceSauce while SHR algorithm was robust in estimating the pitch of creaky instances. The two algorithms were cross-validated. Four more tokens were excluded due to poor estimation. The pitch tracks estimated by the STRAIGHT algorithm were taken for further analysis because of its relatively wider adoption in the field. All pitch tracks ($N = 8259$) were outputted in 13 time-normalized subsections, with the first and last discarded so that any carryover effect or anticipatory effect are avoided. The $F0$ values were then LZ-transformed (Zhu, 2010) by Eq. (1), where mean and SD are talker-specific,

$$F0_x^{LZ} = \frac{\log_{10}(F0_x) - \text{mean of } \log_{10}(F0)}{\text{SD of } \log_{10}(F0)}. \tag{1}$$

The statistical output "Pillai score" of a multivariate analysis of variance (MANOVA) models (Hay *et al.*, 2006) was adopted to examine the degree of overlap between the trajectories of the two tones. The Pillai score has been adopted by researches investigating vowel mergers [see Hall-Lew (2010) and Nycz and Hall-Lew (2014) for a

Table 2. The percentage of participants considered as mergerers of a particular tonal contrast.

	T1-T2	T1-T3	T1-T4	T1-T5	T1-T6	T2-T3	T2-T4	**T2-T5**	T2-T6	T3-T4	T3-T5	**T3-T6**	T4-T5	**T4-T6**	T5-T6
% of perception mergerers	0	0	0	0	0	0	0	**35**	0.8	0	0	**0**	0.8	**18.3**	0
% of production mergerers	0	4.17	0	0.83	0.83	2.50	11.67	**22.5**	0.83	2.50	4.17	**46.67**	2.5	**15**	7.5

discussion]. This study extended this measure to model variations of the distinctiveness of two tones with respect to the 11 time-normalized, LZ-transformed $F0$ values. The Pillai score ranges from 0 to 1. The higher the Pillai score the larger the phonetic distinctiveness between the trajectories of the two tones. The significance of the distinctiveness of the two tones is indicated by the generated p value. The Pillai score of each tone pair of each individual speaker was generated separately using the MANOVA() function in R. The 11 time-normalized pitch points of all valid target syllables produced by a speaker were included as the dependent variable. Tone was included as the independent variable.

The Pillai scores of the 15 AB tonal contrasts are shown by the boxplots in Fig. 3. It is revealed that three tonal contrasts: T3-T6, T2-T5, T4-T6 have the lowest Pillai scores and highest variability. Our result coincides with previous findings that these tone pairs are prone to tone merger.

Next, we will examine the overlap of the tone pairs at an individual level. Note that the Pillai score alone does not confirm whether an individual is a tone merger or not. Hence, we follow the approach by Mayr *et al.* (2019), where participants receiving a Pillai score with statistical significance (i.e., $p < 0.05$) is considered a distinct speaker, while those without statistical significance (i.e., $p > 0.05$) is considered a mergerer. Table 2 shows the percentage of tone mergerer in our study. Note that a high percentage of our participants (46.7%) were unable to produce T3-T6 distinctively, and 22.5% and 15% of our population were not able to discriminate T2-T5 and T4-T6, respectively, in production.

4. Discussion and conclusion

Taking into account the results of the perception and production tasks, three tonal contrasts, T2-T5, T3-T6, and T4-T6 are identified as mergers-in-progress. However, an answer to our research question will not arrive until we solve two challenges. The first challenge: What count as a tone merger if the tonal contrast is not lost in both perception and production? In fact, numerous studies on sound mergers have documented that the collapse of phonemic distinction may not proceed at the same rate in these two language modalities (see DeCamp, 1953; Kleber *et al.*, 2011; Labov *et al.*, 1991). In our study, we found individuals may fall under one of the four types for each of the three problematic tonal contrasts. There were individuals who (1) unmerged in both production and perception; (2) merged in both production and perception; (3) merged in their own productions but were able to perceive the contrast of the stimuli; and (4) unmerged in their own productions but unable to perceive the contrast of the stimuli. These speech patterns have been quite commonly attested in studies of merger-in-progress, as early as in DeCamp (1953). The distribution of the four types of speakers for each problematic tone pair in our study is shown in Fig. 4. Note that a high percentage of our participants (40% for T2/T5, 47% for T3/T6, 32% for T4/T6) failed to maintain the contrast in both perception and production for the three problematic tone pairs. However, most of the merging speakers have problems only in one of the language modalities.

In addition to the asymmetry of production and perception, there is another challenge: How to bridge the gap between the tone variations at the individual level and the tone change at the community level? As an attempt to solve these two challenges, we conducted a correlation between the perception scores and the Pillai scores of all the tonal contrasts for each participant. Figure 5 shows the scatterplots of all 15 tonal contrasts.
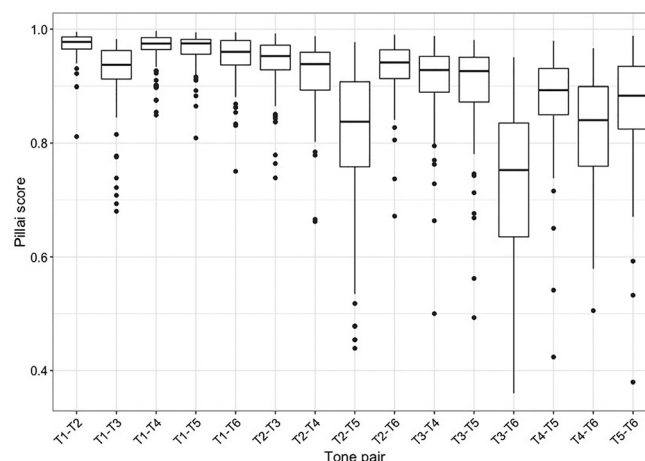
Fig. 3. Boxplots displaying Pillai score of the 15 tonal contrasts in the production task. Dots represent outliers.

(a) T2-T5

Production \ Perception	Unmerged	Merged
Unmerged	60%	17.5%
Merged	5%	17.5%

(b) T3-T6

Production \ Perception	Unmerged	Merged
Unmerged	53.33%	0
Merged	46.67%	0

(c) T4-T6

Production \ Perception	Unmerged	Merged
Unmerged	68.33%	16.67%
Merged	13.33%	1.66%

Fig. 4. Distribution of four types of participants of (a) T2-T5; (b) T3-T6; and (c) T4-T6.

Obviously, the majority of the tonal contrasts display the following pattern: most of the points in the scatterplot are crowded around the top right corner. This indicates that most tonal pairs remain contrastive in both perception and production. However, distinct patterns are found in T3-T6, T4-T6, and T2-T5. The points in the scatterplot of T3-T6 form a horizontal line on the top portion of the plot. It reveals that the tonal contrast is collapsed in production but well maintained in perception. A moderate positive correlation between perception confusion and production confusion is found in T2-T5 (Kendall's T = 0.31, z = 4.63, $p < 0.001$). This suggests that the loss of T2-T5 contrast is actualized in parallel in both language modalities. For the T4-T6 pair, a very weak negative correlation, though not significant, is observed (Kendall's $T = -0.08$, $z = -1.15$, $p = 0.25$). This weak negative correlation seems to point to a tendency that the tonal contrast is maintained in production but collapsed in perception.

The distinct patterns of the scatterplots provide compelling evidence to support the tone-merging phenomenon in contemporary HKC. They also suggest that there may be three different types of mergers-in-progress at the community level: T2-T5 is a full-merger, where contrast is collapsed in both perception and production. T3-T6 is a partial-merger, where contrast is collapsed in production only. T4-T6 is a near-merger, where contrast is collapsed in perception but maintained in production. The phenomenon of near-merger is counterintuitive, which challenges many current speech processing models, but has been widely attested in sociolinguistic studies on sound change (such as DeCamp, 1953; Di Paolo and Faber, 1990; Faber and Di Paolo, 1995; Janson and Schulman, 1983; Labov et al., 1991). The T4-T6 near-merger captured in this study has been reported in a study using neuroimaging technology (Law et al., 2013). Adding further complexity to this phenomenon is the competition of the T4-T6 near-merger with the T3-T6 partial-merger. The low level T6 may merge with the mid level T3 or with the extra-low level T4. It is still unclear to us how the competing mergers are actualized. At the current stage, the T3-T6 merger dominates the production modality whereas the T4-T6 merger dominates the perception modality in the same speech community.

In conclusion, this study provides conclusive evidence for the tone mergers in HKC and breaks new ground by identifying a complex set of tone mergers in HKC of three different types of mergers-in-progress. Based on this foundation, we are going to
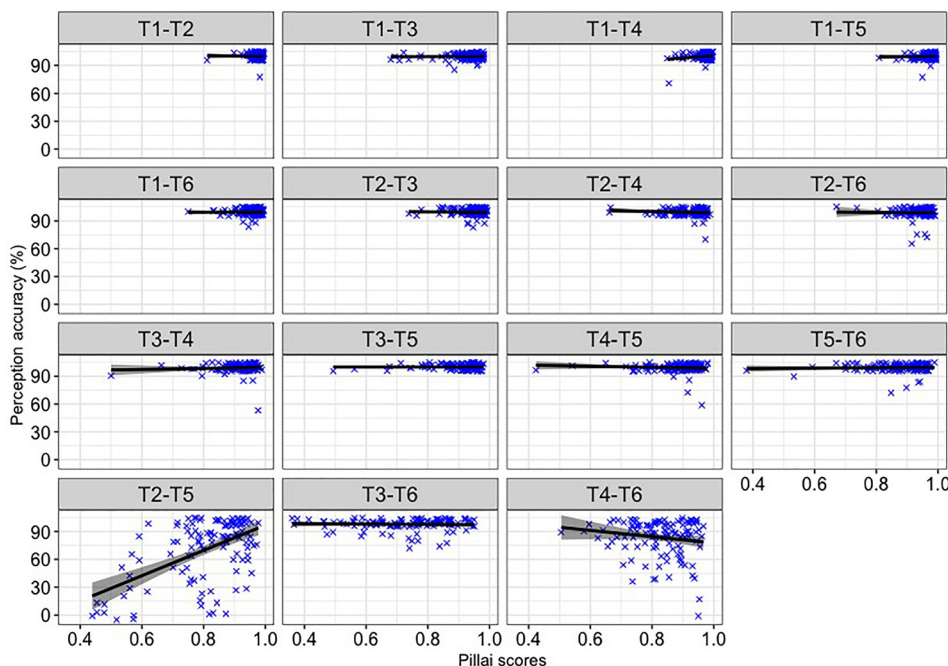

Fig. 5. (Color online) Scatterplots showing the correlations between production (Pillai scores, at the X axis) and perception (percent accuracy, at the Y axis) of the 15 AB tonal contrasts.

report the gender and age group effects of the tone mergers and discuss their implications in the cause, direction, and the mechanism of the mergers in subsequent papers. It is no doubt that studies on HKC tone mergers will shed intriguing light on the long-standing debate on which language modality initiates a sound change.

Acknowledgments

References and links

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Software **67**(1), 1–48.

Bauer, R. S., Cheung, K., and Cheung, P. (**2003**). "Variation and merger of the rising tones in Hong Kong Cantonese," Lang. Variation Change **15**(02), 211–225.

Boersma, P. (**2001**). "PRAAT, a system for doing phonetics by computer," Glot Int. **5**(9/10), 341–347.

DeCamp, D. (**1953**). *The Pronunciation of English in San Francisco* (University of California, Berkeley, CA).

Di Paolo, M., and Faber, A. (**1990**). "Phonation differences and the phonetic content of the tense-lax contrast in Utah English," Lang. Variation Change **2**(2), 155–204.

Faber, A., and Di Paolo, M. (**1995**). "The discriminability of nearly merged sounds," Lang. Variation Change **7**(1), 35–78.

Hall-Lew, L. (**2010**). "Improved representation of variance in measures of vowel merger," Proc. Mtgs. Acoust. **9**, 060002.

Hay, J., Warren, P., and Drager, K. (**2006**). "Factors influencing speech perception in the context of a merger-in-progress," J. Phonetics **34**(4), 458–484.

Janson, T., and Schulman, R. (**1983**). "Non-distinctive features and their use," J. Linguistics **19**(2), 321–336.

Kawahara, H., Cheveigne, D. A., and Patterson, R. D. (**1998**). "An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: Revised TEMPO in the STRAIGHT-suite," in *The 5th International Conference on Spoken Language Processing*, Sydney, Australia (November 30–December 4, 1998), Paper 0659, available at https://pdfs.semanticscholar.org/5295/58bdc6b6fcc880d2242273abc8a0fad4af3b.pdf (Last viewed May 21, 2019).

Kej, J., Smyth, V., So, L. K. H., Lau, C. C., and Capell, K. (**2002**). "Assessing the accuracy of production of Cantonese lexical tones: A comparison between perceptual judgement and an instrumental measure," Asia Pacific J. Speech, Lang., Hear. **7**(1), 25–38.

Kleber, F., Harrington, J., and Reubold, U. (**2011**). "The relationship between the perception and production of coarticulation during a sound change in progress," Lang. Speech **55**(3), 383–405.

Labov, W., Karen, M., and Miller, C. (**1991**). "Near-mergers and the suspension of phonemic contrast," Lang. Variation Change **3**(1), 33–74.

Law, S.-P., Fung, R. S. Y., and Kung, C. (**2013**). "An ERP study of good production vis-à-vis poor perception of tones in Cantonese: Implications for top-down speech processing," PLoS One **8**(1), e54396.

Liang, Y. (**2017**). "The production-perception mechanism in tonal shift: The case of Hong Kong Cantonese," Zhongguo Yuwen **381**(6), 723–732 (in Chinese).

Mayr, R., López-Bueno, L., Vázquez Fernández, M., and Tomé Lourido, G. (**2019**). "The role of early experience and continued language use in bilingual speech production: A study of Galician and Spanish mid vowels by Galician-Spanish bilinguals," J. Phonetics **72**, 1–16.

Mok, P. P. K., Zuo, D., and Wong, P. W. Y. (**2013**). "Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese," Lang. Variation Change **25**(3), 341–370.

Nycz, J., and Hall-Lew, L. (**2014**). "Best practices in measuring vowel merger," Proc. Mtgs. Acoust. **20**, 060008.

Peng, G., and Wang, W. S. Y. (**2005**). "Tone recognition of continuous Cantonese speech based on support vector machines," Speech Commun. **45**(1), 49–62.

R Core Team. (**2019**). "R: A language and environment for statistical computing," https://www.r-project.org/ (Last viewed May 1, 2019).

Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (**2011**). "Voicesauce: A program for voice analysis," in *Proceedings of the 17th International Congress of Phonetic Sciences*, edited by W.-S. Lee and E. Zee, Hong Kong (August 17–21), Vol. 3, pp. 1846–1849.

Sun, X. (**2000**). "A pitch determination algorithm based on Subharmonic-to-Harmonic Ratio," in *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, China (October 16-20), Vol. 4, pp. 676–679, available at https://www.isca-speech.org/archive/icslp_2000/i00_4676.html (Last viewed May 21, 2019).

Wong, Y. W. (**2011**). "Sound changes in Hong Kong Cantonese: A multi-perspective study," doctoral dissertation, The Chinese University of Hong Kong, Hong Kong.

Yiu, C. Y. (**2009**). "A preliminary study on the change of rising tones in Hong Kong Cantonese: An experimental study," Lang. Linguistics **10**(2), 269–291.

Zhang, J. (**2019**). "Tone mergers in Cantonese: Evidence from Hong Kong Macao, and Zhuhai," Asia-Pacific Lang. Variation **5**(1), 28–49.

Zhu, X. (**2010**). *Phonetics* (Commercial Press, Beijing) (in Chinese).