What tones emerge in Cantonese spontaneous speech?*

Roger Yu-Hsiang Lo
University of British Columbia

Molly Babel
University of British Columbia

Abstract: Traditional descriptions of Cantonese identify six tones, in addition to three allotones that occur on checked syllables ending in /ptk/. Since the turn of the century, however, linguists have observed several tone mergers that reshape the Cantonese tone inventory. In the current work, and with inspiration from emergent approaches that consider linguistic data from more bottom-up perspectives, we assess the Cantonese tone space in spontaneous speech using the SpiCE corpus. In line with an emergent approach, we measure f0 trajectories, thereby eschewing point estimates, and duration. We find that while three Cantonese tone pairs are merged at the community level with respect to f0 trajectories, some individuals maintain distinctions. In terms of duration, however, we find evidence at the community and individual levels for distinctions between two of the three merging tone pairs. These findings invite future work that queries how Cantonese listeners navigate the tone space in terms of cross-talker variation and duration cues. In short, what tone grammar emerges from the phonetic variation?

Keywords: Cantonese, tones, emergent phonology, corpus phonetics, spontaneous speech

1 Introduction

In the course of linguistic description, a linguist can assert the number of tones that provide lexically meaningful contrast in a language, but on what do they base their assertion? We can poke at this question and its underlying assumptions from multiple angles. First, there is the basic question of whose language represents *the* language. That is, which user or users of the language constitute one's sample.[1] Wrapped in this decision of the individual(s) whose language is implicitly vaulted as the selected sample is that individual's or those individuals' positioning relative to a perceived standard and whether the language *has* a community standard. Moving to the language material that serves the basis for the description of tone: Is it based on unique patterns observed in spontaneous speech? Is it deduced from patterns observed on elicited target items? Is it based on speakers providing lexically items that contrast based on tone? Is it based on a sample of listeners confirming that patterns that appear distinct in production reliably map to distinct lexical items in word recognition?

Regardless of the means used by the linguist to determine whether a language has lexical tone and, if so, describe the inventory, that method is bound to be rather different from the method implicitly deployed by an early language learner. The early language learner is not in a position to shop around for an "ideal" source of linguistic material, nor will they have deliberated over methodological options from which to ultimately glean the patterns. Archangeli and Pulleyblank (2022)'s emergent phonological framework starts from the hypothesis that a language learner needs only

* Thank you to Doug for your curiosity, kindness, and collegiality. Across your career, you have navigated theoretically transformational work and scholarship that brings value to speakers and communities. We can only aspire to follow in your footsteps. You are an inspiration as a human being and linguist!

[1] Given that the focus of this work is on lexical tone, a property of auditory-vocal languages, from hereafter we use terms that reference users of auditory-vocal languages—that is, listeners and talkers. Our introductory narrative, however, applies to languages of all modalities.

Table 1: Cantonese phonemic tone inventory. Tone numerals following the Chao transcription system characterize the pitch height and shape (Chao 1947); these values are given in square brackets. Jyutping transcription conventions present tones as numbers, which is the system used within this paper.

Contour	Tone	Description	Example word
Level	1 [55]	High-level	衣 *ji1* 'clothes'
	3 [33]	Mid-level	意 *ji3* 'idea'
	6 [22]	Low-level	二 *ji6* 'two'
Rising	2 [25]	High-rising	椅 *ji2* 'chair'
	5 [23]	Mid-rising	耳 *ji5* 'ear'
Falling	4 [21]	Low-falling	疑 *ji4* 'suspicious'

bringing their general cognitive principles to the task of distilling their phonological grammar from the language patterns received. While Archangeli and Pulleyblank's theoretical drive comes more from a re-envisioning of what the learner brings to the task—general cognition and not a universal grammar—emergentist perspectives invite a reconsideration of what is available in the linguistic signal, which in the case of our current inquiry is speech.

Inspired by Doug's recent work on Emergent Phonology and his influential body of work on tone, we take the opportunity to revisit Cantonese tone.

1.1 Cantonese and its tonal inventory

Cantonese is a member of the Yue family of Sino-Tibetan, and is spoken in diaspora, but also in Hong Kong, Macau, and the southern Chinese provinces of Guangdong and Guangxi. We focus on Cantonese as spoken in Hong Kong and in Vancouver, BC, as these are the varieties that predominate in the Speech in Cantonese and English Corpus (SpiCE; Johnson 2021), which provides the data for our investigation.

The Hong Kong Cantonese tone inventory is often described as being comprised of six lexical tones, three of which are level (55, high-level: T1, 33, mid-level: T3, 22, low-level: T6) and three of which exhibit contours (25, high-rising: T2; 23, mid-rising: T5; 21, low-falling: T4). A minimal sextuplet is presented in Table 1 to illustrate these contrasts. Additionally, there are three allotones of the three level tones that occur on syllables with an unreleased obstruent /ptk/ in the coda (Bauer and Benedict 1997).

Traditional linguistic descriptions of the Cantonese tone space recognize the important role of fundamental frequency (f0) in terms of height, contour, and magnitude of said contour, in addition to voice quality and duration (Fok Chan 1974; Gandour 1981; Khouw and Ciocca 2007). Crucially, however, despite apparent differences in duration between Cantonese tones (e.g., Bauer and Benedict 1997), early scholarship found limited evidence for the role of duration in listener behaviour when asked to recognize tones in whispered speech (Fok Chan 1974). More recent work also observes limited use of voice quality by Cantonese listeners. Non-modal phonation increases perception of T4 compared to T6 for low f0 items, but not for higher f0 T6 items (Zhang and Kirby 2020). Francis, Ciocca, Ma, and Fenn (2008:269) characterize the space in stating, "In Cantonese, no non-f0-related properties have been shown to correlate consistently with tone identity, or to be used consistently by listeners even in the absence of f0 information (Ciocca, Francis, Aisha, & Wong, 2002; Fok Chan,

1974; Vance, 1976)."

Alongside these traditional descriptions of the Cantonese tone space, which highlight the value of the f0 signal, scholars have long observed several tone mergers that unmoor the contrasts. One of the mergers, that between T3 (33) and T5 (23), has received less attention in the Hong Kong Cantonese tone literature, though this merger is also extant in other varieties, such as Malayan Cantonese (Matthews and Yip 2011). Three mergers have a growing body of literature, the mergers of T2 (25) and T5 (23); T3 (33) and T6 (22); and T4 (21) and T6 (22). The earliest scholarship on the Cantonese tone mergers refers to these tone mergers as "errors" (Kej, Smyth, So, Lau, and Capell 2002). Now characterized as sound changes, the mergers have been well-studied in single word production (Bauer, Cheung, and Cheung 2003; Yiu 2009) and perception (Lee, Chan, Lam, van Hasselt, and Tong 2015; Mok, Zuo, and Wong 2013; Soo and Babel 2023). Fung and Lee (2019) characterize the T2 and T5 merger as "nearly complete", whereas the T3-T6 merger and the T4-T6 merger are described as "partial" and "near-merger", respectively, pointing to differences in listeners' retaining the ability to reliably identify words associated with particular tones.

Cantonese is spoken in diaspora communities, which invites questions about the state of the tone mergers across different communities. Across a series of experiments that test the relative roles of semantic and phonetic information in guiding the parsing of single words and sentences, Lam (2018) found that Cantonese homeland listeners (i.e., born and raised in Hong Kong) and heritage listeners (i.e., born and raised in Canada) made errors of similar types, but heritage listeners made errors at higher rates. This means, crucially, that both listener groups showed evidence of the mergers in word identification.

In terms of production, recent work by Nagy, Tse, and Stanford (2024) advanced our understanding of the Cantonese tone mergers by crucially looking at the realizations of the tones in spontaneous speech. Using sociolinguistic interviews from the Heritage Language Variation and Change in Toronto Project (HLVC), Nagy and colleagues analyzed the tones of 32 individuals, divided across three generations. Nagy and colleagues refer to their groups as Generation 0 (Hong Kong born, raised, and based), Generation 1 (Hong Kong born and raised, and migrated to Toronto as young adults), and Generation 2 (raised in Toronto). Nagy and colleagues quantified f0 in terms of (i) the mean f0 for a tone in the middle 80% of the vowel, (ii) the f0 as 90% of the vowel, and (iii) the linear slope of the f0 curve from 10% to 90%, and ran models with each of these measures as dependent measure for six tone contrasts: three of merging tones (T2-T5, T3-T6, and T4-T6) and three of tone pairs that are not merging (T1-T4, T1-T6, and T2-T6). Across their 18 models, they find no differences between generations with respect to the merging tone pairs, and suggest that the mergers may be more advanced in spontaneous speech than what had been previously observed in lab-elicited speech.

1.2 Research Questions

While the recent work by Nagy et al. (2024) is the first large-scale study of Cantonese tone from spontaneous speech (note that on a smaller scale, Fok Chan (1974) did include some spontaneously-produced speech in her work), there are aspects of the analysis that preemptively constrain our understanding of the Cantonese tone space. Taking an emergentist approach where we take the f0 signal as a trajectory and not point estimates or coerced linear shapes, we query the tone space of Cantonese speakers in the SpiCE corpus (Johnson 2021).

Moreover, while previous work has observed that Cantonese listeners do not use tone-specific

duration patterns in tone and word identification tasks, we make the decision to include duration in our measures. Duration differences can exist in the speech signal as part of a representation-based duration difference as part of segmental or tonal categories or as part of a psycholinguistic speech production process where less frequent words are produced more slowly than words with higher frequency, rendering apparent homophones not wholly homophonous (e.g., Gahl 2008, 2009; Lohmann 2018).

2 Methodology

2.1 Data

The data comes from the SpiCE corpus (Johnson 2021), which contains Cantonese and English speech from 34 early Cantonese-English bilinguals (17 female and 17 male) from the Cantonese-speaking community in Metro Vancouver. The bilinguals completed three tasks in each language – reading a fixed set of sentences, describing a storyboard, and participating in a structured interview. For this project, we only used the data from the interview portion, although we also used tokens from the sentence part for talker-specific f0 normalization (see Section 2.2). The interviews were first transcribed automatically and then corrected manually.[2] The sound files, along with the transcriptions, were passed through the Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, and Sonderegger 2017) to obtain phone boundaries, which can be used to identify syllable boundaries. A full description of the creation of the SpiCE corpus can be found in Johnson (2021).

2.2 Measurement

We extracted the f0 trajectory and duration of each syllable. For the f0 trajectory, we first used Google REAPER (Talkin 2014) to identify the voiced portion of the syllable. We then estimated and extracted the f0 values of 11 equal-distance points over this voiced portion, using the STRAIGHT algorithm (Kawahara, de Cheveigné, and Patterson 1998) implemented in VoiceSauce (Shue, Keating, Vicenik, and Yu 2011). We chose to extract f0 values over the "acoustically" voiced portion, as opposed to over the "phonological" voiced segments, based on the observation that (1) forced-aligned phone boundaries within a syllable are sometimes inaccurate (so extracting f0 values within these phone boundaries might lead to erroneous values around the beginning and end of a trajectory), and that (2) many phonological voiceless onset segments are lenited (and become voiced) or elided, resulting in an entire syllable being voiced (so extracting f0 based on phone boundaries will result in truncation of the f0 trajectory unexperienced by a hypothetical listener parsing the signal). We also excluded tokens where no voicing was detected. Furthermore, to mitigate errors introduced through pitch-doubling and pitch-halving, we filtered out tokens where the ratio of the maximum f0 (of a trajectory) to the minimum f0 was greater than 2. Altogether this filtering procedure left us with a total of 99,416 syllables. To better compare f0 differences across speakers, we further converted these f0 values to semitones, using as the basis individual talkers' mean f0 of T1 tokens from the read speech portion of the SpiCE corpus.[3]

Following the same principles as those for f0, the duration of a syllable is operationalized as the duration of the acoustically voiced portion of the syllable, making this interval the phonetic

[2] For this project, we excluded any code-switched (Cantonese to English) tokens.

[3] Hertz-to-semitone conversion formula used: $\text{f0}_{\text{st}} = 12 \log_2(\text{f0}_{\text{hz}}/\text{Reference f0}_{\text{hz}})$.

tone bearing unit. To make sure the analysis of duration parallel to that of f0, we only included the durations from tokens that were also included the f0 analysis.

2.3 Analysis

All the analyses were performed using R (R Core Team 2021). We employed a generalized additive model (GAM; Wood 2011) to model f0 trajectories, represented by the 11 equal-distance points, as a function of the linguistic and social variables considered by Nagy et al. (2024). Specifically, we allowed f0 contours to vary in overall height in response to a set of fixed variables—TONE of the syllable in question (T1-6; treatment coding with T1 as the reference level), ONSET (sonorant, obstruent; sum coding with sonorant coded as 1 and obstruent as -1), SYLLABLE TYPE (checked, non-checked; sum coding with non-checked coded as 1 and checked as -1), PRECEDING TONE (T1-6, none; treatment coding with none as the reference level), FOLLOWING TONE (T1-6, none; treatment coding with none as the reference level), POSITION IN UTTERANCE (numeric in $[0, 1]$; 0 indicates the beginning of the utterance and 1 the end)—and random variables—CHARACTER, POSITION IN WORD, GENDER, and TALKER. We also allowed the shape and wigglyness of the trajectory to vary for each tone-talker combination.[4]

For duration, owing to the large number of data points, we fitted an approximate Bayesian mixed-effects model (using the `INLA` package in R; Rue and Martino 2009) that predicts the logarithmic syllable duration based on TONE (T1-6; treatment coding with T1 as the reference level), SYLLABLE TYPE (checked, non-checked; sum coding with non-checked coded as 1 and checked as -1), PRECEDING TONE PRESENT (Yes, No; treatment coding with No as the reference level), FOLLOWING TONE PRESENT (Yes, No; treatment coding with No as the reference level), POSITION IN UTTERANCE (numeric in $[0, 1]$), and SPEECH RATE (phones per second; standardized). The model also contained the following random effects: CHARACTER, POSITION IN WORD, and TALKER.[5]

2.4 Results

Given the exploratory nature of this study, when reporting results, we focus on the patterns revealed in the statistical analysis; statistical significance of these patterns is reported for key comparisons. The predicted f0 trajectories at the community level and for four example talkers are shown in Figure 1. At the population level, although the relative height between the tones follows the patterns

[4] The model was fitted with the `mgcv` package, using the following specifications: `bam(f0 ∼ Tone + Onset + Syllable type + Preceding tone + Following tone + Position in utterance + s(Character, bs = "re") + s(Position in word, bs = "re") + s(Ref point, Gender, k = 10, bs = "fs", m = 2) + s(Ref point, Talker*Tone, k = 10, bs = "fs", m = 2) + s(Ref point, by = Tone, k = 10, bs = "tp", m = 2), method = "fREML", discrete = TRUE, family = scat(link = "identity"))`.

[5] The following formula was used to fit the model: `inla(log(Duration) ∼ 1 + Tone + Syllable type + Preceding tone present + Following tone present + Position in utterance + Speech rate + f(Character, model = "iid", hyper = list(prec = list(prior = "pc.prec", param = c(1, 0.01)))) + f(Position in word, model = "iid", hyper = list(prec = list(prior = "pc.prec", param = c(1, 0.01)))) + f(Talker, model = "iid", hyper = list(prec = list(prior = "pc.prec", param = c(1, 0.01)))) + f(Talker, Tone, model = "iid", hyper = list(prec = list(prior = "pc.prec", param = c(1, 0.01)))), family = "gaussian", control.fixed = list(mean = 0, prec = 1/0.5^2, mean.intercept = 5, prec.intercept = 1/0.5^2))`.

of the citation forms, the overlapping confidence intervals for T2/T5 and T3/T4/T6 indicate that the difference between these tones might not be statistically meaningful. Figure 2 depicts the difference in f0 trajectories for the three merging pairs. For all three pairs, the 95% confidence interval spans zero, suggesting that, in terms of overall f0 contour, the difference within each pair is not significantly different. Overlapping confidence intervals are also a telltale sign that there might be substantial individual variation with respect to how talkers organize the tone space. Indeed, predicted contours at the individual level reveal that talkers exemplify different (non-)merger patterns, suggesting heterogeneity of tones in the speech community. For instance, while Talker M34A has well-separated trajectories among all the tones, Talker M21B has overlap in the latter half of T2 and T2, whereas Talker F21D and Talker M25B show T3-T6 and T4-T6 pairs that are closely tracing each other, respectively.

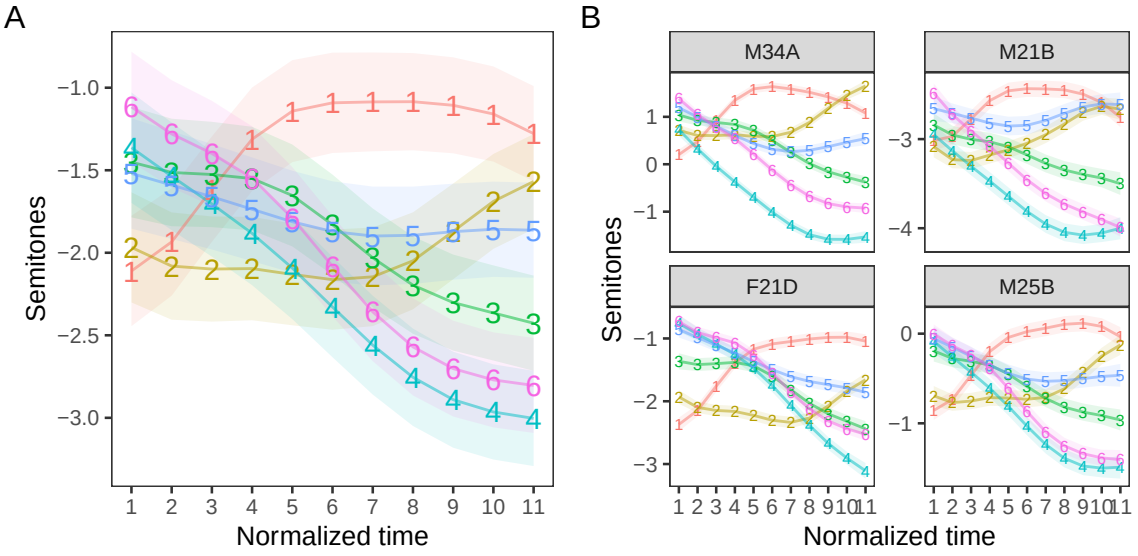

Figure 1: Predicted mean f0 contours for the six Cantonese lexical tones at (**A**) the population level and (**B**) the individual level for four talkers that typify no merger (M34A), the T2-T5 merger (M21B), the T3-T6 merger (F21D), and the T4-T6 merger (M25B). The shaded areas represent the 95% confidence interval.

Figure 3 depicts the predicted syllable duration at the community level and for the same four talkers. Results at both levels suggest that some merging pairs (i.e., T2-T5, T3-T6, T4-T6) might still retain reliable differences with respect to duration. For instance, it appears that there is a robust difference in duration between T2 and T5 (mean $= -0.117$, 95% credible interval [CrI] $= [-0.165, -0.069]$), as well as between T3 and T6 (mean $= 0.051$, 95% CrI $= [0.006, 0.097]$). The difference between T4 and T6, however, seems minimal (mean $= 0.013$, 95% CrI $= [-0.033, 0.059]$). Note, however, that T4 is described as having non-modal phonation, though listeners only appear to use the nonmodal phonation to cue a T4 category in low f0 ranges (Zhang and Kirby 2020). Moreover, these patterns are consistent at both the community and individual levels.

Overall, the combined results of f0 and duration paint a picture where the neutralization between reported merging tone pair appears to be incomplete in natural conversation.

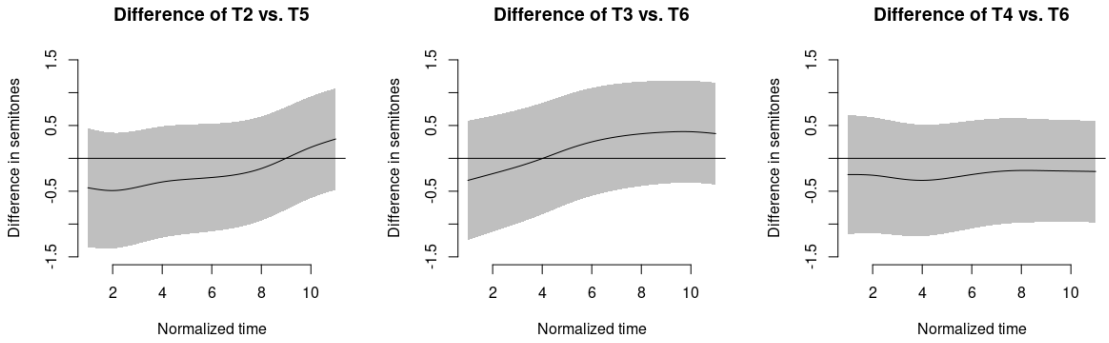

Figure 2: Difference between the two (non-linear) smooths comparing (left) T2 and T5, (middle) T3 and T6, and (right) T4 and T6. Shaded bands show the pointwise 95% confidence interval. All shaded bands encompass zero (represented by the horizontal line $y = 0$), indicating that the f0 trajectories within each merging pair are not significantly different.

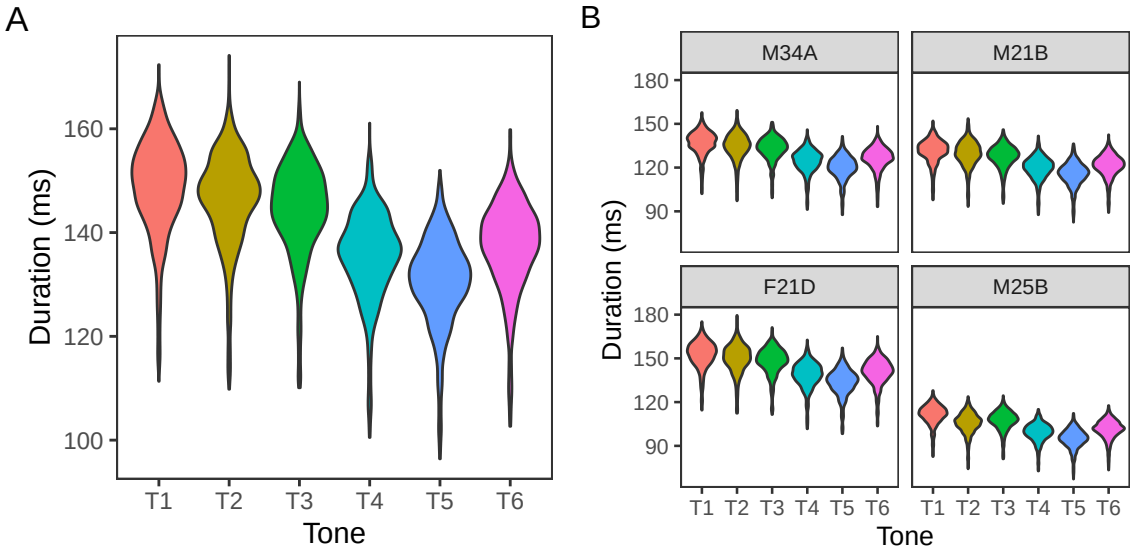

Figure 3: Predicted mean syllable durations for the six Cantonese lexical tones at (**A**) the population level and (**B**) the individual level for the same four talkers as in Figure 1.

3 Discussion

The goal of this project was to see what tone patterns emerge in Cantonese spontaneous speech when one considers both f0 trajectories and duration, as previous work been constrained either in terms of speech type or analytic methods or both. Our approach provided novel insight. When we assess the tone patterns in Cantonese from spontaneous speech and give both the full f0 trajectory and the duration of the tone-bearing unit opportunity, we are presented with a complex picture of the Cantonese tone space. While at the community-level, the data suggest tones T2-T5, T3-T6, and T4-T6 are merged in terms of f0 trajectories, the duration data indicate that T2-T5 and T3-T6 are reliably differentiated by duration. T5 is shorter in duration than T2, and T6 is shorter than T3.

As noted in the introduction, our results are not the first to show differences in the duration of Cantonese tones in speech production (e.g., Fok Chan 1974). Ciocca, Francis, Aisha, and Wong (2002), for example, observes that T3 (mid-level) and T2 (high rising) exhibit longer durations in single word production, and these are the items used as stimuli in their tone discrimination experiment. While there as a trend in their data towards better performance on trials where paired tones had large duration differences, listeners—kids aged 4-9 years with cochlear implants—did not seem to generally capitalize on that this difference. While T4-T6 is not differentiated by duration, it is possible that these tones are distinguished by phonation quality, which we will examine in future acoustic work.

Again, duration differences have been previously observed in descriptions of Hong Kong Cantonese, but studies have yet to find evidence that listeners actively use the duration information. Perhaps methodologies present a challenge. Previous work has used single items, which do not provide opportunity for listeners to normalize for talker-specific or local speech rate. For example, Francis, Ciocca, Wong, Leung, and Chu (2006) find that listeners use f0 information from before and after the target word to normalize and inform the tone category they report hearing, and their experiment 3 shows it needs to be speech material, as humming was ineffective. This highlights the need for appropriate context to normalize and calibrate for f0. Similarly, a study that uses continuous speech with target items that vary in duration may indeed find that listeners do use duration in parsing Cantonese. It might also be necessary to consider how duration affects word recognition, as opposed to only lower level speech categorization behaviour.

While the three tone mergers are observed at the community level in the f0 space, there exist individuals who produce reliable distinctions between the merging pairs in f0. This cross-talker variation in tone patterns does not appear to be a novel feature of Cantonese (Bauer et al. 2003). But, it presents an interesting situation for how a Cantonese language learner will infer a tone inventory. Is the tone grammar that emerges based on what the listener has observed in their broader speech community? Will the grammar be seeded and structured around the tone patterns produced by an infant's caregivers? Regardless of how the cross-talker heterogeneity is established, how do the processes of spoken word recognition and larger language comprehension manage the phonetic and phonological variation? Ultimately, we leave this squib with arguably more questions than answers.

4 Conclusion

By approaching Cantonese tones with an eye towards what emerges from typical language use, we confirm previous descriptions of Cantonese and provide novel insight. In terms of confirming previous descriptions, we, indeed, find evidence of T3-T6, T2-T5, and T6-T4 tone mergers at the

community level and we find variation in the existence and extent of the mergers across individuals when quantifying f0 trajectories. When measuring duration, however, we find evidence of consistent duration differences between two of the merging pairs—T3-T6 and T2-T5—at both the community and individual levels. In future work, we will examine the role of non-modal phonation in Cantonese tones, and, crucially, design speech categorization and recognition experiments that query how listeners encode the duration differences that ultimately appear in their production patterns to better understand the tone grammar that emerges and is created from the phonetic signal.

References

Archangeli, Diana, and Douglas Pulleyblank. 2022. *Emergent phonology*. Number 7 in Conceptual Foundations of Language Science. Berlin: Language Science Press.

Bauer, Robert S., and Paul K. Benedict. 1997. *Modern Cantonese phonology*. New York, NY: Mouton de Gruyter.

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

Chao, Yuen Ren. 1947. *Cantonese primer*. Cambridge, MA: Harvard University Press.

Ciocca, Valter, Alexander L Francis, Rani Aisha, and Lena Wong. 2002. The perception of cantonese lexical tones by early-deafened cochlear implantees. *The Journal of the Acoustical Society of America* 111:2250–2256.

Fok Chan, Yuen-Yuen. 1974. *A perceptual study of tones in Cantonese*. Hong Kong: Centre of Asian Studies, University of Hong Kong.

Francis, Alexander L., Valter Ciocca, Lian Ma, and Kimberly Fenn. 2008. Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics* 36:268–294.

Francis, Alexander L., Valter Ciocca, Natalie King Yu Wong, Wilson Ho Yin Leung, and Phoebe Cheuk Yan Chu. 2006. Extrinsic context affects perceptual normalization of lexical tone. *The Journal of the Acoustical Society of America* 119:1712–1726.

Fung, Roxana S. Y., and Chris K. C. Lee. 2019. Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception. *Journal of the Acoustical Society of America* 146:EL424–EL430.

Gahl, Susanne. 2008. *Time* and *thyme* are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84:474–496.

Gahl, Susanne. 2009. Homophone duration in spontaneous speech: A mixed-effects model. *UC Berkeley PhonLab Annual Report* 5.

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

Johnson, Khia A. 2021. SpiCE: Speech in Cantonese and English.

Kawahara, Hideki, Alain de Cheveigné, and Roy D. Patterson. 1998. An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: Revised TEMPO in the

STRAIGHT suite. In *The proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98)*.

Kej, Joseph, Veronica Smyth, Lydia K.H. So, C.C. Lau, and Ken Capell. 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, and Valter Ciocca. 2007. Perceptual correlates of Cantonese tones. *Journal of Phonetics* 35:104–117.

Lam, Wai Man. 2018. Perception of lexical tones by homeland and heritage speakers of Cantonese. Doctoral Dissertation, University of British Columbia, Vancouver, BC.

Lee, Kathy Y. S., Kit T. Y. Chan, Joffee H. S. Lam, C. A. van Hasselt, and Michael C. F. Tong. 2015. Lexical tone perception in native speakers of Cantonese. *International Journal of Speech-Language Pathology* 17:53–62.

Lohmann, Arne. 2018. *Time* and *thyme* are not homophones: A closer look at gahl's work on the lemma-frequency effect, including a reanalysis. *Language* 94:e180–e190.

Matthews, Stephen, and Virginia Yip. 2011. *Cantonese: A comprehensive grammar*. London: Routledge, 2nd edition.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of INTERSPEECH 2017*, 498–502.

Mok, Peggy P. K., Donghui Zuo, and Peggy W. Y. Wong. 2013. Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change* 25:341–370.

Nagy, Naomi, Holman Tse, and James N. Stanford. 2024. Have Cantonese tones merged in spontaneous speech? In *The phonetics and phonology of heritage languages*, ed. Rajiv Rao, 302–320. Cambridge: Cambridge University Press.

R Core Team. 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rue, Håvard, and Sara Martino. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72:319–392.

Shue, Yen-Liang, Patricia Keating, Chad Vicenik, and Kristine Yu. 2011. VoiceSauce: A program for voice analysis. In *The proceedings of the ICPhS XVII 2011*, 1846–1849.

Soo, Rachel, and Molly Babel. 2023. Perceptual effects of lexical competition on Cantonese tone categories. *Laboratory Phonology* 14.

Talkin, David. 2014. REAPER: Robust Epoch And Pitch EstimatoR.

Wood, Simon N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73:3–36.

Yiu, Carine Yuk-man. 2009. A preliminary study on the change of rising tones in Hong Kong

Cantonese: An experimental study. *Language and Linguistics* 10:269–291.

Zhang, Yubin, and James Kirby. 2020. The role of F_0 and phonation cues in Cantonese low tone perception. *The Journal of the Acoustical Society of America* 148:EL40–EL45.