

Phonetic Accommodation of Cantonese Merging Tones towards AI and Human Talkers

Peggy Pik Ki Mok¹, Yitian Hong¹, Grace Wenling Cao²

¹The Chinese University of Hong Kong, Hong Kong SAR, ²University College Dublin, Ireland

Hong Kong Cantonese (HKC) has a complex tone system. Two tone pairs, the high-rising (T2) vs. low-rising (T5) tones; the mid-level (T3) vs. low-level (T6) tones, are found to be at an initial stage of tone merging [1, 2]. Based on Ohala's theory of listeners being the origin of sound change [3, 4], if HKC listeners perceive a non-target tone realization, they may choose to normalize it or adopt it, which in turn, may affect their own tone production. This perception-production loop can further affect the trajectory of tone change in HKC. Apart from listeners' motivation and sensitivity, perceptual input also significantly influences this listeners' strategies of dealing with the non-target realization. Nowadays, perceptual input can come not only from human talkers but also from Artificial Intelligence (AI) talkers. Will there be differences between these two sources? Moreover, since the accuracy of AI pronunciation is still variable, does the degree of target-like realization affect our predictions?

This study investigates the short-term phonetic accommodation as a synchronic way of understanding tone merging of HKC. In particular, whether and how different perceptual sources (AI vs. human) and speaker's realization of the tones (intermediate vs. opposite realization) affect HKC speakers' accommodation patterns. Eleven monosyllabic minimal pairs were created for each merging tone pair. Among them, four used the opposite tone in the merging pair as priming targets and four used an intermediate realization between the two tones. The remaining three were filler pairs (Table 1). All the stimuli were played to the participants in human voice and AI voice, accompanied by either an image of a human or an AI character. Two native Cantonese-speaking research assistants (RAs) recoded the human voice stimuli. The AI voice stimuli were generated from the RAs' recordings using two customized voice models based on the duration-informed TTS model [5]. Thus, the RAs' voice quality and speech style were preserved in the AI voice with some synthetic characteristics.

Thirty-two native HKC speakers (16F, 16M; mean age = 21.5 ± 3.8) participated in this study which included a shadowing task [6]. Participants were randomly assigned to interact with either an AI face or a human face. Each participant completed five *sessions* with the same face: baseline recording, shadowing with both AI voice and human voice, post-shadowing after each voice. Their production was segmented manually and the STRAIGHT algorithm implemented in VoiceSauce [7] was used to extract the normalized-timepoint F0 of each token. Log z-score transformation was applied. Generalized Additive Mixed Models [8] were fitted to compare the F0 contours between the two merging tones within a pair, and to examine their interaction with *sessions* and *tone realization*. The smooth curves drawn from the model were presented in Figure 1 with significant timepoint windows overlaid on the curves. As shown in Figure 1, in the baseline condition, the two tones within each tone pair remained significantly distinct. When shadowing the intermediate realizations, the contours of the two merging tones became closer to each other, suggesting convergence. When shadowing the opposite realizations, the two tones either *merged* (the red box in Fig.1) or followed the opposite realization and *reversed* their own tone production (the blue box in Fig.1). The similar patterns observed between the AI voice and human voice suggest that AI voice can elicit similar phonetic accommodation patterns of tones as human voice does, conditioned by the degree of target-like realization. However, the *reversed* accommodation pattern is observed in AI face condition for T2-T5 pair while human face elicited more *merged* accommodation patterns. This suggests that visual cues (face difference) may be more robust to trigger the perception of difference between AI and human talkers. Moreover, the *merged* pattern is a more conscious way of accommodation, compared to directly imitating the tone of the model (*reversed* pattern), suggesting more social interest in human interaction. This may potentially give rise to an intermediate tone category, which is less likely to occur when interacting with an AI character.

Table 1. Examples of T2-T5 merging pair stimuli

Tone pair	Original tone	Primed tone	Example (→ means ‘read as’)	No. of Stimuli
T2 – T5	T2	Opposite realization (T5)	柚[jau2] → [jau5]	4
		Intermediate realization	虎[fu2] → intermediate	4
		Filler (T2)	屎[si2] → [si2]	3
	T5	Opposite realization (T2)	誘[jau5] → [jau2]	4
		Intermediate realization	婦[fu5] → intermediate	4
		Filler (T5)	市[si5] → [si5]	3

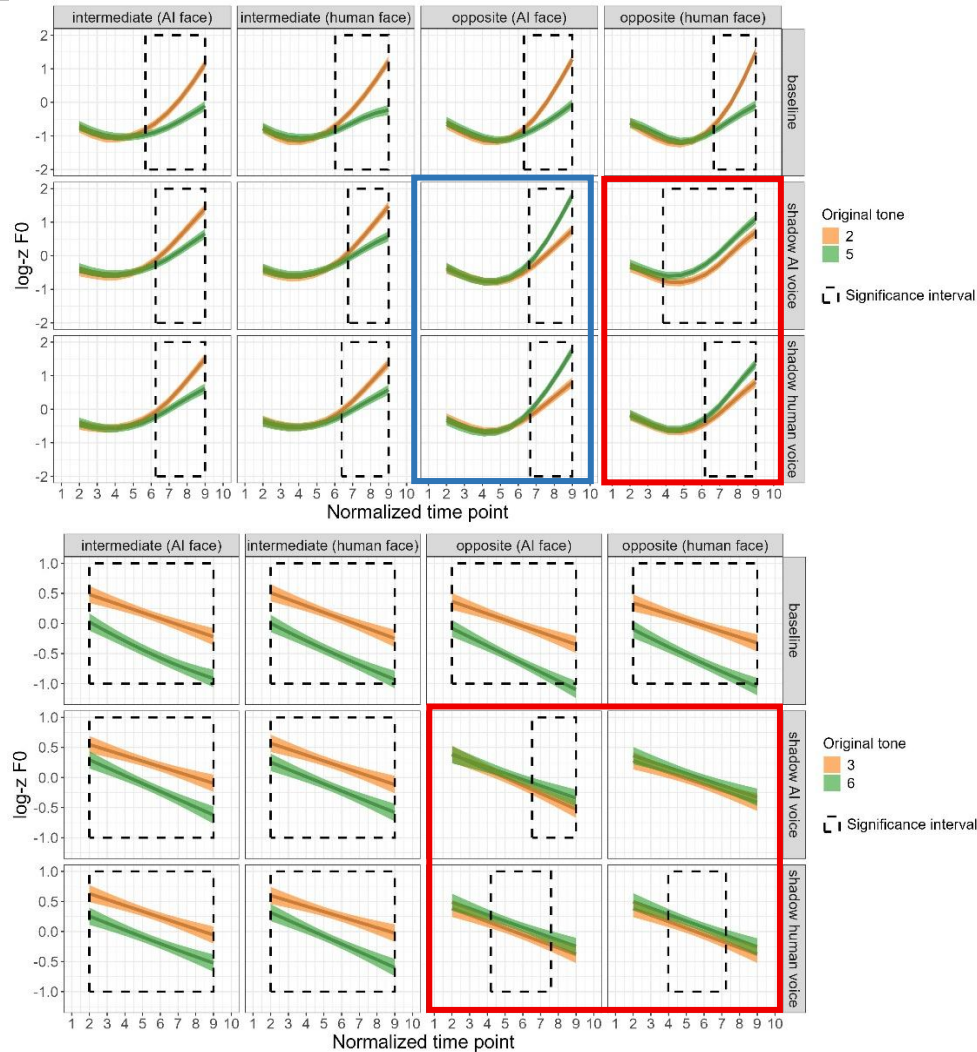


Fig. 1. Smoothed curves of two merging tone pairs (upper: T2-T5; lower: T3-T6).

References

- [1] X. Ren and P. Mok, “The effects of contextual tonal variation on Cantonese tone merging,” *Laboratory Phonology*, vol. 16, no. 1, Oct. 2025. [2] P. P. K. Mok, D. Zuo, and P. W. Y. Wong, “Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese,” *Language Variation and Change*, vol. 25, no. 3, pp. 341–370, Oct. 2013. [3] J. Ohala J., “The listener as a source of sound change,” in *Papers from the Parasession on Language and Behavior*, C. S. Masek, R. A. Hendrick, M. F. Miller (eds.), Chicago: Chicago Linguistic Society, 1981, pp. 178–203. [4] J. J. Ohala, “The listener as a source of sound change: An update,” in *The Initiation of Sound Change*, John Benjamins Publishing Company, 2012, pp. 21–36. Accessed: Apr. 29, 2025. [Online]. [5] C. Yu *et al.*, “DurlAN: Duration Informed Attention Network for Speech Synthesis,” in *Interspeech 2020*, ISCA, Oct. 2020, pp. 2027–2031. doi: 10.21437/Interspeech.2020-2968. [6] S.D.Goldinger, “Echoes of echoes? An episodic theory of lexical access,” *Psychological Review*, vol. 105, no. 2, pp. 251–279, 1998. [7] Y.-L. Shue, P. Keating, and C. Vicenik, “VOICESAUCE : A program for voice analysis,” *The Journal of the Acoustical Society of America*, vol. 126, no. 4_Supplement, pp. 2221–2221, Oct. 2009. [8] M. Wieling, “Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English,” *Journal of Phonetics*, vol. 70, pp. 86–116, Sep. 2018.