

Phonetic convergence toward human and AI interlocutors in three English varieties by Hong Kong English speakers

Xinyi Chen¹, Grace Wenling Cao², Peggy Pik Ki Mok¹

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

Phonetic accommodation reflects speakers' adaptive adjustment toward an interlocutor and varies across linguistic features and target varieties [1]. This phenomenon is well documented in human-human interaction [2] and has attracted increasing attention in human-machine interaction [3,4], but remains underexplored in second language (L2) speech [5]. This study examines L2 phonetic accommodation in Hong Kong English (HKE) speakers, asking whether they converge toward socially prestigious English varieties, specifically Standard Southern British English (SSBE) and General American English (AmE) [6], than toward HKE, and whether convergence differs between human and AI interlocutors. We predict preferential, salience-conditioned convergence toward SSBE and AmE, and reduced accommodation to AI interlocutors relative to humans [3].

To test this hypothesis, following the design in [4], we employed an error-repair paradigm to elicit baseline and post-exposure productions for assessing convergence in interaction (Figure 1). Participants were 22 native Hong Kong Cantonese speakers (13F, 9M; age = 21 ± 2.67 years). We employed a within-subject design in which participants were exposed to three English varieties (SSBE, AmE, HKE). For each variety, they interacted with a human talker (human image and voice) and an AI agent (avatar image and AI voice). The AI voices were generated using copy synthesis from the corresponding human speakers, yielding similar-sounding voices with some synthetic quality.

Target features included nine English monophthongs (/i:, ɪ, u:, ʊ, æ, e, ɔ:, ɒ, ɑ:/), often merged or variably realized by HKE speakers. Postvocalic rhoticity and intervocalic /t/ contrasted SSBE and AmE (non-rhotic vs. rhotic; aspirated vs. flapped), allowing a more direct assessment of accommodation between varieties. The vowel analysis included 925 observations (≈ 42 tokens per participant), plus 350 rhoticity and 80 intervocalic /t/ tokens. Convergence was indexed as the reduction in acoustic distance to the model talker (baseline – post-exposure), computed from F1-F2 Euclidean distance for vowels [1], midpoint F3 differences for postvocalic rhoticity, and absolute duration differences for intervocalic /t/; with positive values indicating increased similarity after exposure.

Vowel convergence was analyzed with a linear mixed-effects model including LanguageVariety (HKE, AmE, SSBE) and SpeakerType (Human vs. AI) as fixed effects, and random intercepts for participants, items and vowels. A significant main effect of LanguageVariety showed greater convergence toward AmE ($\beta = 0.18$; $p = .0049$) and SSBE ($\beta = 0.17$; $p = .005$) than toward HKE, indicating asymmetric accommodation toward exonormative varieties, possibly driven by the more prestigious status associated with these varieties. Pairwise comparisons showed no significant difference between AmE and SSBE ($p = .62$). SpeakerType had no overall effect, though a marginal LanguageVariety × SpeakerType interaction in AmE suggested weaker convergence toward AI than human talkers (Figure 2). Rhoticity and intervocalic /t/ were modelled separately using language-specific (mixed-) effects models, reflecting variety-specific phonological patterns. Rhoticity convergence was significant in AmE, with significantly lower convergence toward the AI talker (Table 1), consistent with weaker adaptation to device interlocutors that was found in other studies [3]. Intervocalic /t/ in HKE showed stronger convergence for flapping in AmE than aspiration in SSBE (Table 2), probably due to the greater perceptual salience of the flapped-t which is uncommon in HKE, with no significant SpeakerType effect. Overall, convergence was weaker toward AI interlocutors, with SpeakerType effects limited to specific features or in LanguageVariety × SpeakerType interactions, suggesting subtle, feature-dependent modulation in L2 speech. Further analysis is underway.

Table 1. Linear model of rhoticity convergence (F3 distance) in AmE.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.21	0.09	2.20	0.03 *
SpeakerTypeAI	-0.29	0.13	-2.15	0.03 *

Table 2. Mixed-effects model of intervocalic /t/ convergence (duration) in HKE.

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	0.02	0.003	55.34	4.95	<0.001 ***
SpeakerTypeAI	-0.01	0.004	57.00	-1.58	0.12
VariantHeardnoflap	-0.02	0.004	57.00	-4.14	<0.001 ***
SpeakerTypeAI:VariantHeardnoflap	0.01	0.006	57.00	1.17	0.25

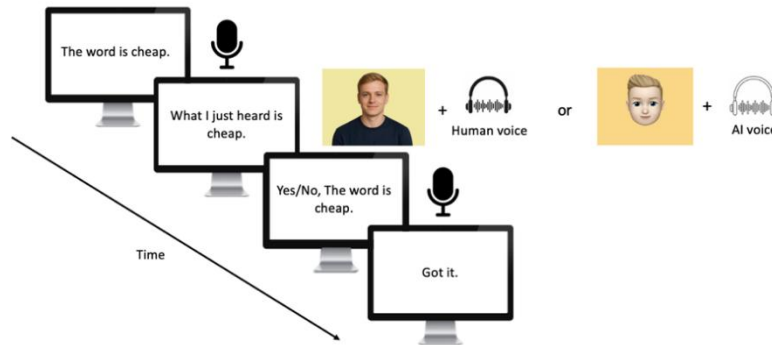


Fig. 1. Error-repair paradigm adapted from [4]. Participants produced a sentence containing the target word, heard the model talker’s reformulation, and then confirmed or corrected it, followed by brief feedback. The study employed a within-subject design: all participants were exposed to three English varieties (SSBE, AmE, HKE), and for each variety interacted with both a human talker and an AI agent (SSBE example shown).

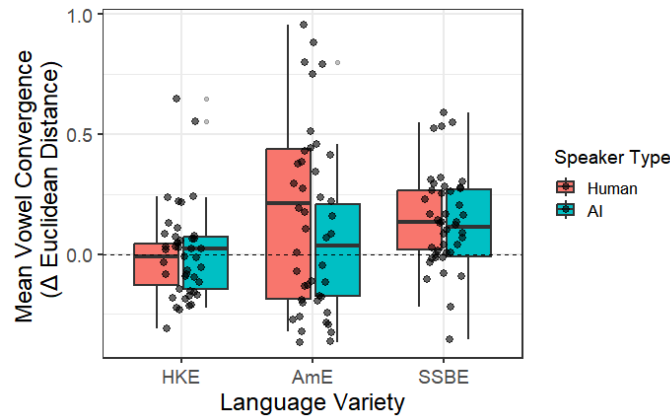


Fig. 2. Mean vowel convergence toward the model talker by Language Variety (HKE, AmE, SSBE) and Speaker Type (human vs. AI). Convergence is measured as change in Euclidean distance in F1–F2 space; positive values indicate increased similarity.

References

- [1] Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, 40(1), 177–189.
- [2] Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, 79(2), 637–659.
- [3] Cohn, M., Segedin, B. F., & Zellou, G. (2019). Imitating Siri: Socially-mediated vocal alignment to device and human voices. In *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)* (Vol. 17). Australasian Speech Science and Technology Association Inc Canberra, Australia.
- [4] Cohn, M., Segedin, B. F., & Zellou, G. (2022). Acoustic-phonetic properties of Siri- and human-directed speech. *Journal of Phonetics*, 90, 101123.
- [5] Gessinger, I., Möbius, B., Andreeva, B., Raveh, E., & Steiner, I. (2020). Phonetic Accommodation of L2 German Speakers to the Virtual Language Learning Tutor Mirabella. *Interspeech 2020*, 4118–4122.
- [6] Hansen Edwards, J. G. (2016). Accent preferences and the use of American English features in Hong Kong: A preliminary study. *Asian Englishes*, 18(3), 197–215.