The impact of initial phonetic distance on L2 speech accommodation in Human- and AI-directed speech

Manson Chun Man Wong [a], Xinyi Chen [a], Yitian Hong [a], Yusheng Tian [a], Grace Wenling Cao [b], Peggy Pik Ki Mok [a]
[a] *The Chinese University of Hong Kong, Hong Kong,*
[b] *University College Dublin, Ireland*

Keywords — phonetic accommodation, speech register, AI vs. Human, second language learning

I. INTRODUCTION

Phonetic accommodation refers to speakers' tendency to modify their speech patterns during interaction, resulting in convergence (becoming more similar to an interlocutor), divergence (becoming more dissimilar), or maintenance (showing no change). Previous studies have shown that convergence is more likely to occur when there is a substantial phonetic difference between a speaker's variety and that of a native model talker speaking another variety of the same language (e.g., [1]). Also, phonetic accommodation has been theorized as a socially mediated behaviour influenced by sociolinguistic factors, such as gender and perceived attractiveness, and other social evaluation of the interlocutor [2]. However, much of the existing studies tended to view speakers within a speech community as a relatively homogeneous group, and comparatively few studies have examined individual variation (e.g., [3], [4]). Overlooking such variability may neglect the effect of phonetic distance on accommodation within the same language variety. As a result, how fine-grained differences in initial phonetic distance shape accommodation patterns is yet to be thoroughly explored. L2 speech, thus, provides an ideal testing ground for investigating the multifaceted nature of phonetic accommodation among non-native speakers, given its high variability.

In Hong Kong, L1 Hong Kong Cantonese (HKC) speakers exhibit a wide spectrum of English pronunciation patterns, ranging from near-native varieties (e.g., British or American English) to localized forms of Hong Kong English [4]. While prior work has attempted to describe English pronunciations in Hong Kong systematically, they consistently noted the considerable variability in speech sound patterns (e.g., [5]). Therefore, treating phonetic distance as a continuous acoustic measure allows for a more precise examination of how initial differences between speakers mediate subsequent accommodation.

In addition to human-directed speech, contemporary language experiences increasingly involve interaction with artificial intelligence (AI) speech systems. While AI voices can be generated to closely match human speech acoustically, they differ fundamentally in social agency. This contrast offers a unique opportunity to examine whether phonetic accommodation is driven by phonetic distance and whether the use of AI technology plays a role. Previous work focused on speech modulation when interacting with AI voices, but little has explored if accommodation occurs (e.g., [6]). Using an error-repair conversational paradigm, the present study examines phonetic accommodation in English vowel produced by L1 HKC speakers. Specifically, this study asks: 1. Is the type of accommodation mediated by the initial phonetic distance between the L2 speaker and the model talker of L1 English? 2. Does accommodation differ between interactions with a human and an acoustically matched AI interlocutor?

II. METHOD

Thirty L1-HKC who speak English as a second language (12M, 18F; age: 21.60±2.62) participated in this study. All are born and raised in Hong Kong, and none has parents who are native English speakers. Age of onset of English learning ranged from one to six years (2.83±1.15). Their IELTS or equivalent score ranges from approximately 6.0 to 8.0 (7.29±0.59). The experiment employed a between-subjects design, with interlocutor type as the factor. Participants were assigned to either a human (human image and human voice) or AI condition (avatar image and AI voice). All participants completed three phases: 1. Pre-immersion dialogue, 2. Immersion using an error-repair paradigm, 3. Post-immersion dialogue. The auditory stimuli were based on the speech of a male speaker of Standard Southern British English. Crucially, in the AI condition, the voice was generated using the same speaker's recordings to match tonal and prosodic characteristics via CosyVoice 2, a zero-shot text-to-speech system [7]. The F0, F1, F2, and F3 values of the vowels did not significantly differ between the voices. While the synthesized voice preserved the prosodic contour of the human speaker, listeners may perceive differences in voice quality attributable to the speech synthesis process.

In both the pre- and post-immersion dialogues, participants engaged in scripted conversations with human or AI speaker. They were given one minute to familiarize themselves with each script prior to interaction. Three target words which were selected to elicit the corner vowels /i, ɑ, u/ respectively appeared three times in each dialogue. Corner vowels were chosen to enable future cross-variety comparisons with this project's L1 speech data. During the immersion phase, participants completed an error-repair conversational task. In each trial, a monosyllabic target word was embedded in a carrier sentence. The interlocutor (human or AI) then responded by repeating the word correctly or with an onset error. Participants were instructed to either confirm or explicitly correct the interlocutor's production. Each interaction ended with a brief closing utterance from the interlocutor. This task was designed to increase exposure to speech produced by the L1-English model speaker. By explicitly directing attention to consonantal features, the task allowed us to evaluate whether vowel accommodation occurs implicitly through phoneme interaction [8].

F1 and F2 values were extracted at the midpoint of all target vowels and normalized using the ΔF method [9]. Acoustic distance between each participant's corner vowel and the corresponding vowel produced by the model talker was calculated using Euclidean distance in the F1-F2 space and subsequently standardized for regression analysis. Phonetic accommodation was operationalized as the change in absolute acoustic distance relative to the L1-English model talker from pre-immersion to post-immersion. As this operationalization inevitably overlooks the difference between maintenance and over-convergence [10], this study excluded potentially affected datapoints through visual screening to prevent miscategorization. Only the /ɑ/ and /u/ tokens from one participant were excluded (6 out of 237). Accommodation was assessed using a linear mixed-effects model with post-immersion acoustic distance as the dependent variable. Fixed effects included pre-immersion acoustic distance, interlocutor type (human vs. AI), while also taking trial number (repetition effects), gender, and vowel category as control. Participant was included as a random intercept.

III. Results and Discussion

To evaluate the overall patterns of phonetic accommodation, the relationship between the pre-immersion and the post-immersion acoustic distance was first examined. The model revealed a significant positive slope ($m = 0.58$, $SE = 0.05$, $t(208) = 10.29$, $p < .0001$). Importantly, this slope was also significantly lower than the maintenance line of 1 ($m = 0.58$, $SE = 0.05$, $t(208) = -7.52$, $p < .0001$). This finding reflects a degree of consistency in individual speakers' vowel production before and after immersion, while simultaneously indicating that phonetic accommodation occurred, as illustrated in Figure 1. Also, the predicted regression line is not parallel to the maintenance line (y = x, shown as a black dashed line), thereby creating a convergence threshold (Human: 0.30; AI: -0.33). For participants whose pre-immersion acoustic distance exceeded this threshold, the model predicts a reduction in distance in post-immersion task (convergence). Conversely, when the pre-immersion acoustic distance is smaller than this threshold, the model predicts a shift away from the model talker (divergence). This pattern suggests when speakers are already highly similar, further convergence may be inhibited. These findings align with previous research showing that phonetic convergence is more likely to occur when an initially large acoustic discrepancy exists between the speaker and the model talker (e.g., [3]).

In addition, the model revealed a significant main effect of interlocutor ($\beta = 0.27$, $SE = 0.10$, $t(29.94) = 2.65$, $p = .01$), suggesting that participants' productions were overall closer to the AI interlocutor than to the human following the immersion phase. This indicates that participants were more influenced by AI interlocutor, mirroring its production more closely. Furthermore, the interlocutor type also significantly affected the convergence threshold. Due to the lower intercept for the AI interlocutor, the threshold was shifted toward the negative end of the scale compared to the human interlocutor. In practical terms, participants required a substantially smaller initial acoustic distance to initiate convergence toward the AI speaker, supported by bootstrapping showing a significant gap between two thresholds ($\bar{x} = 0.63$, 95% *CI* [0.20, 1.09]. This means that at a certain range of distance, speakers were more likely to diverge from human while they would have already begun converging toward the AI interlocutor.

Taken together, these results demonstrate that phonetic accommodation is mediated by the initial acoustic distance between the speaker and the model talker, resulting in convergence or divergence. Moreover, the observed interlocutor effect suggests that speakers may engage differently with human and AI interlocutors. One possible interpretation is that AI interlocutors may reduce socially evaluative pressures, thereby facilitating greater phonetic adjustment. This highlights the importance of considering both acoustic and social factors in phonetic accommodation, particularly in the era of increasing human-AI speech interaction.

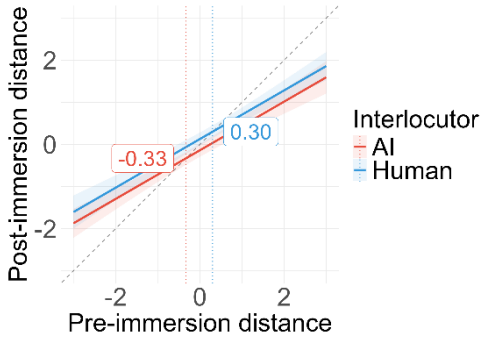

Fig. 1. Comparison of phonetic accommodation towards AI and Human interlocutors.

References

[1] A. Walker and K. Campbell-Kibler, "Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task," *Front. Psychol.*, vol. 6, May 2015.

[2] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *J. Phon.*, vol. 40, no. 1, pp. 177–189, Jan. 2012.

[3] G. W. Cao, "Phonetic Dissimilarity and L2 Category Formation in L2 Accommodation," *Lang. Speech*, vol. 67, no. 2, pp. 301–345, Jun. 2024.

[4] N. Lewandowski and M. Jilka, "Phonetic Convergence, Language Talent, Personality and Attention," *Front. Commun.*, vol. 4, May 2019.

[5] D. Deterding, J. Wong, and A. Kirkpatrick, "The pronunciation of Hong Kong English," *Engl. World-Wide*, vol. 29, no. 2, pp. 148–175, Jan. 2008.

[6] L. Bradshaw, V. Perepelytsia, and V. Dellwo, "Vocal effort in human interactions with voice-AI," in *Proceedings of ICPhS*, 2023.

[7] Z. Du *et al.*, "CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models," Dec. 25, 2024.

[8] J. Padgett, "Consonant–Vowel Place Feature Interactions," in *The Blackwell Companion to Phonology*, John Wiley & Sons, Ltd, 2011, pp. 1–26.

[9] K. Johnson, "The ΔF method of vocal tract length normalization for vowels," *Lab. Phonol.*, vol. 11, no. 1, Jul. 2020.

[10] U. C. Priva and C. Sanker, "Limitations of difference-in-difference for measuring convergence," *Lab. Phonol.*, vol. 10, no. 1, Sep. 2019.