# Articulatory-acoustic dynamics in naïve listener imitation of Cantonese vowels

Jonathan Havenhill[1], Madeleine Oakley[2], and Ming Liu[1]

*[1]The University of Hong Kong, [2]North Carolina State University*

While theories of non-native phonological acquisition differ regarding the theoretical status of articulatory gestures vs. acoustic phonetic categories [1, 2], most work on non-native speech production is based on acoustic rather than articulatory evidence [but see 3, 4]. The present study is an articulatory and acoustic analysis of non-native Cantonese vowel production. We ask whether native English speakers recruit L1 acoustic or articulatory vowel targets in the implementation of non-native /y-u/ (and /œ-ɔ/) vowel contrasts, and whether such targets are static or dynamic.

Numerous studies have examined the perception and production of non-native /y-u/ contrasts by Anglophones, for whom the overlap of both vowels with English /u/ presents a challenge [5, 6]. In many English varieties, /u/ is acoustically fronted to [y] or [ʉ], although the articulatory configuration of fronted /u/ is potentially variable [7, 8]. Lip rounding and tongue backing have similar acoustic effects on F2, so English speakers may adopt either round [y] or unround [ɨ/ɯ] articulatory configurations for fronted English /u/ and thereby vary in their realization of non-native /y/. The situation is further complicated by the often-diphthongal phonetic quality of English /u/; realizations of /u/ as [iu] may overlap with both [y] and/or [u] during separate vowel phases [8]. The extent to which speakers will transfer dynamic articulatory-acoustic trajectories to non-native speech is unclear, especially as the articulatory basis for L1 vowel inherent spectral change has not been widely studied.

Participants were five native American English speakers (4F/1M) and five native Hong Kong Cantonese (HKC) speakers (4F/1M). English speakers, who had no prior HKC exposure, completed two perception (discrimination, categorization + goodness) and two production (non-native imitation, L1 production) tasks. HKC speakers completed an L1 production task only. Materials comprised 147 HKC words with the vowels /i, y, u, ɛ, œ, ɵ, ɔ, a, ɐ/ and 136 English words containing the full English monophthong + diphthong inventory. Words were monosyllabic and balanced for phonological context, including coronal and non-coronal onsets and codas. L1 items were presented orthographically while non-native HKC items were presented via auditory prompts produced by a native speaker and resynthesized to remove tone contours. Items were presented in pseudorandom order and repeated 3 times in isolation. Simultaneous ultrasound (81 fps), lip video (60 fps), and audio were recorded in AAA, with the transducer and lip cameras held in place by a stabilizing headset [9]. Recordings were force aligned with MFA [10] and dynamic acoustic measurement was performed using FastTrack [11]. Tongue and lip movement were tracked using DeepLabCut [12, 13].

Figure 1 shows representative results for three native English speakers. Acoustically, all participants produce English /u/ as a diphthong, with a high F2 onset and low F2 offglide. Non-native HKC /y/ and /u/, however, are monophthongal and differ significantly from one another in F2. F2 corresponds well to tongue backness; non-native HKC /y/ is fronter than English /u/, HKC /u/ backer than English /u/, and both HKC /y/ and /u/ show less tongue body movement than for English /u/. Lip rounding is variable across participants; some produce non-native HKC /y/ or /u/ with less rounding than English /u/ (VEN001), while others produce /y/ and /u/ with English /u/-like rounding (VEN005). For some speakers (VEN001 and VEN003), the acoustic (but not articulatory) targets for HKC /y/ and /u/ approximate the nucleus and offglide, respectively, of English /u/. More generally, however, the tongue position for HKC /y/ differs from both English front and back vowels, suggesting the formation of a new static central target for /y/. These novel data suggest that perceptual sensitivity to the differing dynamic qualities of HKC and English vowels may lead learners to form new phonetic categories for the HKC round vowels rather than transfer L1 /u/ targets wholesale. Static acoustic measurement alone obscures such patterns and is likely to misinform theories of L2 category formation.

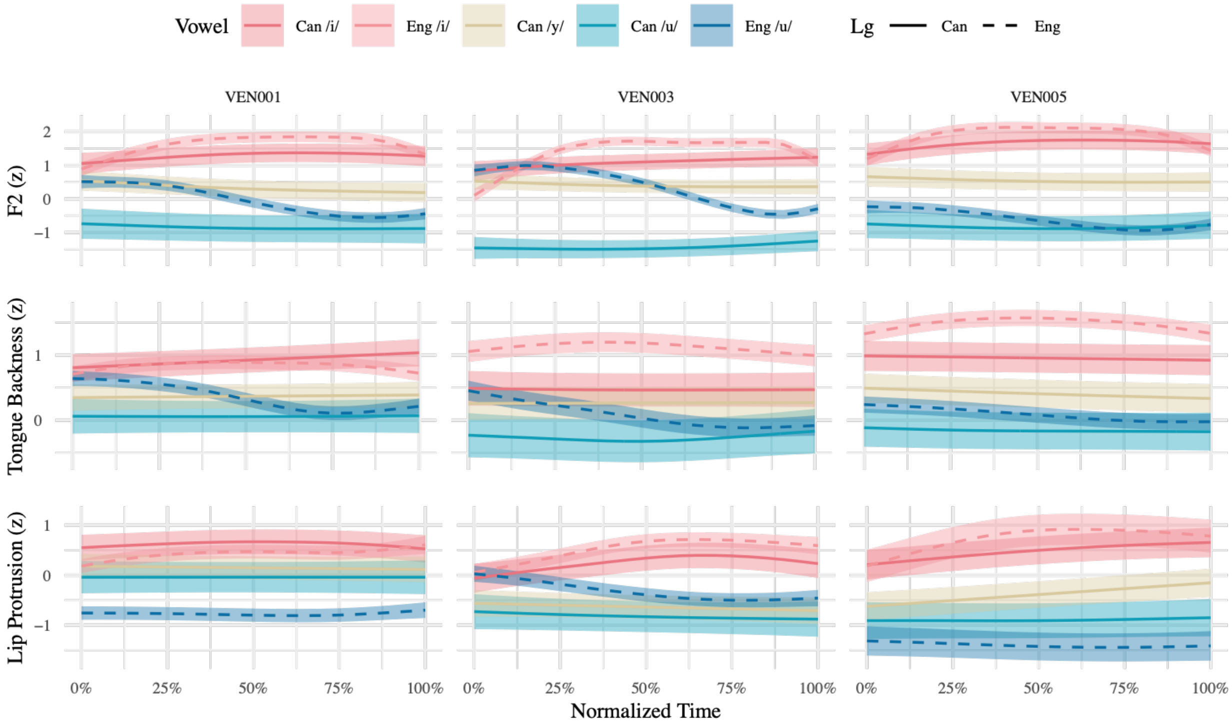

**Figure 1.** Fitted acoustic and articulatory GAMM smooths for native English speakers. Shading indicates 95% confidence interval; overlap indicates no significant difference. For lip protrusion, higher values indicate less rounding. For tongue backness, higher values indicate fronter position.

## References

[1] J. E. Flege and O.-S. Bohn. “The revised Speech Learning Model (SLM-r)”. In: *Second Language Speech Learning*. Ed. by R. Wayland. Cambridge: Cambridge University Press, 2021.

[2] C. T. Best and M. D. Tyler. “Nonnative and second-language speech perception: Commonalities and complementarities”. In: *Language experience in second language speech learning: In honor of James Emil Flege*. Ed. by O.-S. Bohn and M. J. Munro. Philadelphia, PA: John Benjamins, 2007, pp. 13–34.

[3] M. Oakley. “Articulating non-native vowel contrasts”. Georgetown University doctoral dissertation, 2021.

[4] M. Oakley, J. Mielke, and J. Matthews. “Transfer of articulatory targets in production of second language Korean sibilants.” Poster presentation at LabPhon19, Seoul, South Korea (= this conference, 2024).

[5] T. L. Gottfried. “Effects of consonant context on the perception of French vowels”. In: *Journal of Phonetics* 12.2 (1984), pp. 91–114.

[6] E. S. Levy and F. F. Law. “Production of French vowels by American-English learners of French: Language experience, consonantal context, and the perception-production relationship”. In: *The Journal of the Acoustical Society of America* 128.3 (2010), pp. 1290–1305.

[7] E. Lawson, J. Stuart-Smith, and L. Rodger. “A comparison of acoustic and articulatory parameters for the GOOSE vowel across British Isles Englishes”. In: *The Journal of the Acoustical Society of America* 146.6 (2019), pp. 4363–4381.

[8] J. Havenhill. “Articulatory and acoustic dynamics of fronted back vowels in American English”. In: *The Journal of the Acoustical Society of America* (in press).

[9] L. Spreafico, M. Pucher, and A. Matosova. “UltraFit: A speaker-friendly headset for ultrasound recordings in speech science”. In: *Proceedings of the 19th Conference of the International Speech Communication Association (Interspeech 2018)*. (2018), pp. 1517–1520.

[10] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. “Montreal Forced Aligner: Trainable text-speech alignment using Kaldi.” In: *Proceedings of the 18th Conference of the International Speech Communication Association (Interspeech 2017)*. (2017), pp. 498–502.

[11] S. Barreda. “Fast Track: Fast (nearly) automatic formant-tracking using Praat”. In: *Linguistics Vanguard* 7.1 (2021).

[12] A. Mathis et al. “DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning”. In: *Nature neuroscience* 21.9 (2018), pp. 1281–1289.

[13] A. Wrench and J. Balch-Tomes. “Beyond the Edge: Markerless pose estimation of speech articulators from ultrasound and camera images using DeepLabCut”. In: *Sensors* 22 (2022).