What is similarity? Approaches to the quantification to voice similarity

Suyuan Liu, Molly Babel, and Jian Zhu
Department of Linguistics, University of British Columbia (Canada)

Phonetic imitation or convergence is often defined as talkers or tokens becoming more "similar-sounding" as a result of auditory exposure [e.g., 2]. This presumes an understanding of what makes utterances more or less similar, in addition to having a method of quantifying that (dis)similarity. Listener judgments as the means of assessing phonetic imitation are the gold standard [11], as listeners are able to make global judgments that consider the multidimensional voice signal in a way that targeted acoustic measures (e.g., f0; [3]) and global acoustic measures (e.g., [1]) may not. One aspect of what makes similarity a complex issue is that talkers and tokens can exhibit *linguistic similarity* or *voice similarity*. With respect to linguistic similarity, this can be informed and operationalized by phonological theory the established acoustic-phonetic cues [10; 8]. Voice similarity is a more challenging concept to coherently wrangle, but we approach it with the psycho-acoustic model of voice [6], and understand the voice as a rich signal that delivers biological, physiological, psychological, social, and linguistic meaning [12]. Crucially, voices have structure that can be queried and compared [9; 5].

The goal of our paper is to better understand similarity by comparing similarity metrics across acoustic analysis, perceptual judgments by human listeners, and automatic speaker verification systems. We focus on spontaneous speech from the English portion of the Speech in Cantonese and English (SpiCE) corpus [4]. In our comparison of vocal similarity, we compare (i) similarity scores generated from 24 acoustic dimensions [7]; (ii) speaker verification scores generated by seven pretrained speaker verification models using Wespeaker [13]; (iii) perceptual similarity from human listeners in an AX discrimination task, and (iv) perceptual (dis)similarity from an independent group of human listeners in a rating task.

The output of our Bayesian regression models suggests that when controlling for the specific talkers being compared, the speaker verification models correlate with the psychoacoustic similarity scores, but not with either listener-based measure. When the pairs of voices being compared are not controlled, there is a relationship between listeners and speaker verification models. We take this to suggest that assessments of similarity manifest differently when the focus is on the gross- versus fine-phonetic levels. We discuss these results in the context of quantifying similarity for measuring phonetic imitation and understanding it as a linguistic process.

References

[1] Abel, J., and Babel, M. Cognitive load reduces perceived linguistic convergence between dyads. *Language and Speech 60*, 3 (2017), 479–502.

[2] Babel, M. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics 40*, 1 (2012), 177–189.

[3] Babel, M., and Bulatov, D. The role of fundamental frequency in phonetic accommodation. *Language and speech 55*, 2 (2012), 231–248.

[4] Johnson, K. A. SpiCE: Speech in Cantonese and English, 2021.

[5] Johnson, K. A., and Babel, M. The structure of acoustic voice variation in bilingual speech. *The Journal of the Acoustical Society of America 153*, 6 (2023), 3221–3221.

[6] Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., and Zhang, Z. Toward a unified theory of voice production and perception. *Loquens 1*, 1 (2014), e009.

[7] Kreiman, J., and Sidtis, D. *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Wiley-Blackwell, 2011.

[8] Kwon, H. The role of native phonology in spontaneous imitation: Evidence from seoul korean. *Laboratory Phonology 10*, 1 (2019).

[9] Lee, Y., Keating, P., and Kreiman, J. Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America 146*, 3 (2019), 1568–1579.

[10] Nielsen, K. Specificity and abstractness of vot imitation. *Journal of Phonetics 39*, 2 (2011), 132–142.

[11] Pardo, J. S., Urmanche, A., Wilman, S., and Wiener, J. Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics 79* (2017), 637–659.

[12] Podesva, R. J., and Callier, P. Voice quality and identity. *Annual Review of Applied Linguistics 35* (2015), 173–194.

[13] Wang, H., Liang, C., Wang, S., Chen, Z., Zhang, B., Xiang, X., Deng, Y., and Qian, Y. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023), IEEE, pp. 1–5.