Revisiting prosodic boundary-induced voicing categorization in Whisper across model size and L1/L2 English speech

Richard Hatcher, Jiyoung Jang
Hanyang Institute for Phonetics and Cognitive Sciences of Language, Hanyang University

Voice onset time (VOT), a primary cue to the English voiced–voiceless stop contrast (Lisker & Abramson, 1964), is interpreted relative to prosodic context. Domain-initial consonants are typically strengthened, including through increased VOT, particularly at intonational phrase (IP) onsets (Fourgeron & Keating, 1997). Human listeners use this regularity when judging stops, shifting voicing judgments according to preceding prosodic context (Kim & Cho, 2013). Recent work asks whether comparable context-sensitive normalization emerges in AI-based automatic speech recognition (ASR) systems trained on large-scale speech corpora. Jang & Hatcher (2026) tested whether Whisper exhibits human-like prosodic normalization during categorization of a /b-p/ VOT continuum. Unlike human listeners, however, Whisper showed little evidence of the expected boundary-induced category shift. In some conditions, the effect was even reversed, with stronger voiceless responses in the phrase-medial word-initial position than in the IP-initial position. It was suggested that Whisper may rely less on abstract prosodic structure and more on distributed acoustic information elsewhere in the speech signal (Jang & Hatcher, 2026).

The present study extends this previous work in two ways. First, we compare different variants of Whisper, specifically Whisper-small and Whisper-large-v3. Although both models share the same general architecture and training approach, they differ substantially in model size and representational capacity (Radford et al., 2023). Larger models may therefore show greater sensitivity to broader contextual dependencies or finer-grained acoustic regularities, yielding different patterns of sensitivity to prosodically conditioned phonetic variation. Second, we examine both native English speech and L2 English speech produced by a Korean speaker. Because Korean-accented L2 English may differ from L1 English in the phonetic realization of both VOT and prosodic context, this comparison provides a first test of whether Whisper's categorization patterns are stable across phonetically different data sets. By comparing categorization behavior across model size and speaker condition, the present study aims to clarify the sources of contextual information that contemporary ASR systems use when interpreting ambiguous phonetic cues.

Stimuli consisted of carrier sentences containing word-initial English stops along a voiced–voiceless VOT continuum, following Jang & Hatcher (2026). Target stops from voiced and voiceless source tokens are manipulated across 15 VOT steps and embedded in two prosodic contexts: IP-initial position and phrase-medial word-initial position, hereafter IP and Wd conditions, respectively. Stimuli were recorded by a male native English speaker and a female Korean L2 speaker of English. Because the L1 and L2 tokens differ in speaker and sex as well as language background, Speaker effects should be interpreted as a first-pass comparison rather than an isolated L1/L2 contrast. Two variants of Whisper, Whisper-small and Whisper-large-v3, received each stimulus as input, and the resulting transcription outputs were analyzed to determine how each model categorized the target stop across VOT steps and experimental conditions. Statistical analyses using generalized linear models implemented in R (R Core Team, 2025) examine whether categorization patterns vary as a function of VOT, Boundary, Speaker, Model, and Source Voicing.

Scaling from Whisper-small to Whisper-large-v3 did not shift categorization toward the human prosodic-boundary effect: the VOT × Boundary × Model interaction was not significant (β = -0.011, SE = 0.007, p = .124). Across models, Whisper showed little evidence of the prosodic boundary effect observed in human listeners (Figure 1). Two interactions clarified what categorization did depend on. A significant VOT × Boundary × Speaker interaction (β = 0.068, SE = 0.008, $p < .001$) indicated that categorization patterns differed across the L1 and

L2 data sets. This interaction was driven in part by a steeper VOT-based shift for the L1 data set than for the L2 data set, especially in the IP condition (red lines in Figure 1a). In addition, a significant VOT × Boundary × Source Voicing interaction (β = -0.063, SE = 0.008, p < .001) indicated that categorization patterns varied systematically across boundary conditions and source-token voicing. Specifically, stronger voiceless responses in the Wd condition suggest that Whisper drew on residual acoustic information preserved by greater coarticulatory continuity in Wd contexts (blue lines in Figure 1b). By contrast, stronger IP junctures may reduce such continuity, making the manipulated VOT continuum more influential (red lines in Figure 1b). This pattern was strongest for the L1 data set as reflected in a significant Boundary × Speaker × Source Voicing interaction (β = 0.898, SE = 0.296, p = .002). In the L2 data set, post-hoc inspection suggests that following-vowel f0 differences may also have cued voicing (cf. Kingston & Diehl, 1994). (Seoul) Korean speakers may produce them with greater magnitude given that f0 on the following vowel is a cue to the laryngeal contrast in their L1 (Kang, 2014). Because the data set in the present study consists of one participant per Speaker condition, whether the observed pattern reflects L1 transfer, speaker sex, or both cannot be resolved.

These findings suggest that scaling from Whisper-small to Whisper-large-v3 is not sufficient to produce human-like prosodic normalization, at least across the two variants tested. Instead, categorization appears to depend on residual acoustic cues whose availability varies across boundary context and data set.

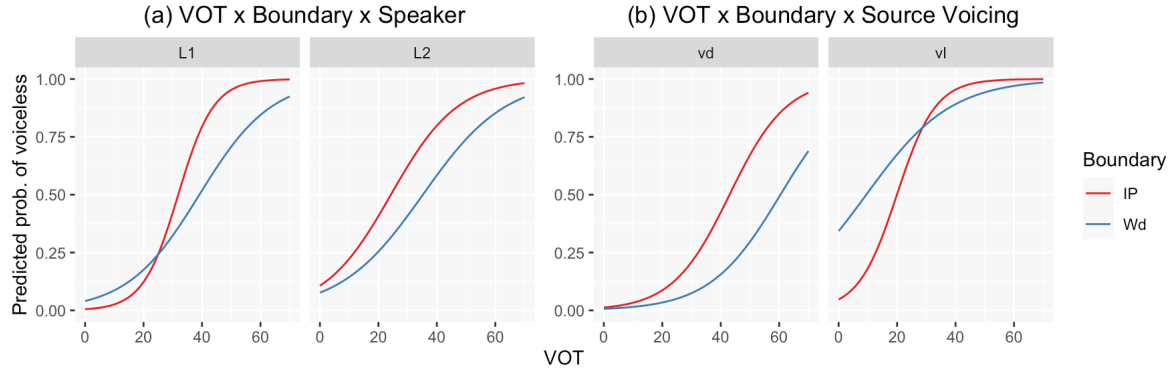

Figure 1. Predicted probability of voiceless categorization across the VOT continuum as a function of prosodic boundary condition. Panel (a) shows categorization patterns for the L1 and L2 data sets, while panel (b) shows patterns by source-token voicing (vd: voiced vs. vl: voiceless).

References

Bang, H. Y., Sonderegger, M., Kang, Y., Clayards, M., & Yoon, T. J. (2018). The emergence, progress, and impact of sound change in progress in Seoul Korean: Implications for mechanisms of tonogenesis. *Journal of Phonetics*, *66*, 120-144.

Fougeron, C., & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America, 101*(6), 3728-3740.

Jang, J., & Hatcher, R. (2026). Testing prosodic boundary-induced phonetic categorization in AI-based automatic speech recognition. *Phonetics and Speech Sciences*, *18*(1), 65-73.

Kim, S., & Cho, T. (2013). Prosodic boundary information modulates phonetic categorization. *The Journal of the Acoustical Society of America, 134*(1), EL19-EL25.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), 419-454.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384-422.

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 28492-28518). PMLR.