

Does Model Choice Matter? Comparing Whisper medium.en and large-v3 on Korean-Accented L2 English

Keechan Hong & Hayeun Jang
Sungkyunkwan University
keechan14@gmail.com, hyjang@skku.edu

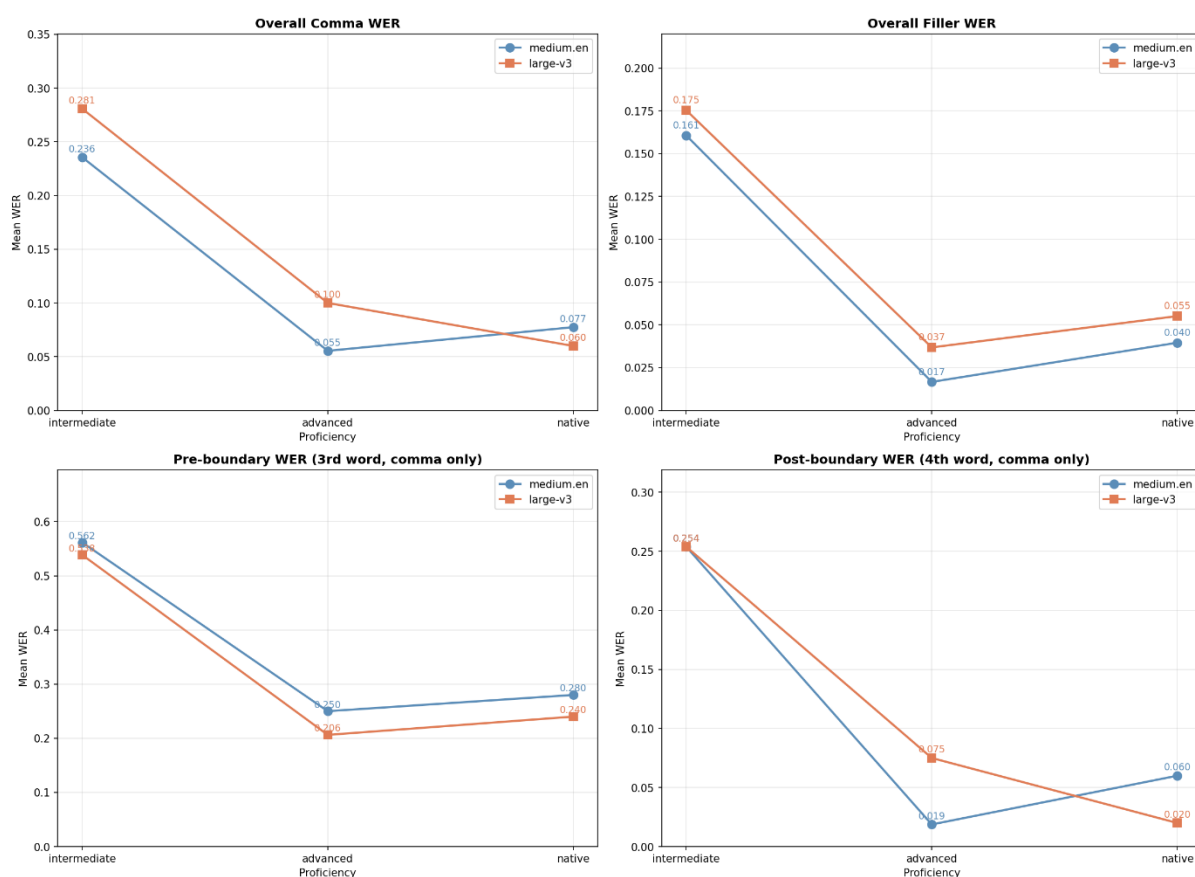
This study investigates how Whisper ASR responds to prosodic boundary cues in Korean learners' L2 English, assessing its robustness and fairness for non-native speech in language-learning contexts. Prosodic phrase boundaries cue syntactic structure for both human listeners and ASR systems, and misrecognitions at structurally critical positions are known to degrade transcription accuracy (Biron et al., 2021). Korean learners differ from native speakers in preboundary lengthening, break, and boundary tones at intonational phrase boundaries (Baek, 2021; Jun, 2004). The present study asks (a) whether Whisper's WER at comma-induced boundaries differs across proficiency levels, (b) whether boundary-adjacent words are loci of punctuation errors, and (c) whether model architecture modulates these effects.

Sentences with preposed adverbial clauses (e.g., *"Before Claire followed, her brother climbed the stairs."*) and fillers were read by 13 intermediate, 16 advanced, and 5 native speakers of English. Transcriptions from two Whisper models (medium.en and large-v3) were aligned to targets to compute sentence-level and boundary-local WER and comma placement accuracy.

Results reveal robust proficiency effects, with substantially higher WER for intermediate learners (medium.en: $M = .24$; large-v3: $M = .28$) than advanced (medium.en: $M = .06$; large-v3: $M = .10$) and native speakers (medium.en: $M = .08$; large-v3: $M = .06$), aligning with recent Whisper evaluations on non-native English (Graham & Roll, 2024; McGuire, 2025). Boundary-local analyses show consistently higher pre-boundary WER (intermediate: $M = .56$; advanced: $M = .25$; native: $M = .28$) than post-boundary (intermediate: $M = .25$; advanced: $M = .02$; native: $M = .06$), and comma placement accuracy also varies by proficiency. Critically, medium.en yields lower overall WER than large-v3 on this corpus of Korean-accented speech, likely reflecting negative transfer from large-v3's multilingual training across 99 languages reducing English-specific phonetic sensitivity.

Taken together, proficiency-related differences in prosodic boundary marking are reflected in Whisper's accuracy at sentence and boundary-local levels and model choice is a critical methodological consideration in ASR-based L2 research. These results point to the potential of ASR measures as indirect indicators of learners' prosodic features in production, while cautioning that multilingual large models may underperform English-only models in fairness-sensitive contexts involving accented non-native speech.

WER Comparison: medium.en vs large-v3 by Proficiency



References

- Baek, H. (2021). Korean speakers' use of prosody for the resolution of English PP attachment ambiguity. *Studies in Phonetics, Phonology and Morphology*, 27(3), 403-424.
- Biron, T., Baum, D., Freche, D., Matalon, N., Ehrmann, N., Weinreb, E., Biron, D., & Moses, E. (2021). Automatic detection of prosodic boundaries in spontaneous speech. *PLOS ONE*, 16(5), e0250969. <https://doi.org/10.1371/journal.pone.0250969>
- Graham, C., & Roll, N. (2024). Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2). <https://doi.org/10.1121/10.0024876>
- Jun, S. (2004). Intonational phonology of Seoul Korean revisited. *UCLA Working Papers in Phonetics*, 104, 14-25.
- McGuire, M. (2025). Automatic speech recognition for non-native English: Accuracy and disfluency handling. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2503.06924>