AI-Assisted Assessment of L2 (Dys)Fluency: Automated Measures vs. Native-Speaker Judgments

Yahya Aldholmi
Department of Linguistics, King Saud University, Saudi Arabia
yaldholmi@ksu.edu.sa

Fluency is a central dimension of second-language (L2) speech proficiency, shaped by the temporal organization of speech production, including articulation rate and pausing, that native listeners heavily rely on when judging L2 fluency and comprehensibility, often independently of segmental accuracy (Flege, 1995; Munro & Derwing, 1995; Derwing & Munro, 2009). With the growing use of automatic speech recognition (ASR) in language learning and assessment, recent work has begun to examine whether ASR-based measures can approximate human judgments of L2 speech. However, ASR systems continue to show reduced reliability with accented and non-native speech, largely due to limited representation of such varieties in training data (Inceoglu, Chen, & Lim, 2023). Whether transcript-free acoustic pipelines can nonetheless capture fluency dimensions meaningful to native listeners remains an open question, particularly for L2 Arabic, where empirical work on AI-based fluency assessment is still scarce.

This study investigates fluency in L2 Arabic and explores the potential of artificial intelligence (AI) as a tool for fluency assessment. Speech samples were collected from twenty-two L2 Arabic speakers representing three first-language backgrounds — Chinese/Mandarin (n = 7), Russian (n = 8), and African-language/Swahili (n = 7) — at an advanced (approximately, B2–C1) proficiency level, following 1–1.5 years of Arabic study, mostly in Saudi Arabia. Speech samples were elicited through a reading task using passages similar in length to the Rainbow Passage (Arabic version). Fluency was first analyzed manually by a phonetician using Praat (Boersma, Weenink, & Shchupak, 2026), then automatically through a modified version of the transcript-free Praat script for syllable nuclei and speech rate detection (de Jong & Wempe, 2009), using a minimum pause duration threshold of 400 ms, and finally compared against two AI-based ASR systems — Whisper and TurboScribe — which transcribed the speech samples, marked pauses according to duration, and ranked the three L1 groups by fluency using speech rate and pausing as the deciding measures. Speech rate was calculated over the full recording, articulation rate over speech-only intervals identified by the script, and pause percentage as the proportion of each recording occupied by silence-like intervals exceeding the 400 ms threshold. In addition, ten native speakers of Arabic were recruited to judge the fluency of the L2 speech samples using a 7-point Likert scale.

The results revealed systematic differences across speaker groups. The Russian group showed the highest mean speech rate (M = 5.6 syllables/sec), the African-language group occupied an intermediate position (M = 4.8 syllables/sec), and the Chinese group showed the lowest overall rate (M = 4.5 syllables/sec). Across the dataset, pause duration percentage displayed a strong negative relationship with speech rate, indicating that pause behavior played a particularly important role in shaping fluency. Native-speaker judges showed substantial agreement with the patterns from the manual vs. AI-based transcript, yielding broadly similar impressions of fluency across the speech samples. Importantly, the acoustic measures (both by human and AI) aligned with these human judgments, suggesting that AI-based analysis can successfully capture dimensions of fluency that are meaningful to native listeners.

Given the modest overall sample size, these findings are presented as preliminary evidence supporting the feasibility of AI-assisted, transcript-free acoustic analysis as a practical tool for assessing fluency in L2 Arabic, warranting confirmation in a larger study.

Keywords: L2 Arabic; fluency assessment; acoustic analysis; artificial intelligence; speaking rate

References

Boersma, P., Weenink, D., & Shchupak, A. (2026). Praat: Doing phonetics by computer [Computer program]. Version 6.6.30, retrieved 30 June 2026 from https://praat.org

Derwing, T. M., & Munro, M. J. (2009). Comprehensibility as a factor in listener interaction preferences. Applied Linguistics, 30(4), 565–585.

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods, 41(2), 385–390.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), Speech perception and linguistic experience (pp. 233–277). York Press.

Inceoglu, S., Chen, W.-H., & Lim, H. (2023). Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition. ReCALL, 35(1), 89–104.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language Learning, 45(1), 73–97.