AI–HUMAN SCORE DISCREPANCIES IN AUTOMATED JAPANESE SPEAKING ASSESSMENT: THE ROLE OF ASR ERRORS

Safarov Abdulatif
Senmon Kyoiku Publishing Co., Ltd., Japan
abdulatif.safarov92@mail.ru

Introduction. Automated speaking assessment offers a scalable alternative to exclusively human scoring, but its reliability depends on multiple processing stages. In transcript-based systems, automatic speech recognition (ASR) converts a learner's utterance into text before an artificial intelligence (AI) component assigns a score. Errors in transcription may therefore influence the evidence available for assessment. Previous research stresses automated scoring validation [1, 2]. This pilot study examines score agreement and explores whether recognition errors plausibly contribute to large differences.

Materials and methods. Eighty Japanese spoken responses from an operational assessment system were analyzed. The human rater listened to the original audio, whereas the automated system used an ASR transcript together with the question context. Mean absolute error (MAE) and signed differences were calculated. Spearman correlation measured score association. Twelve high-discrepancy cases were purposively selected. A Japanese teacher prepared reference transcriptions from the audio without seeing the AI score. These were compared with the ASR output to identify meaningful substitutions, deletions, or semantic distortions.

Results and discussion. The average human rating was 8.15 (SD = 3.34), compared with 6.81 (SD = 4.06) for AI. The MAE was 2.29 points, and the mean signed difference was −1.34 points. Spearman's correlation was moderately positive ($\rho = 0.462$, $p < 0.001$). Exact equality occurred in 34 responses (42.5%); the AI score was lower in 30 (37.5%) and higher in 16 (20.0%). Thus, although the two scoring approaches showed a meaningful rank association, the size of individual disagreements limits any assumption that their scores are interchangeable.

In the qualitative subset, meaningful ASR discrepancies were identified in 5 of 12 cases (41.7%), while the remaining 7 (58.3%) showed no meaningful recognition discrepancy. In one example, the teacher transcribed a request containing the word

"dictionary," but the ASR output replaced its central lexical content with a different polite expression; the human score was 10 and the AI score was 1.9. Such distortion could plausibly alter downstream scoring. Conversely, in another case the teacher and ASR transcriptions matched, yet the human score was 10 and the AI score was 0. These contrasts motivate separate recognition and scoring diagnostics. The 12 cases were deliberately selected; their proportions are not estimates of the overall system-wide ASR error rate.

Conclusion. The results indicate partial alignment between human and AI ratings but substantial disagreement for some spoken responses. ASR errors may contribute to certain large score differences, yet accurate transcripts do not guarantee consistent scoring. Validation must examine recognition and scoring separately. Limitations include the small pilot sample, one human rater, a concentration of maximum human scores (54 of 80 responses), and qualitative ASR analysis restricted to 12 selected cases. Future work should use independent raters, broader reference transcription coverage, and controlled re-scoring to distinguish the effects of recognition and scoring.

References

[1] Xu J., Jones E., Laxton V., Galaczi E. Assessing L2 English speaking using automated scoring technology: Examining automarker reliability. Assessment in Education. 2021;28(4):411–436. DOI: 10.1080/0969594X.2021.1979467.

[2] van Doremalen J., Cucchiarini C., Strik H. Optimizing automatic speech recognition for low-proficient non-native speakers. EURASIP J. Audio Speech Music Process. 2010:973954. DOI: 10.1155/2010/973954.