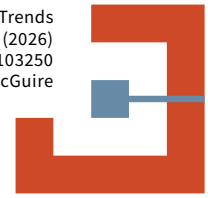
Trends

JALTCALL Trends
ISSN 2982-3749, 2(1), 103250 (2026)
https://doi.org/10.29140/jct.v2n1.103250



Castledown

Comparing ASR system accuracy on non-native English read speech across L1 backgrounds

Michael McGuire
Doshisha University, Japan

This study assesses the accuracy of five cutting-edge ASR systems in the recognition of non-native accented English read speech from six different L1 backgrounds. Speech samples were taken from the L2-ARCTIC non-native English speech corpus, containing recordings of 24 speakers from Arabic, Chinese, Hindi, Korean, Spanish, and Vietnamese L1s. Four US English speakers were included as a control. Five ASR systems were tested: AssemblyAI's Universal-2, Deepgram's Nova-2, RevAI's V2, Speechmatics' Ursa-2, and Whisper large-v3 by OpenAI. ASR accuracy was measured using the Match Error Rate (MER) metric. Overall, the study found much lower error rates for all five selected ASR systems than systems used in previous studies. Results showed that Whisper and AssemblyAI performed significantly better than other models, with no significant difference between them. Hindi speakers showed lower error rates while Chinese and Vietnamese showed consistently higher error rates across all systems.

Keywords: automatic speech recognition, learner English, non-native speech, CALL

Introduction

Automatic speech recognition (ASR), or the automated conversion of spoken language into text, has been considered an essential component of computer assisted language learning (CALL) and computer assisted language testing (CALT) for many years (Ehsani & Knodt, 1998). It has been widely adopted for both assessment and student practice, particularly for developing pronunciation and oral proficiency skills. ASR has developed rapidly, achieving unprecedented accuracy in recent years. This study assesses five cutting-edge ASR systems' accuracy in the recognition of non-native accented English read speech from six different L1 backgrounds to help language instructors and researchers understand the current state of ASR and which system might be appropriate for their use case.

The early development of ASR was language-specific, but in the 90s there was

JALTCALL Trends
Vol. 2 No. 1
2025 JALTCALL Conference Papers Special Issue

discussion about creating multilingual ASR systems (Byrne et al., 2000). However, ASR training data comes overwhelmingly from native speakers of each language, not from non-native accented speakers. Accented speech has been an important issue in ASR for many years (Neri et al., 2003) because, as numerous studies have shown, ASR transcriptions of accented speech show higher error rates than non-accented speech (Graham & Roll, 2024; Hirai & Kovalyova, 2024; Koenecke et al., 2020). This is likely due to the sourcing of accurate supervised training data, which is much more difficult and expensive for accented speech (Hinsvark et al., 2021).

To see where the technology currently stands, this study assesses the transcription accuracy of five cutting-edge, publicly available ASR systems on read L2 English learner speech from different L1 backgrounds with the following research question:

How do the selected automatic speech recognition systems compare in accuracy when transcribing L2 English read speech, and how do the accuracy rates vary by L1 background?

Methods

The speech samples used in this study were taken from the L2-ARCTIC non-native English speech corpus (Zhao et al., 2018), which contains laboratory recordings and transcriptions of 24 speakers from six L1 backgrounds–Arabic, Chinese, Hindi, Korean, Spanish, and Vietnamese–with two male and two female speakers each.

Each speaker (n = 24) read sentences from CMU's ARCTIC prompts (Kominek & Black, 2004) – sentences extracted from Project Gutenberg books. To represent more prototypical learner speech, the full list of sentences was filtered to remove any sentences that contain words that do not appear in the New General Service List (NGSL) (Browne, 2014). From this filtered list, 100 sentences were randomly selected, yielding 2,400 recordings ranging from 0.8 to 7.9 seconds. Recordings of the same sentences by four native speakers of American English taken from the CMU ARCTIC database (Kominek & Black, 2004) were used as a control for L1 analysis.

Five of the newest ASR systems at the time of writing were chosen for this study: AssemblyAI's Universal-2, Deepgram's Nova-2, RevAI's V2, Speechmatics' Ursa-2, and Whisper large-v3 by OpenAI. Only Whisper is open source; the other four systems are commercial. All systems were accessed through APIs using Python to ensure consistent testing conditions.

ASR accuracy was measured using the Match Error Rate (MER) metric following Morris et al. (2004), which accounts for substitutions, deletions, and insertions. MER was calculated using the jiwer Python package (Vaessen, 2018) by comparing the ASR transcription to the original human transcription of each speech sample.

Results

The mean MER across all 2400 non-native English speech samples for each ASR was calculated to establish overall accuracy, seen in Table 1 and Figure 1. Most transcriptions had a MER of zero (perfect accuracy), but those with errors had a wide range. Because of this skew, a Friedman test was conducted with the R programming

JALTCALL Trends
Vol. 2 No. 1
2025 JALTCALL Conference Papers Special Issue

language (R Core Team, 2024) in RStudio (RStudio Team, 2024) to compare the effect of ASR system on MER, finding a statistically significant difference with a small effect size ($\chi^2 = 471.93$, $df = 4$, $p < .001$, Kendall's $W = 0.039$). Post-hoc comparisons using Nemenyi's test indicated significant differences between ASR systems, and effect sizes (r) between systems were calculated based on mean rank differences, also seen in Table 1. Whisper and AssemblyAI performed significantly better than other models, with no significant difference between them. RevAI showed the highest error rate, performing significantly worse than both Whisper and AssemblyAI. Deepgram and Speechmatics showed intermediate performance levels with no significant difference between them. While the differences between models were statistically significant, the effect sizes based on mean rank differences were generally small, indicating limited practical significance. However, Whisper and Assembly AI demonstrated the best accuracy, and these two systems should be considered by CALL practitioners when dealing with non-native English read speech.

Table 1. ASR system performance comparison and mean rank effect size *r*

ASR ssystem	Mean MER (SD)	AssemblyAI	Deepgram	RevAI	Speechmatics	Whisper
AssemblyAI	0.056 (0.112)	—	0.137*	0.182*	0.118*	0.018
Deepgram	0.080 (0.138)	0.137*	—	0.045	0.019	0.155*
RevAI	0.086 (0.136)	0.182*	0.045	—	0.064*	0.200*
Speechmatics	0.074 (0.132)	0.118*	0.019	0.064*	—	0.136*
Whisper	0.054 (0.117)	0.018	0.155*	0.200*	0.136*	—

Note. * Nemenyi's test p < 0.05.

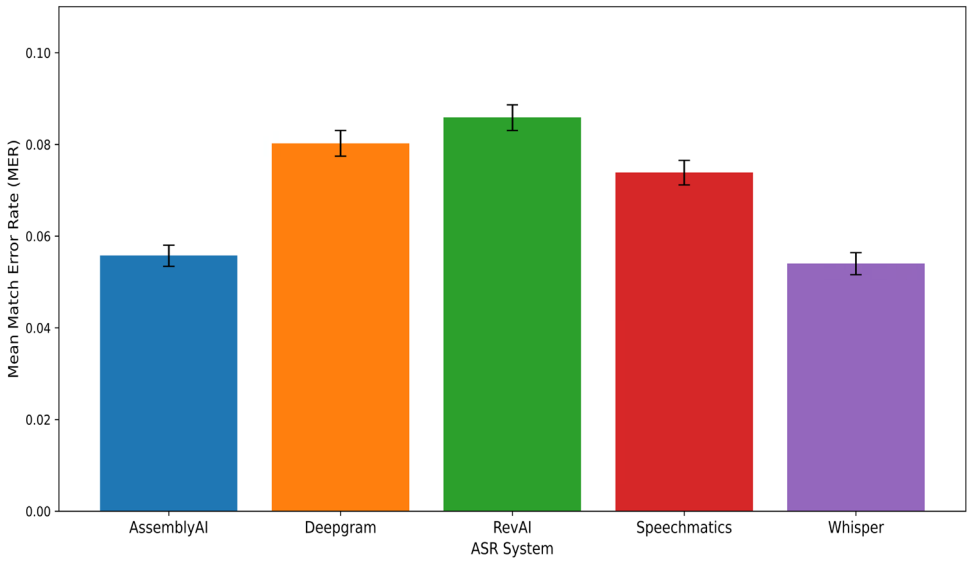


Figure 1. Mean MER by ASR system with standard error

Mean MER was calculated for each L1 background to examine potential differences in ASR performance across speaker background. Figure 2 shows the mean MER by

L1 background for each ASR system, also listed in Table 2. The pattern of relative performance among ASR systems remained consistent across L1 groups, with Whisper and AssemblyAI showing generally lower error rates than other systems. However, magnitude varied substantially across L1 groups, with US English speakers showing the lowest error rates (M = 0.007, SD = 0.036) and Vietnamese speakers showing consistently higher error rates (M = 0.143, SD = 0.186) across all systems. These variations in accuracy across L1s are presumably due to the lack of similarly accented training data used for the ASR systems.

Table 2. Mean MER and SD by L1 and ASR system

L1	AssemblyAI (SD)	Deepgram (SD)	RevAI (SD)	Speechmatics (SD)	Whisper (SD)
Arabic	0.051 (0.104)	0.076 (0.131)	0.082 (0.129)	0.064 (0.114)	0.047 (0.113)
Chinese	0.054 (0.096)	0.086 (0.132)	0.083 (0.120)	0.078 (0.125)	0.053 (0.099)
Hindi	0.02 (0.066)	0.034 (0.085)	0.039 (0.082)	0.031 (0.071)	0.021 (0.063)
Korean	0.042 (0.088)	0.058 (0.107)	0.061 (0.111)	0.058 (0.112)	0.033 (0.077)
Spanish	0.046 (0.094)	0.066 (0.120)	0.085 (0.132)	0.068 (0.122)	0.046 (0.098)
US English	0.007 (0.034)	0.008 (0.041)	0.007 (0.035)	0.006 (0.031)	0.007 (0.040)
Vietnamese	0.121 (0.171)	0.161 (0.194)	0.165 (0.187)	0.143 (0.191)	0.124 (0.183)

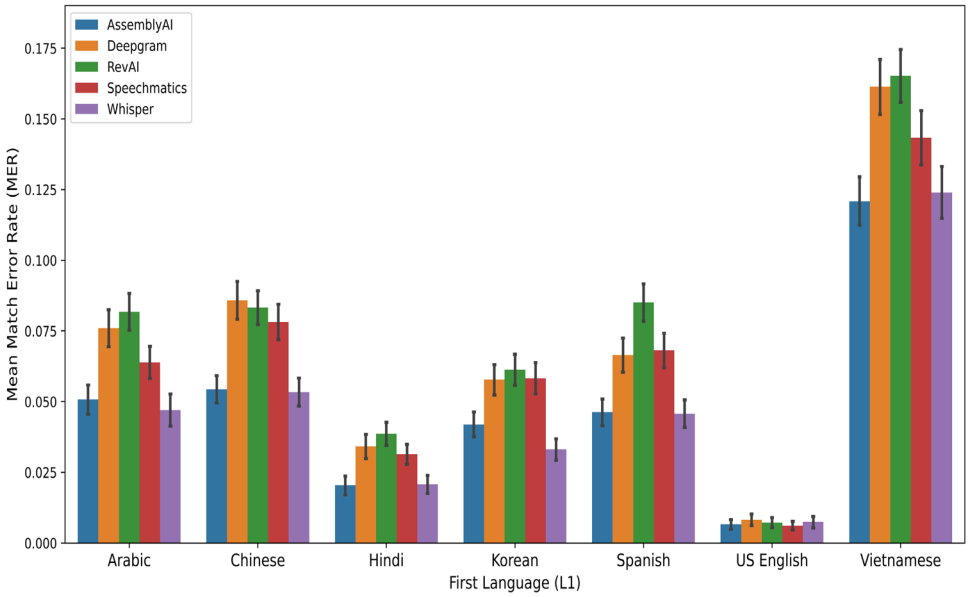


Figure 2. Mean MER by L1 and ASR system

Discussion

This study aimed to compare the accuracy of five cutting-edge ASR systems across different L1 backgrounds. The results found lower error rates for all five selected ASR systems than the systems used in previous studies. While no peer-reviewed studies have assessed the performance of AssemblyAI, Deepgram, RevAI, or Speechmatics on non-native speech, a few studies have examined Whisper. Notably, Graham

and Roll (2024) tested Whisper-large-v3 on English read speech from 14 different L1 backgrounds and found a mean MER of 0.157, much higher than the present study, likely due to differences in the data sets. Their study used recordings from the Speech Accent Archive (Weinberger, 2015) and likely calculated MER against the original paragraph rather than the actual transcriptions, which contained many errors. Recent studies of other commercial ASR systems also found lower accuracy levels for non-native English. McCrocklin et al. (2019) reported an accuracy rate of 74.44% for Windows Speech Recognition on read speech, and 88.61% for Google Voice Typing. Hirai and Kovalyova (2024) found accuracy rates between 48.91% and 67.04% across five different ASR systems with non-native speakers.

The present study shows that the five cutting-edge systems perform better on non-native English than those used in previous studies, with Whisper and AssemblyAI performing exceptionally well, with accuracy comparable to human transcription. Accuracy varied by L1 background, with Hindi speakers showing lower error rates and Chinese and Vietnamese showing higher rates. L1 background had a substantial impact ($F(6, 14) = 8.71$, $p < .001$, $ges = 0.778$), suggesting the need to adjust error thresholds for specific L1 backgrounds. Even so, these ASR tools are powerful enough for language teachers to use for a variety of purposes.

It is hoped that more ASR researchers and developers turn their attention to non-native learner speech. The most critical element for improvement is training data. Creating high quality training data is expensive, which explains why many of the best systems are commercial. However, creating and openly sharing non-native speech data with high quality transcriptions can greatly benefit both ASR training and testing.

References

Anthropic. (2024). Claude (Versions 3-5-sonnet-20241022) [Computer software]. *Anthropic*. https://claude.ai

Browne, C. (2014). The new general service List version 1.01: Getting better all the time. *Korea TESOL Journal, 11*(1), 35–50.

Byrne, W., Beyerlein, P., Huerta, J. M., Khudanpur, S., Marthi, B., Morgan, J., Peterek, N., Picone, J., Vergyri, D., & Wang, T. (2000). Towards language independent acoustic modeling. *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings.* https://doi.org/10.1109/ICASSP.2000.859138

Ehsani, F., & Knodt, E. (1998). Speech technology in computer-aided language. *Language Learning & Technology, 2*(1), 54–73.

Graham, C., & Roll, N. (2024). Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters, 4*(2), 025206. https://doi.org/10.1121/10.0024876

Hinsvark, A., Delworth, N., Rio, M. D., McNamara, Q., Dong, J., Westerman, R., Huang, M., Palakapilly, J., Drexler, J., Pirkin, I., Bhandari, N., & Jette, M. (2021). Accented speech recognition: A survey (No. arXiv:2104.10747). *arXiv*. https://doi.org/10.48550/arXiv.2104.10747

Hirai, A., & Kovalyova, A. (2024). Speech-to-text applications' accuracy in English language learners' speech transcription. *Language Learning & Technology, 28*(1), 1–22.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Kominek, J., & Black, A. W. (2004). The CMU ARCTIC speech databases. *Fifth ISCA Workshop on Speech Synthesis,* 223–224. https://www.isca-archive.org/ssw_2004/kominek04b_ssw.pdf

McCrocklin, S., Humaidan, A., & Edalatishams, I. (2019). ASR dictation program accuracy: Have current programs improved? In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference* (pp. 191–200).

Morris, A. C., Maier, V., & Green, P. (2004). From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. *Interspeech 2004,* 2765–2768. https://doi.org/10.21437/Interspeech.2004-668

Neri, A., Cucchiarini, C., & Strik, W. (2003). Automatic speech recognition for second language learning: How and why it actually works. *15th International Congress of Phonetic Sciences,* 1157–1160. http://www.internationalphoneticassociation.org/icphs/icphs2003

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.1 (2024-06-14)) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

RStudio Team. (2024). *RStudio: Integrated development environment for R* (Version 2024.04.2 Build 764) [Computer software]. Posit Software, PBC. https://posit.co/products/open-source/rstudio/

Vaessen, N. (2018). *Jiwer* (Version 3.1.0) [Computer software]. https://github.com/jitsi/jiwer

Weinberger, S. (2015). *Speech accent archive* [Dataset]. http://accent.gmu.edu

Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., & Gutierrez-Osuna, R. (2018). L2-ARCTIC: A non-native English speech corpus. *Interspeech 2018,* 2783–2787. https://doi.org/10.21437/Interspeech.2018-1110

Author biodata

Michael McGuire is an associate professor of second language acquisition in the Department of English at Doshisha University. His research interests are oral proficiency assessment, listening assessment, spoken fluency, and computational linguistics..