УДК 004.932.2:81'25
DOI https://doi.org/10.32782/tps2663-4880/2025.44.1.38

ACCENT INFLUENCE ON AUTOMATIC SPEECH RECOGNITION ACCURACY IN INTERPRETING CONTEXTS

ВПЛИВ АКЦЕНТУ НА ТОЧНІСТЬ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВИ В СИТУАЦІЇ УСНОГО ПЕРЕКЛАДУ

Litvinyak O.V.,
orcid.org/0000-0002-1378-9851
PhD in Philology, Associate Professor,
Associate Professor at the Hryhoriy Kochur Department
of Translation Studies and Contrastive Linguistics
Ivan Franko National University of Lviv

This paper presents a description of a pilot study examining the accuracy of automatic speech recognition (ASR) systems in transcribing interpreted, non-native English speech, specifically English speech produced by Ukrainian native speakers. This domain has been underexplored in automatic speech recognition publications, as most of them focus on original (even if non-native) speech. However, interpreted speech differs from original speech in fluency, prosody, and structure. It is more prone to contain pauses, reformulations, and disfluencies. The hypothesis is that these may cause certain problems for automatic speech recognition systems, since they are mainly trained on fluent, mostly native-speaker data. The paper describes the comparison of two automatic speech recognition tools, namely OpenAI's Whisper and Microsoft's Azure Speech Services. Since this is a pilot study, the research is based on a small corpus of only four simultaneous interpretations of speeches delivered in Ukrainian into the English language. The small corpus was intended to verify whether the hypothesis is worth analyzing and researching based on a larger sample. Each segment was manually transcribed to serve as a reference and then automatically transcribed by both automatic speech recognition systems. Accuracy was evaluated using Word Error Rate (WER). Both systems appeared to struggle with emotionally charged or semantically dense passages. The findings highlight the limitations of current automatic speech recognition models in capturing the unique characteristics of interpreted speech; however, further research is needed to draw more reliable conclusions. Prospects of further research include the use of an expanded corpus of speeches interpreted into English from different source languages by non-native English interpreters.

Key words: automatic speech recognition (ASR), word error rate (WER), accent, interpreting, accuracy.

У статті представлено опис пілотного дослідження, в рамках якого вивчено точність систем автоматичного розпізнавання мови (ASR) при транскрибуванні промов, перекладених англійською мовою з української усними перекладачами, що є носіями української мови, і для яких англійська мова – не рідна. Ця сфера недостатньо досліджена в публікаціях про автоматичне розпізнавання мовлення, оскільки більшість з них зосереджується на оригінальних промовах (навіть якщо вони виголошені не рідною мовою). Однак мовлення в усному перекладі відрізняється від оригінального мовлення плавністю, просодією та структурою. Воно частіше містить паузи, переформулювання та брак плавності. Гіпотеза полягає в тому, що це може спричинити певні проблеми для систем автоматичного розпізнавання мовлення, оскільки вони в їх в основному навчають на даних плавного мовлення, переважно носіїв мови. У статті описано порівняння Whisper від OpenAI та Azure Speech Services від Microsoft. Оскільки це пілотне дослідження, воно базується на невеликому корпусі, що складається лише з чотирьох синхронних перекладів українських промов англійською мовою. Невеликий корпус мав на меті перевірити, чи варто аналізувати та досліджувати гіпотезу на основі більшої вибірки. Кожен сегмент був вручну транскрибований для використання як еталон, а потім автоматично транскрибований обома системами ASR. Точність оцінювалася за допомогою показника частоти неправильно розпізнаних слів (Word Error Rate, WER). Обидві системи мали труднощі з емоційно навантаженими або семантично насиченими фрагментами. Результати дослідження підкреслюють обмеження сучасних моделей автоматичного розпізнавання мовлення у фіксації унікальних характеристик перекладеного мовлення; однак для отримання надійніших висновків необхідні подальші дослідження. Перспективи подальших досліджень включають використання розширеного корпусу промов, перекладених на англійську мову з різних вихідних мов усними перекладачами, для яких англійська мова не рідна.

Ключові слова: автоматичне розпізнавання мови (ASR), коефіцієнт помилок слів (WER), акцент, переклад, точність.

Statement of the problem. The main aim of this study is to assess how accurately automatic speech recognition (ASR) systems transcribe interpreted speech. The underlying assumption is that interpreted speech is less fluent, with more pauses, false starts, and disfluencies, while ASR tools are trained on fluent, native speech, or at least original (not interpreted) speech.

The study investigates how two widely used ASR systems – Whisper by OpenAI and Microsoft Word Online transcription tool, powered by Azure Cognitive Services – perform when applied to interpreted speech delivered in non-native English.

ASR functions are increasingly used in various domains, including interpreting. In recent years, we have heard various ideas about how ASR could be

used in and for interpreting. For instance, one of the widely discussed applications is CAI tools, where ASR could help prompt term lookup or figure recognition. Some (e.g. [2]) claim that ASR could completely change the face of consecutive interpreting turning it from a memory and analysis exercise into an on-the-spot translation of automatically generated text. A famous example of AI and ASR use is the interview of Ukraine's Volodymyr Zelenskyy to American podcaster Lex Friedman, which was AI-dubbed into Ukrainian, English, and Russian.

Literature review. Fantinuoli distinguishes the following challenges for ASR: use of spoken language; speaker variability; ambiguity; continuous speech; background noise; speed of speech; body language [2]. Koenecke et al. in their paper *Racial Disparities in Automatic Speech Recognitio*n have identified "1) a performance gap in the "language models" (models of lexicon and grammar) underlying modern ASR systems; and 2) a performance gap in the acoustic models underlying these systems" [5].

Thai researchers Tammasrisawat and Rangponsumrit in their 2023 paper *The Use of ASR-CAI Tools and their Impact on Interpreters' Performance during Simultaneous Interpretation* claim that the use of the ASR-CAI tool improves the quality of terminology rendition and decreases error rate in trainee interpreters [9]. In a domain-specific evaluation, the Svarah project compared Whisper, Azure, and Google ASR on Indian-accented English [7]. Hinsvark et al. identified accent bias as a persistent issue in commercial ASR [3]. McGuire evaluates five commercial ASR systems on the L2-ARCTIC corpus, which contains both read and spontaneous speech from non-native speakers with various L1 backgrounds. Reported mean error rates (MERs) on read speech were between 5–15%, with all systems performing worse on spontaneous speech [6].

Materials and methods. This study investigates the transcription accuracy of automatic speech recognition (ASR) systems on interpreted, non-native English speech. The corpus comprises four speeches originally delivered in Ukrainian and simultaneously interpreted into English by professional interpreters with English as their B language (active non-native language). The selected speeches include high-profile, public-facing addresses covering diverse domains such as politics, human rights, historical reflection, and military support. Each interpreted speech was manually transcribed and then processed using two commercial ASR systems. The test material is the simultaneous interpretation of speeches from Ukrainian into English by interpreters who have English as their B language (i.e., active non-native language).

Since this is a pilot study, aiming to understand further feasibility of doing research in this area, four short videos (of 4–5 min each) have been selected. The videos feature different speakers and different interpreters, all of them male (not intentionally).

For each speech, three aligned transcript versions were compiled:

Manual Reference: A human-verified transcription of the interpreted English speech, capturing all spoken words, including disfluencies (e.g., "uh", "eee"), repetitions, sentence fragments, and prosodic pauses.

ASR1 (Microsoft Azure): The output of Microsoft Azure's cloud-based speech-to-text API (Word Online or Azure Speech Services), representing a proprietary commercial ASR system with punctuation and formatting automatically applied.

ASR2 (Whisper): The transcription generated by OpenAI's Whisper (Large v3) model, an open-source multilingual ASR system trained on a large-scale, noisy, and varied audio-text corpus. Transcriptions were produced using default decoding settings and included punctuation.

Deletions (D): Words missing from ASR output but present in the reference.

These were used to calculate Word Error Rate (WER) as follows:

$$\mathrm{WER} = \frac{S + D + I}{N}$$

where N is the total number of words in the manual transcript.

Before comparing the ASR-generated transcripts to the human-generated ones, we investigated the matter of normal WER for human transcription. In controlled conditions using high-quality, multi-pass transcription workflows – such as those applied in the Switchboard Conversational Telephone Speech (CTS) corpus – inter-transcriber disagreement typically yields a WER between 4.1% and 4.5% [10]. However, when time constraints are introduced, as in "quick transcription" tasks, WERs can rise to approximately 9.6%. These values provide a practical benchmark for evaluating ASR system accuracy. Professional transcriptionists operating in real-world scenarios generally achieve WERs around 4%, though variability increases with factors like speaker accent, background noise, and emotional delivery. In recognition of these nuances, Apple introduced the Humanizing WER (HEWER) metric, which separates minor transcription deviations from those that affect semantic meaning. Using this measure, average ASR WERs around 9.2% on naturalistic

podcast audio translated to only 1.4% of errors being classified as "major" [1]. These findings reinforce that while human transcription remains a high standard, it is not perfect. For ASR systems to be considered comparable to human performance, a WER approaching 4–5% is often regarded as the benchmark [1].

Results. Word Error Rate (WER) Comparison

Transcript Set	ASR System	S	I	D	WER (%)
Zelenskyy to EU Parliament [14]	Microsoft	15	12	9	7.4
	Whisper	8	5	5	3.8
Lviv BookForum (Matviichuk) [11]	Microsoft	28	19	16	7.7
	Whisper	12	9	7	3.5
Confronting 1938 Mindset [12]	Microsoft	34	25	21	15.3
	Whisper	10	5	7	4.4
Veterans Peer Support (PTSD) [13]	Microsoft	20	18	15	7.3
	Whisper	11	6	8	3.4

Microsoft Azure produced WERs between 7.3% and 15.3%, with the highest error rate occurring in the emotionally charged and historically rich speech "Confronting 1938's Mindset." In contrast, Whisper maintained consistently lower WERs across all transcript sets, ranging from 3.4% to 4.4%. In all four cases, Whisper yielded a reduction in total word-level errors compared to Azure.

Error types were also analyzed. Microsoft Azure exhibited more frequent substitution and insertion errors, particularly with semantically complex terms and disfluency-laden segments. For example, "Kharkiv" was misrecognized as "hierarchy," and "Babi Yar" appeared as "Bobby years." Whisper's transcriptions featured fewer such misrecognitions and better semantic alignment with the manual reference.

Named entity recognition varied by system. Whisper accurately transcribed key proper nouns such as "Donetsk," "Luhansk," and "Tribunal for Putin." Azure misrecognized these terms more frequently or replaced them with phonetically similar but incorrect alternatives. Terminology associated with legal or historical contexts (e.g., "war crimes," "accountability architecture") was more faithfully preserved in Whisper's outputs.

Overall, the results suggest that both ASR systems face challenges when transcribing interpreted, non-native English speech, especially in emotionally charged, disfluent, or semantically dense contexts.

Discussion. The WER analysis revealed measurable differences across transcript sets and systems. Both ASR tools showed higher error rates in semantically dense or emotionally charged content, particularly in the "Confronting 1938's Mindset" speech, where historical and cultural references were frequent. While the WERs varied, both systems exhibited a similar pattern: increased errors in segments with higher disfluency density, rapid speech, or topic shifts. This suggests that interpretation-specific speech patterns, such as syntactic compression, hesitations, and reformulations, may be critical sources of error, regardless of ASR model.

One notable observation concerns disfluency handling. Interpreted speech naturally includes a high frequency of fillers (e.g., "uh," "eee"), repetitions, and sentence fragments, especially during real-time rendering. Both systems occasionally misrepresented these features. In some instances, literal transcriptions of false starts or hesitations produced syntactically incoherent outputs. Conversely, when disfluencies were omitted, the resulting output sometimes sacrificed fidelity to the interpreter's phrasing. This presents a tradeoff between fluency and transcription accuracy that warrants further exploration, particularly when disfluency carries semantic or pragmatic value.

Sentence segmentation also posed challenges. Interpretation often involves incomplete clauses or delayed corrections, which can result in fragmented or redundant constructions when transcribed. Both systems occasionally introduced segmentation errors – either by prematurely punctuating mid-thought or failing to mark clause boundaries – especially in passages with complex or emotive delivery. These errors complicate the readability and interpretability of ASR output in contexts where accurate documentation of speech is crucial.

Named entity recognition (NER) was another area of difficulty. Proper nouns – such as "Kharkiv," "Donetsk," and "Babi Yar" – were sometimes misrecognized, particularly when embedded in rapid or accented speech. In some cases, these errors affected the semantic integrity of the transcript, especially when the substitution produced a phonetically similar but contextually unrelated term (e.g., "Bobby years" for "Babi Yar"). This limitation may stem from ASR

models being optimized for general spoken content rather than multilingual, geopolitically specific discourse.

Conclusions. The interpreted nature of the source speech adds a layer of complexity that differentiates this study from evaluations using native, scripted, or read speech. Interpreters introduce subjective filtering, compressed syntax, and real-time adaptation, which amplify challenges for ASR systems trained predominantly on native, fluent speech samples. Consequently, these findings highlight a need for further research into ASR performance on interpreted corpora, including the development of domain-specific tuning or hybrid human-in-the-loop approaches.

In sum, while both ASR systems demonstrate competent baseline performance, the variability across transcript types and speech conditions underscores the importance of context in evaluating ASR reliability. Interpreted speech remains a challenging domain, and future evaluations should consider expanding to broader datasets and incorporating real-time constraints to better reflect applied settings such as live translation, legal proceedings, or multilingual broadcasting.

Overall, this study emphasizes the need for further evaluation of ASR tools on interpreted speech, a context that remains underexplored in ASR research. Future work should consider real-time evaluation, additional languages, and expanded datasets that reflect the variability of interpreter styles and speech conditions. Enhancing ASR systems to accommodate the nuances of interpreted and non-native speech is a critical step toward their effective deployment in multilingual and live communication environments.

REFERENCES:

1. DubSmart.ai. Understanding Word Error Rate in Speech Models. 2023. URL: https://dubsmart.ai/blog/understanding-word-error-rate-in-speech-models

2. Fantinuoli, C. Speech Recognition in the Interpreter Workstation, 2017. URL: https://www.staff.uni-mainz.de/fantinuo/download/publications/Speech%20Recognition%20in%20the%20Interpreter%20Workstation.pdf

3. Hinsvark, J., Cucerzan, S., & Chintalapudi, K. A survey of accented speech recognition: Challenges, methods, and opportunities . *arXiv*, 2021. URL: https://arxiv.org/pdf/2104.10747

4. Humanizing word error rate for ASR transcript readability and accessibility. (n.d.). Apple Machine Learning Research. URL: https://machinelearning.apple.com/research/humanizing-wer

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

6. McGuire, M. Evaluating automatic speech recognition systems on spontaneous and read non-native English speech. *arXiv*, 2025. URL: https://arxiv.org/pdf/2503.06924

7. Mehta, D., Ramesh, K., & Singh, A. Svarah: Evaluating English ASR systems on Indian accents. *ResearchGate*, 2023. URL: https://www.researchgate.net/publication/371040550_Svarah_Evaluating_English_ASR_Systems_on_Indian_Accents

8. Stolcke, A., & Droppo, J. Comparing Human and Machine Errors in Conversational Speech Transcription. *Redmond, WA, USA: Microsoft AI and Research*, 2017. URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/paper-revised2.pdf

9. Tammasrisawat, P., & Rangponsumrit, N. The use of ASR CAI tools and their impact on interpreters' performance during simultaneous interpretation. New Voices in Translation Studies, 2023, № 28(2). DOI: https://doi.org/10.14456/nvts.2023.27

10. Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., & Zweig, G. The Microsoft 2016 Conversational Speech Recognition System. *arXiv*. URL: https://arxiv.org/abs/1609.03528.

11.Materials analyzed

12. Oleksandra Matviichuk at Lviv BookForum 2023 [Video]. YouTube. https://www.youtube.com/watch?v=2l1cOxGjlbs

13. To fear evil or stop evil: Confronting 1938's mindset today [Video]. YouTube. https://www.youtube.com/watch?v=-uVtVHM1Aa0

14. Veterans peer support on coping with PTSD – Robin Imthorn and Artem Denysov [Video]. YouTube. https://www.youtube.com/watch?v=pLKdO4bF1ls

15. PRESIDENT ZELENSKY TODAY'S FULL SPEECH AT EUROPEAN PARLIAMENT 3/01/22, TUESDAY [Video]. YouTube. https://www.youtube.com/watch?v=3kqoIrNkV0E

Дата першого надходження рукопису до видання: 17.11.2025
Дата прийнятого до друку рукопису після рецензування: 19.12.2025
Дата публікації: 31.12.2025