Evaluating OpenAI's Whisper ASR: Performance Analysis Across Diverse Accents and Speaker Traits

Calbert Graham[1] and Nathan Roll[2]

[1)] *Phonetics Laboratory, University of Cambridge, CB3 9DP, UK*[a)]

[2)] *UC Santa Barbara, Santa Barbara, CA 95405, USA*

crg29@cam.ac.uk,

nroll@ucsb.edu

(Dated: 29 January 2024)

Abstract:

This study investigates Whisper's ASR system performance across diverse native and non-native English accents. Results reveal superior recognition in American compared to British and Australian English accents, with similar performance in Canadian English. Overall, native English accents demonstrate higher accuracy than non-native ones. Exploring connections between speaker traits (sex, L1 typology, and L2 proficiency) and word error rate uncovers notable associations. Furthermore, Whisper exhibits enhanced performance in read speech over conversational speech, with modifications based on speaker gender. The implications of these findings are discussed.

[a)] Author to whom correspondence should be addressed.

1. Introduction

The widespread adoption of video conferencing and virtual meeting platforms in recent years, particularly during the COVID-19 pandemic, has amplified the need for inclusive and accessible communication systems. To facilitate comprehension for diverse audiences, ASR technology has been leveraged to introduce live transcription services offered by multiple providers. While some studies have reported advancements in state-of-the-art ASR models (e.g., (Xiong *et al.*, 2018)), research has consistently shown that the performance of ASR systems varies as a function of speaker characteristics such as dialect ((Wheatley and Picone, 1991); (Meyer *et al.*, 2020)), gender background ((Tatman and Kasten, 2017); (Tatman, 2017)), racial backgrounds ((Koenecke *et al.*, 2020); (Martin and Tang, 2020)), and the linguistic structure of their native language (L1) ((Chan *et al.*, 2022)).

For instance, (Koenecke *et al.*, 2020) reviewed ASR systems sold by Amazon, Google, IBM, Apple, and Microsoft and highlighted vast disparities in the word error rates, the standard measure of ASR performance, calculated for black speakers of American English as compared to white speakers. Moreover, recognition rates of non-native speech are generally poor ((Bouselmi *et al.*, 2005); (Baykal and Erdogan, 2013); (Knill *et al.*, 2018)). In large part, this is due to speech recognisers being trained on predominantly native, standard varieties of the target language. The evaluation of ASR performance on non-native English accents remains a critical yet understudied area. This is of paramount importance, considering the difficulty in generating accurate transcriptions due to the unique challenges posed by signal degradation or unfamiliarity with specific English variants. (Markl, 2022, p.1) examined commercial ASR systems and suggested that they may 'reproduce structural oppression when

they perform worse for marginalised language communities'. The author states that speakers of stigmatised variants of English spoken in the British Isles are particularly vulnerable to such representational harms. Furthermore, the impact of speakers' first language (L1) on the comprehensibility of their L2 English, combined with variables such as their proficiency, adds an additional layer of complexity to the understanding of speakers with diverse language backgrounds.

It is clear from previous studies that the challenges ASR systems encounter in addressing accent diversity in speech are intricate and multifaceted. Whilst complete elimination of errors in ASR systems may not be attainable; an understanding of why those errors happen will allow us to create better mitigation strategies and improve system performance (Knill *et al.*, 2018). This is particularly relevant in safety-critical contexts such as telemedicine, air traffic control, among many other ASR applications.

The current study aims to address these challenges by examining the performance of Whisper (Radford *et al.*, 2022) on various English accents, with the main focus being on non-native accents. Whisper is an advanced automatic speech recognition (ASR) system developed by OpenAI, boasting high accuracy in transcribing audio into text. Trained on an extensive 680,000-hour of multilingual and multitask data collected from the web, the developers claim that Whisper demonstrates improved robustness to accents, background noise, and technical language. Whisper was chosen as the focal point of evaluation due to being open-source, its prominent usage in ASR research and its potential applications involving live transcription and related ASR services. Moreover, evaluating the performance of Whisper allows us to explore the generalisability of the results to other deep learning-

based ASR systems and provide insights into the broader landscape of non-native accent recognition.

The present study set out to achieve three primary research objectives (ROs): RO1a involves a comparison of Whisper's performance across native English accents of four countries: The US, Canada, Britain, and Australian. The dependent variable (MER) was also regressed against speakers' age and sex, with the goal of establishing a performance baseline for these phonetically distinct variants of native English accents. In RO1b we report the median scores for all accent variants represented in the dataset and rank them in order of increasing MER.

Research shows a link between L2 speech intelligibility and the characteristics of the L1, including the vowel system (e.g., (Bohn and Flege, 1990), (Munro and Derwing, 2008)) and the prosodic system (see (Graham *et al.*, 2006) for an overview). Research (e.g., (Iverson and Evans, 2009)) also shows that the vowel inventory size (i.e., the number of vowels) of the L1 of a speaker may influence their performance on vowel perception tasks in English. (Iverson and Evans, 2009) found that the larger the vowel inventory, the better the performance. RO2 therefore explores the association between Whisper's performance and the speakers' English proficiency and the linguistic structure of their L1 (prosodic typology: stress-accent vs pitch-accent vs tone; vowel inventory: the number of vowels in the L1). We also evaluate the influence of demographic information such as the age and biological sex of the speaker.

RO3 evaluates Whisper's performance using a conversational speech corpus, characterised by its spontaneity compared to the read speech corpus within the Speech Accent

Archive. This assessment aims to provide a more authentic representation of Whisper's capabilities in real-time transcription scenarios.

2. Method

2.1 Dataset

There are two data sources used in this study: the Speech Accent Archive (hereafter SAA) and the in-house Cambridge English Corpus (hereafter CEC) for non-native English.

2.1.1 Speech Accent Archive

SAA (Weinberger, 2015) is an online corpus which consists of recordings of native and non-native speakers of English with different accents. The recordings are roughly 30 seconds in length and consist of individual recordings of the same English paragraph (see supplementary material 1). The fact that all speakers read the same speech makes it possible to evaluate the acoustic model without having to grapple with variability in lexical choice and syntactic structure. There was roughly an even split between male and female speakers.

Available demographic information, including the speakers' age, L1, and age of onset of English learning, makes it possible to do a detailed analysis of the influence of speaker background on the performance of the ASR system. The number of speakers in each L1 is as follows: American English (442), British English (61), Canadian English (48), Australian English (45), Arabic (102), Russian (48), Dutch (46), French (26), German (28), Italian (32), Japanese (27), Korean (52), Mandarin (65), Polish (34), Spanish (16), Swedish (20), Vietnamese (22), Thai (15).

2.1.2 Cambridge English Corpus

This is a relatively small corpus of semi-spontaneous English. These were recorded in scenarios where an English exam candidate is prompted by an examiner to discuss various topics. A total of 133 speakers met the conditions for analysis (e.g., having an equal distribution of speakers across target L1s): Arabic (21 speakers), Dutch (24 speakers), French (21 speakers), Polish (24 speakers), Thai (21 speakers), Vietnamese (22 speakers). This matches six of the languages analysed in the SAA. The age range of 20–67 years, with a roughly equal split between males and females. The proficiency scores of these non-native speakers were determined by trained assessors based on the CEFR scale (Council of Europe (2001)). The dataset has been manually annotated by two trained phonetic transcribers with a 95% agreement. Disagreements between the annotators were resolved by the first author, who is a trained phonetician. The annotations and data processing were conducted using (Boersma and Weenink, 2010), a software tool for speech analysis

2.2 Data processing

2.2.1 SAA

During the process of transcribing spoken words into written text, various errors can arise. Word Error Rate (WER) is a widely used metric in Automatic Speech Recognition (ASR) that measures the accuracy of transcriptions by considering substitutions, deletions, and insertions. It is calculated as the sum of substitutions, deletions, and insertions divided by the total number of reference words. However, WER has limitations, as it lacks an upper bound and can exceed 100% in certain conditions. Match Error Rate (MER) is an alternative metric that focuses on the proportion of input/output matches that are correct. The crucial difference is that whereas the WER can exceed 100%, the MER is a probability distribution

between 0 and 1 and can therefore not exceed 100% (see (Morris *et al.*, 2004) for a review of these methods). MER is calculated according to the following formula:

$$\text{MER} = \frac{S + D + I}{N = H + S + D + I} = 1 - \frac{H}{N} \tag{1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, N is the number of words in the reference sentence, and H is the total number of 'hits'.

For each selected file, the audio (in the form of MP3) is extracted using the SAA online corpus. The BeautifulSoup package ((Richardson, 2007) was used to parse the data to HTML. We aggregated the sound file from the Speech Accent Archive and initialised the most recent version of Whisper (version 3) on an NVIDIA A100 GPU to automatically transcribe each audio file.

We created a function that takes in the raw audio and outputs Whisper's transcription, then iterated that over all the audio files (saving the output as a list). From that list we created a column of the Pandas dataframe ((Reback *et al.*, 2020). Next we used the Jiwer ((Vaessen, 2022) package to calculate the WER and MER, although only the latter is reported in this study.

2.2.2 CEC

The procedure was the same as for the SAA, with the main exception being that the sound files are in `.wav` format and the accompanying Praat Textgrids transcription was different for each recording and had to be read into the parser each time.

2.3 Quantifying English Experience

Assessing the English proficiency of speakers in the study using dedicated proficiency metrics was deemed infeasible due to the large number of speakers and the limited amount of speech data available for each individual. Additionally, all speakers in the study read the same set text, which further constrained the ability to gauge language proficiency based on a standard scale such as the CEFR scale (Council of Europe (2001)). In this research, we devised an algorithm to evaluate the 'English experience' (i.e., proficiency) of speakers in the SAA. The algorithm comprised the following steps:

(a) Each speaker was initially assigned a score of 100, from which we subtract their age of English learning onset. For example, after this step native speakers were assigned a score of 100 ($100 - 0 = 100$), while a speaker who began learning English at the age of 20 received a score of 80 ($100 - 20 = 80$).

(b) To account for the impact of age of onset on language proficiency, we implemented penalties. A penalty of -5 was applied for an age of onset falling between 11-20 years, -10 for 21-30, -15 for 31-40, and so on. This is in accordance with findings in second language acquisition theory, which establishes an inverse relationship between earlier language learning and higher proficiency. For instance, a speaker who commenced English study at 20 with an initial score of 80 would incur a penalty of -10, leading to a final score of 70. No penalty was applied for the first decade (0-10).

2.4 Variables and coding

In the analyses discussed in this study, MER is the dependent variable. The variables used as predictors include English experience (a numerical variable), and various characteristics of the speakers, such as their native language (L1), typology of the speakers' L1, sex, age,

and the number of vowels in the speakers' L1. Categorical variables, like typology, have been one-hot-encoded into binary format, where the presence of a feature is denoted as 1 and absence as 0. For instance, stress-accent is represented as 0, 0; pitch-accent as 1, 0; and tone as 0, 1.

3. Results

3.1 RO1A: MER of four Native English Accents

RO1 examines whether there are any statistically significant differences in Whisper's performance among four primary English accent groupings: American (reference level), Australian, British, and Canadian. We conducted a linear mixed effects model to investigate the impact of native English accents, speaker sex, and age on Whisper's performance, measured by Match Error Rate (MER). Speakers were treated as the random variable: (i.e., (MER $\sim$ English accents * age * sex + $(1|speaker)$).

We found that compared to American speakers, MER was significantly higher for both Australian English (β = 0.013, SE = 0.005, z = 2.401, p = 0.016) and British English speakers (β = 0.031, SE = 0.007, z = 4.571, p = 0.000). Notably, Canadian speakers did not demonstrate a statistically significant difference in MER compared to the American speakers ((β = 0.005, SE = 0.009, z = 0.596, p = 0.551). Additionally, age did not exhibit a statistically significant association with MER (β = 0.000, SE = 0.000, z = -1.139, p = 0.255). Sex also did not show a significant association with MER (β = -0.003, SE = 0.004, z = -0.825, p = 0.409). There were no significant interactions between any of the variables. The results are shown in Figure 1.

RO1b computes Whisper's average MER across the 18 accent variants of English. The ranking by median MER of the accents is shown in in Figure 2.

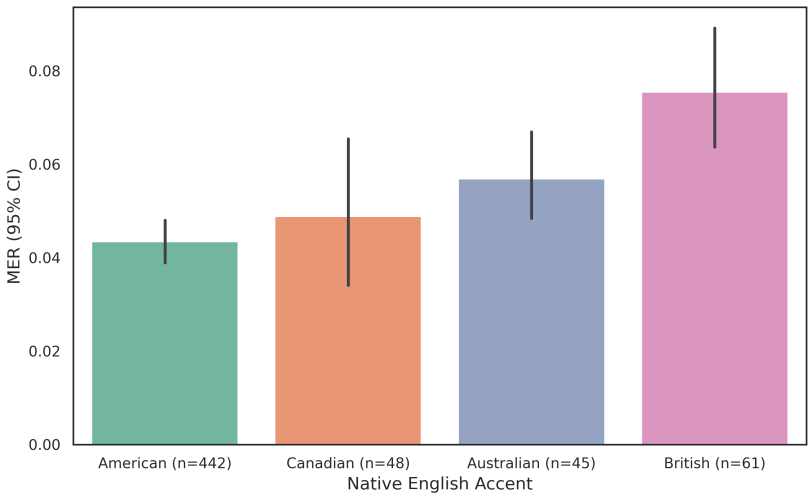

Fig. 1. Whisper's mean MER for four native English accents, with bars representing a 95% confidence interval of the mean.

3.2 RO2: MER and speaker characteristics

In RO2 we conducted an Ordinary Least Squares (OLS) regression analysis to examine the relationship between the response variable, MER, and a combination of continuous and categorical predictor variables. This is based on the SAA dataset.

The analysis revealed the following key findings:

1. English experience showed a statistically significant negative relationship with MER (β = -0.0036, STE = 0.001, z = -5.391, p = 0.000). An increase in English experience was associated with a decrease in MER. This result is shown in Figure 3.

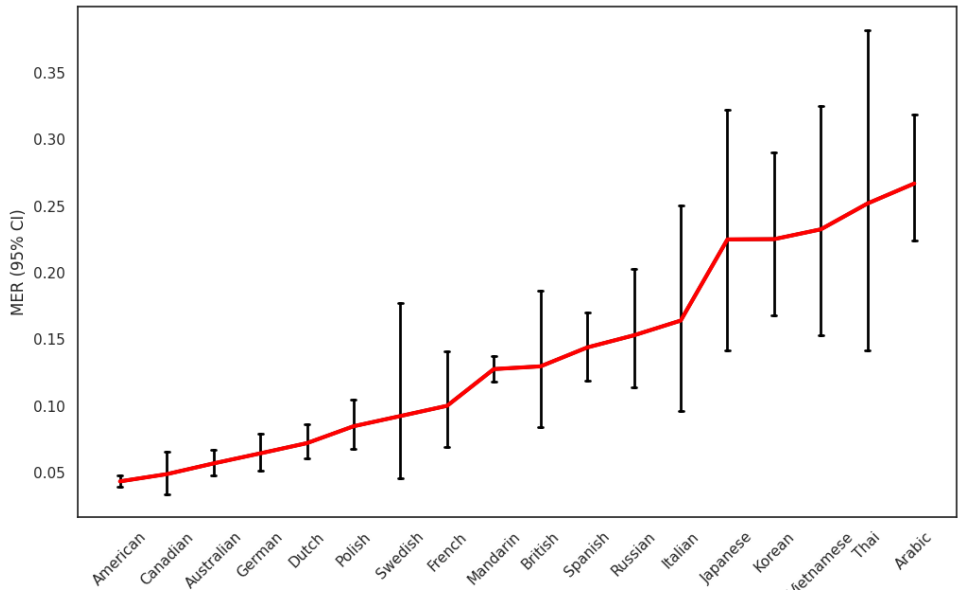

Fig. 2. Whisper's mean MER performance across English accents (red) and 95% confidence interval (black)

2. Vowels exhibited a statistically significant negative relationship with MER (β = -0.0054, STE = 0.002, z = -3.564, p = 0.000). Lower values of vowels were associated with higher MER.

3. In relation to speaker sex, female speakers exhibited a statistically significant lower MER level than male speakers (β = -0.0310, STE = -0.009, z = -3.289, p = 0.001). This indicates that Whisper was better at recognising English spoken by female speakers than male speakers.

4. In relation to the L1 typology of speakers, the MER of pitch-accent did not exhibit a statistically significant difference compared to the reference level (stress-accent) (β

= 0.040, SE = 0.035, z = 1.156, p = 0.248). In contrast, speakers of tone languages showed a significantly higher MER than those of stress-accent languages (β = 0.031, SE = 0.014, z = 2.189, p = 0.029). This relationship is visually supported by Figure 2, where Mandarin, Vietnamese, and Thai (all tone languages) cluster together as languages with the highest median MER values. This observed trend persisted even when native English speakers were excluded from the dataset, including only non-native English speakers. In this case, the comparison between stress-accent and pitch-accent yielded (β=0.037, SE = 0.036, z = 1.033, p = 0.302), while the difference between stress-accent and tone was statistically significant (β=0.0316, SE = 0.014, z = 3.938, p = 0.000).

The results for vowels, speaker sex and L1 typology are summarised in Table 1.

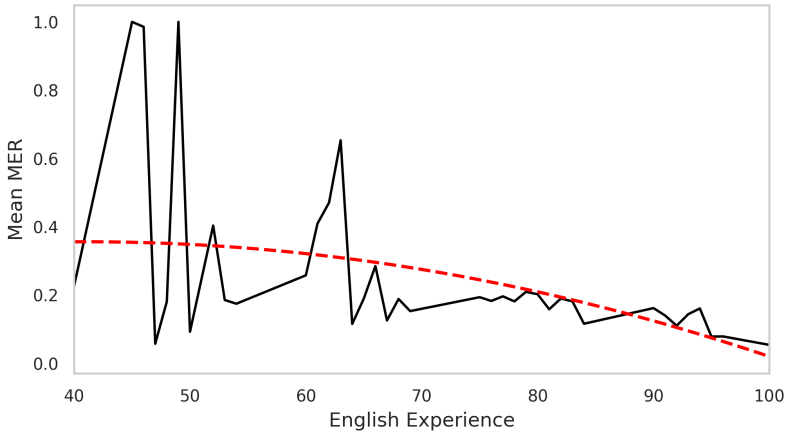

Fig. 3. Whisper's mean MER based on speaker characteristics (black) and quadratic trend line (red)

Table 1. Mean MER and WER by background

		MER	WER
Sex	Female	0.115	0.118
	Male	0.140	0.145
Typology	Pitch-accent	0.180	0.181
	Stress-accent	0.120	0.124
	Tone	0.161	0.166
No. of Vowels	0-5	0.257	0.261
	5-10	0.175	0.180
	10-15	0.085	0.089
	15-20	0.110	0.112

5. Other predictor variables did not show statistically significant relationships with MER ($p > 0.05$).

3.3 RQ3: Whisper's performance on conversational data

We examined the performance of Whisper on the conversational Cambridge English corpus with six languages and two typologies: stress-accent languages (Arabic, Dutch, French, Polish) and tone languages (Thai and Vietnamese). A linear regression model was fitted to

examine the relationship between MER and the predictor variables: L1 typology, sex, and (proficiency) score. As in previous analyses, the predictor variables were one-hot-encoded.

Among the individual predictor variables, proficiency score (β = -0.011, STE = 0.003, z = -4.144, p = 0.000) and age (β = -0.0045, STE = 0.002, z = -2.711, p = 0.007) showed significant associations with MER. Specifically, a one-unit increase in the score was associated with a decrease of approximately 0.011 units in MER, while a one-year increase in age was associated with a decrease of approximately 0.0045 units in MER. The difference between the tone languages and stress-accent languages, while not significant, approached the 5% level of significance (β = -0.058, STE = 0.033, z = -1.781, p = 0.075). There were no significant interactions between any of the variables.

Finally, we conducted a linear mixed model analysis to investigate the relationship between MER and speech type (read or conversational) and speaker sex. The result revealed that overall the conversational speech type had a significantly higher MER (β = 0.020, STE = 0.042, t = 4.723, p = 0.000) compared to the read speech type. MER for male speakers was also significantly higher than female speakers (β = 0.107, STE = 0.024, t = 4.371, p = 0.000). There was also a statistically significant interaction between sex and speech type (β = -0.139, STE = 0.047, t = -2.949, p = 0.003). The model accounted for the variability in MER among speakers by including speakers as a random effect. The estimated variance of the random intercepts for speakers was 0.004, indicating some variability in MER among different speakers. The results are shown in figure 4.

4. Discussion and Conclusion

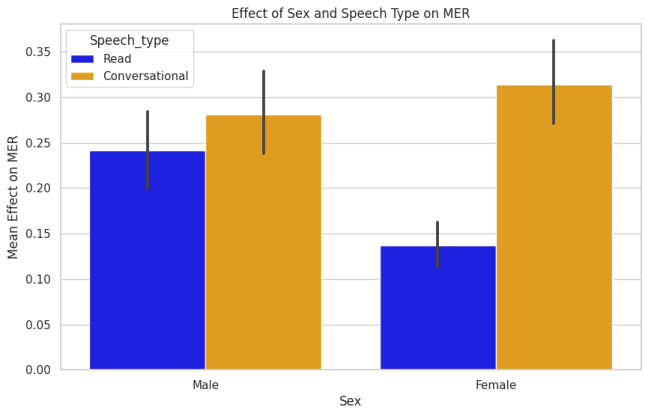

Fig. 4. MER by speakers' sex conversational vs read speech

RO1a examined Whisper's performance across major native English accents (American, Australian, British, and Canadian), revealing superior performance in North American English compared to British and Australian accents, with no significant difference in Canadian English. These findings emphasise the impact of regional accent variations on Whisper's ASR performance, aligning with prior research ((Markl, 2022)). Notably, the recognition of several non-native English accents (Swedish and German) surpassing British English is linked to the diverse nature of British accents and the more consistent pronunciation patterns among speakers of the same L1 background. This obscures the recognition disparities among British accents; for instance, native English speakers with the Leeds accent had the highest MER (close to 100%). This underscores the importance of considering regional accent diversity when developing and testing ASR systems for optimal communication experiences across regions.

RO1 (b) computed the average match error rate (MER) across 18 English variants using Whisper's ASR system. Notably, North American English (American and Canadian)

exhibited the lowest mean MERs, indicating superior ASR performance, while Vietnamese and Thai had the highest mean MERs, suggesting lower accuracy. These variations underscore the challenges in accurately recognising non-native accents, emphasising the need for inclusive ASR technologies. Addressing these performance gaps is vital, given the scarcity of English speech data from certain L1 backgrounds. Strategies like accent-agnostic meta-learning ((Winata *et al.*, 2020)) and transfer learning ((Meng *et al.*, 2020); (Viglino *et al.*, 2019)) show promise in adapting to unseen accents.

The second research objective (RO2) explored in greater detail the impact of speaker characteristics on Whisper's performance. The analysis revealed several significant relationships between speaker characteristics and MER. English experience demonstrated a significant negative relationship with MER, indicating that higher English proficiency was associated with lower MER values. Whilst this finding is unsurprising, it confirms that the proficiency of speakers influences the intelligibility of their speech, which in turn impacts on the performance of an ASR system. This corroborates earlier findings establishing a link between pronunciation accuracy and intelligibility in L2 speech (e.g., (Deshmukh *et al.*, 1996); (Loukina *et al.*, 2015); (Raymond *et al.*, 2002)). More broadly, this further underscores the need for other sources of speech variation, including atypical or dysarthric speech to be considered in the deployment of ASR systems (e.g., (Christensen, 2021); (Xue *et al.*, 2023))

The analysis revealed a significant negative association between the vowel variable and MER, indicating that speakers of languages with smaller vowel inventories tend to produce higher MERs. This aligns with findings reported by (Iverson and Evans, 2009) in the context of second language (L2) vowel perception. It is plausible to speculate that

individuals whose native languages feature a limited vowel system, in contrast to English with its relatively larger set of approximately fourteen vowels, may encounter increased challenges when learning a more extensive and perceptually complex vowel system.

The study identified a significant negative correlation between speaker sex and MER, suggesting higher MERs for male speakers compared to females. This aligns with prior research indicating that various ASR systems exhibit lower accuracy in recognising male speech, attributing this difference to male speakers' increased disfluencies and longer filled pauses ((Adda-Decker and Lamel, 2005); (Goldwater *et al.*, 2010); (Shariah and Sawalha, 2013)). Notably, these challenges persist in a large pre-trained system like Whisper, a point we will revisit in the discussion of results for RO3.

The study found that the native language (L1) typology of the non-native speakers exhibited a positive correlation with MER. Stess-accent languages exhibited the lowest MER, whereas tone languages exhibited the highest MER values. This finding suggests that Whisper may face challenges when trying to process the speech of speakers whose native language prosodic structure is typologically distant from the target language. Previous studies such as (Graham and Post, 2018) have reported on the influence of prosodic typology in the intelligibility of non-native speech in other ASR systems. Taken together, these findings underscore the importance of examining the rich interplay between different variables, and not just a few in isolation. An understanding of the relationships between speaker characteristics and MER is crucial in aiding developers to improve ASR models, in tailoring them to specific user groups, and in reducing biases inherent in this technology. However, it is important to note that due to the limited transparency regarding OpenAI's training datasets, there is

a possibility that the Speech Accent Archive might have been integrated into the training of the Whisper model. While this aspect is not a drawback of the present paper, it is a relevant factor to take into account when assessing the obtained results. This is because if utilised in training, it might contribute to the model's ability to handle various accents and linguistic nuances, which would obscure our understanding of the factors that have shaped its performance on diverse accents.

RO3 compared Whisper's performance on conversational vs read speech for selected L1s, aligning with prior studies highlighting ASR systems' challenges with spontaneous speech ((Horii *et al.*, 2022); (Gabler *et al.*, 2023)). The inherent disfluencies in spontaneous speech contribute to ASR errors, a phenomenon less prevalent in read speech. The significant interaction between sex and speech type indicates a modified effect of being male on 'MER' in conversational speech, suggesting a reduced impact compared to read speech. Understanding these dynamics is pivotal for optimising ASR technologies in realistic, conversational settings, ensuring effective communication across diverse user populations.

Future research could leverage advanced machine learning techniques, such as neural networks, to enhance ASR system robustness across diverse accents and speech types. Investigating the effectiveness of different approaches to ASR training could yield more accurate and inclusive ASR technologies. Investigating contextual factors, like environmental noise ((Dua *et al.*, 2023; Kumalija and Nakamoto, 2022)), can offer insights into real-world challenges and inform strategies for noise-robust speech recognition. A comprehensive examination of the interplay between speaker characteristics, linguistic features, and ASR outcomes using large speech corpora may reveal complex relationships, paving the way for personalized

and context-aware systems. Lastly, exploring biases in ASR technologies can guide efforts to address disparities and ensure equitable human-machine communication experiences.

Acknowledgments

We wish to thank the Leverhulme Trust for a Research Fellowship to the first author and the Cambridge Language sciences for an incubator grant to fund aspects of this research.

References and links

Adda-Decker, M., and Lamel, L. (**2005**). "Do speech recognizers prefer female speakers?," in *Proc. INTERSPEECH*.

Baykal, M., and Erdogan, H. (**2013**). "Non-native speech recognition using multi condition training," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6988–6992.

Boersma, P., and Weenink, D. (**2010**). "Praat: Doing phonetics by computer" http://www.praat.org/.

Bohn, O., and Flege, J. (**1990**). "Interlingual identification and the role of foreign language experience in l2 vowel perception," Applied Psycholinguistics **11**, 303–328.

Bouselmi, G., Fohr, D., and Haton, J. (**2005**). "Fully automated non-native speech recognition using confusion-based acoustic model integration," in *Proc. of Eurospeech, Lisboa*, pp. 1369–1372.

Chan, M., Choe, J., Li, A., Chen, Y., Gao, X., and Holliday, N. (**2022**). "Training and typological bias in asr performance for world englishes," in *Proc. Interspeech 2022*, pp. 1273–1277, doi: 10.21437/Interspeech.2022-10869.

Christensen, H. (**2021**). "Towards automatic speech recognition for people with atypical speech," in *Proc. Interspeech 2021*.

Council of Europe (**2001**). *Common European Framework of Reference for Languages: Learning, teaching, assessment* (Cambridge: CUP / Council of Europe).

Deshmukh, N., Duncan, R., Ganapathiraju, A., and Picone, J. (**1996**). "Benchmarking human performance for continuous speech recognition," in *Proc. of 4th International Conference on Spoken Language Processing. ICSLP '96*, Vol. 4, pp. 2486–2489.

Dua, M., Akanksha, and Dua, S. (**2023**). "Noise robust automatic speech recognition: Review and analysis," Springer-Verlag **26**(2), doi: 10.1007/s10772-023-10033-0.

Gabler, P., Geiger, B. C., Schuppler, B., and Kern, R. (**2023**). "Reconsidering read and spontaneous speech: Causal perspectives on the generation of training data for automatic speech recognition," Information **14**(2), doi: 10.3390/info14020137.

Goldwater, S., Jurafsky, D., and Manning, C. (**2010**). "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," Journal of Speech Communication **52**(3), 181–200.

Graham, C., Buttery, P., and Nolan, F. (**2006**). "Vowel characteristics in the assessment of l2 english pronunciation," in *Interspeech*.

Graham, C., and Post, B. (**2018**). "Second language acquisition of intonation: Peak alignment in american english," Journal of Phonetics **66**, 1–14, doi: https://doi.org/10.1016/j.wocn.2017.08.002.

Horii, K., Fukuda, M., Ohta, K., Nishimura, R., Ogawa, A., and Kitaoka, N. (**2022**). "End-to-end spontaneous speech recognition using disfluency labeling," in *Proc. Interspeech 2022*, pp. 4108–4112, doi: 10.21437/Interspeech.2022-281.

Iverson, P., and Evans, B. G. (**2009**). "Learning english vowels with different first-language vowel systems ii: Auditory training for native spanish and german speakers," The Journal of the Acoustical Society of America **126**(2), 866–877, https://doi.org/10.1121/1.3148196, doi: 10.1121/1.3148196.

Knill, K., Gales, M., Kyriakopoulos, K., Malinin, A., Ragni, A., Wang, Y., and Caines, A. (**2018**). "Impact of asr performance on free speaking language assessment," in *Interspeech 2018*, Hyderabad, India, pp. 1641–1645.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. (**2020**). "Racial disparities in automated speech recognition," Proceedings of the National Academy of Sciences **117**(14), 7684–7689.

Kumalija, E., and Nakamoto, Y. (**2022**). "Performance evaluation of automatic speech recognition systems on integrated noise-network distorted speech," Front. Sig. Proc. **2**, Art. no. 999457, doi: 10.3389/frsip.2022.999457.

Loukina, A., Lopez, M., Evanini, K., Suendermann-Oeft, D., Ivanov, A. V., and Zechner, K. (**2015**). "Pronunciation accuracy and intelligibility of non-native speech," in *Proc. Interspeech 2015*, pp. 1917–1921, doi: 10.21437/Interspeech.2015-423.

Markl, N. (**2022**). "Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition," in *Proceedings of 2022 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*, pp. 521–534, doi: 10.1145/3531146.3533117.

Markl, N. (**2022, p.1**). "Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition," in *Proceedings of 2022 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*, pp. 521–534, doi: 10.1145/3531146.3533117.

Martin, J. L., and Tang, K. (**2020**). "Understanding racial disparities in automatic speech recognition: The case of habitual 'be'," in *INTERSPEECH*, pp. 626–630.

Meng, Z., Hu, H., Li, J., Liu, C., Huang, Y., Gong, Y., and Lee, C. (**2020**). "L-vector: Neural label embedding for domain adaptation," in *ICASSP 2020*, IEEE, pp. 7389–7393.

Meyer, J., Rauchenstein, L., Eisenberg, J. D., and Howell, N. (**2020**). "Artie bias corpus: An open dataset for detecting demographic bias in speech applications," in *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6462–6468.

Morris, C., Maier, V., and Green, P. (**2004**). "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition," 4.

Munro, M., and Derwing, T. (**2008**). "Segmental acquisition in adult esl learners: A longitudinal study of vowel production," Language Learning **58**, 479–502, doi: 10.1111/j.1467-9922.2008.00448.x.

Radford, A., Kim, J., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (**2022**). "Robust speech recognition via large-scale weak supervision," ArXiv preprint arXiv:2212.04356.

Raymond, W., Pitt, M., Johnson, K., Hume, E., Makashay, M., Dautricourt, R., and Hilts, C. (**2002**). "An analysis of transcription consistency in spontaneous speech from the buckeye corpus," in *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pp. 1125–1128, doi: 10.21437/ICSLP.2002-371.

Reback, J., McKinney, W. *et al.* (**2020**). "pandas-dev/pandas: Pandas" How published: Zenodo. Version: 1.1.3. URL: https://doi.org/10.5281/zenodo.3509134. DOI: 10.5281/zenodo.3509134.

Richardson, L. (**2007**). "Beautiful soup documentation" How published: April. URL: https://www.crummy.com/software/BeautifulSoup/.

Shariah, M., and Sawalha, S. (**2013**). "The effects of speakers' gender, age, and region on overall performance of arabic automatic speech recognition systems using the phonetically rich and balanced modern standard arabic speech corpus," in *Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2*.

Tatman, R. (**2017**). "Gender and dialect bias in youtube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pp. 53–59.

Tatman, R., and Kasten, C. (**2017**). "Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions," in *Interspeech*, pp. 934–938.

Vaessen, N. (**2022**). "Jiwer: Similarity measures for automatic speech recognition evaluation" Version: 2.5.1. How published: https://pypi.org/project/jiwer/. Retrieved: 3 January 2024.

Viglino, T., Motlicek, P., and Cernak, M. (**2019**). "End-to-end accented speech recognition," in *INTERSPEECH*, pp. 2140–2144.

Weinberger, S. (**2015**). "Speech accent archive," Http://Accent.gmu.edu.

Wheatley, B., and Picone, J. (**1991**). "Voice across America: Toward robust speaker-independent speech recognition for telecommunications applications," Digital Signal Processing **1**(2), 45–63, https://linkinghub.elsevier.com/retrieve/pii/1051200491900953, doi: 10.1016/1051-2004(91)90095-3.

Winata, G., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., Xu, P., and Fung, P. (**2020**). "Learning fast adaptation on cross-accented speech recognition," arXiv preprint arXiv:2003.01901 .

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., and Stolcke, A. (**2018**). "The microsoft 2017 conversational speech recognition system," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5934–5938.

Xue, W., Cucchiarini, C., van Hout, R., and Strik, H. (**2023**). "Measuring the intelligibility of dysarthric speech through automatic speech recognition in a pluricentric language," Speech Communication **148**, 23–30, https://doi.org/10.1016/j.specom.2023.02.004.