Training and typological bias in ASR performance for world Englishes

Conference Paper · September 2022

DOI: 10.21437/Interspeech.2022-10869

6 authors, including:

May Pik Yu Chan
University of Pennsylvania
7 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Yiran Chen
University of Pennsylvania
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Interspeech 2022
18-22 September 2022, Incheon, Korea

Training and typological bias in ASR performance for world Englishes

May Pik Yu Chan[1], *June Choe*[1], *Aini Li*[1], *Yiran Chen*[1], *Xin Gao*[1], *Nicole Holliday*[1]

[1]Department of Linguistics, University of Pennsylvania, USA

{pikyu, yjchoe, liaini, cheny39, kauhsin, nholl}@sas.upenn.edu

Abstract

The use of automatic speech recognition (ASR) has been increasing to promote inclusion and accessibility. Nonetheless, prior work on ASR finds performance gaps conditioned by specific gender and racial groups, revealing systematic biases in modern ASR systems. However, work has focused on native varieties of English, glossing over its impact on a wider range of ASR users, namely second language speakers of English. The present work compares the performance of the transcription system Otter, on 24 varieties of English, 21 of them are non-native varieties. We compare the word and phone error rate (WER/PER) of accent varieties that are claimed to be supported by Otter and those that are unsupported. Results show that English varieties that are supported have lower WERs compared to that of unsupported varieties. However, there are still systematic differences in performance conditioned by linguistic structure in both supported and unsupported Englishes. Specifically, Otter performs better on English varieties from non-tonal first language speakers. We conclude that while inclusion of more varieties of English in the training data set for ASR may promote inclusivity, there may still be biases inherent to the linguistic structure that should not be overlooked.

Index Terms: speech recognition, natural language processing, speech-to-text, bias in transcription

1. Introduction

The past few years saw the rise in the use of video conferencing and virtual meeting platforms. Live transcriptions of meeting content have been introduced to promote inclusivity and accessibility for diverse audiences. While some works have highlighted recent successes of speech recognition (e.g., [1]), other studies find that performance of live captioning systems varies by users' dialect [2, 3] and gender background [4, 5, 6, 7], as well as racial backgrounds [8, 9]. Furthermore, such systematic biases have been found beyond English varieties, including varieties of Dutch [10] and Arabic [11].

A number of works have also found that ASR systems consistently underperform for individuals speaking their second language (e.g., [12, 13]). Despite an increasing amount of work on dialects of L1-English, such work on world Englishes remains limited. This is however an important question of interest, because meeting transcriptions are most often needed when some attendees have difficulty understanding a speaker, either due to a degraded signal or lack of familiarity with an English variety. In addition, for L2 English speakers, the influence from their L1 coupled with factors such as their L2 age of onset of acquisition, further adds to the difficulty of understanding speakers of different language profiles. Furthermore, there has been an increased reliance on transcription systems, following the rise of video conferencing. This paper expands on prior work on biases in automated speech recognition systems by evaluating the effectiveness of automatic captioning for world English varieties. We explore the accuracy of live-transcription services for online video meetings. Specifically, we evaluate the performance of Otter, a speech recognition transcription platform that claims to work on multiple varieties of English including "(southern) American, Canadian, Indian, Chinese, Russian, British, Scottish, Italian, German, Swiss, Irish, Scandinavian, and other European accents" [14].[1] Otter was chosen as the platform of interest because (1) it seeks to support a wide range of English varieties, and (2) Otter collaborates with popular video conferencing platforms (e.g., Zoom) that reach broad international audiences.

Existing proposals for mitigating bias in ASR performance include diversifying the training data to ensure that the technology is inclusive of underrepresented language varieties [8]. As Otter claims that their service supports various Englishes, we infer that Otter is one of few companies that do include a more diverse training data set for their model. We ask, however, whether the mere inclusion of a larger data training set is sufficient to combat bias, or are there other sources of bias that exist dependent on linguistic structure (e.g. tonal vs non-tonal languages). In the present work, we are interested in whether transcription services that do consider a relatively more diverse training set achieve this goal. Specifically, we focus our analysis on the performance of Otter's transcription services on varieties of English spoken in the US, UK, Canada, and samples of world Englishes. We seek to address two main research questions: (1) Whether Otter performs better on English language varieties that are claimed to be supported compared to unsupported varieties of English, regardless of the speakers' demographic information. (2) Whether system performance is also driven by differences in linguistic structure, such as whether a language is tonal or non-tonal. In this work, we use word error rates and phone error rates as metrics for the effectiveness of the ASR system.

2. Materials and Methods

2.1. Corpus

The data analyzed in this study are drawn from the Speech Accent Archive at "http://accent.gmu.edu/howto.php" [15]. This corpus is established to uniformly exhibit a large set of English varieties from a wide range of language backgrounds. Native and non-native speakers of English are recorded individually to read the same English paragraph.

The reading passage, which is around 30 seconds long, was carefully designed to contain practically all of the sounds of English, making it tangible to include more detailed phonetic and acoustic analysis in our current study if necessary. In addition, all the recordings were conducted in a quiet room with participants sitting at a table and being approximately 8-10 inches

[1]Note that while Otter uses the term "accents", we will be adopting the word "varieties" to refer to Englishes from both native and non-native speakers in this paper.

10.21437/Interspeech.2022-10869

from the microphone. The data quality therefore can be guaranteed. Most importantly, this corpus makes it possible to compare English speakers of different demographic and linguistic backgrounds, while controlling for complexities involved in speech style and recording environment. The fact that all speakers read the same passage ensure that any errors in system performance would not be driven by the language model (model of grammar and word choice) but instead would mainly reflect a performance gap in the acoustic model. In other words, none of the speakers would have had syntactic structures or lexical choices that were more or less predictable than other speakers. As such, the dataset is ideal for analysing the performance of Otter's ASR system. In addition to the speech files, each speakers' recording has been systematically coded with the following demographic information: the speaker's birthplace, age (at the time of recording), sex, native language, other languages (besides English and their native language), English learning residence, age of English onset, English learning method (naturalistic vs. academic), English residence country, and length of living in an English-speaking country. Information about this corpus and their data collection protocol can be found online.

Out of 2979 recordings available from the Speech Accent Archive, we chose a subset of recordings to enter our analysis based on the following criteria: (1) Only varieties of English that are either listed as supported regional varieties by Otter, or (2) have recording entries from at least 10 speakers. Regarding supported varieties of English, we only included speakers whose native language is English if their birth place matches their English country of residence. Furthermore, while Irish English is a variety of English that Otter supports, we did not include speakers of Irish English as only one recording was available from the Speech Accent Archive. This leaves us with eleven English varieties that are Otter-supported. For the purpose of comparison, we then chose another eleven English varieties that are not supported by Otter to enter the analysis; all eleven varieties include recordings from at least 10 speakers. This amounted to a total of 1227 recordings. The chosen English varieties and the number of speakers per language variety is summarized in Table 1. In cases where the "Dialect entry" is left empty, we did not filter the specific dialect column for that language variety.

Table 1: *English varieties that were chosen to enter the analysis. "Likely" refers to a variety that Otter does not explicitly claim to support, but may fall into a general category that it does support (e.g. "European/Scandinavian accents")*

Otter Supported	**Birth place entry**	**Language entry**	**Dialect entry**	**N (of speakers)**
Yes	USA	English	English	416
Yes	China	Chinese	Mandarin	135
Yes	Russia	Russian	Russian	54
Yes	UK	English	English	44
Yes	Canada	English	English	40
Yes	Italy	Italian	Italian	36
Likely	France	French	French	36
Yes	Germany	German	German	34
Yes	India	Hindi		33
Likely	Sweden	Swedish		18
Likely	Spain	Spanish	Spanish	18
Yes	Switzerland	Swiss German	Swiss German	9
No	South Korea	Korean		95
No	Japan	Japanese		42
No	Vietnam	Vietnamese		38
No	Ethiopia	Amharic		27
No	Hong Kong	Cantonese / Chinese	Cantonese	25
No	Philippines	Tagalog		24
No	Morocco / Egypt	Arabic		24 (13+11)
No	Thailand	Thai		22
No	Pakistan	Urdu		21
No	Bangladesh	Bengali		14
No	Indonesia	Bahasa		12
No	Afghanistan	Dari	Dari	10

2.2. Data Processing

After file selection, the list of recordings was pseudo-randomized to order the recordings passed to Otter, such that the likelihood that consecutive speakers sound too similar for Otter to recognize a speaker change was minimized. Each sound file was then re-sampled to 22050 Hz, and concatenated based on the psuedo-randomized order, with a silent pause of one second inserted between each recording. The start and end times of each recording in the concatenated file was recorded. The concatenating procedure was done in Praat using a script [16]. The total recording time amounted to almost 9.5 hours. As we used an Otter Pro account to generate the transcriptions, the maximum transcription duration per conversation is 4 hours. Consequently, the concatenated audio files were separated into three sound files, which were uploaded onto Otter separately. To test for potential training effects in the language or acoustic model in Otter, one of the three concatenated sound files was uploaded to Otter twice. The resulting transcription was identical, suggesting no training effect between recording transcriptions.

After Otter transcription generation, the transcript for each speaker was re-merged based on the concatenation timestamps. There were occasions where Otter merged multiple passages as coming from a single speaker. In these cases, two human annotators checked through the recordings independently to separate the transcripts to the respective speakers, reaching 99.8% agreement. Where needed, a third annotator resolved the conflict. Resulting transcriptions entered the error rate analysis.

2.3. Word Error Rate

The Word Error Rate (WER) was calculated for each speaker. Each individual's observed string of words (transcript output) was compared to the reading passage ("truth") via a minimal edit distance algorithm implemented at the word-level which identified locations of deletions, substitutions, and insertions [8]. The WER was calculated from the sum of deletions (D), substitutions (S) and insertions (I) of entire words, divided by the total number of words (N) in the passage.

$$WER = \frac{D + S + I}{N} \quad (1)$$

The same edit distance algorithm was also applied at a narrower phone level to calculate the phone error rate (PER), which is the number of deletions, substitutions, and insertions of phones divided by the total number of phones in the passage. Words were translated into their ARPABET pronunciations

by referencing The CMU Pronouncing Dictionary prior to PER calculation. The CMU Pronunciation Dictionary was used for this conversion as the speakers from US/Canada performed the best on the WER metric, suggesting that Otter's acoustic model may be assuming American pronunciations as the standard. In cases when the dictionary was missing entries for words generated by Otter (a total of 129 words which amounts to roughly 10% of all unique words across transcriptions), their ARPABET pronunciations were coded by two human transcribers.

3. Results

To recapitulate, the two goals of the present study are to compare Otter's performance on supported and unsupported languages, and to examine whether its performance is modulated by typological differences. We begin by computing the average word error rate of Otter transcriptions across all chosen varieties of Englishes. Across the board, Otter performed better with varieties that were supported (WER: mean = 0.071, sd = 0.085; PER: mean = 0.060, sd = 0.083) compared to unsupported varieties (WER: mean = 0.135, sd = 0.112; PER: mean = 0.118, sd = 0.112). Detailed mean WERs are summarized in Table 2. Figure 1 illustrates the distribution of WER by Otter supported and unsupported varieties respectively. Overall descriptive results show that varieties that are supported by Otter have lower error rates than varieties that are not explicitly stated as supported. Since the WER and PER distributions are largely comparable, the following analyses will focus on WER.

Table 2: *Variable coding and mean WER by speaker language background.*

	Supported	Tonal	Mean WER
English	+		0.035
Hindi	+		0.057
Swedish	+		0.059
German	+		0.065
Swissgerman	+		0.084
French	+		0.098
Italian	+		0.114
Spanish	+		0.115
Russian	+		0.136
Mandarin	+	+	0.157
Cantonese		+	0.162
Thai		+	0.202
Vietnamese		+	0.214
Urdu			0.052
Japanese			0.109
Tagalog			0.109
Arabic			0.114
Korean			0.116
Indonesian			0.122
Dari			0.123
Bengali			0.129
Amharic			0.167

In order to determine whether the WER between English varieties that are supported and unsupported show meaningful differences, we fitted a linear mixed-effects model to predict WER with fixed effects of OtterSupported (Supported vs Not Supported, sum coded with Supported as 1), speaker Sex (Male vs Female, sum coded with Female as 1) and speaker Age (scaled and centered to mean). Additionally, the model included an interaction effect between OtterSupported and Sex, given prior works showing that ASR systems exhibit biases at the intersection of gender and language varieties [4]. Lastly, the model fitted random intercepts by language.

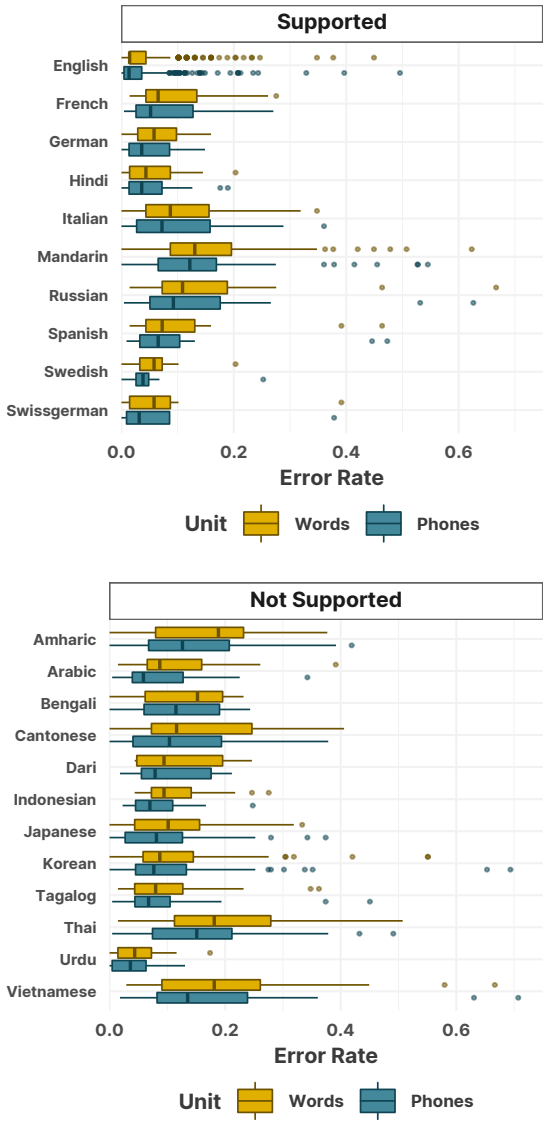

Figure 1: *Summary of Word Error Rates in sampled data.*

Results show a significant main effect of OtterSupported (β = -2.16e-02, SE = 9.49e-03, t = -2.27, p = 0.03) as well as a significant main effect of age (β = 5.49e-04, SE = 1.60e-04, t = 3.44, $p < 0.001$). No significant effects of gender (β = -3.75e-03, SE = 2.65e-03, t = -1.42, p = 0.16) nor interaction effect of OtterSupported and gender (β = -1.11e-04, SE = 2.64e-03, t = -0.04, p = 0.97) were found.

After gathering results for Otter's performance across supported and unsupported varieties of English, we looked for structural similarities between languages that might explain differences in performance. One typological grouping that emerged was the split between tonal vs non-tonal L1s. To fur-

ther examine this point of difference, we divided speakers' native languages into tonal and non-tonal varieties and found that Otter performs better on speakers with a non-tonal L1 (WER: mean = 0.071, sd = 0.083) than that of tonal varieties (WER: mean = 0.172, sd = 0.119).

To investigate what linguistic and demographic factors modulate WER among the non-native Englishes, we fitted a linear mixed-effects model to a subset of the data including only non-native speakers, again with WER as the response variable. In addition to the previously included fixed effects — OtterSupported, Sex, Age — the model also included the age of English onset and IsTone (Tonal vs. Non Tonal language, sum-coded with Tonal as 1) and the interaction between OtterSupported and IsTone. The random effects included a random intercept and a random slope of age of English onset by language.

We find significant main effects of OtterSupported (β = -1.57e-02, SE = 6.33e-03, t = -2.48, p = 0.04) and IsTone (β = 3.81e-02, SE = 6.37e-03, t = 5.98, p = 0.0002). In other words, when we consider nonnative speech alone, Otter is still more accurate on languages that they explicitly state they support. Furthermore, performance is systematically worse on speakers with a tone language background, compared to non-tone language speakers. We also find a significant main effect of age of English onset (β = 4.19e-03, SE = 7.89e-04, t = 5.30, p = 0.0003), but not age of recording (β = 1.64e-04, SE = 3.13e-04, t = 0.53, p = 0.60). As can be seen in Figure 2, the age at recording and age of English onset are not necessarily correlated. In other words, the earlier the speaker's age of English onset, independently of their age at recording, the better Otter recognizes their speech. The main effect of gender (β = -4.26e-03, SE = 3.72e-03, t = -1.15, p = 0.253) and the interaction effect between OtterSupported and IsTone (β = -5.05e-03, SE = 6.33e-03, t = -0.80, p = 0.45) did not reach significance.

Taken together, these results suggest that although Otter performed better on the varieties they claimed that are supported, ASR performance is still significantly modulated by languages' phonological structures as well as speakers' social demographic groups.

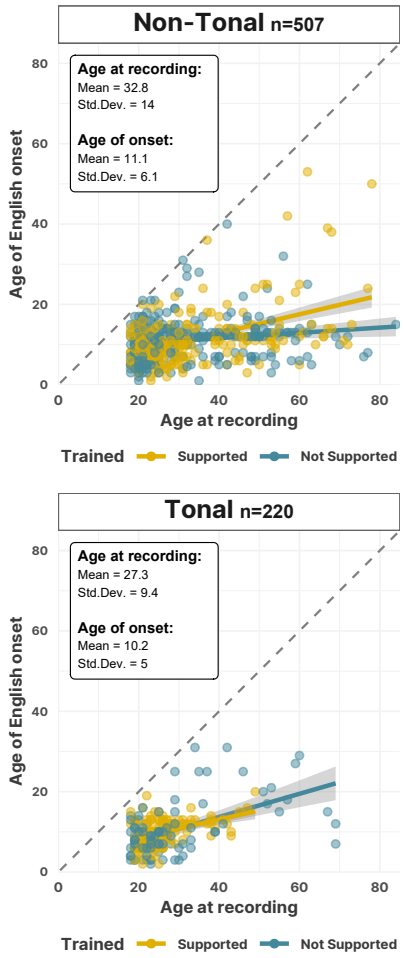

Figure 2: *Summary of speaker demographic information by tonal and non-tonal L1 backgrounds.*

4. Discussion and Conclusion

Overall, the results of the present study reveal systematic biases in modern automated speech recognition systems. There are two main findings in the present work. The first concerns the effect of training. As of June 2021, Otter claims to be able to handle a wide variety of native and non-native varieties of English [14], we thus infer that English varieties that are claimed to be supported have been trained by Otter and should perform better than untrained English varieties. These results are largely borne out, where the mean error rates for varieties of English that are not supported are higher than that of supported varieties, revealing a performance gap. By design of our study, all inputs from each of the 1227 speakers have the identical ground truth as it is the same passage that was being read. Therefore, our findings indicate that the performance gap between the supported and unsupported data set are unlikely to be driven by the underlying language model in the ASR, but from the acoustic model instead. Future work using acoustic features to predict model performance should take this finding into account.

A second main finding is that systematic bias by language structure in ASR performance exists. Namely, in both the supported and unsupported varieties of English, speech from speakers whose first language is a tone language performs worse compared to speech from speakers who speak a non-tonal language. One potential explanation to this would be that lexical pitch differences in speakers' L1 may surface in their English variety (e.g. [17]), which may cause pitch contours to deviate from the prosodic structure that ASR systems are trained to recognize. Further work exploring differences between WER and PER, which reveal ASR performance gaps at a segmental level may provide more conclusive explanations.

Furthermore, our results also show that speakers' age of onset is a significant predictor for Otter's performance as measured by the WER. One way to interpret these results is that age of onset is a proxy for not only the speaker's amount of exposure to English, but also how native-like their English (e.g., prosody) sounds, which is highly correlated with speakers' access to English learning resources in their early years.

In sum, our findings highlight the biases in speech recognition systems that exist not only in speakers' social demographic background, but also in speakers' individual backgrounds and in typological differences. We believe that the mere inclusion of more training data itself containing different English varieties may not be sufficient to eliminate bias in ASR systems, if the linguistic structure of the varieties is not taken into account. With the increasing use of ASR systems globally, we hope that the effort to eliminate bias in ASR will include non-native varieties of English.

5. References

[1] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5934–5938.

[2] B. Wheatley and J. Picone, "Voice across america: Toward robust speaker-independent speech recognition for telecommunications applications," *Digital Signal Processing*, vol. 1, no. 2, pp. 45–63, 1991.

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

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

[5] R. Tatman, "Gender and dialect bias in youtube's automatic captions," in *Proceedings of the first ACL workshop on ethics in natural language processing*, 2017, pp. 53–59.

[6] M. Sawalha and M. Abu Shariah, "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*. Leeds, 2013.

[7] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" in *Ninth European Conference on Speech Communication and Technology*, 2005.

[8] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

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

[10] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *arXiv preprint arXiv:2103.15122*, 2021.

[11] G. Droua-Hamdani, S.-A. Selouani, and M. Boudraa, "Speaker-independent asr for modern standard arabic: effect of regional accents," *International Journal of Speech Technology*, vol. 15, no. 4, pp. 487–493, 2012.

[12] X. Wang, N. Kanda, Y. Gaur, Z. Chen, Z. Meng, and T. Yoshioka, "Exploring end-to-end multi-channel asr with bias information for meeting transcription," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 833–840.

[13] Y. Alotaibi, S.-A. Selouani, and D. O'shaughnessy, "Experiments on automatic recognition of nonnative arabic speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, pp. 1–9, 2008.

[14] A. Lai, "Supported languages," 2021. [Online]. Available: https://help.otter.ai/hc/en-us/articles/360047247414-Supported-languages

[15] S. Weinberger, "Speech accent archive. george mason university," *Online:¡ http://accent. gmu. edu*, 2015.

[16] P. Boersma, "Praat: doing phonetics by computer," *http://www. praat. org/*, 2006.

[17] H. Ding, R. Hoffmann, and D. Hirst, "Prosodic transfer: A comparison study of f0 patterns in l2 english by chinese speakers," in *Speech Prosody*, vol. 2016, 2016, pp. 756–760.