Performance Disparities between Accents in Automatic Speech Recognition (Student Abstract)

Alex DiChristofano [1], Henry Shuster [2], Shefali Chandra [3,4], Neal Patwari [1,2,5]

[1] Division of Computational & Data Sciences, Washington University in St. Louis
[2] Department of Computer Science & Engineering, Washington University in St. Louis
[3] Department of Women, Gender, and Sexuality Studies, Washington University in St. Louis
[4] Department of History, Washington University in St. Louis
[5] Department of Electrical & Systems Engineering, Washington University in St. Louis
a.dichristofano@wustl.edu, henryshuster@wustl.edu, sc23@wustl.edu, npatwari@wustl.edu

Abstract

In this work, we expand the discussion of bias in Automatic Speech Recognition (ASR) through a large-scale audit. Using a large and global data set of speech, we perform an audit of some of the most popular English ASR services. We show that, even when controlling for multiple linguistic covariates, ASR service performance has a statistically significant relationship to the political alignment of the speaker's birth country with respect to the United States' geopolitical power.

Materials and Methods

We evaluate three top cloud-based ASR services, defined in terms of evaluation performance in Koenecke et al. (Koenecke et al. 2020): Microsoft, Amazon, and Google.

Speech Accent Archive

Our recordings come from *The Speech Accent Archive*, a collection of recordings of speakers born across the world and with different first languages all reading the same text. This passage was crafted by linguists to include many of the sounds and most of the consonants, vowels, and clusters that are common to English (Weinberger 2015).

Speaker Information Collected The information on speakers collected at the time of recording includes their age, sex, country of birth, first language, age of onset of English speaking, and whether the speaker's English learning environment was academic or naturalistic. Age of onset is particularly useful, as it has been shown to be correlated with perceived accent (Flege, Munro, and MacKay 1995; Moyer 2007; Dollmann, Kogan, and Weißmann 2019). This speaker-level information is integrated into the regression discussed in the Results.

Data Description Our data set includes 2,713 speakers with an average age of 32.6 years, and an average age of onset of English speaking of 8.9. Our speakers represent 212 first languages across 171 birth countries. The top first languages represented in the data by number of speakers were English (620), Spanish (212), Arabic (164), Mandarin (128), Korean (94), Russian (79), French (74), Portuguese (62), Dutch (51), and German (40).

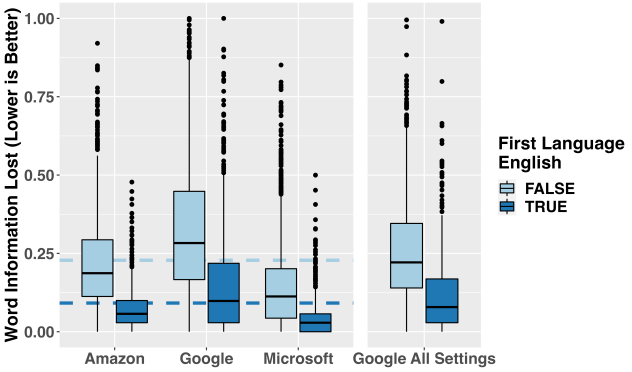

Figure 1: Word Information Lost (WIL) by ASR service and English as the speaker's first language. For each service, WIL is significantly lower when English is a first language.

We note that at the time of recording, 2,023 (74.6%) speakers were either current or previous residents of the United States. By default, most ASR services that would be used on and by these speakers while they are in the United States would likely be configured to use the "United States English" setting for transcription. For the results presented here, we also use this setting as the default. In the Discussion, we discuss the effects of using the most accurate transcription from all language settings.

Results

Group-Level Analysis

In Figure 1, we compare Word Information Lost (WIL) across ASR services grouped by whether a speaker's first language was English. Lower WIL means the ASR service is more accurate. All services performed significantly better ($P < 0.001$) for speakers whose first language was English. On average across all services, WIL was 0.14 lower for first language English speakers. By service, the disparities followed overall performance, with differences of 0.17, 0.14, and 0.10 for Google, Amazon, and Microsoft, respectively.

In Figure 2, we highlight mean ASR service performance for the ten first languages for which we have the most data. The order of performance found in Figure 1 is maintained

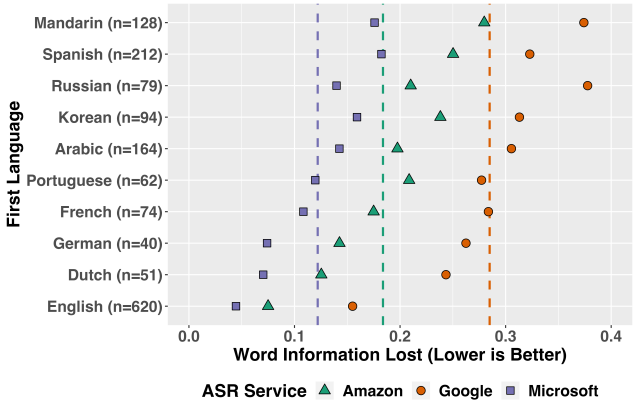

Figure 2: Mean Word Information Lost (WIL) for ASR Services vs. First Language (showing the top 10 first languages by number of observations, sorted by total mean WIL). Mean WIL performance across all speakers for a specific ASR service is shown as a vertical dashed line.

across services — across all ten first languages, Microsoft performs the best, followed by Amazon, and then by Google. We find that all the ASR services perform best for those whose first language is English, followed by Dutch and German. The worst performance is on speakers whose first language is Mandarin, Spanish, or Russian.

Speaker-Level Regression

Motivated by the previous results, we construct a linear regression to understand what factors have a significant effect on the performance of ASR services. We want to know if ASR performance is correlated with how the speaker is perceived from a lens of United States global political power. As a broad single measure for this political power, we encode if the speaker's birth country is a part of the North Atlantic Treaty Organization (NATO) as of January 2022. This avoids encoding each country as a separate regression variable (of which there are 171), and avoids having to develop a measure of United States political power in each country.

Specifically, we include about each speaker in our analysis:

- Age and age of onset of English speaking;
- Sex;
- English learning environment;
- If their first language is Germanic, as a measure of first language similarity to English;
- If their birth country is in NATO.

We create a nested covariate for English and the United States in the Germanic first language and birth country in NATO covariates, respectively.

Regression Results We find multiple covariates that have a significant effect across all three services:

1. WIL significantly increases with a later age of onset of English speaking. As described in the Materials and Methods, age of onset is correlated with perceived accent.
2. Speakers who learned English in a naturalistic environment have a significantly lower WIL over speakers who learned in an academic one.
3. WIL significantly decreases with speaking a Germanic first language, having controlled for the effect on WIL of English as a first language.
4. Finally, being born in a country that is a part of NATO is associated with a significantly lower WIL, having controlled for the effect of being born in the United States.

The final result suggests that a person's birth in a country proximate to the United States' geopolitical power is related to the performance of ASR on their speech, even controlling for the other covariates. This holds for all services tested.

Discussion

For reasons argued in the Materials and Methods, we use the "United States English" setting of the ASR services tested. However, for the service which demonstrated the greatest performance disparity, Google, we evaluate the performance when we take the most accurate recording from all available English transcription settings. The performance is shown in Figure 1 under "Google All Settings". We note that, while performance improves for both first language and non-first language English speakers under this method, a significant disparity in WIL still exists (0.14), and that the same significant covariates discussed in earlier regression analysis remain significant. This suggests that current efforts to offer different global transcription options do not reduce performance disparities as we might expect.

English is a language spoken around the world with many variations, in part due to colonization and globalization, and no speaker can claim to speak English without an accent. These results indicate that major ASR providers provide better service to those born in nations aligned with US political power.

References

Dollmann, J.; Kogan, I.; and Weißmann, M. 2019. Speaking Accent-Free in L2 Beyond the Critical Period: The Compensatory Role of Individual Abilities and Opportunity Structures. *Applied Linguistics*, 41(5): 787–809.

Flege, J. E.; Munro, M. J.; and MacKay, I. R. A. 1995. Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, 97(5): 3125–3134.

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.

Moyer, A. 2007. Do Language Attitudes Determine Accent? A Study of Bilinguals in the USA. *Journal of Multilingual and Multicultural Development*, 28(6): 502–518.

Weinberger, S. 2015. Speech Accent Archive. https://accent.gmu.edu. Accessed: 2021-04-01.