How Well Can an Agent Understand Different Accents?

Divya Tadimeti[1,2], **Kallirroi Georgila**[2], **David Traum**[2]
[1]Electrical Engineering & Computer Science, University of California, Berkeley
[2]Institute for Creative Technologies, University of Southern California
dtadimeti@berkeley.edu [kgeorgila,traum]@ict.usc.edu

Abstract

We evaluate several state-of-the-art automatic speech recognition systems on dialogue agent-directed English speech from speakers with General American vs. non-American accents. Our results show that the performance of the speech recognizers for non-American accents is considerably worse than for General American accents, with $\sim$20% higher word error rate on average (relative difference). This work indicates a need for more diligent collection of and training on non-native English speaker data in order to narrow this performance gap. There are performance differences across recognizers, and while the same general pattern holds, with more errors for non-American accents, there are some accents for which the best recognizer is different than in the overall case. We expect these results to be useful for dialogue system designers in developing more robust inclusive dialogue systems, and for speech recognition providers in taking into account performance requirements for different accents.

1 Introduction

Automatic speech recognition (ASR) systems are being used for an increasing number of speech-to-text applications. With this proliferation, it is increasingly important for this technology to serve all subgroups of consumers. Recent work has shown that speech recognizers have a much higher error rate on speakers of African American Vernacular English (AAVE) than on rural White Californians engaging in sociolinguistic interviews (Koenecke et al., 2020). Recent evaluation of ASR systems on speech directed at computer agents (Georgila et al., 2020) shows that speech recognizers have been getting better recently on agent-directed speech compared to previous years (Yao et al., 2010; Morbini et al., 2013), but leaves open the question of whether this performance is equivalent for different speakers or whether the pattern observed by (Koenecke et al., 2020) also holds for other kinds of accents and for agent-directed speech. To this end, we evaluate the performance of popular ASR platforms — Google, Microsoft, Apple, Amazon, and IBM — on English speech from populations with different accents. We begin with a high-level distinction between general American accents and non-American accents, and then focus on more specific categories of non-American accents including French, Indian, British, and East Asian accents. We report on a re-analysis of a subset of the data examined by Georgila et al. (2020), with new annotations for speaker identity and accent.

2 Data and Methods

We evaluated the ASR systems on a dataset of 2500 utterances collected from conversation between human participants and SGT Blackwell, a virtual agent developed by the USC Institute for Creative Technologies. SGT Blackwell (Leuski et al., 2006) is a question-answering character who answers general questions about the Army, himself, and his technology. Speech comes from visitors to the Cooper-Hewitt Museum in New York from December 2006 to March 2007, who interacted with SGT Blackwell at his booth as part of the National Design Triennial exhibition (Robinson et al., 2008). In this museum setting open to the general public, it was assumed that the majority of visitors would be American native

Annotators and labelling setup	Krippendorff's alpha	Absolute agreement (%)
Annotators 1, 2, 3 (American & rest of categories)	0.719	76.43
Annotators 1, 2, 3 (American & Else)	0.879	95.33
Annotators 1, 2 (all 8 distinct categories)	0.672	71.34
Annotators 1, 2 (General & Northeast American & Else)	0.8	91.72
Annotators 1, 2 (American & rest of categories)	0.712	75.80
Annotators 1, 2 (American & Else)	0.9	96.18
Annotators 1, 3 (American & rest of categories)	0.719	76.43
Annotators 1, 3 (American & Else)	0.835	93.63
Annotators 2, 3 (American & rest of categories)	0.725	77.07
Annotators 2, 3 (American & Else)	0.901	96.18

Table 1: Krippendorff's alpha values and absolute agreement percentages for different comparisons. "American" means that the General American and Northeast American accents are merged into one category. "Else" means that all non-American accents are merged into one category.

speakers of English. Thus, the ASR component of SGT Blackwell used acoustic models for standard American English. Similar to this setup, in our experiments below we use commercial ASR systems with default settings for standard American English accents.

Speakers were anonymous and not identified in the data. In order to categorize the speech by accent, we listened to every audio file. Using this method, we classified the audio files into two main groups: General American English and non-American English accents. We use the term "General American" to encompass the utterances in our dataset lacking distinct regional and social characteristics (Wells, 1982; Van Riper, 1986). This includes mostly Western and Midwestern English accents and excludes noticeably Northeastern accents (i.e., New York, Boston), Southern American accents, and distinct dialects such as AAVE. Next, we segmented the non-American subset further into subcategories of non-American accents, the most common of which were French, British, Indian, and East Asian. In some cases, it was not possible to distinguish the precise accent, so we also included an "uncategorized" class. For each non-American subset of files, we grouped utterances by individual speakers for additional analysis. These categories were then used to compute category-specific error rates for the recognizer results reported in (Georgila et al., 2020). Word Error Rate (WER) is a standard measure of ASR performance, used for example by both Koenecke et al. (2020) and Georgila et al. (2020).

To assess inter-annotator reliability of accent classification, the three authors listened to a subset of 157 audio files and annotated the accent in each file as General American, Northeast American, British, Indian, French, East Asian, European uncategorized, and non-American uncategorized (8 distinct categories). Two of the annotators (Annotators 1 and 2) were American native speakers of English, and the third annotator was a non-native but fluent speaker of English (Annotator 3). Agreement results between annotators are shown in Table 1. Krippendorff's alpha between Annotators 1 and 2 was measured at 0.672 (with absolute agreement at 71.34%) when all 8 distinct categories were considered. Krippendorff's alpha among all 3 annotators was measured at 0.719 (with absolute agreement at 76.43%) when General American and Northeast American accents were merged into one "American" category. We also calculated pairwise inter-rater agreement scores after merging the General American and Northeast American accents into one "American" category, and after merging all non-American accents into one large category "Else". The results shown in Section 3 are based on the annotations of Annotator 1.

3 Results

In Figure 1, we see that ASR performance is worse with non-American accents for all recognizers. Notably, for both General American and non-American accents, Google performs the best and IBM performs the worst. In Figure 2, we see that the general pattern does not always hold for each accent category. Apple performs slightly better than Google for British and East Asian accents, while IBM is

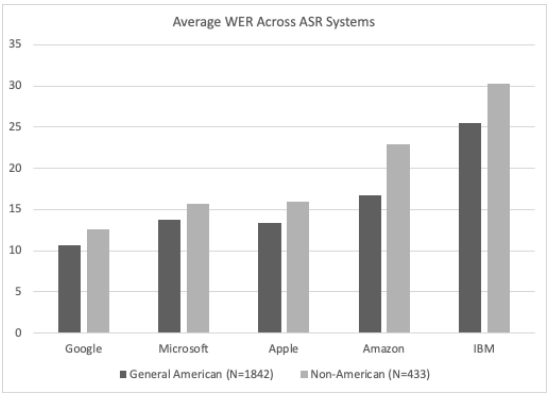

Figure 1: Average WERs across ASR systems.

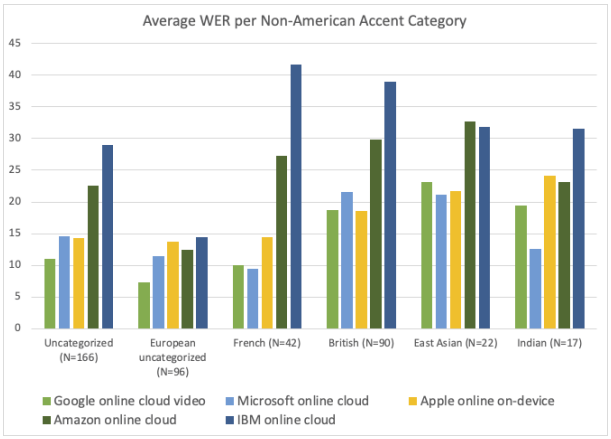

Figure 2: Average WERs for non-American accents.

better than Amazon for East Asian accents. Microsoft is also better than Google for French, East Asian, and Indian accents, and is the best among all ASR systems for Indian accents. All differences in WERs shown in Figure 1 are statistically significant (two-sided Wilcoxon test with Bonferroni correction, each sample is a list of WERs, one value per utterance).

We also calculated WERs for regional American accents (105 audio files) and the results were as follows: 12.06% for Google, 18.27% for Microsoft, 22.86% for Apple, 16.52% for Amazon, and 24.2% for IBM. We can see that the performance of the Google, Amazon, and IBM recognizers for regional American accents is similar to their performance for General American accents. It is interesting that this is not the case for the Microsoft and Apple recognizers. For these two recognizers, WERs for regional American accents are even higher than WERs for some non-American accents, although it is unclear if this effect generalizes or it is just a result of the small size of our dataset.

4 Discussion and Future Work

All ASR systems perform fairly well for General American accents, but do considerably worse for non-American accents. The performance gap suggests that consumers with non-American English accents may find it considerably harder to take advantage of speech recognition technology. It is an open research question and an active area of research whether speech recognizers should be expected to perform equally well for native and non-native speakers of a language, in our case American English (Le et al., 2007; Ghorbani and Hansen, 2018; Jain et al., 2018; Viglino et al., 2019; Ahamad et al., 2020). Nevertheless, the performance gap shown in our results is wide enough to suggest that there is potential for improvement. To improve performance, ASR systems should be trained on more diverse speaker data (Fukuda et al., 2018). This requires more diligent collection of non-American English speaker data.

The above analysis is still preliminary in several respects. We are currently analyzing the errors of individual speakers and also calculating the impact of these speech errors on agent response selection. We would like to enhance the analysis by looking at additional domains of agent-directed speech and additional demographic groups (such as regional American accents, gender, age, etc.). Additionally, we would like to attempt a more objective approach to accent categorization, e.g., using databases such as eWAVE to make more linguistically-informed data categorizations, or analyzing or collecting data with demographic information about the speakers.

Acknowledgements

The first author was supported by the NSF REU program, award 1852583, during her internship at the USC Institute for Creative Technologies. The second and third authors were sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Afroz Ahamad, Ankit Anand, and Pranesh Bhargava. 2020. AccentDB: A database of non-native English accents to assist neural speech recognition. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 5351–5358, Marseille, France.

Takashi Fukuda, Raul Fernandez, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, Alexander Sorin, and Gakuto Kurata. 2018. Data augmentation improves recognition of foreign accented speech. In *Proceedings of the 19th Annual Conference of the Speech Communication Association (INTERSPEECH)*, pages 2409–2413, Hyderabad, India.

Kallirroi Georgila, Anton Leuski, Volodymyr Yanov, and David Traum. 2020. Evaluation of off-the-shelf speech recognizers across diverse dialogue domains. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 6469–6476, Marseille, France.

Shahram Ghorbani and John H.L. Hansen. 2018. Leveraging native language information for improved accented speech recognition. In *Proceedings of the 19th Annual Conference of the Speech Communication Association (INTERSPEECH)*, pages 2449–2453, Hyderabad, India.

Abhinav Jain, Minali Upreti, and Preethi Jyothi. 2018. Improved accented speech recognition using accent embeddings and multi-task learning. In *Proceedings of the 19th Annual Conference of the Speech Communication Association (INTERSPEECH)*, pages 2454–2458, Hyderabad, India.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 117(14):7684–7689.

Jennifer T. Le, Catherine T. Best, Michael D. Tyler, and Christian Kroos. 2007. Effects of non-native dialects on spoken word recognition. In *Proceedings of the 8th Annual Conference of the Speech Communication Association (INTERSPEECH)*, pages 1589–1592, Antwerp, Belgium.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. Building effective question answering characters. In *Proceedings of the 7th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 18–27, Sydney, Australia.

Fabrizio Morbini, Kartik Audhkhasi, Kenji Sagae, Ron Artstein, Doğan Can, Panayiotis Georgiou, Shri Narayanan, Anton Leuski, and David Traum. 2013. Which ASR should I choose for my dialogue system? In *Proceedings of the 14th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 394–403, Metz, France.

Susan Robinson, David Traum, Midhun Ittycheriah, and Joe Henderer. 2008. What would you ask a conversational agent? Observations of human-agent dialogues in a museum setting. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 1125–1131.

William R. Van Riper. 1986. General American: An ambiguity. *Dialect and Language Variation*, pages 123–135.

Thibault Viglino, Petr Motlicek, and Milos Cernak. 2019. End-to-end accented speech recognition. In *Proceedings of the 20th Annual Conference of the Speech Communication Association (INTERSPEECH)*, pages 2140–2144, Graz, Austria.

John C. Wells. 1982. *Accents of English, Volume 3: Beyond the British Isles*. Cambridge University Press.

Xuchen Yao, Pravin Bhutada, Kallirroi Georgila, Kenji Sagae, Ron Artstein, and David Traum. 2010. Practical evaluation of speech recognizers for virtual human dialogue systems. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 1597–1602, Valletta, Malta.