Automatic speech recognition: system variability within a sociolinguistically homogenous group of speakers

Lauren Harrington, Vincent Hughes

Department of Language and Linguistic Science, University of York, UK
lauren.harrington@york.ac.uk, vincent.hughes@york.ac.uk

ABSTRACT

Orthographic transcription is a time-consuming task that is increasingly being automated for various applications. A growing body of work on algorithmic bias has highlighted variability in automatic speech recognition performance across demographic groups, but little research has focused on variability within a sociolinguistically homogenous group.

This paper considers this issue and explores the extent to which phonetic properties of the voice (e.g. f0, formants) can predict how well an automatic system performs. Recordings of 99 young, male Southern British English speakers were automatically transcribed using Amazon Transcribe, and word error rate was calculated.

A wide range of error rates (between 11-33%) is observed. However, of all acoustic, temporal and voice quality measures analysed, only articulation rate has a significant impact on error rate. System performance therefore remains difficult to predict within demographic groups.

Keywords: automatic transcription, automatic speech recognition

1. INTRODUCTION

Automatic speech recognition (ASR) technology is used increasingly for a variety of applications, including live captioning, virtual assistants, and transcribing professional meetings. The field of ASR has received an increasing amount of interest in recent years (see [1] for an overview) and is now a huge area of research, with modern systems, integrating state-of-the-art machine learning techniques, demonstrating consistent improvements in overall performance year-on-year. Systems tend to be evaluated on the basis of overall word error rate (WER; the percentage of errors in a transcript relative to words spoken in the audio recording), with recent reports of WERs as low as 5% [2] on the Switchboard corpus, a dataset commonly used for ASR performance analysis which contains General American English conversational telephony data.

There is now a growing focus on understanding how and why ASR systems perform in the way that they do, which has led to the development of a body of work exploring algorithmic bias in ASR systems [3,4]. Some studies have investigated system performance as a function of speaker demographic factors such as accent [5,6] and gender [7], and have demonstrated significant differences in performance across groups. The issue of variability across demographic groups is an area of concern within the field and work on mitigating the effects of performance disparities is gaining popularity [8,9].

The present study explores an issue which has received relatively little attention within the field of ASR: the extent of individual variability in system performance across a sociolinguistically homogeneous set of speakers, matched for age, gender, level of education and regional accent. Here, we keep confounding factors such as speech content and audio quality consistent, in order to isolate differences in performance due to characteristics of the speech and speaker. This work explores the extent to which linguistic properties of a speaker's voice may predict speech recognition performance, using a state-of-the-art, and widely used, commercial system, Amazon Transcribe. To our knowledge, no work has yet considered how the phonetic properties of speech may be contributing to system variability.

In this paper we address two research questions: (1) what is the range of variability in ASR performance across a set of demographically homogeneous speakers? And (2) is any of the variability predictable based on well-understood phonetic properties of a speaker's speech production?

2. METHODOLOGY

2.1. Stimuli

Data from the Dynamic Variability in Speech (DyViS) [10] project was used for this study. DyViS is a large database of speech collected under simulated forensic conditions, consisting of recordings of 100 university-educated young adult male speakers of Standard Southern British English (SSBE) taking part in a range of tasks. The speakers are demographically homogeneous and are generally very similar sounding [11]; indeed, speakers with atypical features for SSBE were removed in a screening phase when collecting the corpus [10]. DyViS task 2 was used for the present study. It

consists of a telephone conversation with an 'accomplice', played by a Research Assistant, in which the participant discusses his experience of a police interview (a previous task) so that the accomplice can 'get their story straight'. Details of the crime discussed in both the police interview and telephone conversation were presented to participants in the form of visual prompts, therefore the content of the speech is relatively controlled across speakers.

Studio recordings with a sampling rate of 44.1 kHz and minimal background noise were used for this study. We decided to use high quality samples in order to maximise the performance of the ASR system and, therefore, minimise the confounding effects of channel and background noise. Speech of the 'accomplice' and non-speech sounds, e.g. coughs and inhales/exhales, were removed such that each file contained only the speech of the participant. One speaker was removed due to issues with his data, resulting in 99 stimuli between 4 and 11 minutes in length.

2.2. Transcripts

Ground truth reference transcripts were based on orthographic transcriptions completed by researchers on the DyViS project and provided with the database in the form of Praat [12] TextGrids for each audio file. The content of the TextGrids was extracted using a Praat script and each resulting transcript was saved in .txt format and manually checked for spelling errors. Minor adjustments were made to correct spelling errors and to ensure consistency in the representation of certain words or phrases, e.g. some transcripts contained "DIY" while others contained "D I Y".

The audio files were automatically transcribed using a popular commercially-available ASR service, Amazon Transcribe via Amazon Web Services. The "en-GB" (British English) language model was used for all transcriptions. As a standard variety, we expected that the system should perform optimally with SSBE compared with non-standard varieties of British English. Again, the intention was to maximise the potential performance of the system under favourable conditions. The Amazon output for each file was saved in .txt format and minor adjustments were made to the representation of filled pauses to ensure consistency between both sets of transcripts (these are referred to below as 'hypothesis' transcripts).

2.3. Performance evaluation

Python package JiWER [13], which computes the minimum edit distance between two strings of text, was used to compare the reference and hypothesis transcripts for each audio file and to calculate the word error rate (WER) for each speaker. A lower WER demonstrates that fewer errors were made. Microsoft Azure documentation [14] posits that a WER of 5-10% is considered good quality, while error rates of over 30% signal bad quality and that the system requires further training and customisation.

A more detailed evaluation of the speakers that produced the best and worst WERs, DyViS speakers 30 and 26 respectively, was carried out to explore the types of errors produced. Reference and hypothesis transcripts were aligned on a word-level basis and word pairs were marked as a match or as one of three types of error: deletion, substitution or insertion.

2.4. Auditory and acoustic measurements

To explore the effects of phonetic properties of speech production on automatic transcription performance, a range of acoustic and auditory measurements for the speakers were collected from existing sources. These were f0, long term formant distributions (F1~F4), and articulation rate from [15] and Vocal Profile Analysis of laryngeal and supralaryngeal voice quality from [11]. One speaker was excluded from further analysis due to an issue with the audio file and inconsistency between the speech content and corresponding transcription, resulting in a total of 99 speakers.

Mean values (and standard deviation for f0) for each speaker were calculated for the acoustic measures. The VPA data provides an array of information about voice quality features from the whole vocal tract. Speakers were assigned a score from 0 to 3 for each setting, e.g. "fronted tongue body" and "raised larynx", to represent the extent to which that setting was present in the speaker's voice. For the purposes of this study, we collapsed the features to produce two voice quality categories, supralaryngeal and laryngeal, and calculated the Euclidean distance for each speaker's profile from the group mode for each category. This provides a single value indicating how unusual a speaker's laryngeal and supralaryngeal profiles are from the average for the group.

2.5. Statistical analysis

A multiple linear regression model was fitted using the *lm* function in R [16] to predict word error rate using each of the phonetic measures as fixed effects. Mixed effects regression was not conducted as there was only one data point per speaker.

3. RESULTS

3.1. Variability in system performance

Performance was measured by calculating WER for each speaker, where a higher percentage indicates a higher proportion of errors to correctly transcribed words. WERs ranged from 11.2% to 33% with a mean error rate of 20% across speakers. Figure 1 displays the variability of the word error rates, and shows a relatively normal distribution across the range. The variability itself is extremely large, especially given the favourable conditions in terms of audio quality and accent. Since the speakers are matched for multiple demographic factors and the content of the utterances is relatively controlled, the variability of WER within the group suggests that characteristics related to a specific speaker's voice must, to some extent, be responsible for the wide range of error rates.

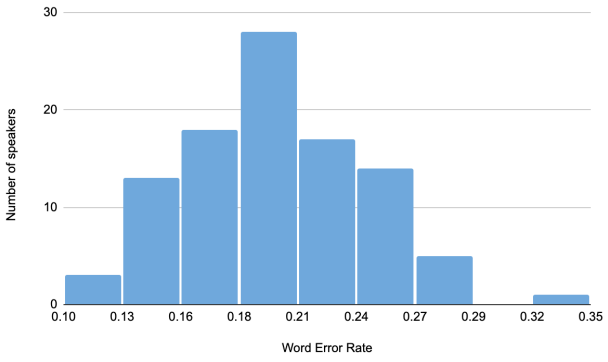

Figure 1: Distribution of word error rate across 99 DyViS speakers.

The number and types of errors most frequently made by Amazon Transcribe also varied across speakers. These can be grouped as (i) deletions (e.g. *going to the steak house → going to steak house*), (ii) substitutions (e.g. *hope avenue → pope avenue*) and (iii) insertions (e.g. *about telephone booth → about the telephone booth*). Table 1 displays the number of each type of error produced for speakers 30 and 26, representing those for whom the system performed best and worst. The total number of errors for speaker 26 is over double that of speaker 30, and the proportion of substitution errors is much higher. A higher number of the substitution errors were related to proper nouns (e.g. place names) for speaker 26 (42 compared with 23), and the system transcribed many names incorrectly for speaker 26 but correctly for speaker 30.

Speaker	WER	DEL	SUB	INS	Total
30	11.2%	115	69	13	197
26	33.0%	216	215	28	459

Table 1: Number of each error type for the speakers for whom Amazon Transcribe performed best (11.2%) and worst (33.0%).

3.2. Phonetic properties

Multiple linear regression was used to test if the phonetic properties listed in 2.4 significantly predicted word error rate. No significant effects were found for mean f0 (β = -0.00055, p = 0.14), f0 standard deviation (β = 0.00136, p = 0.22) or long-term formant distributions for any of the first four formants (F1: β = 0.00007, p = 0.55; F2: β = -0.00012, p = 0.18; F3: β = 0.00004, p = 0.42; F4: β = -0.00004, p = 0.16). The VPA scores of distance from the group mode did not significantly predict WER for either supralaryngeal (β = 0.00599, p = 0.29) or laryngeal (β = 0.00465, p = 0.44) features.

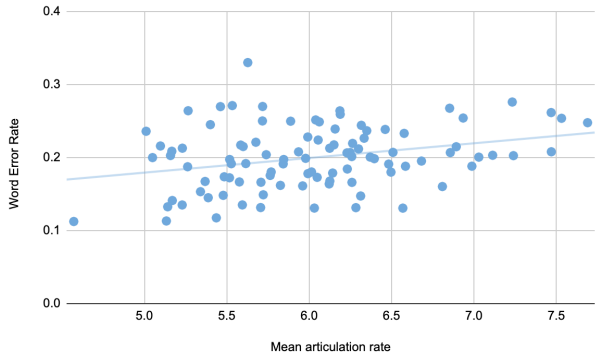

Figure 2: Relationship between word error rate and mean articulation rate (syllables uttered per second).

The only phonetic feature to significantly predict WER was articulation rate (β = 0.01844, $p < .01$) for which there was a positive correlation, demonstrated in Figure 2. Speaker 30, for whom Amazon Transcribe achieved the lowest WER, had the lowest articulation rate in the group (4.6 syllables per second) which was 0.4 lower than the second lowest articulation rate observed within the group and 3.1 lower than the highest.

4. DISCUSSION

Not much is known about why systems may perform better for some speakers than others. This study has provided an initial exploration into the amount of variability observed within a group of speakers matched for demographic factors such as age, sex and level of education, whilst controlling for audio quality and content. We find that there is a high level of variability across speakers, with word error rates

ranging from 11% to 33% and the quality of transcripts varying across speakers. The range of variability itself is worrying given the favourable conditions in the present study and raises issues about the general utility of ASR systems for many applications. Such variability also raises a question about the value of a single WER as a measure of overall system performance across lots of samples and speakers, as is typical in benchmarking systems. Minimally, it would be useful for developers to report the variability in system performance across speakers.

Furthermore, WER is unable to provide information about the types of errors made by the system and how impactful they may be to the meaning conveyed. By more closely inspecting the reference and hypothesis transcripts, it is clear that some transcripts contain many more critical errors than others. For example, the hypothesis transcript for speaker 26, for whom Amazon Transcribe achieved the highest WER, contained several substitution errors which could completely change the meaning, such as *Rose has picked me up* in place of *Rozzers* [police officers] *picked me up*. Meanwhile, the majority of substitutions for speaker 30, for whom Amazon performed best, were related to minor representational issues such as grammatical corrections (e.g. *'cause* → *because*) or contractions/expansions (e.g. *she's* → *she is*). Proper nouns, such as names of colleagues or place names, were a reliable source of errors for Amazon Transcribe. Spelling mistakes and other substitutions were common, although there were many cases where a particular name proved problematic for some but not all speakers. For example, *Weasley* was uttered in 91 stimuli and correctly transcribed at least once in 71, but for speaker 26 this name was substituted with *easily* and *weezy*.

The results in 3.2 reveal that only articulation rate is significantly correlated with WER, and other phonetic properties such as mean f0 cannot significantly predict WER. The lack of acoustic correlates with WER is worrying because there is only limited phonetic evidence as to why a given speaker may be 'easy' or 'difficult' to transcribe. State-of-the art ASR systems extract abstract features and implement deep learning within their training, contributing to the 'black box' problem (i.e. the extent to which the processing is opaque and scrutinizable). This in turn raises the general issue of interpretability; without knowing the cause of variability within a group, it is challenging to improve systems and interpret their results.

5. CONCLUSION

The results of this study show a large amount of ASR performance variability across a set of homogenous speakers, despite keeping several possible confounding factors consistent. Variability is found in overall error rates as well as the number and types of errors produced across speakers. Performance variability across speakers cannot be predicted by well-known acoustic phonetic properties, such as fundamental frequency or long-term formant distribution, or by properties related to voice quality. Only the temporal property, articulation rate, is able to significantly predict error rate, whereby more syllables uttered per second is significantly related to the ratio of errors produced relative to words in the reference transcript. In future work it would be interesting to explore whether other temporal properties, such as disfluency, may be responsible for an increase in error rates.

ACKNOWLEDGMENTS

This research was funded by the White Rose College of Arts and Humanities, a Doctoral Training Partnerships supported by the Arts and Humanities Research Council (AHRC). Thanks also to Phil Harrison for support with Praat scripting, and to Ryan Willis for technical support with Amazon Transcribe and JiWER.

6. REFERENCES

[1] Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Almojil, M. 2021. Automatic speech recognition: Systematic literature review. *IEEE Access* 9, 131858–131876.

[2] Tüske, Z., Saon, G., Audhkhasi, K., Kingsbury, B. 2020. Single headed attention based sequence-to-sequence model for state-of-the-art results on switchboard. arXiv preprint arXiv:2001.07263.

[3] Shah, D., Schwartz, H. A., Hovy, D. 2019. Predictive biases in natural language processing models: A conceptual framework and overview. arXiv preprint arXiv:1912.11078.

[4] Suresh, H., Guttag, J. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and access in algorithms, mechanisms, and optimization*, 1–9.

[5] Markl, N. 2022. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* Seoul, 521–534.

[6] DiChristofano, A., Shuster, H., Chandra, S., Patwari, N. 2022. Performance Disparities Between Accents in Automatic Speech Recognition. arXiv preprint arXiv:2208.01157.

[7] Tatman, R., Kasten, C. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. *Proc. Interspeech* Stockholm, 934–938.

[8] Dheram, P., Ramakrishnan, M., Raju, A., Chen, I.-F., King, B., Powell, K., Saboowala, M., Shetty, K., Stolcke, A. 2022. Toward Fairness in Speech Recognition: Discovery and mitigation of performance disparities. *Proc. Interspeech* Incheon, 1268–1272.

[9] Zhang, Y., Zhang, Y., Halpern, B., Patel, T., Scharenborg, O. 2022. Mitigating bias against non-native accents. *Proc. Interspeech* Incheon, 3168–3172.

[10] Nolan, F., McDougall, K., de Jong, G., Hudson, T. 2009. The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language & the Law* 16(1), 31–57.

[11] San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V., Kavanagh, C. 2019. The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals. *Journal of the International Phonetic Association* 49(3), 353–380.

[12] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 10 December 2022 from http://www.praat.org/.

[13] Vaessen, N. 2022. JiWER: Similarity measures for automatic speech recognition evaluation. Version 2.5.1, retrieved 16th November 2022 from https://pypi.org/project/jiwer/.

[14] Microsoft Azure. 2022. Test accuracy of a Custom Speech model. https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data?pivots=speech-studio.

[15] Gold, E. 2014. Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters. PhD thesis, University of York.

[16] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.