BIASES AND SPEECH-TO-TEXT EFFICACY FOR BRITISH ENGLISH VARIETIES

Kyra Hung[a], Amanda Cardoso[b], Devyani Sharma[c], Erez Levon[d]

[a,b]University of British Columbia, [c]Queen Mary University of London, [d]University of Bern
[b]amanda.cardoso@ubc.ca, [c]d.sharma@qmul.ac.uk, [d]erez.levon@unibe.ch

ABSTRACT

Multiple factors impact speech-to-text (STT) performance. Previous work demonstrated poorer performance for North American marginalised English varieties [1] and faster speech rates [2]. We examine STT efficacy of British English varieties, considering: i) the varieties' social statuses; ii) individual repertoires; and iii) speech rate to better understand how potential biases relate to STT efficacy. Speech rate differences impact STT efficacy most with faster (compared to slower) speech rates performing worse. These effects interact with varieties' status and individual repertoires. STT performs better for non-marginalised compared to marginalised varieties with similar speech rates. Marginalised variety speakers who use more accent features have worse STT performance compared to speakers who use fewer accent features. For these speakers, slower speech rates also do not consistently improve performance, which is not found for non-marginalised speakers. While speech rate impacts STT efficacy, we also find accent bias through the differential performance for varieties and individuals.

Keywords: accent bias, speech rate, google speech-to-text, non-standard varieties, accentedness

1. INTRODUCTION

Investigations of differences in automatic speech recognition and speech-to-text performance, which is termed "efficacy" in industry discussions, have found evidence of social biases (e.g., [1, 3, 4]). These investigations suggest that performance differences relating to a number of social variables, such as gender, race, and social status of North American varieties, demonstrate social biases that are upheld in the datasets used in the back-end of the speech recognition systems (e.g., in training) through inclusion and/or exclusion of particular speakers. For example, STT performs worse with speakers of marginalised varieties (often also referred to as non-standard) compared with speakers of non-marginalised ones (often also referred to as standard).

However, some inconsistencies are reported in the results across investigations, either in the magnitude or presence of STT efficacy differences relating to these social variables (e.g., [4]). These inconsistencies across studies may, in part, be due to different speakers being used in the investigations when assessing STT efficacy, as individuals who speak the "same" variety will produce different features or have different levels of use of features that are generally associated with those varieties. In other words, individual repertoires will mean that every speaker of a variety sounds different, to some extent, from every other speaker of that variety. These differences could relate to STT efficacy, in particular, for speakers that use many features that are stigmatised, associated with lower status, or associated with marginalised groups. These speakers are more likely to be excluded in the training data due to social biases and access to speakers.

STT efficacy is also affected by a range of other factors, such as speech rate [2] (see [5] for a review). It has been found that faster speech impacts speech recognition efficacy. Speech rate differences for individuals and varieties have so far not been considered when assessing performance differences in speech recognition systems and social biases. However, there may be differences in the typical speech rate across varieties and across individuals, which could contribute to the STT efficacy differences that are found comparing varieties.

Our investigation considers these three factors (variety social status, individual repertoires, and speech rate) by exploring STT efficacy differences for a content-controlled (i.e., same lexical content) set of recordings by speakers of British English varieties. We examine whether the type of accent bias found for North American varieties hold in this sample (i.e., worse performance for marginalised varieties), what the effect of individual repertoires is

on STT performance (i.e., differential use of accent features) and what the effect of speech rate is on its own and when considering variety social status and individual repertoires. Based on previous work, we would predict that STT efficacy will be better for: i) non-marginalised varieties compared with marginalised ones; ii) speakers whose individual repertoires include fewer marginalised features compared with those that include more marginalised features; iii) slower speech rates compared with faster ones.

2. METHODS

STT outputs were generated using Google API STT (used by, e.g., YouTube's auto-captions) for scripted responses to 15 interview questions performed by 10 male speakers of 5 British English varieties (2 speakers per variety) which are high quality recordings created for the Accent Bias Britain project (https://accentbiasbritain.org/; for more details about the data see [6]). These scripted responses are the same for all speakers, which limits content or lexical differences and allows for a more controlled comparison of STT performance differences. The five varieties represent intersections of different socially-relevant contrasts in the UK, such as region and social status. Estuary English (EE), Multicultural London English (MLE) and Received Pronunciation (RP; aka Queen's English) are all southern varieties with the first two being marginalised. General Northern English (GNE) and Urban West Yorkshire English (UWYE) are northern varieties with the first being non-marginalised. MLE is also considered a multiethnic urban variety [7], while all the others are associated with white ethnicity.

The STT outputs are compared to the spoken transcripts to calculate word error rate (WER) for each speaker and for each question by speaker. WER is used in academic and industry assessments of STT efficacy (see, e.g., [1]) and is calculated by summing the number of deletions (i.e., a word appears in the spoken transcript but not in the STT output), insertions (i.e., a word appears in the SST output but not in the spoken transcript), and substitutions (i.e., a different word appears in the spoken transcript compared with the STT output) divided by the number of words in the full spoken transcript. Higher WERs indicate worse performance.

A dialect density measure (DDM) is used to quantify differences in individual repertoires by calculating the extent of use of accent features. Accent features here refer to ways of producing sounds or combinations of sounds that differ from careful speech and are associated with, at least, one British English variety. Two question responses (Q6 and Q13) for each speaker were coded by two trained phoneticians for presence of 61 accent features, which encompassed a wide variety of features ranging from very localisable (i.e., associated with one variety) to completely non-marginalised (i.e., occurs regularly in all British English varieties). These features were chosen as they are reported in previous literature to occur in one or more of the varieties (see [8] for more details about accent features chosen). For example, we coded for /k/-backing which occurs only in MLE, merged *and* split FOOT-STRUT vowels which differentiate northern and southern varieties, /ð/-fronting which occurs in the northern and southern marginalised varieties, and intrusive-r which occurs regularly in all British English varieties (see [8] for more information about DDMs, where we also discuss relative stigmatisation of features and their effect on perceived accentedness). The DDM value is the sum of all accent features used by the speaker divided by the number of times that an accent feature could have been used based on the known patterns for these accent features. This measure correlates well with differences in perceived accentedness [8] and quantifies differences in individual repertoires within and across varieties. Using a proportional measure accounts for differences in the likelihood of accent features occurring across the varieties, as all features do not occur equally likely in all varieties. Furthermore, DDMs are scaled to allow us to better compare across varieties as well as across speakers within a variety.

Finally, for a measure of speech rate, the average phone duration for each utterance for the two interview responses used in the DDM calculations (Q6 and Q13) was calculated. Speech was divided into utterances using pause length. Pauses greater than 300 ms were taken to indicate a new utterance. Then average phone duration is taken across utterances by speaker and by question for each speaker. Average phone duration across utterances as a proxy for speech rate has been used in previous work and is termed inverse speaking rate (ISR). Higher numbers indicate slower speech and lower numbers indicate faster speech, which is different from many other speech rate measures. ISR compared with other speech rate measures have also been used in discussions of text-to-speech modeling and their improvement (e.g., [9]).

3. RESULTS

Many speakers vary considerably in the STT WERs and DDMs across different questions, while ISR is more consistent across speakers. RP1 is the only speaker that has a difference of more than 0.003 (3 ms) in ISR across the questions (0.066, 0.076). Table 1 indicates the word error rate (WER) and inverse speaking rate (ISR) for each speaker, and dialect density measures (DDM) for questions by speaker and provides the social contrasts (labeled "SC"; i.e., region and social status) associated with the variety for each speaker. Standard deviations are provided for WER and are calculated from the WER values for each question by the speaker. The highest value in each of the different measures is indicated through bold font and the lowest value is indicated through underlining. Recall that high values indicate poorer performance for WER, slower speech for ISR, and more accent feature use for DDMs. Therefore, according to our predictions based on previous work, the best SST performance (i.e., lowest WER) should occur with higher ISR, lower DDMs, and non-marginalised varieties. Note that marginalised speakers are listed at the top of table and non-marginalised are at the bottom, and within each group northern speakers are listed first.

ID	SC	WER(sd)	DDM	ISR
UWYE1	NM	0.09(0.03)	0.88,**1.40**	**0.072**
UWYE2	NM	0.10(0.05)	0.23, 1.36	**0.072**
EE1	SoM	0.22(0.10)	-0.14,0.50	0.067
EE2	SoM	0.12(0.05)	-0.37,-0.28	0.069
MLE1	SoM	0.15(0.07)	-0.81,-0.15	0.064
MLE2	SoM	**0.31**(0.11)	0.94,1.17	0.062
GNE1	NSt	0.11(0.04)	-0.87,0.33	0.071
GNE2	NSt	0.18(0.08)	0.75,1.0	0.063
RP1	SoSt	0.16(0.06)	-1.71,-0.17	0.070
RP2	SoSt	0.18(0.12)	-1.71	0.068

Table 1: Summary of word error rates, scaled dialect density measure, and inverse speaking rate for each speaker ("ID"). "SC" indicates main social contrasts for each variety and is coded as So=southern, N=northern, St=non-marginalised, M=marginalised. Col. 3 has speaker word error rate (WER) with standard deviation in brackets (higher=worse & lower=better performance). Dialect density measures (DDM) for each question (higher = more & lower = fewer accent features used) is in col. 4. Average speaker inverse speaking rate (ISR) is provided in col. 5 (higher = slower & lower = faster speech rate). Bold font = highest and underlined = lowest WER, DDM and ISR.

Speech rate has the strongest effect of all the variables under consideration (see Figure 1). UWYE1, UWYE2, GNE1 are the speakers with consistently slower speech rates and they also have the lowest WERs. Therefore, they obtain the best STT performance. On the other hand, MLE2, the speaker with the fastest speech rate, obtains the worst STT performance.

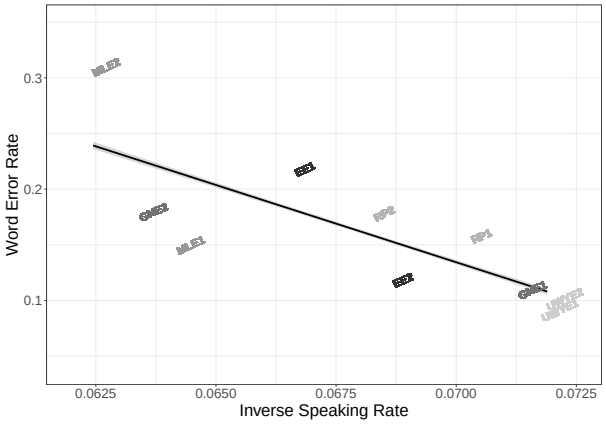

Figure 1: Speech-to-text word error rate plotted against average inverse speaking rate (ISR) for each speaker using the ggplot lm smooth function. Speaker labels represent exact values.

Variety does not appear to have a clear effect on STT efficacy (see Figure 2). The non-marginalised varieties do not have better STT performance than marginalised varieties overall. UWYE speakers, who speak with a northern marginalised variety, have the lowest WER (i.e., best performance) and RP speakers, who speak with a southern non-marginalised variety have the 3rd and 4th highest WERs (i.e., relatively poor performance). In the case of the UWYE speakers, these are also the speakers with the slowest speech rates, which may partially explain this result. However, the RP speakers have slower speech rates as well and have worse STT performance than speakers with similar speech rates, so it cannot be entirely due to speech rate effects. For example, RP2 has a similar ISR to both EE speakers, but has a much higher WER than EE2 and lower one than EE1.

Differences across varieties only seem to appear when individual repertoires are considered. This can be seen through the inconsistent STT performance for speakers of the same variety, who have large differences when comparing their DDMs to each other (see Table 1). For example, EE1 and EE2 both speak EE, which is a southern marginalised variety, and have similar speech rates. However, EE1 has a relatively high WER (i.e., poorer performance) and

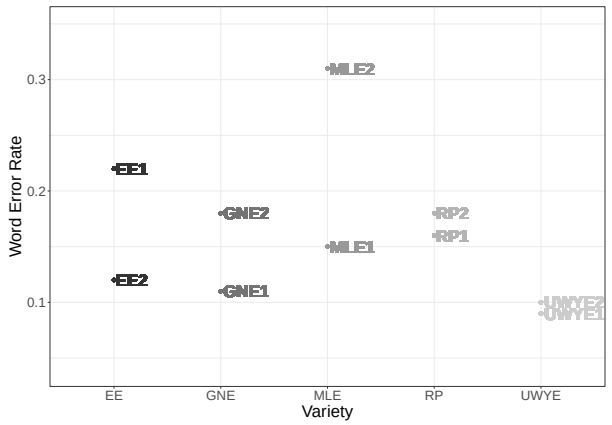

Figure 2: Speech-to-text word error rates (WER) for each speaker plotted by variety. Speaker labels indicate WER for that speaker within the variety.

EE2 does not. If we take the individual repertoires into account then we see that EE1 uses more accent features than EE2. A similar finding occurs when comparing the MLE speakers, who have similar speech rates, but differ in their accent feature use and have substantially different WERs. Again the speaker who uses fewer accent features (MLE1) has better STT performance.

Considering the non-marginalised varieties (GNE and RP), it is difficult to tease apart the effects of individual repertoire compared with speech rate, as these often point in the same direction. The RP speakers have similar speaking rates and use of accent features and also have similar WERs. On the other hand, STT performance differs for the GNE speakers, but this aligns with both the DDM and ISR results. GNE1 who has better performance also has a much slower speech rate and uses fewer accent features than GNE2.

4. DISCUSSION

Previous work has indicated that non-marginalised North American varieties have better STT performance than marginalised ones, but this does not appear to be straightforwardly the case when considering these British English varieties. We do find the worst STT performance for a speaker of a marginalised variety (MLE2), but the best STT performance also occurred for speakers of a marginalised variety (UWYE). This result seems to be, at least, partly due to the difference in speech rate, as the UWYE speakers also have the slowest speech rate. Furthermore, it seems that at the slowest speech rates differences in accent feature use does not equate to differences in STT performance. The UWYE speakers also have the highest DDMs. Non-marginalised varieties, such as RP, do not have very low WERs despite both speakers having relatively slow speech rates and the lowest accent feature use.

Differences in the individual repertoires of speakers within an accent are able to account for some of the differences in STT performance that we find and this occurs most obviously for EE and MLE, which are southern marginalised varieties. Speakers who use more accent features (EE1, MLE2) also have worse STT performance. These performance differences do not seem to be related to speech rate differences as both of the MLE speakers and both of the EE speakers have similar speaking rates compared to each other. Therefore, individual repertoires rather than variety has a more substantial effect on the STT performance in the direction that was predicted.

Overall speech rate is found to strongly affect WER, but again this is not found consistently across all speakers and varieties. The three speakers with the slowest speech rates (UWYE1, UWYE2, and GNE1) also have the best STT performance and the speaker with the fastest speech rate (MLE2) has the worst STT performance. Other than these speakers, we need to consider more than just speech rate to understand the differences in STT performance.

While more work on STT performance for a wider range of speakers of different varieties needs to be done, our findings suggest a complicated type of accent bias might occur in training and language models for STT systems. Even when marginalised varieties may be included and those varieties or speakers use slower speech rates, there may continue to be unequal outcomes for speakers of marginalised varieties who use more accent features. Building on the suggestions from other work, such as [3, 1], more diverse datasets are required when training these speech recognition systems. These datasets need to be more inclusive and should represent multiple varieties (marginalised and non-marginalised) *and* speakers with a wide range of individual repertoires within these varieties. Only then it may be possible to obtain more equal STT performance for everyone and limit the effects of social biases.

5. REFERENCES

[1] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 53–59. [Online]. Available: https://aclanthology.org/

W17-1606
[2] S. Oviatt, M. MacEachern, and G.-A. Levow, "Predicting hyperarticulate speech during human-computer error resolution," *Speech Communication*, vol. 24, no. 2, pp. 87–110, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639398000053
[3] 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. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1915768117
[4] R. Tatman and C. Kasten, "Effects of Talker Dialect, Gender Race on Accuracy of Bing Speech and YouTube Automatic Captions," in *Proc. Interspeech 2017*, 2017, pp. 934–938.
[5] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007, intrinsic Speech Variations. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639307000404
[6] *Journal of English Linguistics*, 2021.
[7] J. Cheshire, P. Kerswill, S. Fox, and E. Torgersen, "Contact, the feature pool and the speech community: The emergence of multicultural london english," *Journal of Sociolinguistics*, vol. 15, no. 2, pp. 151–196, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9841.2011.00478.x
[8] *Language and Speech*, in preparation.
[9] G.-T. Liou, C.-Y. Chiang, Y.-R. Wang, and S.-H. Chen, "Estimation of hidden speaking rate," in *9th International Conference on Speech Prosody*, 2018, pp. 592–596.