Individual Repertoires and Efficacy (WER) of Automatic Captions: A Case study of British English Varieties

Kyra Hung[1], Katreena Abernathy[1], Amanda Cardoso[1], Devyani Sharma[2], Erez Levon[3] & Ryan Taylor[1]
University of British Columbia[1], Queen Mary University of London[2], University of Bern[3]

Index Terms: captions, speech-to-text, word error rate, phonetic features, language attitudes, accentedness, judgement

1. Introduction

Performance differences in auto-captioning or speech-to-text (STT) for North American English varieties relate to systematic biases in some cases [1]. Namely, the exclusion of some varieties in training of STT models may result in poorer performance for those varieties (i.e., marginalized varieties) than those with sociopolitical power, which are more well-represented in the public sphere (i.e., standard varieties). However, individual repertories using different accent features within varieties occur, which may lead to STT performance differences for individuals using different accent features, but who speak the "same" variety. Moreover, attitudes to standard and marginalized varieties lie on a continuum [2], which may have an effect on training data biases. Performance differences for STT for individual repertoires within varieties and how those efficacy differences relate to systemic biases for varieties require further examination. The current study uses controlled British English stimuli to: i) examine if conclusions for North American Englishes regarding standardness and STT efficacy hold for British Englishes; ii) explore performance differences for STT between varieties; and iii) investigate performance differences for STT whilst considering individual repertoires.

2. Methodology

Efficacy: We assess the efficacy of Google API STT (used, e.g., by You Tube for auto-caption generation) for UK English, through word error rate (WER). WER has been used in academic discussions of STT efficacy [1] and in industry product comparison. STT outputs were generated for 15 scripted responses produced by 10 individuals speaking 5 British English varieties (Estuary=EE, General Northern=GNE, Leeds=LE, Multicultural London=MLE, Received Pronunciation=RP). These stimuli were developed for a separate study [3]. Transcripts as spoken are compared to STT outputs. WER is a sum of the number of deletions (i.e., a word appears in the script but not in the STT output), insertions (i.e., a word appears in the SST output but not in the script), and substitutions (i.e., a different word appears in the script and the STT output) divided by the number of words in the original transcript. Higher WERs indicate worse performance.

Individual Repertoires: Dialect density measures (DDM) quantify differences in individual repertoires by calculating the number of times accent features are used.

Perceived Accentedness: 80 British listeners rated the stimuli for perceived accent strength on a 7-point scale (z-scored for comparability). Table 1 provides a summary of the speakers and varieties, WER means and standard deviations (sd), and means for DDMs, and accentedness ratings.

3. Results

STT performance differences do not correspond to the expected patterns, as WERs are lower for some marginalized varieties (e.g., LE) compared with standard varieties, as shown in Table 1. Unexpectedly, those varieties that are generally perceived as more accented do not consistently have higher WERs. We also find that STT for standard varieties (e.g., RP) may be both less accurate (i.e., have higher WERs), and less precise (i.e., inconsistent WER; shown by sd), which indicates very poor performance. Finally, WER in the current sample is affected by individual repertoires, so that individuals with higher DDMs and are perceived as more accented within the same variety have higher WERs (e.g., MLE1 compared to MLE2).

Table 1. *Summary of speaker information and results. Col. 2 So=southern, N=northern, St=standard, M=marginalized. Col. 5 indicates accentedness (positive=more & negative=less accented). Bold font = highest WER, DDM and accentedness ratings.*

ID	Variety	WER mean(sd)	DDM mean	Accent mean
EE1	SoM	0.24 (0.10)	18.5	0.26
EE2		0.13 (0.05)	13.5	-0.25
GNE1	NSt	0.13 (0.05)	12.5	-0.26
GNE2		0.19 (0.08)	21.5	-0.23
LE1	NM	*0.11 (0.04)*	**42**	0.35
LE2		0.13 (0.06)	37.5	**0.48**
MLE1	SoM	0.18 (0.07)	17.5	0.09
MLE2		**0.34 (0.11)**	41	**0.48**
RP1	SoSt	0.18 (0.05)	1	-0.43
RP2		0.20 (0.13)	*0*	*-0.48*

4. Discussion

While it's clear that systematic biases are reflected in the efficacy of STT for British English varieties similar to the findings for North American varieties, these results also demonstrate that individual accent feature use affects the efficacy of STT. Furthermore, STT did not perform poorly for all marginalized varieties and did not perform well for all standard varieties. A potential reason for this could be the types of data and varieties which have been included in the STT British English training models, which could have included Urban West Yorkshire English speakers (LE) and therefore, this variety does particularly well. If this is the case, it further points to the need for purposeful inclusion of varieties with a range of speakers in STT training data to ensure that STT is not another area where marginalized language users experience inequality.

5. References

[1] Tatman, R. "Gender and Dialect Bias in YouTube's Automatic Captions." In Proc. of the First ACL Workshop on Ethics in Natural Language Processing, 53–59, 2017.

[2] Coupland, N. and Bishop, H. "Ideologised values for British accents." Journal of Sociolinguistics, 11: 74-93, 2007 doi:10.1111/j.1467-9841.2007.00311.x

[3] Anonymous, "Anonymous". Journal of English Language and Linguistics.