Exploring the variable efficacy of Google speech-to-text with spontaneous bilingual speech in Cantonese and English

Nikolai Schwarz (nschwa01@student.ubc.ca), Khia A. Johnson (khia.johnson@ubc.ca), & Molly Babel (molly.babel@ubc.ca)

Linguistics, University of British Columbia, Vancouver, BC, Canada | 5pSC | Friday, Dec. 3, 2021, 1:00 pm

Introduction

- As Automatic Speech Recognition (ASR) grows in use, it is important to test for biases in the efficacy of said systems.
- Recent work assessing ASR efficacy and bias implicates factors like race, gender, dialect, and age as leading to different efficacy rates [6, 4, 2, 5, 1]
- While lots of research highlights biases based on speaker information, there is research missing looking at the potential effects on efficacy on bilingual populations.

Research Question

Question: How well does Google Cloud Speech-to-Text (STT) transcribe the English and Cantonese speech in the SpiCE corpus? What factors contribute to successful transcription?

With this experiment, we explore factors in the efficacy of Google STT. We focus on the effects of code switching, place of birth, language dominance, and gender.

Data

- **SpiCE: Speech in Cantonese and English** is a sizable open-access corpus of conversational bilingual speech [3]
- Heterogeneous group of 34 early Cantonese-English bilinguals in Vancouver, BC (19–34; 17 male, 17 female)
- More information: `https://spice-corpus.rtfd.io`

Methods

- Using Google STT API, the SpiCE corpus was run through the system for both Cantonese and English.They were run through the Cantonese model and the English model, respectively.
- Google STT transcriptions were compared to the manual transcriptions using fuzzy string matching.
- Fuzzy string matching then returned a "matching score" that is the percentage the SST transcription matches with the manual transcriptions.
- More information on fuzzy string matching: `https://github.com/seatgeek/fuzzywuzzy`
- Blank transcriptions were taken out. Transcriptions were also altered to align with transcription standards in Google STT (i.e., "fifty fifty" changed to 50/50).
- We consider efficacy alongside the following language use and demographic variables: code-switching, place of birth, language dominance, and gender.

Results

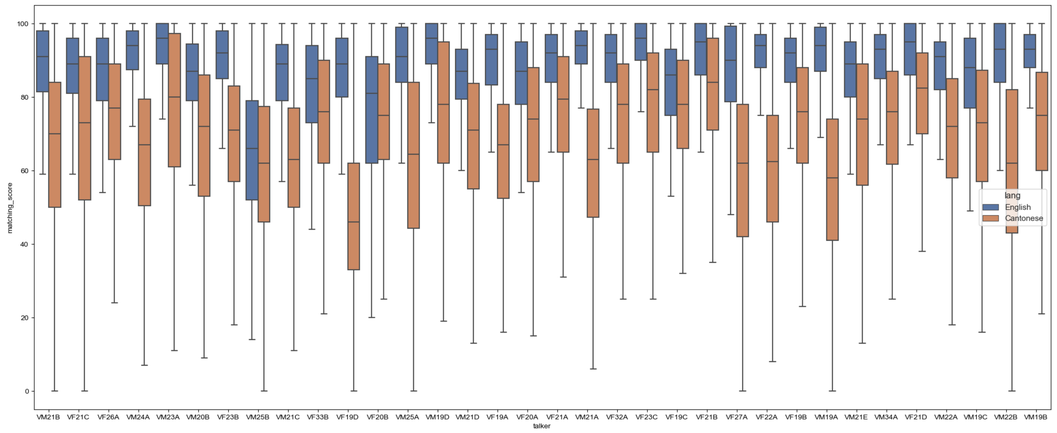
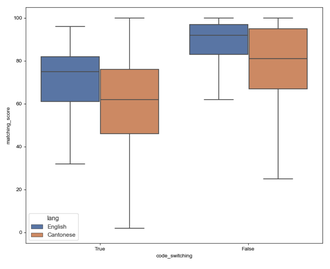
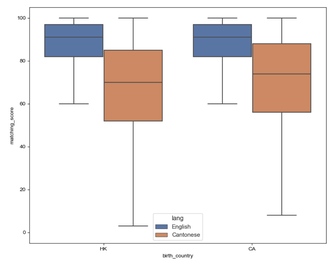


Figure: The distribution and median of data per speaker in both English and Cantonese. Medians –> Cantonese: 71, English: 91.

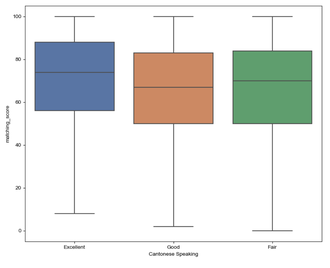
Matching Score of Google STT by language and demographic factors



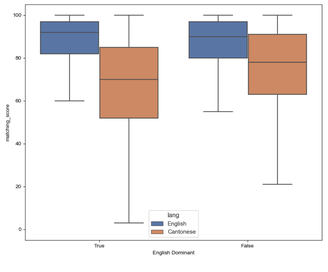
(a) Relationship between code-switching and matching score



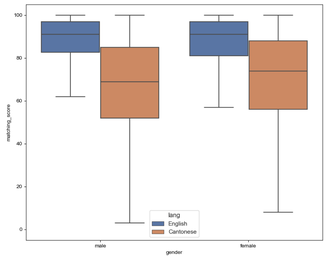
(b) Relationship between code-switching and birth country



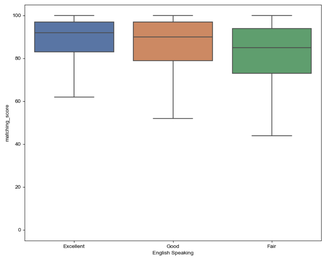
(c) Relationship between participant's self-evaluated speaking ability of Cantonese and matching score



(d) Relationship between English Dominant Speakers and matching score



(e) Relationship between gender and matching score



(f) Relationship between participant's self-evaluated speaking ability of English and matching score

Statistical Analysis

Below shows a generalized linear mixed-effects model with a beta-distributed response variable. The only two parameters that showed significance are below.
**No significance tests were performed for speaking ability*

Parameter	Estimate	SE	p
Intercept	1.613842	0.064663	<0.002
Cantonese	-0.271844	0.080539	0.000737
Code-switching	-1.004711	0.112315	<0.002

Discussion & Conclusion

- Utterances with code-switches have lower accuracy.
- There seems to be more variability in Cantonese non-code switching than in English non-code switching. However, Cantonese reached higher matching scores in code-switching than English
- No clear relationship between gender and matching score, contrary to findings in previous research.
- The median for English is higher than that of Cantonese, could this be because there are more English dominant speakers in the corpus, or because of the location of the corpus recording, or another reason?
- How could this affect people using ASR systems in bilingual households?
- Could self reported speaking ability affect these scores?

Take Home Point

Although ASR machines have shown to be very useful, this study shows that more work needs to be done to ensure that there is no bias when recognizing speech. In addition to this, code switching seems to have a great effect on the efficacy of the Google ASR, this could affect numerous people who are bilingual and are already accustomed to code switching in everyday speech.

References

[1] M. Costa-jussà, Christine Basta, and Gerard I. G'allego. Evaluating gender bias in speech translation. *ArXiv*, abs/2010.14465, 2020.

[2] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. Quantifying bias in automatic speech recognition, 2021.

[3] K. A. Johnson. SpiCE: Speech in Cantonese and English [V1], 2021. Scholars Portal Dataverse.

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

[5] Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *EthNLP@EACL*, 2017.

[6] Rachael Tatman and Conner Kasten. Effects of talker dialect, gender and race on accuracy of Bing speech and YouTube automatic captions. In *Proceedings of Interspeech 2017*, pages 934–938, 2017.