Sociolinguistic Variation and Automatic Speech Recognition: Challenges and Approaches

Dr. Rachael Tatman

Who am I?

- Dr. Rachael Tatman
- PhD in Linguistics (2017): Modeling the Perceptual Learning of Novel Dialect Features
  - Commercial automatic speech recognition systems were less accurate for some demographic groups
  - Humans use non linguistic information when adapting to a new dialect
  - Machine learning systems that do the same show a human-like pattern of errors
- Afterwards:
  - 2017 - 2019: Data scientist at Kaggle
  - 2020 - now: Senior Developer Advocate at Rasa

Gustav the Hedgehog 🦔

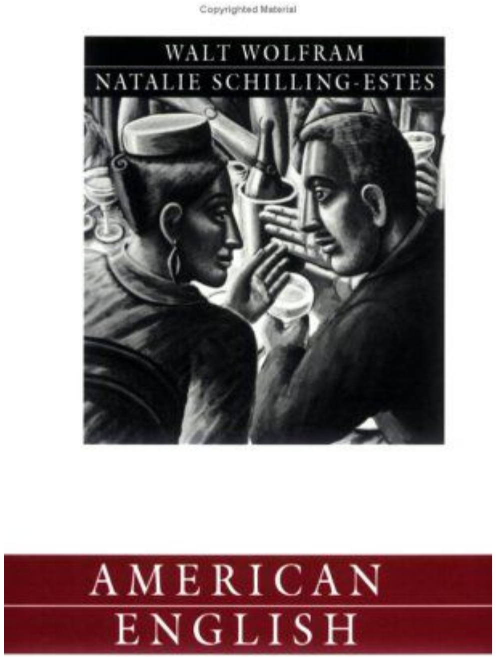
@rctatman

Twitter! :)

- **Why do automatic speech recognition (ASR) systems struggle with language variation?**
- **What are some ways of accounting for it?**

Language Variation

- All language use is shaped by its social context
- Many demographic factors are linked to systematic variation in speech, including:
    - Gender
    - Regional Origin
    - Age
    - Socio-economic status/Social class
    - Race/ethnicity
- Failure to account for these differences results in different system performance across groups

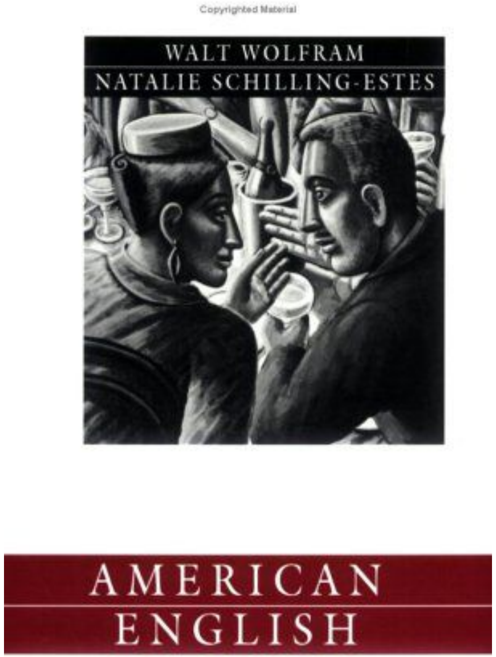


"American English" by Wolfram and Schilling-Estes is a nice introduction

@rctatman

Language Variation

- All language use is shaped by its social context
- Many demographic factors are linked to systematic variation in speech, including:
  - Gender
  - Regional Origin
  - Age
  - Socio-economic status/Social class
  - Race/ethnicity
- **Failure to account for these differences results when building ASR systems in different system performance across groups**

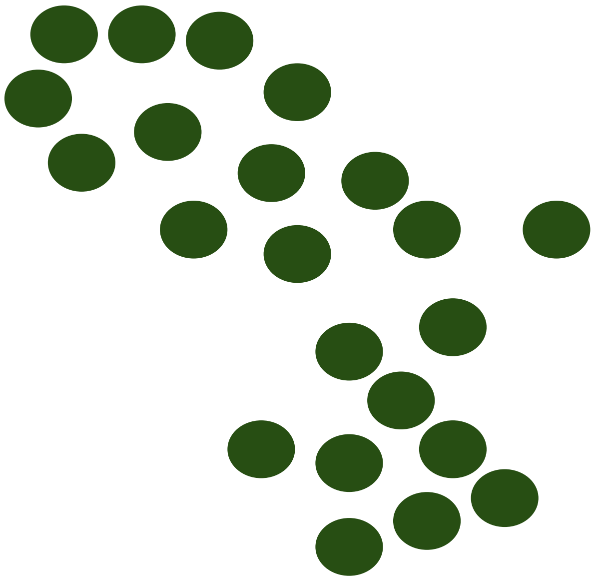


"American English" by Wolfram and Schilling-Estes is a nice introduction

@rctatman

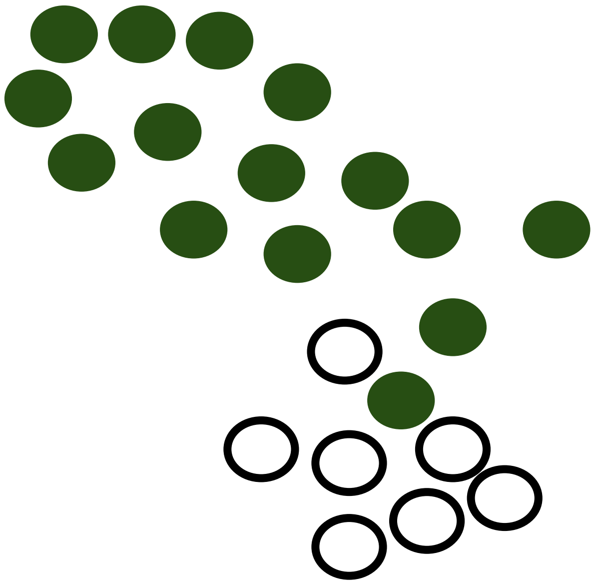
How does this happen?

- Most modern language technology is built using machine learning
  - Rule-based methods = learning from hand-built rules
  - Machine learning methods = learning from lots of examples
- If you have fewer examples from a specific group then your model won't be as accurate for them
  - Where is the center of this cluster of dots?



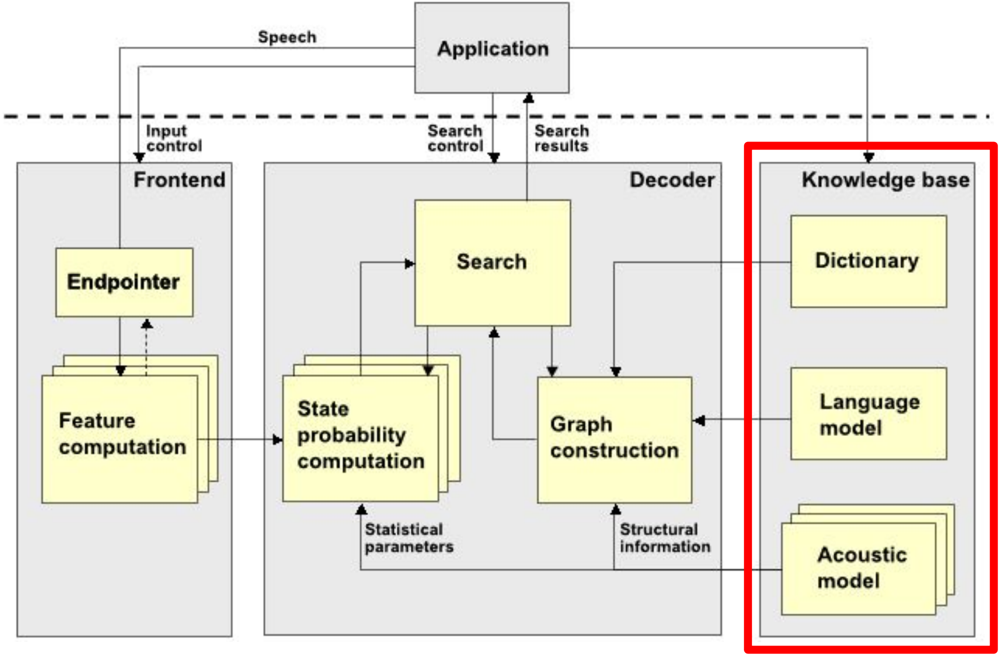
How does this happen?

- Most modern language technology is built using machine learning
    - Rule-based methods = learning from hand-built rules
    - Machine learning methods = learning from lots of examples
- If you have fewer examples from a specific group then your model won't be as accurate for them
    - Where is the center of this cluster of dots?



How does automatic speech recognition work?

- Dictionary:
  - A hand-written guide to what sounds are in each word
- Language model:
  - A statistical model of how common words & phrases are
  - Currently a very fast-moving area of research
- Acoustic model:
  - Statistical model mapping signal to speech (sounds or words)

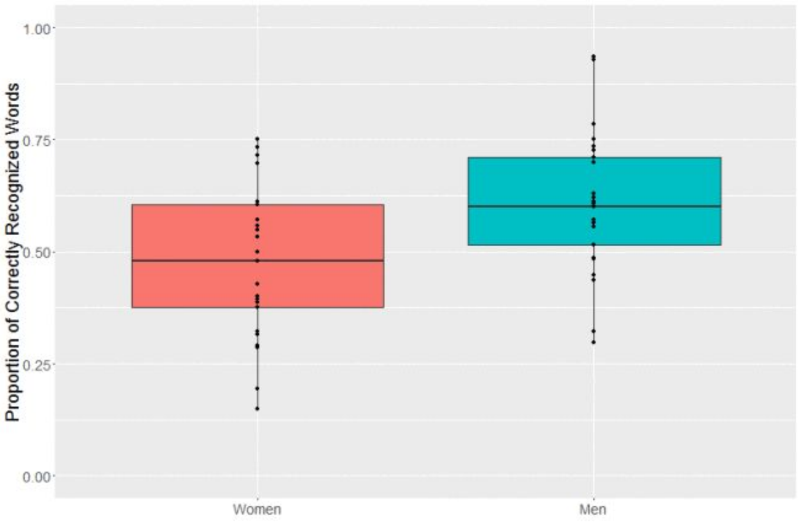


Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., ... & Wolf, P. (2003, April). The CMU SPHINX-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong* (Vol. 1, pp. 2-5).

@rctatman

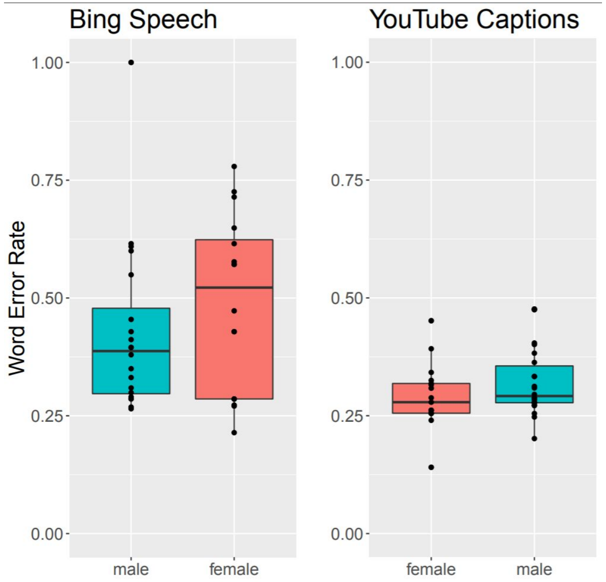
What demographic factors matter?

There's a difference in accuracy for men and women (Tatman 2017)...



Tatman, R. (2017, April). Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 53-59).

But only when signal quality is not controlled for (Tatman & Kasten 2017)

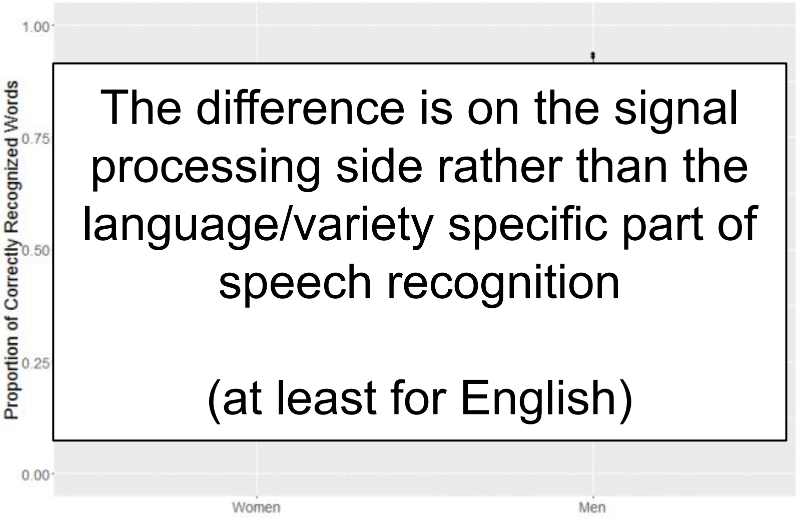


Tatman, R., & Kasten, C. (2017, August). Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *INTERSPEECH* (pp. 934-938).

@rctatman

What demographic factors matter?

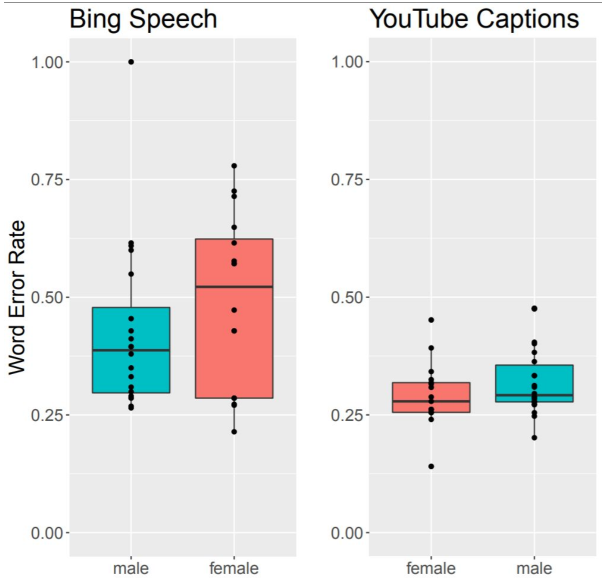
There's a difference in accuracy for men and women (Tatman 2017)...



The difference is on the signal processing side rather than the language/variety specific part of speech recognition

(at least for English)

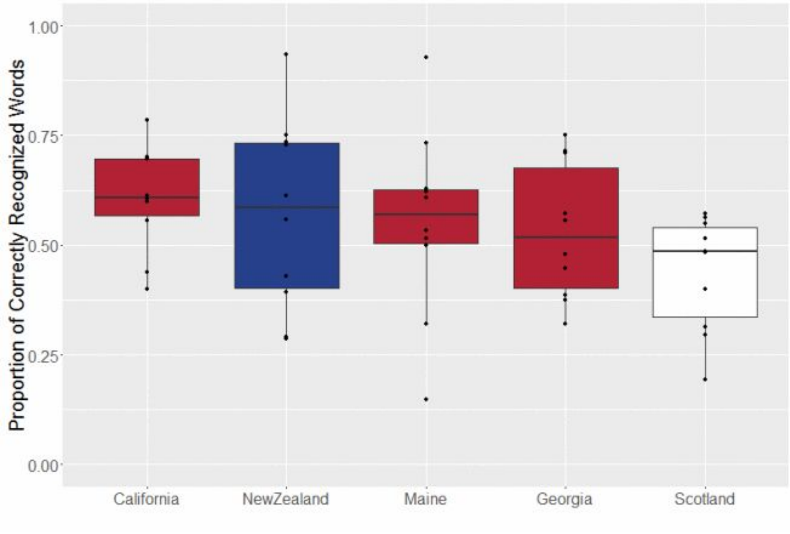
But only when signal quality is not controlled for (Tatman & Kasten 2017)



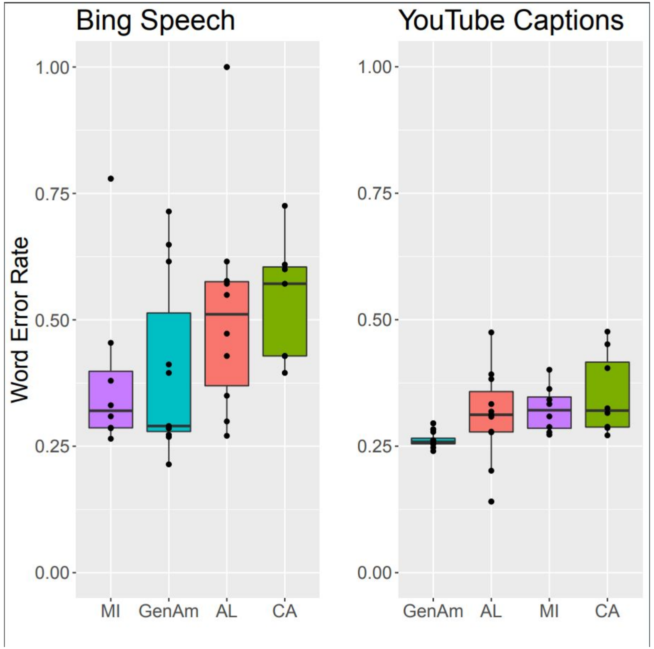
@rctatman

What demographic factors matter?

Dialect region



Tatman 2017, higher is better

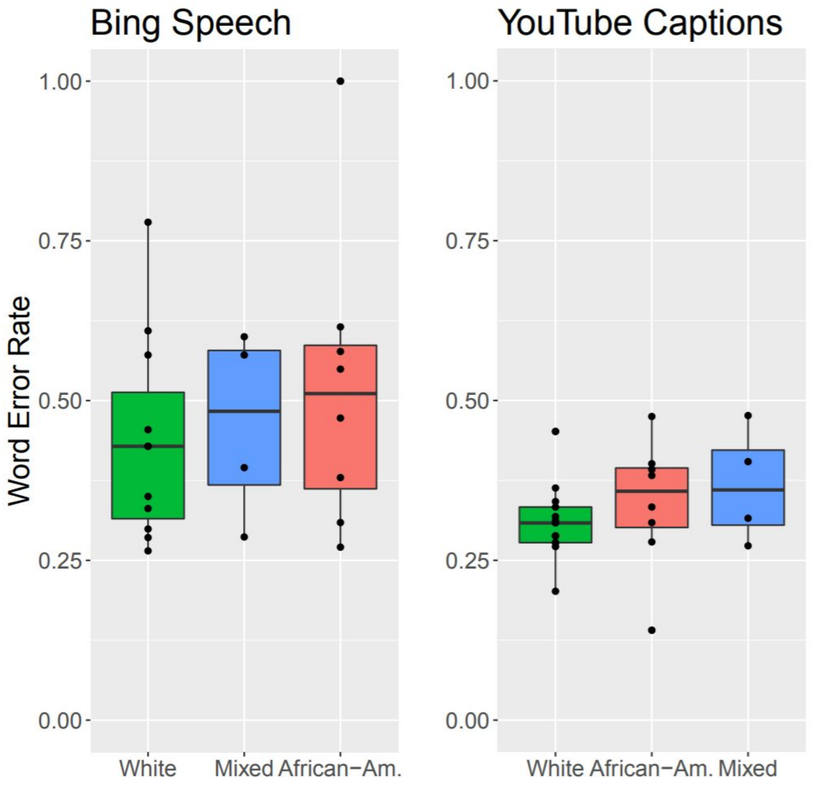


Tatman & Kasten 2017, lower is better

What demographic factors matter?

Ethnicity?

- African American English consistently has a higher error rate when systems are trained only on Standard American English (Tatman & Kasten 2017, Dorn 2019)
- Systems trained on AAE had more than a 16.6% improvement in error rate for AAE speech (Dorn 2019)



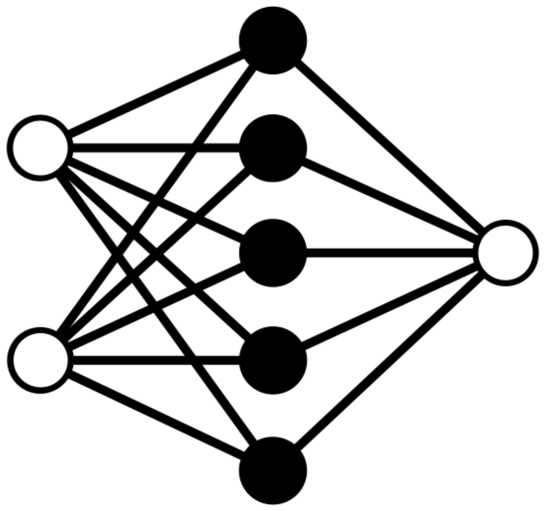
@rctatman

Language varieties vary systematically. Any automated system trained predominately on one variety will not work as well for other varieties.

- **Why do automatic speech recognition (ASR) systems struggle with language variation?**
- **What are some ways of accounting for it?**

Some Approaches

- Training multiple models
- Multi-accent models
- Adapting a single model
- Adding more data



Created by Product Pencil
from Noun Project

@rctatman

Training multiple models

- Train a separate model for each dialect & select the correct model for the talker
    - Accent-specific pronunciation modelling (Humphries et al., 1996)
    - Unsupervised model selection for recognition of regional accented speech (Najafian et al., 2014)
- Downsides:
    - Using extra-linguistic data requires collecting potentially sensitive personal data
    - Basically a social category detector

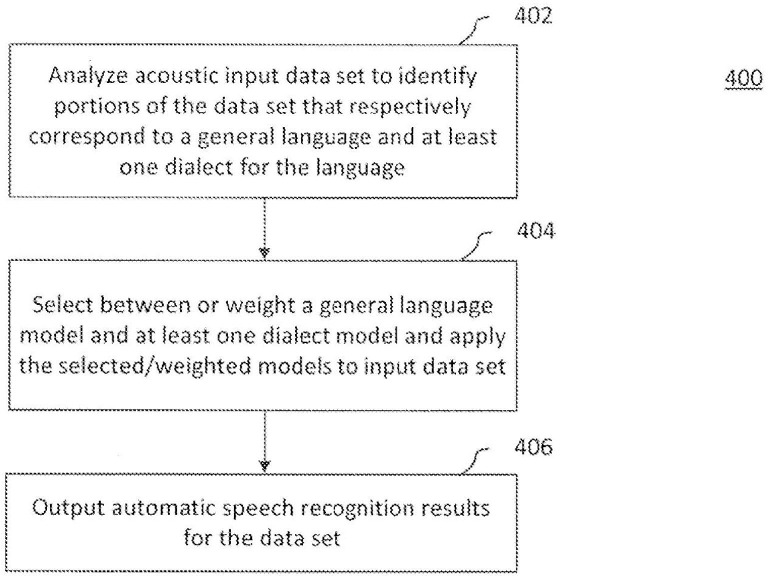


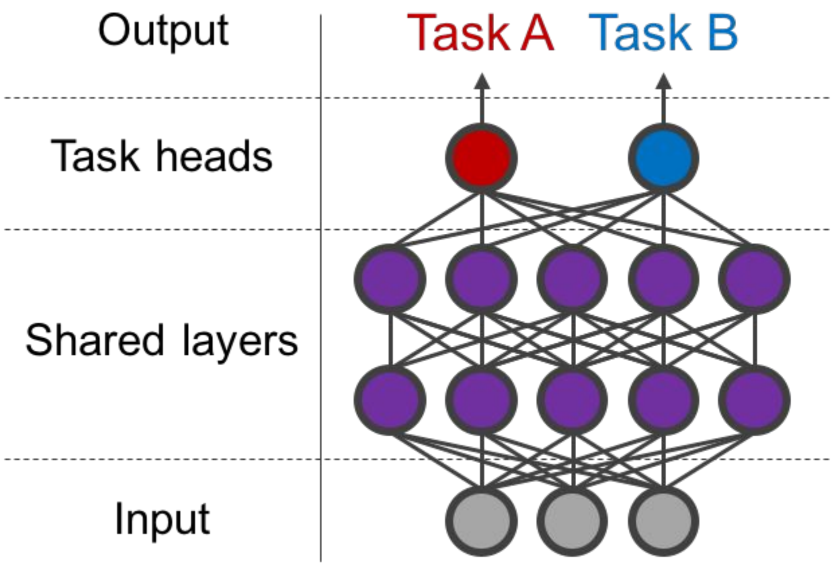
FIG. 4

Biadsy, Fadi, Lidia Mangu, and Hagen Soltau. "Dialect-specific acoustic language modeling and speech recognition." U.S. Patent Application No. 15/972,719.

@rctatman

Multi-accent models

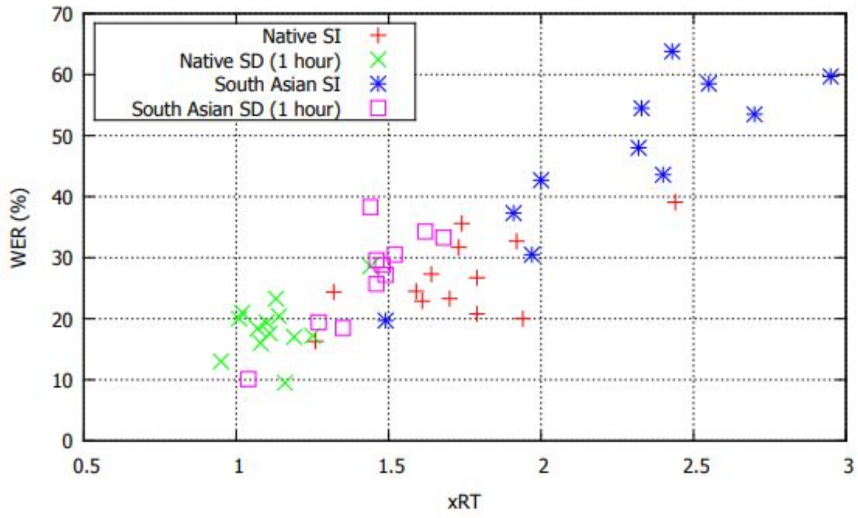
- Multitask learning (jointly training both an accent identifier and acoustic model)
  - Towards acoustic model unification across dialects (Elfeky et al 2016)
  - Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning (Jain et al, 2018)
- Mixture of experts (one classifies speech sounds, one classifies accents)
  - A Multi-Accent Acoustic Model using Mixture of Experts for Speech Recognition (Jain, Singh & Rath, 2019)
- You're still building an accent detector



Ratner, Hancock & Ré, Emerging Topics in Multi-Task Learning Systems

Speaker Adaptation

- Adapt the acoustic model for each speaker
- Examples:
  - MAP (Gauvain and Lee, 1994)
  - MLLR (Anastasakos et al., 1997)
  - Eigenvoices(Botterweck, 2000)
  - i-Vectors for neural nets (Saon et al, 2013)
- Downsides:
  - Expensive & slow
  - Need to correctly identify the speaker
  - If initial model is poor fit for group, adapted models will also be less good for that group :(

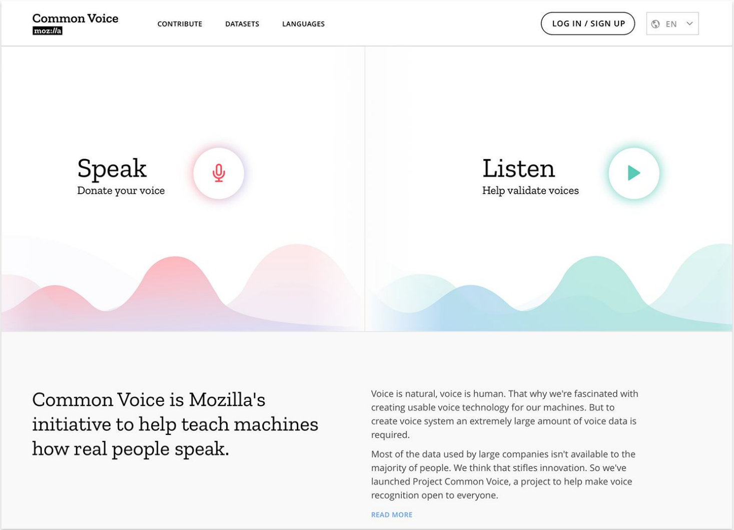


Nallasamy, U. (2016). *Adaptation techniques to improve ASR performance on accented speakers* (Doctoral dissertation, Carnegie Mellon University).

@rctatman

More data!

- Corpus of Regional African American Language (Kendall & Farrington, 2018)
    - Audio & transcriptions of 140 sociolinguistic interviews
    - Free & open source (CC-BY-NC-SA 4.0)
- Common Voice (Mozilla foundation)
    - 4,257 hours of speech in 40 languages, (many recordings include demographic metadata like age, sex, and accent
    - Free & open source (CC-0)
    - Crowd-powered: **you can help** by donating recordings or checking transcriptions



@rctatman

How not to do it 😬

GOOGLE TECH

Google contractors reportedly targeted homeless people for Pixel 4 facial recognition

They need facial scans of people with darker skin

By Sean Hollister | @StarFire2258 | Oct 2, 2019, 8:46pm EDT

SHARE

https://www.theverge.com/2019/10/2/20896181/google-contractor-reportedly-targeted-homeless-people-for-pixel-4-facial-recognition

- **Why do automatic speech recognition (ASR) systems struggle with language variation?**
- **What are some ways of accounting for it?**

Questions?

For Conversational AI Q's:
r.tatman@rasa.com

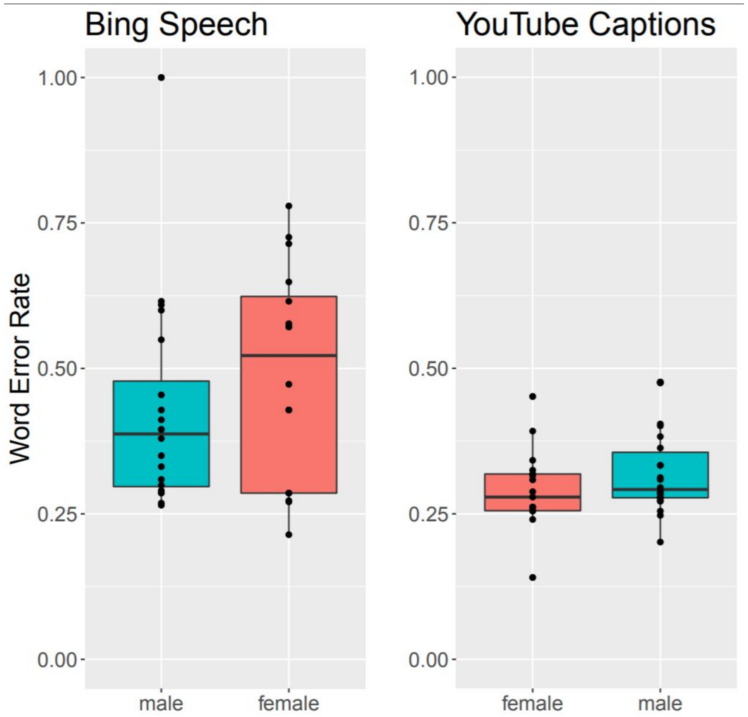
Dialect	WER	Levenshtein Distance
AAVE	16.6%	18.5%
SAE	7.5%	13.7%

Table 3: Improvements in Error Rate Between Dialect-Specific Model and Combined Model

Dorn, R. (2019). Dialect-Specific Models for Automatic Speech Recognition of African American Vernacular English. In *Student Research Workshop* (pp. 16-20).

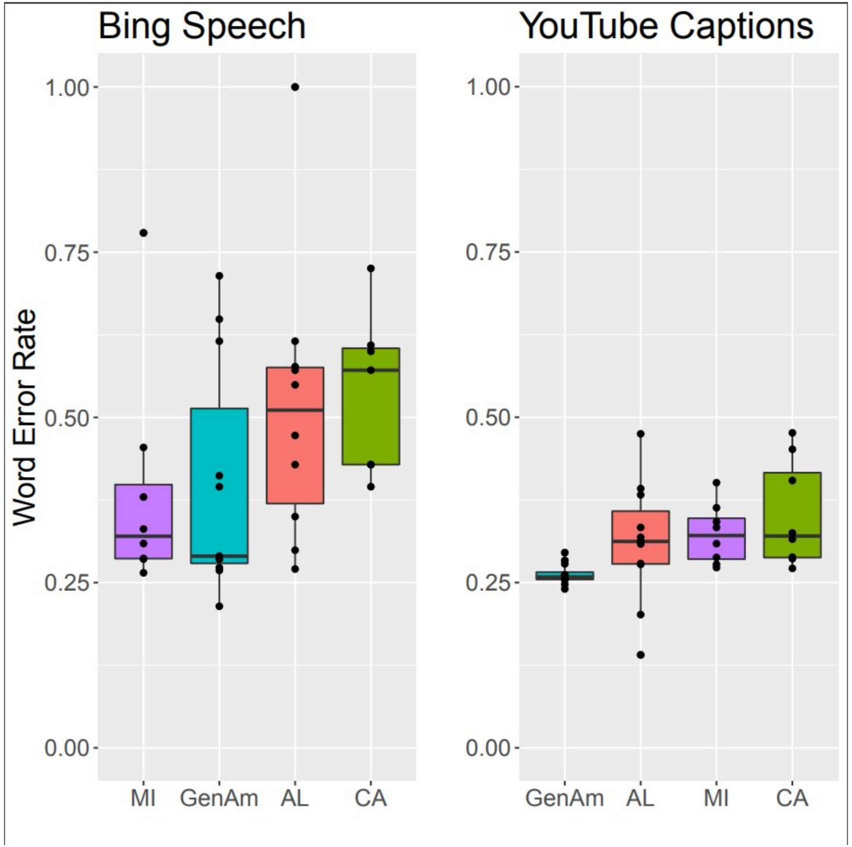
Systems evaluation -- Gender

Tatman, R., & Kasten, C. (2017, August). Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *INTERSPEECH* (pp. 934-938).



Neither Bing ($F[1, 34] = 1.13$, $p = 0.29$), nor YouTube's automatic captions ($F[1, 37] = 1.56$, $p = 0.22$) had a significant difference in accuracy by gender.

Systems evaluation -- Dialect

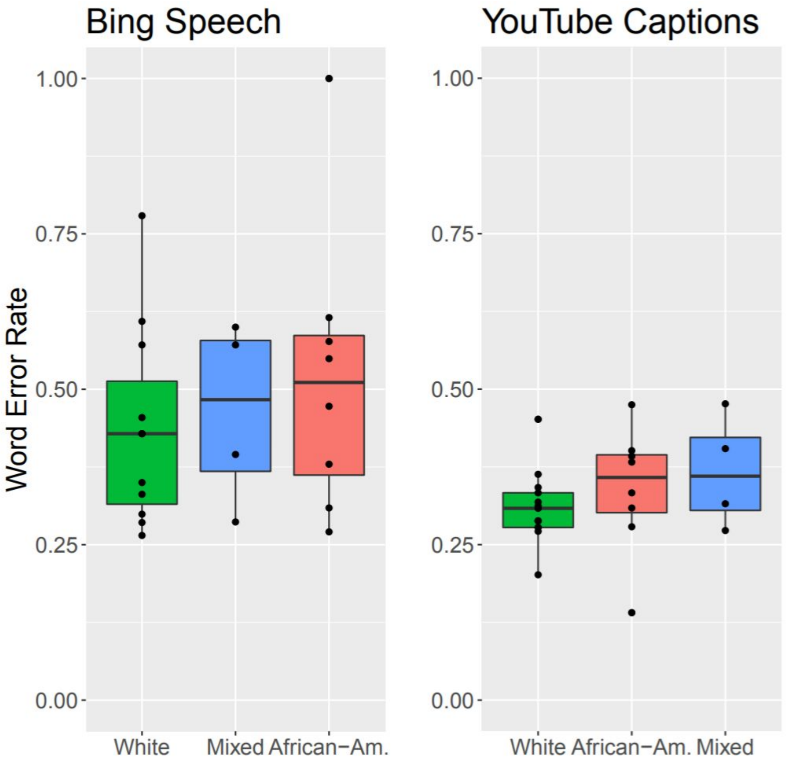


Tatman, R., & Kasten, C. (2017, August). Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *INTERSPEECH* (pp. 934-938).

Differences in WER by dialect were not robust enough to be significant for Bing (under a one way ANOVA) ($F[3, 32] = 1.6$, $p = 0.21$), but they were for YouTube's automatic captions ($F[3, 35] = 3.45$, $p < 0.05$).

Systems evaluation -- Ethnicity

Tatman, R., & Kasten, C. (2017, August). Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *INTERSPEECH* (pp. 934-938).



As with dialect, differences in WER between races were not significant for Bing ($F[4, 31] = 1.21$, $p = 0.36$), but were significant for YouTube's automatic captions ($F[4, 34] = 2.86$, $p < 0.05$).

@rctatman