Automatic Speech Recognition, Word Error Rate, and Candidate Characteristics

Samuel A. Feuer, Quantitative Analysis Center, Wesleyan University
Faculty Advisors: Markus Neumann and Erika Franklin Fowler

DELTA LAB

WESLEYAN UNIVERSITY

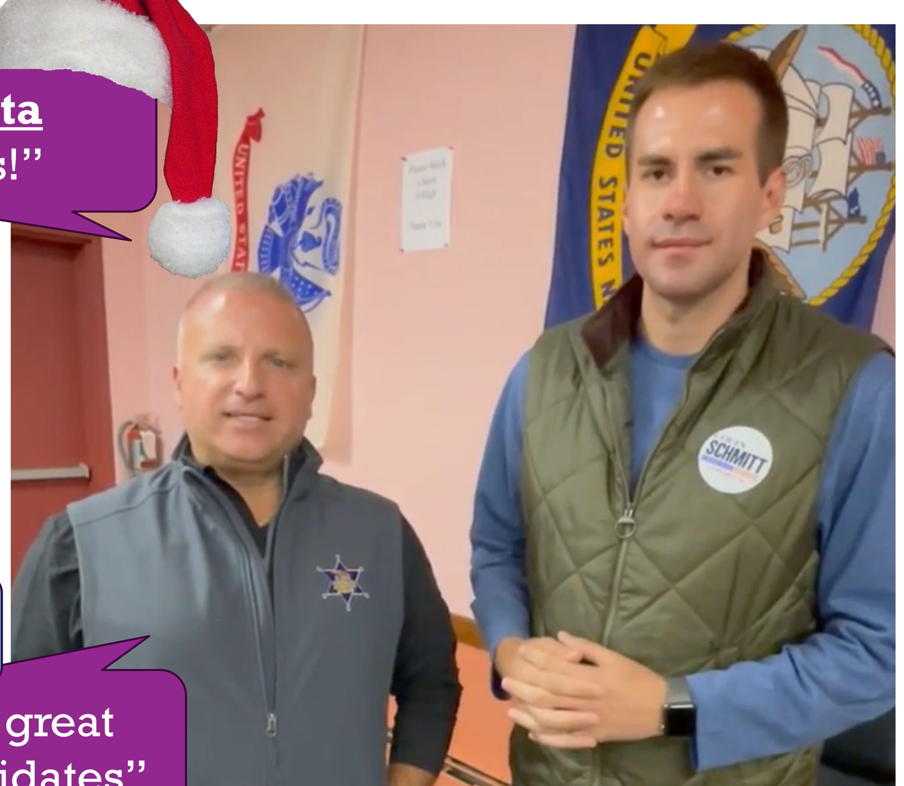
QAC

Introduction

- Automatic speech recognition (**ASR**) converts audio into text (e.g. automatic YouTube captions)
  - Has gained popularity among political scientists to analyze large audio datasets
  - Proksch et al. have validated its general reliability in this context [**1**]
  - Methods are improving, but **transcription quality impeded** by background music, uncommon words/pronunciations, accents, poor quality audio, etc.
- **Correlation of transcription errors with candidate/ad-level info** could threaten statistical inference made with ASR
  - Many researchers [**2,3**] use ASR results as proxies for manual transcripts to make analysis feasible
- These errors could also have **implications for downstream text applications** of ASR
  - Examples: structural topic modelling (**STM**), named entity recognition (NER)

Google ASR

"Great <u>**Santa**</u> Candidates!"



Actual Speech

"We have a great <u>**span of**</u> candidates"

Data

- 8,892 video advertisements with detected speech run on Facebook by 392 general election candidates for U.S. House in the two months before the 2022 midterm elections
- For each ad, we used the **Google Speech API's video model** to obtain ASR transcriptions
- We sampled **200 candidates** from this set, sampling **non-incumbent candidates with higher probability**
- From each sampled candidate, we sampled 3 unique Google ASR transcriptions (or less if they have less than 3), then removed near-duplicates (with text similarity > .98)
  - Final dataset: **478** unique ads
- Coders **hand transcribed** each of these ads and noted type of speaker (candidate, non-candidate) and non-English words
- Also used candidate- and ad-level metadata:
  - Candidate-level data from WMP and OpenSecrets (race, party, gender, incumbency, total spend, etc.)
  - Cook Political Report race competitiveness scores [**4**]
  - Ad-level spend data from Facebook
- After removing third party and Indigenous candidates (small sample size), ads with non-English words, and ads missing data, we had **439** ads

Methods: Regression

- Transcription error measured using Word Error Rate (**WER**)
  - Notable data processing: converted numbers to words, manually correcting special cases
    - Fractions and dates (3/4), and dollars/cents ($56.85)
- To test for transcription error correlations, we fit a **beta regression model with random intercepts for candidates**

Word Error Rate (WER) ~ **Candidate-level Data** + **Ad-level Data**

<u>**Candidate-level Data**</u>
- Party
- Incumbency
- Race
- Gender
- Competitiveness*
- log(Total Spend)

<u>**Ad-level Data**</u>
- log(Spend)
- CandidateSpeaking**
- NonCandidateSpeaking**

Results: Regression

- **Ad spend**, **candidate party**, and **presences of non-candidate and candidate voices** seem to correlate with WER
- **Other ad and candidate data do not** correlate with WER

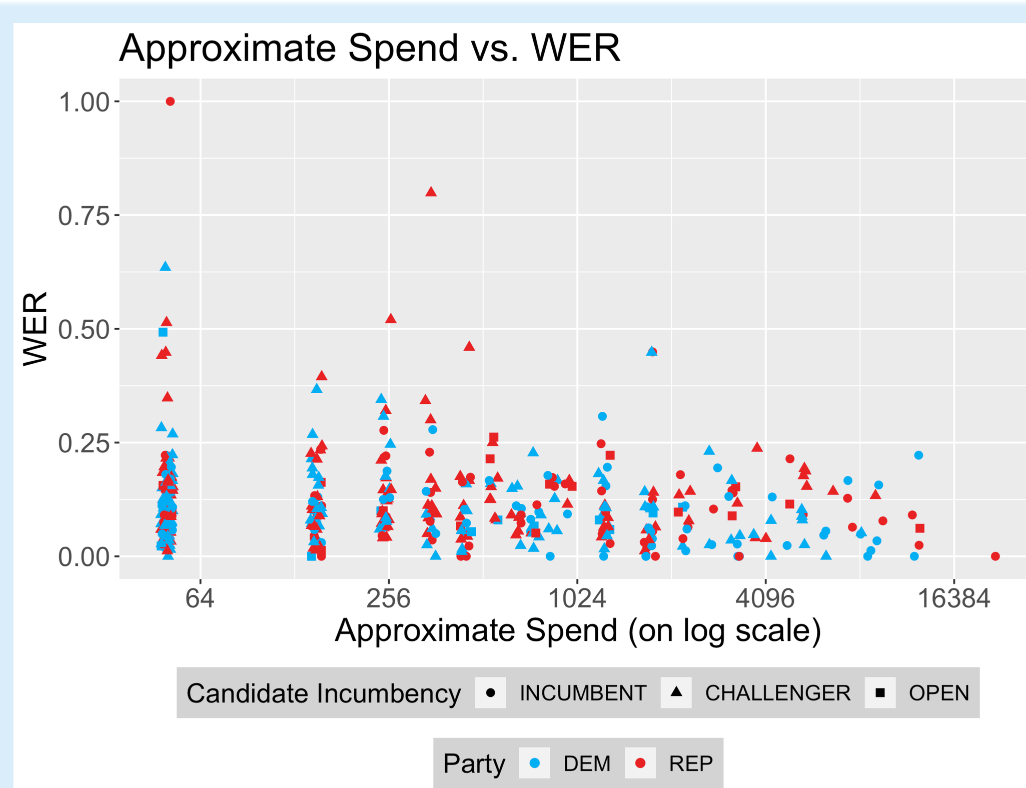


Figure 1. As ad-level approximate spend increases, average WER decreases.

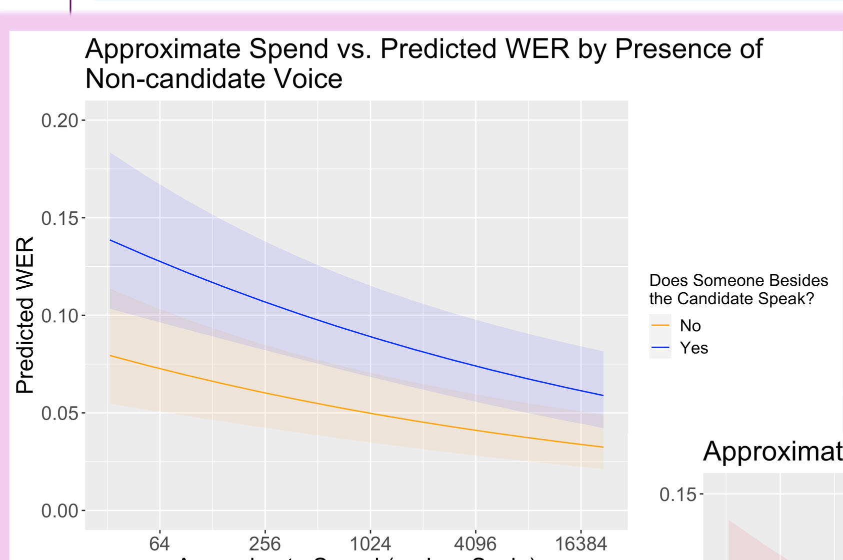


Figure 2. Ads that include speech from someone besides the candidate tend have higher WER on average.

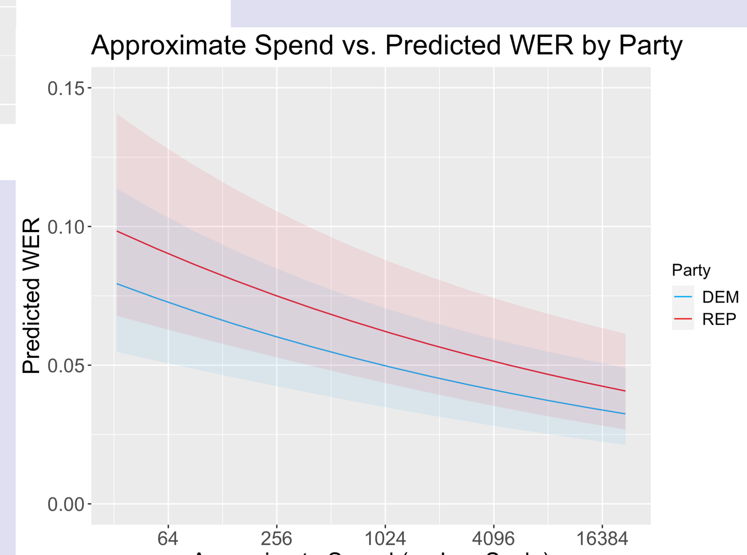


Figure 3. Republicans have higher WER on their ads than Democrats even when controlling for other candidate characteristics.

Methods: Structural Topic Models

- Fit **structural topic models** for manual and ASR transcripts, stemmed and with rare words removed
- Used all above variables as prevalence predictors, trying both with and without WER
- K = 14 topics chosen by held-out likelihood & residuals
- Manually labelled topics based on highest likelihood words

Results: Structural Topic Models

- Despite these correlations, the effect of transcription errors on topic models and their interpretation was minor
  - Topic ideas were very similar between STMs created based on Google and manual transcriptions

Google Topic (highest prob. words)	**Hand Topic** (highest prob. words)
Vote (vote, elect, congress, day)	**Vote** (vote, elect, get, novemb)
Donate (can, dollar, help, district)	**Donate** (can, dollar, help, district)
Crime (new, polic, crime)	**Crime** (new, joe, polic, peopl)
America (american, time, chang)	**America** (american, time, work)
Economy (inflat, tax, vote, price)	**Economy** (gas, economi, say, price)
Drug Prices/Immigration (drug, border, secur, colorado, lower)	**Immigration** (border, secur, work, drug)
Small Business (uh, know, go, just, busi)	**Small Business** (uh, know, people, just, busi)
Working (work, fight, care, make)	**Working** (work, tax, district)
<u>**Abortion** (abort, right, ban, woman)</u>	<u>**Abortion 1** (abort, right, ban)</u>
Generic (us, one, go congressman)	<u>**Abortion 2** (take, right, even)</u>
Generic (run, district, vote)	**Generic** (repress, district, work)
Generic (messag, approv, fight)	**Generic** (approv, messag, fight)
Unclear (people, differ, district)	**Unclear** (differ, well, work)
Unclear (us, take, can, help)	**Unclear** (us, people, one, come)

Figure 4. Both STMs had topics about crime, abortion, voting, donations, immigration, and America, among others. The biggest difference was that the manual STM had two abortion topics.

- While a few predictors for issue-related topics changed, many stayed the same (reflecting randomness of STMs)

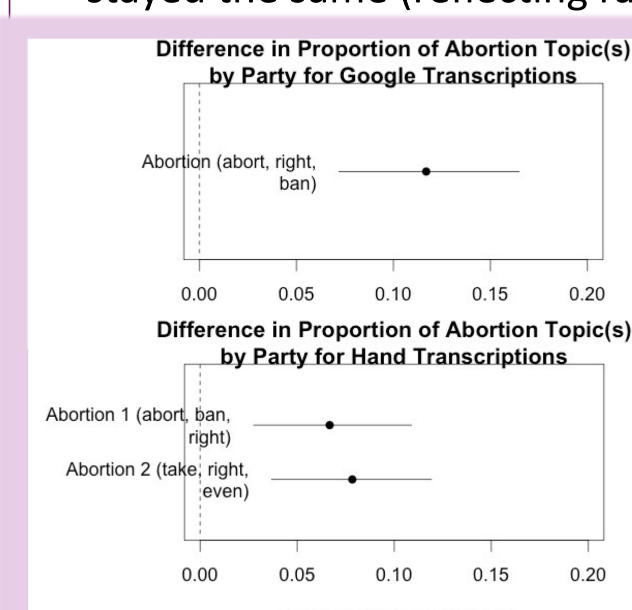


Figure 5. In both STMs, party correlated with proportion of abortion topics, with Democrats much more represented than Republicans.

- When using WER as a prevalence predictor, changes to effects of other variables on issue topic prevalence were minimal

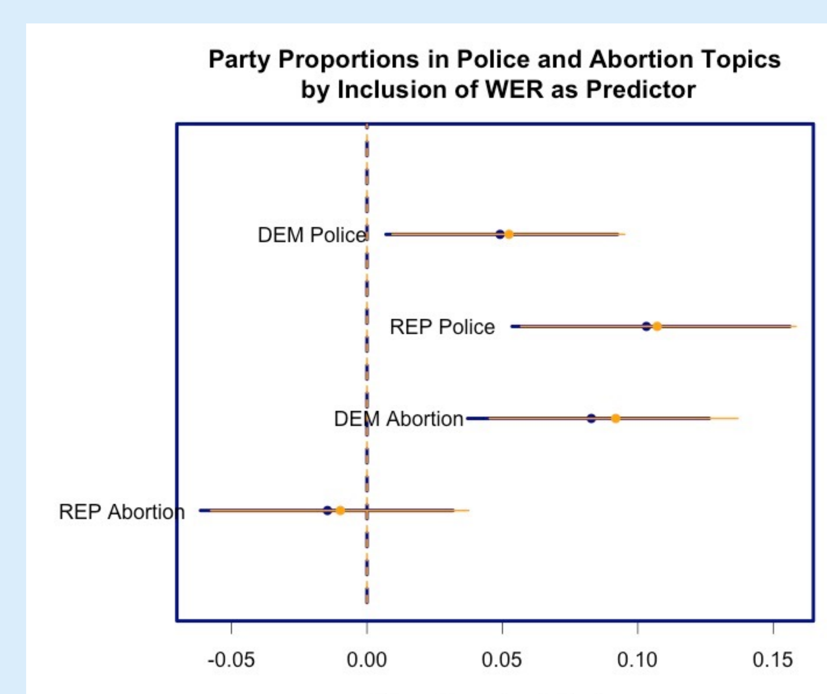


Figure 6. Adding WER as a prevalence predictor had almost no effect on others, such as party as a predictor of police & abortion topic proportions.

Discussion

- Spend-WER correlation should be considered by researchers using ASR transcriptions as proxies for hand transcriptions
- Candidate and non-candidate voices in ads are more difficult to study, but do seem to effect WER
- Regardless of these correlations, topic models and their interpretations are very resilient to ASR transcription errors
- Next Steps
  - Account for randomness in STMs with repetition
  - More downstream applications: is **Named Entity Recognition (NER)** robust to ASR errors?
  - More study of types of voices found within ads
  - Can we predict within-candidate WER variance?

Notes

*Based on Cook Political Report **competitiveness** scores, without regard to party. 1 = "solid"/non-competitive, 2="likely"/slightly competitive,3="lean"/fairly competitive, 4="toss-up"/very competitive.
**CandidateSpeaking is a 3-level, categorical hand-coded variable indicating candidate speech: 0=candidate did not speak at all, 1=candidate said one phrase/an endorsement, 2=candidate spoke for a good portion of the ad
NonCandidateSpeaking is a binary categorical hand-coded variable indicating speech by non-candidate people: 0=no non-candidate people spoke, 1=some of the ad had non-candidate speech

Acknowledgements

Special thanks to **Jielu Yao** and **Breeze Floyd**, who both helped tremendously in numerous ways.
Project was funded in part by the National Science Foundation under award #2235006 (PIs: Fowler and Zimmeck).

NSF

References

1. Proksch, S., Wratil, C., & Wäckerle, J. (2019). Testing the Validity of Automatic Speech Recognition for Political Text Analysis. *Political Analysis, 27*(3), 339-359. doi:10.1017/pan.2018.62
2. Müller, S., Kennedy, G., & Maher, T. (2023). Reactions to experts in deliberative democracy: the 2016–2018 Irish Citizens' Assembly. *Irish Political Studies*, 1-22.
3. van der Vegt, I., Mozes, M., Gill, P. *et al.* Online influence, offline violence: language use on YouTube surrounding the 'Unite the Right' rally. *J Comput Soc Sc* 4, 333–354 (2021). https://doi.org/10.1007/s42001-020-00080-x