Speaker Recognition from Audio for the Social Sciences: Application to Campaign Advertising Strategy.*

Dominic Valentino†

February 8, 2024

1 Introduction

In recent years, technological advancement has allowed social scientists to begin tapping into the vast collection of video and audio data from political debates, speeches, advertisements, and other communications made by politicians and the public. Typically this type of data is analyzed by transcribing the audio into text (e.g., Tarr, Hwang and Imai, 2023) or used for estimating the emotional arousal/vocal pitch of the speaker (e.g., Dietrich, Hayes and OâĂŹbrien, 2019; Dietrich, Enos and Sen, 2019). I propose to extend the political scientist's toolbox by automatically detecting speaker identity in audio, thereby allowing researchers to treat the speaker, rather than the audio file, as the unit of analysis.

To demonstrate its utility, I apply the tool to a corpus of presidential campaign advertising videos to assess whether candidates themsleves, as opposed to their surrogates, have become more vitriolic in their advertising. Recent work finds that, whereas there was once a strong norm against making explicit racial appeals in their ads for fear of backlash (Valentino, Hutchings and White, 2002), audiences have since become more receptive to those same explicit racial appeals (Valentino, Neuner and Vandenbroek, 2018). I argue this is part of a broader downward trend in adherence to norms

*

†Department of Government, Harvard University. email: dvalentino@g.harvard.edu

of civility in politics. Especially with the recent success of Donald Trump's bombastic and aggresive style, I predict that candidates are now making more vitriolic statements in their advertisements than in the past, and that they are more likely to be making these statements themselves as opposed to a surrogate.

This paper makes several contributions. First, combining speaker identification with other methods for analyzing audio commonly used in the social sciences (e.g. transcription, vocal pitch, emotional arousal) unlocks the ability to apply these methods to individual speakers rather than entire audio clips. As such, the method described in this paper enhances existing audio analysis techniques, rather than replaceing them. Second, this paper produces a publicly-available set of speaker embeddings for all major presidential candidates dating to the 2000 election. Researchers interested in presidential audio can use these embeddings without having to produce their own training set, or they may extend the database beyond presidents to other politicians or public figures. Third and substantively, the paper advances our understanding of campaign advertising by applying issue detection and sentiment analysis to utterances from the candidates themselves compared with the advertisement as a whole.

2 Audio Data in Political Science

This section will review the use of audio data in political science in more detail and describe how the proposed methodology could expand upon it.

3 Declining Norms of Civility

This section will review what we know from political science about changing norms of civility.

4 Data

There are two main data sources involved in this study: presidential primary and general election debate video, used for training models to detect presidential candidates' voices in audio, and pres-

idential campaign advertising videos, which are the substantive focal point of the study. I describe each in turn below.

Presidential Debate Data

The goal of this study is to build a model that can take campaign advertising audio and detect when a candidate is speaking versus when they are not. The first step towards this goal is to compile a set of labeled audio data where the labels indicate who is speaking. I chose presidential debate video for this purpose because of its ready availability on YouTube, its relatively high audio quality, the extended and largely uninterrupted blocks of speech by each candidate, and the fact that only a single file per election is needed to capture all the candidates' voices.

From YouTube, I collected the video of one debate per presidential election beginning with 2000, including the Democratic and Republican primaries to ensure that all major candidates are included even if they did not receive their party's nomination. Armed with this data, I extracted the audio from the video files and used speaker change detection (SCD) to identify segments of the audio corresponding to separate speakers (Bredin and Laurent, 2021). With the audio segmented, I then manually label each segment with the speaker's identity, including an "other" category for speakers not of interest[1].

Finally, I combine segments from the same speaker and extract speaker embeddings (Snyder, 2020; Bredin et al., 2020; Coria et al., 2020). This process uses a deep neural network (DNN) to take audio snippets of varying length and return a numeric vector of fixed length that represents the speaker's vocal fingerprint (voiceprint?). Because these embedding vectors have constant length, they can be compared using cosine similarity.

Presidential Campaign Advertising Data

Substantively, this paper seeks to study how candidate strategy has changed over time regarding who says what in their advertisements. To that end, I am in the process of collecting presidential campaign

[1]This is the current stage of the project. All of the following steps will be completed subsequently.

advertising data dating back to 2000[2].

Very similarly to the presidential debate data, I will first extract and segment the advertising audio, create target speaker embeddings, and compare each target embedding to the ground truth embeddings using cosine similarity. Each target embedding will then be labeled according to the closest ground truth embedding.

5 Models

This section will contain more detailed information regarding the segmentation and embedding models.

References

Bredin, Hervé and Antoine Laurent. 2021. "End-to-end speaker segmentation for overlap-aware resegmentation." *arXiv preprint arXiv:2104.04045* .

Bredin, Hervé, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz and Marie-Philippe Gill. 2020. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*. Barcelona, Spain: .

Coria, Juan M., Hervé Bredin, Sahar Ghannay and Sophie Rosset. 2020. A Comparison of Metric Learning Loss Functions for End-To-End Speaker Verification. In *Statistical Language and Speech Processing*, ed. Luis Espinosa-Anke, Carlos Martín-Vide and Irena Spasić. Springer International Publishing pp. 137–148.

Dietrich, Bryce J, Matthew Hayes and Diana Z OâĂŹbrien. 2019. "Pitch perfect: Vocal pitch and the emotional intensity of congressional speech." *American Political Science Review* 113(4):941–962.

[2] I currently have direct access to presidential advertising in 2000, 2004, and 2008 as part of a project with Kosuke Imai and Adam Breuer.

Dietrich, Bryce J, Ryan D Enos and Maya Sen. 2019. "Emotional arousal predicts voting on the US supreme court." *Political Analysis* 27(2):237–243.

Snyder, David. 2020. X-Vectors: Robust neural embeddings for speaker recognition PhD thesis Johns Hopkins University.

Tarr, Alexander, June Hwang and Kosuke Imai. 2023. "Automated coding of political campaign advertisement videos: An empirical validation study." *Political Analysis* 31(4):554–574.

Valentino, Nicholas A, Fabian G Neuner and L Matthew Vandenbroek. 2018. "The changing norms of racial political rhetoric and the end of racial priming." *The Journal of Politics* 80(3):757–771.

Valentino, Nicholas A, Vincent L Hutchings and Ismail K White. 2002. "Cues that matter: How political ads prime racial attitudes during campaigns." *American Political Science Review* pp. 75–90.