A Framework for Studying Causal Effects of Speech Style: Application to U.S. Presidential Campaigns[*]

Taylor J. Damann [†]   Dean Knox [‡]   Christopher Lucas [§]

August 17, 2023

Abstract

How do voters evaluate political candidates? Qualitative research has long highlighted the importance of vocal style in public speaking, yet quantitative research has largely ignored this component of speech in favor of easier-to-measure text. We collect a new audiovisual corpus of U.S. presidential campaign speech, computationally measure vocal style, and descriptively study variation across candidates and topics. We advance a unified causal framework for studying effects of text, audio, and visual speech components, forming the basis for: (1) a naturalistic experiment, exploiting subtle variation in campaign "catchphrases" with identical or near-identical wording, identified with new automated phrase-clustering methods; and (2) an audio conjoint experiment with nearly 1,000 recordings manipulating specific vocal mechanisms, produced with professional voice actors and audio editing software. We find strong evidence that candidates are evaluated not just on the positions they express, but how they express them. We also find suggestive evidence that the penalty for less desirable vocal styles is larger for women than for men. Throughout, we lay methodological foundations for a broad agenda on campaign speech.

Keywords: causal inference, elections, speech analysis, unstructured data

---

[†]PhD Candidate, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130; taylordamann.com, tjdamann@wustl.edu

[‡]Assistant Professor, Wharton School of the University of Pennsylvania, University of Pennsylvania; http://www.dcknox.com/

[§]Assistant Professor, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130; christopherlucas.org, christopher.lucas@wustl.edu

1   Introduction

Politicians spend tremendous amounts of time and money communicating with constituents, and a large body of research examines the textual content of this communication (e.g., Cohen, 1995; Canes-Wrone, 2001; Baum, 2004; Rule et al., 2015). In practice, however, political communication is more than words alone; the way in which words are spoken meaningfully shapes voter perceptions (Klofstad, 2016; Klofstad, 2017). Candidates rehearse stump speeches, prepare for public debates, and painstakingly record countless statements and advertisements. The scale of this effort suggests that not only does the content of speech matter, but the way it is delivered also matters. If not, why would parties and candidates invest precious resources on what amounts to mere acting skills?[1] And yet, despite the strategic importance that this investment implies, research on political communication has almost entirely ignored vocal style—*how* candidate speech is delivered to audiences—instead focusing only on the textual content of speech.

In this article, we lay the foundation for a new approach to studying how candidates use speech to mobilize voters. To do so, we collect an original corpus of audiovisual campaign speech recordings and conduct three analyses. We begin with a large-scale descriptive analysis of vocal delivery over a presidential campaign, examining how Barack Obama and Mitt Romney appealed to voters over the course of the 2012 presidential election. Second, we use those recordings to construct a naturalistic experiment, exploiting subtle variation in how real-world candidates delivered similar policy statements, which we identify using new computational methods for clustering speeches into approximately recurring "catchphrases"—sentences that candidates repeated often on the campaign trail, but with varying vocal style. We show that this variation in delivery has large effects on real-world voters' affect toward and evaluations of these candidates. Third, we conduct a controlled experiment to evaluate

---

[1]For an example of organizations and candidates emphasizing the speaking ability of political candidates, Run for Something—a progressive political organization that recruits young candidates for down-ballot elections—notes that they are looking for candidates who communicate well, both in person and online; it provides training in public speaking for selected candidates (Run for Something Website, 2021). In fact, every candidate training program that we identified provides instruction in public speaking: Run for Something, Emerge America, She Should Run, Wellstone, Running Start, VoteRunLead, and Emily's List.

the auditory mechanisms behind these perceptual effects using a combination of professional voice actors and audio editing software to produce nearly 1,000 audio recordings that manipulate specific elements of vocal delivery. Respondents' affect and willingness to vote for the fictional candidates varies depending on their vocal qualities. Interestingly, we also report suggestive evidence that the penalty for less desirable vocal styles is larger for women than for men. In sum, we find that candidates' vocal styles have a sizeable impact on substantively important political outcomes.

To arrive at these substantive findings, we develop a suite of methods for quantifying non-textual speech behaviors and relating them to textual measures. Our observational analysis of the election-speech corpus (Section 3) demonstrates a broadly applicable rubric for future speech research: we extract automated, timestamped transcriptions, then pair these with audio features to construct a more faithful representation of campaign speech than mere text alone can provide. We then develop a new causal framework for studying the effects of textual, auditory, and visual speech components (Section 4). Next, in our first experiment, we show how to identify and exploit natural variation in speech corpora (Section 5), providing a guide for how researchers can quantify its effects. Finally, we develop a second conjoint-style audio experimental design to isolate the effects of specific speech elements (Section 6). To our knowledge, this is the first conjoint experiment to vary speech features and offers a easy-to-use template for future work.

2   What We Miss When We Ignore Audio Data

The considerable amount of research on political campaigns, which spans many disciplines, relies heavily on textual analyses of speech (Benoit, 2017; Bligh et al., 2010; Carlson and Montgomery, 2017; Conway III et al., 2012; Degani, 2015; Franz et al., 2016; Fridkin, Kenney, et al., 2007; Fridkin and Kenney, 2011a; Fridkin and Kenney, 2011b; Schroedel et al., 2013; Sides and Karch, 2008; Spiliotes and Vavreck, 2002) This contrasts sharply with popular analysis of political campaigns, which frequently emphasizes candidates' vocal style. Prominent examples include commentators nicknaming Hillary Clinton "Shrillary" for

the sound of her voice (Khazan, 2016), Margaret Thatcher artificially lowering her voice to appeal to audiences (Moore, 2013), and former U.K. Prime Minister Edward Heath training to create a public voice distinct from his private voice (Rosenbaum, 2016). This anecdotal evidence prompts our argument that vocal style—the way in which words are delivered—uniquely affects the listener. In this section, we argue that voters receive strong information from audio cues that is not conveyed by text alone.

Audio data conveys words—that is, textual information—but also has the unique ability to convey information that is not directly represented by the linguistic content of these words alone.[2] Listeners' inferences about speaker qualities may be divided into two categories: time-invariant traits and time-varying status. The former is what a listener takes away from the speaker as a person. For example, non-textual components of speech can project a facade of dominance and power (Kalkhoff et al., 2017; Carney et al., 2005; Gregory and Gallagher, 1999). Vocal cues can also communicate levels of intelligence to the listener. Qualities of speech such as rate, pitch, pronunciation and use of dysfluencies indicate to the listener whether the speaker is not only competent on the subject of the speech, but competent as an individual (Klofstad et al., 2012; Tigue et al., 2012; Surawski and Ossoff, 2006). Vocal characteristics are also the primary way that viewers interpret charisma of a speaker (Niebuhr et al., 2017; Novák-Tót et al., 2017). Qualities such as intelligence, charisma and dominance do not change and thus will stay with the listener as an important impression.

The voice also offers unique insight into the extemporaneous parts of speech, which indicate the dynamic—or time-varying—characteristics of the speaker. Perhaps the most relevant dynamic trait of a speaker is their expressed emotion (Knox and Lucas, 2021). Several measurable qualities of speech can identify a speaker's use of emotion (Banse and Scherer, 1996; Johnstone and Scherer, 2000; Scherer, 2003; Dietrich, Hayes, et al., 2019). For example, tone of voice and intonation patterns can indicate which emotion a speaker is experiencing and projecting (Bänziger and Scherer, 2005). Qualities such as breathiness and

---

[2]Due to space constraints, we are only able to highlight a fraction of the vast literature linking speakers' vocal cues to the specific perceptions and inferences formed by listeners. A Google Scholar search for "paralinguistic," referring to non-textual components of human communication, returned over 94,000 results in May 2023.

meekness can also affect the communication of emotions (Gobl and Chasaide, 2003). These aspects of speech style are highly correlated, meaning that it can be difficult to disentangle the effect of a single vocal characteristic in observational data. For the same reason, a single speech can affect evaluation on many aspects of perceived personality.

In sum, there is a glaring disconnect between (i) the wealth of research establishing that vocal qualities like pitch, rate of speech, emotional intensity and volume affect evaluations of a speaker, and (ii) the lack of research on vocal style in political campaigns. This omission is not, presumably, due to theoretical considerations. Rather, it likely originates from a lack of established methods for studying this phenomenon.

To illustrate the complexity of speech analysis, Figure 1 presents a quantitative summary of one specific campaign speech: then-candidate Barack Obama's acceptance speech at the 2012 Democratic National Convention. Panels A1, A2, and A3 display raw features of a single utterance in the speech (approximately a sentence). Panel A1 is the waveform of the sentence, the measured displacement of air over the duration of the sentence. Panel A2 shows the pitch, along with the timestamped words of the sentence.[3] Panel A3 shows the same words, but with the sentence's loudness instead of pitch. Note that Obama uses pitch modulation and long pauses to emphasize certain terms as he speaks. In each of these plots, the x-axis is time.

However, a speech is of course composed of many sentences. Panels B1–4 display a summary of the full speech. Instead of time, the x-axis in these plots is simply an index of the utterance, and the displayed feature is a summary of the audio in that sentence (either the mean or the variance of the pitch or the loudness). While it is possible to observe general patterns (for instance the variation in volume appears to steadily increase toward the end of the speech), note that these sentence-level summaries discard much of the information displayed in panels A1-3.

We now turn to the collection of a corpus of campaign speeches before laying out our

---

[3]We only plot pitch where it can be confidently estimated. Note that pitch is only well-defined during voiced speech, such as /a/ or /z/; for unvoiced speech, such as /t/, pitch is difficult to quantify. If the estimates are implausibly large or small, or if results from two pitch-estimation methods diverge substantially, we record the value as missing.

causal framework for studying textual, auditory and visual speech components.
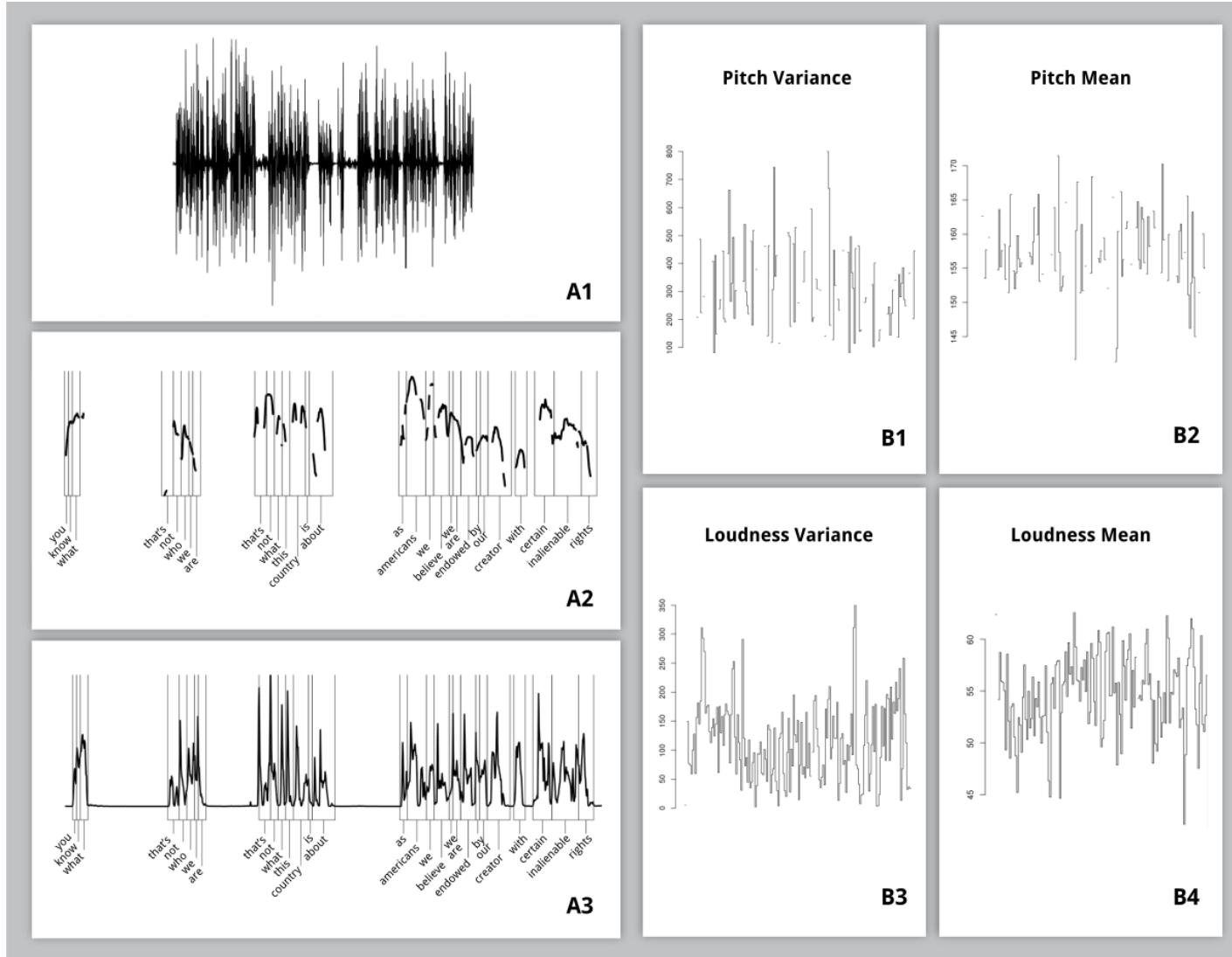
Figure 1: Panel A1 is the waveform of the sentence, the measured displacement of air over the duration of the sentence. Panel A2 shows the pitch, along with the timestamped words of the sentence. Panel A3 shows the same words, but with the sentence's loudness instead of pitch. Panels B1–4 show summary measures of these features for a full speech.

3   A New Corpus of Campaign Speech

We first establish that political speech delivery varies by creating and analyzing a new corpus of campaign speeches from the 2012 Presidential Election. To construct our corpus, we scrape 100 recorded campaign speech videos from the nonpartisan website `ElectAd`: 38 of Obama and 62 of Romney. All speeches were given in 2012, with most occurring in the three months before election day. For example, from September 10–17, the data include Romney's campaign rallies in Mansfield and Painesville, Ohio; his address to the National Guard Association Convention in Reno, Nevada; a question and answer session in Jacksonville, Florida; a campaign rally in Fairfax, Virginia; and an address to the Hispanic Chamber of Commerce in Los Angeles, California. Our corpus also includes Obama's victory speech and Romney's concession speech from November 6 of that year.

In raw form, each video contains multiple speakers, only one of which is the candidate of interest. To circumvent this problem, we use human coders to manually identify the start and stop points of the candidate's speech and drop the remaining video. Next, to construct text transcripts corresponding with these recordings, we use the `Google Speech-to-Text API`. Because campaign speeches may be unlike data on which `Speech-to-Text` is trained, we provide a series of "hints"—n-grams that are likely to be common in campaigns but rare in other contexts—based on frequent phrases used by Obama and Romney in speeches recorded in the American Presidency Project (Woolley and Peters, 2008).

Figures 2 and 3 show how Obama and Romney respectively varied their vocal styles depending on the topic.[4] To measure the topic of speech, we used the Lexicoder policy agendas dictionary (Albaugh et al., 2013). Results show that Obama uses rhetorical flourishes to draw attention to issues of religion and the economy—issues emphasized during his presidency, illustrating the face validity of the audio features we employ—while speaking less emphatically when discussing national defense. However, audio analysis can reveal more than just engagement: measures of vocal behavior indicate that Romney appears exceptionally monotonous when discussing technology.[5]

---

[4]Points plotted in red are significant after multiple testing correction (Benjamini and Hochberg, 1995).

[5]Appendix Figure 12 compares Obama and Romney's styles across four audio features. We find substantial

7
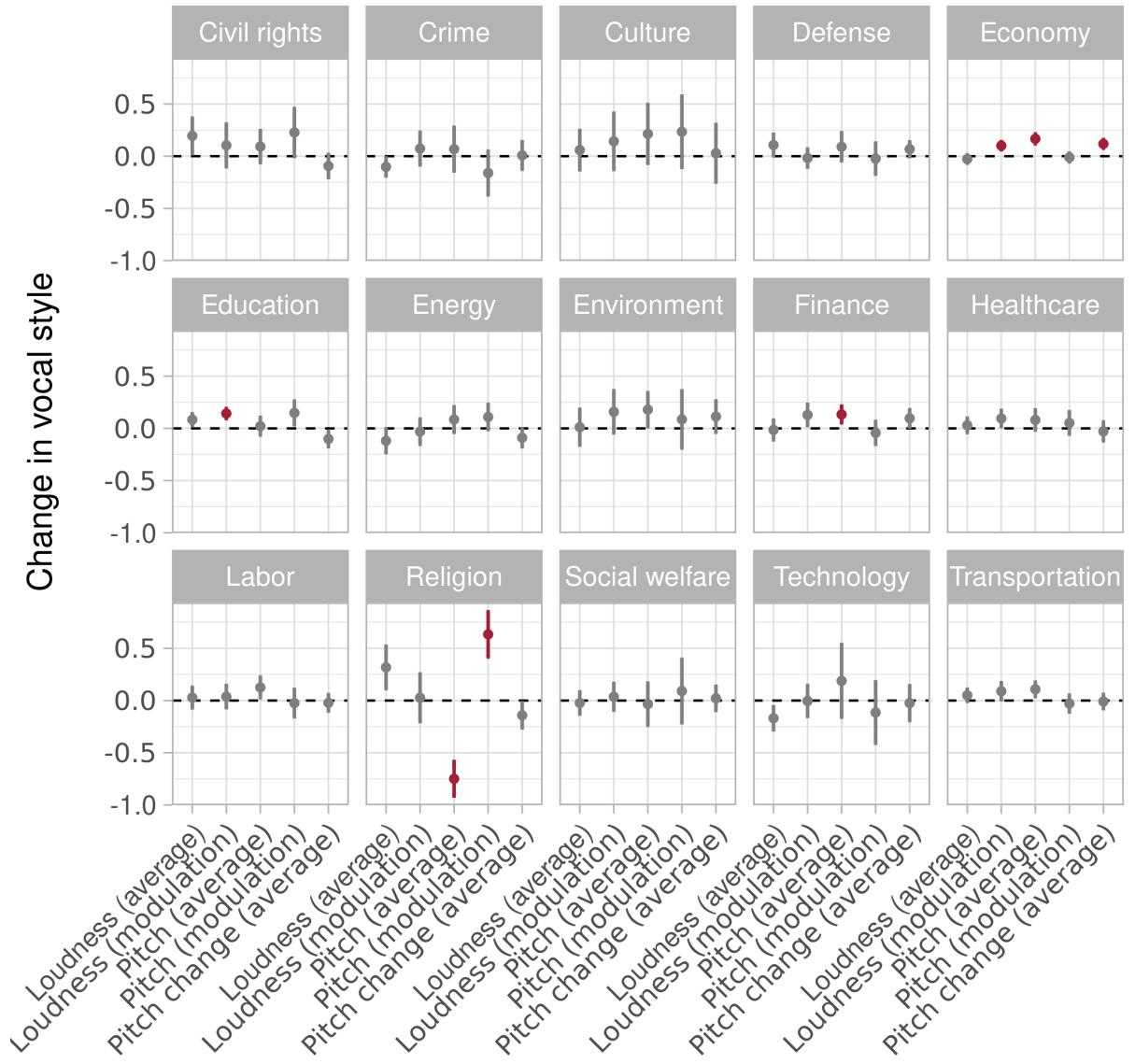
Figure 2: Change in vocal style by Obama conditional on the topic of speech. Red estimates are those which remain significant after a multiple testing correction. Table 9 in the appendix presents these results in tabular form.

variation in modulation, both in pitch and volume, but seemingly less variation in the mean of these features. These results also show that Obama modulates both volume and pitch to a greater degree than does Romney, while Romney has a higher average volume than Obama—differences that are broadly consistent with the numerous popular accounts suggesting that Obama is a more talented public speaker than Romney.
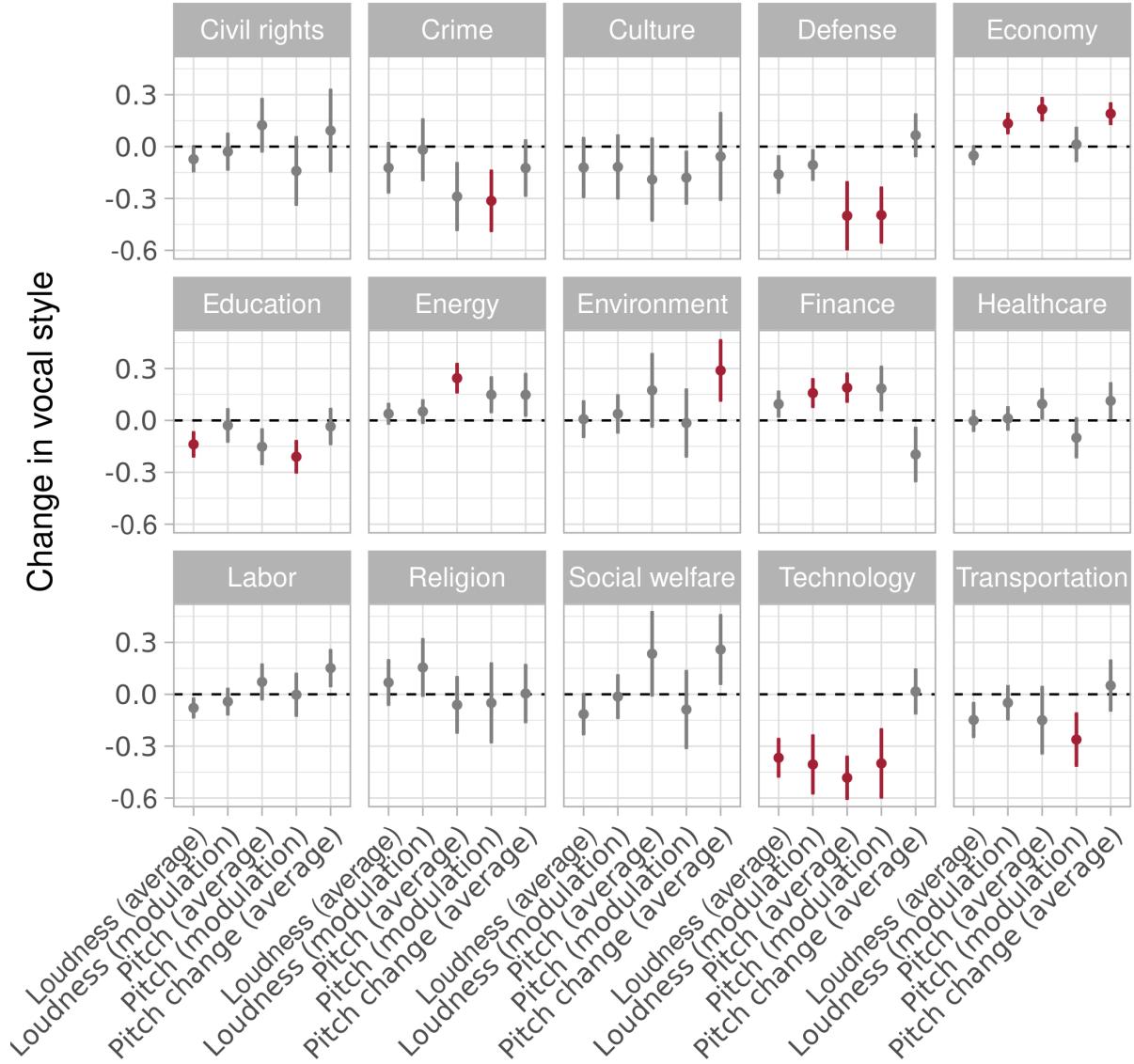
Figure 3: Change in vocal style by Romney conditional on the topic of speech. Red estimates are those which remain significant after a multiple testing correction. Table 10 in the appendix presents these results in tabular form.

4 A Causal Framework for Studying Effects of Textual, Auditory, and Visual Speech Components

In this section, we introduce a formal causal framework for studying the effects of audiovisual treatments, such as recorded campaign speech. Our approach draws on prior work on causal inference in text (Egami et al., 2018) and conjoint experiments (Hainmueller et al., 2014). We consider a sample of N voters, indexed by $i \in \{1, \ldots, N\}$, who consume a series of J candidate utterances, indexed by $j \in \{1, \ldots, J\}$, where an utterance is approximately a sentence-length statement. We denote the j-th utterance consumed by respondent i with the triple $\boldsymbol{U}_{ij} = \{\boldsymbol{T}_{ij}, \boldsymbol{A}_{ij}, \boldsymbol{V}_{ij}\}$, respectively corresponding to the textual, auditory, and visual components of the utterance.[6] In what follows, we will denote the collection of J utterance transcripts observed by respondent i as $\bar{\boldsymbol{T}}_i = \{T_{i1}, \ldots, T_{iJ}\}$; similarly, the collection of utterance audio and visual recordings will be denoted $\bar{\boldsymbol{A}}_i = \{A_{i1}, \ldots, A_{iJ}\}$ and $\bar{\boldsymbol{V}}_i = \{V_{i1}, \ldots, V_{iJ}\}$. After consuming the candidate's j-th utterance, the i-th voter forms a K-dimensional evaluation, indexed by $k \in \{1, \ldots, K\}$, that we collect in an outcome vector $\boldsymbol{Y}_{ij} = [Y_{ij1}, \ldots, Y_{ijK}]$. This multidimensional evaluation includes the voter's evaluation of the candidate's competence and trustworthiness, as well as the respondent's willingness to vote for the candidate.

In studying the causal effects of candidate speech, researchers are interested in understanding how voters would have evaluated a candidate, counterfactually, if voters had been exposed to an utterance with different textual, auditory, or visual components. We denoted the components of this counterfactual utterance as $\boldsymbol{t}$, $\boldsymbol{a}$, and $\boldsymbol{v}$, respectively. Quantifying these effects in speech data, which is highly unstructured, is a challenging task. To do so, we rely on the notion that a complex or high-dimensional treatment can be represented with a "sufficient reduction" that summarizes all possible aspects of the treatment that can influence the outcome. In the text-analysis setting, Egami et al. (2018) refers to such sufficient reductions as "codebook functions" which map a high-dimensional sequence of words, $\boldsymbol{t}$, into a low-dimensional representation, $g_T(\boldsymbol{t})$, such as the presence or absence of

---

[6]Throughout, we will use {} to denote ordered sets.

a topic (see also Fong and Grimmer, 2016). This broad formulation encapsulates numerous analytic approaches used to study the effects of text dictionary-based classification, bag-of-words representations, as well as topic models (Roberts, Stewart, Tingley, et al., 2014; Roberts, Stewart, and Airoldi, 2016) and text-embedding models (Rodriguez and Spirling, 2022) learned from the data. Sufficient-reduction assumptions are commonly used in network studies of "peer effects," or the contagion of behavior, where scholars often suppose that a focal individual's decisions are driven by the number or proportion of peers adopting a particular behavior, a simple-to-analyze scalar, rather than the specific identities of those peers, a vector that can take on combinatorially many values (Eckles et al., 2016; Bramoullé et al., 2020).

This concept of a sufficient reduction can be extended to non-textual components of speech. For example, Dietrich, Enos, et al. (2019) employs an audio reduction in which $\boldsymbol{a}$ is a Supreme Court justice's utterance and $g_A(\boldsymbol{a})$ is defined as the average vocal pitch of that utterance, which is shown to correlate with their voting. Knox and Lucas (2021), also in a study of Supreme Court Oral Arguments, model $g_A(\boldsymbol{a})$ with a supervised hidden-Markov-model classification of each speaker's vocal tone, mapping justice utterances into domain-relevant categories—"skeptical" or "neutral" questioning. These sufficient reductions are then used to study the flow of conversation in judicial deliberations. In this paper, we represent the vocal characteristics of each candidate utterance with a multidimensional $g_A(\boldsymbol{a})$ that covers a plethora of auditory summary statistics, including speech rate along with levels and variation in pitch and volume. In principle, analysts can employ visual reductions (Torres, 2018), $g_V(\boldsymbol{v})$, to represent elements of visual style, such as facial expressions and head movements as in Boussalis et al., 2021; Reece et al., 2022. We do not pursue this approach in this study of candidate vocal expression, due to the difficulty of manipulating candidate facial expressions while holding audio fixed. However, recent computational advances in the creation of "deepfakes"—fabricated videos synthesized by deep learning—may make it possible to conduct experiments of this sort (Barari et al., 2021).

Finally, completing our notation, for simplicity we will use $\bar{g}_X()$ to denote the repeated application of the sufficient reduction function to multiple utterances, so that $\bar{g}_X(\bar{\boldsymbol{X}}_i) =$

$\{g_X(\boldsymbol{X}_{i1}),\dots,g_X(\boldsymbol{X}_{iJ})\}$, .

We are now ready to introduce a potential-outcome framework (Neyman, 1923; Rubin, 1974) for studying the effects of speech. Let $Y_{ijk}(\bar{\boldsymbol{u}})$ denote the potential evaluation by respondent i on candidate characteristic k that would be observed after the j-th utterance, if they were counterfactually assigned to the sequence of candidate utterances represented by $\bar{\boldsymbol{u}}$, which is comprised of components $\{\bar{\boldsymbol{t}}, \bar{\boldsymbol{a}}, \bar{\boldsymbol{v}}\}$ (respectively, the sequence of transcripts, audio recordings, and silent video recordings).

Assumption 1 (Sufficiency of reduced representation). *$Y_{ijk}(\bar{\boldsymbol{u}}) = Y_{ijk}(\bar{\boldsymbol{u}}')$ for $\boldsymbol{u} = \{\bar{\boldsymbol{t}}, \bar{\boldsymbol{a}}, \bar{\boldsymbol{v}}\}$ and $\boldsymbol{u}' = \{\bar{\boldsymbol{t}}, \bar{\boldsymbol{a}}, \bar{\boldsymbol{v}}\}$ if $\bar{g}_T(\bar{\boldsymbol{t}}) = \bar{g}_T(\bar{\boldsymbol{t}}')$, $\bar{g}_A(\bar{\boldsymbol{a}}) = \bar{g}_A(\bar{\boldsymbol{a}}')$, and $\bar{g}_V(\bar{\boldsymbol{v}}) = \bar{g}_V(\bar{\boldsymbol{v}}')$.*

This assumption allows us to rewrite $Y_{ijk}(\bar{\boldsymbol{u}})$ in terms of the sufficient reductions for each utterance, $Y_{ijk}(\bar{g}_T(\bar{\boldsymbol{t}}), \bar{g}_A(\bar{\boldsymbol{a}}), \bar{g}_V(\bar{\boldsymbol{v}}))$.

This formulation is without loss of generality for two reasons: (1) text, audio, and visual reduction functions are allowed to be arbitrarily complex, and (2) because this formulation does not restrict interference between successive utterances, such as gradual updating by a voter over the course of a campaign speech. We now discuss each of these in turn. By justifying assumptions about $g_T(\cdot)$, $g_A(\cdot)$, and $g_V(\cdot)$, analysts can use domain expertise to place more assumed structure on the way that voters respond to campaign speech. When these are taken to be the identity function, so that no reduction is made at all, analysts effectively assume that even the slightest deviation—a stray "uh," the slightest pause, or a miscolored pixel—can produce entirely different potential evaluations. In contrast, when analysts make more restrictive assumptions about sufficient reductions, this notation implicitly makes a stable unit treatment value assumption (SUTVA, Rubin, 1980) that any variation in $\boldsymbol{t}$, $\boldsymbol{a}$, or $\boldsymbol{v}$ is causally irrelevant as long as they have the same sufficient reduction, i.e. that $g_T(\boldsymbol{t}) = g_T(\boldsymbol{t}')$, $g_A(\boldsymbol{a}) = g_A(\boldsymbol{a}')$, and $g_V(\boldsymbol{v}) = g_V(\boldsymbol{v}')$.[7] When $g_T(\cdot)$ counts the number of words in an utterance that appear in a keyword dictionary, analysts assume that word ordering and non-dictionary words have no causal effect on opinion formation. Similarly, when $g_A(\cdot)$ measures only the average pitch, analysts assume that a monotonous

---

[7]Formally, $Y_{ijk}(g_T(\bar{\boldsymbol{t}}), g_A(\bar{\boldsymbol{a}}), g_V(\bar{\boldsymbol{v}})) = Y_{ijk}(g_T(\bar{\boldsymbol{t}}'), g_A(\bar{\boldsymbol{a}}'), g_V(\bar{\boldsymbol{v}}'))$ if $g_T(\bar{\boldsymbol{t}}) = g_T(\bar{\boldsymbol{t}}')$, $g_A(\bar{\boldsymbol{a}}) = g_A(\bar{\boldsymbol{a}}')$, and $g_V(\bar{\boldsymbol{v}}) = g_V(\bar{\boldsymbol{v}}')$.

drone is interchangeable with a highly modulated utterance centered on the same value. Analysts' context-specific assumptions about the nature of these sufficient-reduction functions therefore play an essential role in causal inference about the effects of speech (Egami et al., 2018).

In this paper, we will make the simplifying assumption—defined formally in Assumption 2—that a respondent's potential evaluation in one task does not depend on the candidate speech that they have been exposed to in the past.

Assumption 2 (No cross-utterance interference). $Y_{ijk}(\bar{\boldsymbol{t}}, \bar{\boldsymbol{a}}, \bar{\boldsymbol{v}}) = Y_{ijk}(\bar{\boldsymbol{t}}', \bar{\boldsymbol{a}}', \bar{\boldsymbol{v}}')$ for all i, k and for all speech component pairs $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}')$ differing only in the j-th position, i.e. with $\{g_X(\bar{\boldsymbol{x}}_{1:(j-1)}), \boldsymbol{x}, g_X(\bar{\boldsymbol{x}}_{(j+1):J})\}$ and $\bar{\boldsymbol{x}}' = \{g_X(\bar{\boldsymbol{x}}'_{1:(j-1)}), \boldsymbol{x}, g_X(\bar{\boldsymbol{x}}'_{(j+1):J})\}$.

This states that an individual's potential responses after being exposed to utterance j will be the same regardless of what they have been exposed to in the past or will be exposed to in the future. This is closely related to the "no interference" component of SUTVA, as well as the "no carryover effect" and "no profile-order effect" assumptions commonly employed in the conjoint literature (Hainmueller et al., 2014). We note that this is a strong assumption in the campaign speech setting, where voters form opinions about candidates gradually by consuming hundreds or even thousands of utterances over a campaign season. However, it may *approximately* hold in the settings of Experiments 1 and 2, to the extent that respondents learn only a small amount about a candidate from each campaign-speech utterance. With this simplifying assumption, we can eliminate past and future utterances from our potential outcomes, dropping the j subscript to obtain the simplified notation $Y_{ik}(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v}))$. However, we emphasize that developing experimental designs for studying the accumulated effects of campaign speech exposure remains an important direction for future work.

Next, we formalize and discuss a core assumption in prior text-based research. Scholars using transcripts to study the effects of campaign speeches—extracting and analyzing only $\boldsymbol{t}$—are effectively assuming that paralinguistic cues are causally irrelevant. That is, analysts discard the auditory and visual components of speech, $\boldsymbol{a}$ and $\boldsymbol{v}$, setting them equal to the empty set, $\varnothing$. Thus, analysts can only elicit $Y_{it}(g_T(\boldsymbol{t}), \varnothing, \varnothing)$ from respondents. In essence,

this past work implicitly assumes that any other way of delivering the same words would have produced the same audience reaction.

Assumption 3 (Irrelevance of paralinguistic cues)**.**

$$Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right) = Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), g_V(\boldsymbol{v}')\right) \text{ for all } \boldsymbol{a}, \boldsymbol{a}', \boldsymbol{v}, \boldsymbol{v}'.$$

In many settings, this can be weakened to require only equality in expectations.[8]

We are now ready to formally define the experiments presented in Sections 5 and 6. In Experiment 1 (Section 5), we test Assumption 3 and find that it is entirely incompatible with actual candidate evaluations. We use a novel phrase-clustering method to identify instances of a candidate recycling a well-worn campaign "catchphrase," $\boldsymbol{u} = \{\boldsymbol{t}, \boldsymbol{a}, \boldsymbol{v}\}$ and $\boldsymbol{u}' = \{\boldsymbol{t}', \boldsymbol{a}', \boldsymbol{v}'\}$ in two differing styles, so that $\boldsymbol{t} = \boldsymbol{t}'$ but $\boldsymbol{a} \neq \boldsymbol{a}'$ and $\boldsymbol{v} \neq \boldsymbol{v}'$. Respondents are exposed to videos of both catchphrase variants, then asked to select the variant that leads to a more positive evaluation—that is, identifying whether $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right)$ or $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), g_V(\boldsymbol{v}')\right)$ is larger—and test the null hypothesis that this choice probability is equal to $\frac{1}{2}$, as Assumption 3 suggests. To ensure respondents are influenced by vocal style, we then repeat this experiment with audio recordings only, eliciting comparisons between $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right)$ or $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right)$ is larger. Finally, we expand our analyses to the common scenario where wording differs slightly, so that $\boldsymbol{t} \neq \boldsymbol{t}'$. We develop a novel "difference in differences" design that compares the text-only contrast, $Y_{ik}\left(g_T(\boldsymbol{t}), \varnothing, \varnothing\right)$ versus $Y_{ik}\left(g_T(\boldsymbol{t}), \varnothing, \varnothing\right)$, to the audio contrast, $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right)$ versus $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right)$. Finally, we formalize a key assumption under which the difference in differences can be used to test the null hypothesis of Assumption 3.

While Experiment 1's use of actual U.S. presidential candidate speech allows us to evaluate the impact of vocal style in a highly naturalistic setting, this experimental approach also constrains the types of questions that can be asked. We therefore design Experiment 2 (Section 6) to address two specific limitations. First, the real-world recordings used in Experiment 1 are constrained by the fact that vocal style for a particular catchphrase will

---

[8]Note that our formulation is stronger than the sufficiency assumption of Egami et al. (2018), which states only that $\mathbb{E}_i\left[Y_{ijk}(g_T(\bar{\boldsymbol{t}}), g_A(\bar{\boldsymbol{a}}), g_V(\bar{\boldsymbol{v}}))\right] = \mathbb{E}_i\left[Y_{ijk}(g_T(\bar{\boldsymbol{t}}'), g_A(\bar{\boldsymbol{a}}'), g_V(\bar{\boldsymbol{v}}'))\right]$. This is due to a subtle issue in our use of a paired-utterance forced-choice design in Experiment 1 (Section 5), which requires distributional equality rather than merely equality in expectation.

only vary within a narrow window—perhaps slightly more sluggish after several tiring days of campaigning or slightly more energetic before a boisterous crowd, but all within the range of a candidate's baseline speaking style. In Experiment 2, we use a combination of voice actors and audio-editing manipulations to examine more substantively meaningful dimensions of variation in campaign speech. We examine realistic interventions on two dimensions—speech rate and vocal modulation—corresponding to common aspects of real-world training in public speaking. Voice actors are encouraged to read scripts quickly, slowly, monotonously, and dynamically. We demonstrate how these encouragements manifest in our audio summary statistics and show that despite the fact that encouragements are targeted to specific elements of $g_A(\boldsymbol{a})$, it is difficult even for professional actors to modify one dimension of voice (e.g., speed) in isolation from others (e.g., loudness, pitch, and modulation). To examine the contribution of individual vocal elements, we therefore edit the audio to artificially modify pitch and loudness while holding other aspects of speech constant. Second, while the paired-utterance, forced-choice design of Experiment 1 is useful for maximizing statistical power, it is ill-suited for quantifying the magnitude of a vocal style shift on candidate evaluations. Therefore, in Experiment 2, we present respondents with one audio recording at a time, $\boldsymbol{u} = \{\boldsymbol{t}, \boldsymbol{a}\}$, then ask them to report $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right)$.

5 Experiment 1: Real Campaign Speech

We now design a naturalistic experiment that leverages variation in how candidates deliver campaign catchphrases in the corpus described in Section 3. We first use a new computational technique for "substring clustering" to identify frequently repeated catchphrases. Then, we locate pairs of utterances with identical or near-identical wording but differing vocal style. These matched pairs are used to test the null hypothesis that vocal style has no effect on listener perception. We use this approach to study the impact of vocal style in a maximally faithful setting: using real-world campaign messages, delivered in real-world campaign vocal styles, tested on a sample of real-world voters.

We find strong evidence that variation in candidate vocal delivery has an effect on voter

15

evaluations. Importantly, the differences in vocal style that we exploit are extremely subtle. Candidates for the U.S. presidency are selected in part for being skilled public speakers, and they have strong incentives to perform optimally throughout their campaign. Experiment 1 is therefore an especially conservative test, as most plausible real-world interventions—for example, professional speech coaching or focus-group evaluation of speech styles—are likely to create larger shifts in vocal style than the slight deviations that we study here.

5.1   Designing the Naturalistic Experiment

To design our experiment, we first identify instances in which Obama or Romney uttered identical or near-identical statements on the campaign trail. We began by comparing every 10-word sequence in the corpus to every other 10-word sequence in the corpus. This is an extremely computationally intensive procedure involving roughly 90 billion pairwise comparisons. Accomplishing this task in an efficient manner required the development of a new text-matching algorithm. Briefly, we (1) propose a new distance metric based on the correlation in letter frequencies between each pairwise comparison; (2) use this metric to reformulate the string-search problem as a convolution problem; and (3) exploit the Fourier convolution theorem to sweep a single phrase over an entire target document with only a handful of mathematical operations. Details are provided in Appendix Section A. The chief benefit of this approach is that it is much faster—by up to 60 times, in our testing—than the current state-of-the-art computational technique for fuzzy substring matching, `agrep`.

The speed of this approach allows us to compute similarity scores for every pair of k-word sequences in the corpus. We then construct a network of phrases and apply network clustering techniques to identify sets of approximately matched 10-word sequences, extend sequences to complete sentences, and identify recurring "catchphrases." Next, human coders inspected raw video for each group of catchphrases, qualitatively assessing both the cohesion of transcripts and the divergence of vocal delivery for utterances in a catchphrase group. They identified catchphrase clusters with a relatively large degree of naturally-occurring variation in spoken delivery, then noted the start and stop times of the complete sentences (rather than the k-word sequence) for each recording in the community. From these, we

selected 29 matched pairs of utterances with identical or near-identical phrasing.

From each pair of matched recordings, we created three conditions: textual transcripts, audio recordings, or full video of the utterance pairs. We asked respondents to evaluate the utterances on $K = 8$ dimensions. Respondents selected the versions that made them feel more angry, afraid, hopeful, and proud; as well the versions that were more consistent with a statement made by a strong, knowledgeable, moral, and inspiring leader. (Respondents were assumed to answer randomly when they are indifferent.) We adopt this paired-utterance approach to allow within-respondent comparisons, with the goal of addressing potential power issues due to the relatively subtle $\boldsymbol{a}$-$\boldsymbol{a}'$ and $\boldsymbol{v}$-$\boldsymbol{v}'$ differences. The forced-choice design avoids the potentially confusing scenario of asking a respondent to evaluate a candidate twice after being exposed to two similar recordings.

We fielded the experiment on a sample of actual voters in the 2016 U.S. presidential election, using Amazon Mechanical Turk. The 29 catchphrases were divided into three batches, in which subjects were sequentially shown nine or ten catchphrases (paired utterances), with utterance modality (text, audio, or video) randomly assigned at the pair level. Subjects were permitted to participate in more than one batch but could complete each batch once, ensuring that no individual was assigned the same catchphrase more than once. After dropping subjects who failed an audio-based attention check and/or had duplicated IP addresses, 773 voters participated in the first experiment. On average, more than 250 voters coded each phrase. Appendix Section C displays screenshots depicting exactly what respondents saw, and Appendix Figure 13 plots the means for each of these conditions, for each of the paired recordings, demonstrating substantial variability across these conditions.[9]

5.2 Null Hypothesis Testing in the Naturalistic Experiment

We begin by introducing a stock phrase that Obama repeats verbatim in back-to-back campaign appearances on November 1, 2012: "Let's put Americans back to work doing the work that needs to be done." When campaigning in Boulder, CO, Obama speaks deliberatively ($\boldsymbol{u} = \{\boldsymbol{t}, \boldsymbol{a}, \boldsymbol{v}\}$); in contrast, when appearing in Green Bay, WI, he delivers the same message

---

[9]Attention checks and IP filtering resulted in slight variation in sample size across phrases.

emphatically ($\boldsymbol{u}' = \{\boldsymbol{t}', \boldsymbol{a}', \boldsymbol{v}'\}$). We first develop the basic logic of the experimental design, using notation introduced in Section 4, before extending the approach to account for the slight variations in phrasing that appear in other identified catchphrase pairs.

In this utterance pair, the two transcripts are a perfect match, so that $\boldsymbol{t} = \boldsymbol{t}'$. However, vocal and nonverbal delivery differ in the two appearances, so that $\boldsymbol{a} \neq \boldsymbol{a}'$ and $\boldsymbol{v} \neq \boldsymbol{v}'$. We are therefore able to directly evaluate Assumption 3 by assessing whether $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right)$ and $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), g_V(\boldsymbol{v}')\right)$ are equal. To do so, we exposed one third of respondents to videos of both $\boldsymbol{u}$ and $\boldsymbol{u}'$ in randomized order.

Assumption 3 implies that the two evaluations will be equal.

$$\mathbf{H1}: \quad \Pr\left(Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right) - Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), g_V(\boldsymbol{v}')\right) > 0\right)$$
$$+ \frac{1}{2}\Pr\left(Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right) - Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), g_V(\boldsymbol{v}')\right) = 0\right) = \frac{1}{2}$$

We reject this null hypothesis at $p < 0.001$ for each of the $K = 8$ evaluation criteria. For example, 72% of respondents found the emphatic variant of the utterance to be more consistent with strong leadership, and 76% said it made them feel more proud.

Next, to rule out the possibility that respondent evaluations are driven by the visual channel, rather than the audio component that is the focus of this work, we exposed another third of respondents to audio recordings alone, so that evaluations were based on a comparison between $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right)$ and $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right)$, eliminating the visual channel. We then test the following null hypothesis.

$$\mathbf{H2}: \quad \Pr\left(Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right) - Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right) > 0\right)$$
$$+ \frac{1}{2}\Pr\left(Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right) - Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right) = 0\right) = \frac{1}{2}$$

Again, we find strong evidence that vocal style matters. In every criterion, respondents exhibited a preference for one video over the other by more than 35 percentage points, and the null hypothesis is rejected at $p = 0.005$ or less for each of the $K = 8$ evaluation criteria.

While this simple approach allows us to reject Assumption 3—the assumed irrelevance of vocal delivery that is implicit in much prior work—it is not directly applicable to many of the catchphrases that we identify. Most catchphrases are repeated with slight variations,

which can range from the minor as the insertion of a stray "so" or the contraction of "I will" to "I'll."

For example, on September 12, 2012, after Ansar al-Sharia's attack on the U.S. consulate in Benghazi, Obama stated, "We still face threats in this world, and we've got to remain vigilant. But that's why we will be relentless in our pursuit of those who attacked us yesterday. But that's also why, so long as I'm commander in chief, we will sustain the strongest military the world has ever known." His speech was highly modulated, with punctuated bursts of loudness and well-timed pauses. The next day, however, Obama delivered a listless and halting variant on this theme, stumbling over many of the same words. However, a direct comparison between two audio recordings does not allow us to test Assumption 3, because we cannot rule out the possibility that differences in respondent evaluations were due to minor differences in wording—his use of "There are still threats" instead of "We still face threats," or "we have to be relentless in pursuing" instead of "we will be relentless in our pursuit." To deal with this issue, we develop a "difference in differences" design that compares the pair of audio recordings, $\{t, a\}$ and $\{t', a'\}$, to the pair of utterance transcripts alone, t and t'. Intuitively, the goal of doing so is to measure the gap in evaluations for two audio utterances (differing in both transcript and vocal delivery), measure the gap in their textual versions (differing only in transcript), and subtract the textual gap from the audio gap to estimate the portion due to vocal delivery alone. Formally justifying this procedure requires one additional assumption, which we make explicit below. Assumption 4 is only used in the context of Experiment 1.

Assumption 4 (Additive separability of potential evaluations)**.**
$Y_{ik}(g_T(t), g_A(a), g_V(v)) = \alpha_{ik} + h_{ik}^T(g_T(t)) + h_{ik}^A(g_A(a)) + h_{ik}^V(g_V(v))$, *where* α_{ik} *represents respondent* i*'s baseline evaluation on metric* k*, and* $h_{ik}^X(\cdot)$ *denotes deviations from that baseline evaluation based on sufficient reductions of component* X*.*

This states that candidate speech text and speech audio do not interact in terms of how they contribute to a respondent's potential evaluations.[10] It is closely related to the parallel

---

[10]In many settings, Assumption 4 can be weakened to an assumption about additive separability of the

19

trends assumption in conventional difference-in-differences analyses. An important special case that automatically satisfies Assumption 4 is when the $\boldsymbol{t}$-to-$\boldsymbol{t}'$ manipulation, the $\boldsymbol{a}$-to-$\boldsymbol{a}'$ manipulation, or both manipulations have constant treatment effects. Due to the complexity of candidate speech and voter evaluations, this assumption is unlikely to be generally satisfied for all $\boldsymbol{t}$, $\boldsymbol{a}$, and $\boldsymbol{v}$. However, because the manipulations studied in Experiment 1 are generally extremely subtle, it may hold approximately for the specific variations in transcript and vocal style that we study.

Under Assumption 4, the forced-choice probability between audio recording pairs $\{\boldsymbol{t}, \boldsymbol{a}\}$ and $\{\boldsymbol{t}', \boldsymbol{a}'\}$ can be rewritten

$$
\begin{aligned}
&\Pr\left(Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right) - Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right) > 0\right) \\
&\quad + \frac{1}{2}\Pr\left(Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right) - Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right) = 0\right) \\
&= \Pr\left(h_{ik}^T(g_T(\boldsymbol{t})) + h_{ik}^A(g_A(\boldsymbol{a})) - h_{ik}^T(g_T(\boldsymbol{t}')) - h_{ik}^A(g_A(\boldsymbol{a}')) > 0\right) \\
&\quad + \frac{1}{2}\Pr\left(h_{ik}^T(g_T(\boldsymbol{t})) + h_{ik}^A(g_A(\boldsymbol{a})) - h_{ik}^T(g_T(\boldsymbol{t}')) - h_{ik}^A(g_A(\boldsymbol{a}')) = 0\right)
\end{aligned}
$$

Assumption 3—irrelevance of paralinguistic cues, an unstated assumption in prior text-based work—then implies that $h_{ik}^A(g_A(\boldsymbol{a})) = 0$, suggesting that forced-choice proportions between utterances $\boldsymbol{u}$ and $\boldsymbol{u}'$ should be equal regardless of whether respondents are exposed to the text or audio variants of the utterance pair. This is given formally in the following null hypothesis:

$$
\begin{aligned}
\textbf{H3}: \quad &\Pr\left(Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right) - Y_{ik}\left(g_T(\boldsymbol{t}'), g_A(\boldsymbol{a}'), \varnothing\right) > 0\right) \\
&\quad + \frac{1}{2}\Pr\left(Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right) - Y_{ik}\left(g_T(\boldsymbol{t}'), g_A(\boldsymbol{a}'), \varnothing\right) = 0\right) \\
&= \Pr\left(Y_{ik}\left(g_T(\boldsymbol{t}), \varnothing, \varnothing\right) - Y_{ik}\left(g_T(\boldsymbol{t}'), \varnothing, \varnothing\right) > 0\right) \\
&\quad + \frac{1}{2}\Pr\left(Y_{ik}\left(g_T(\boldsymbol{t}), \varnothing, \varnothing\right) - Y_{ik}\left(g_T(\boldsymbol{t}'), \varnothing, \varnothing\right) = 0\right)
\end{aligned}
$$

We use this approach to examine voter evaluations in the text-based contrast and find that mild wording variations in Obama's response to the Benghazi attack—his catchphrase about

conditional expectation function, rather than the individual-level potential-outcome function itself. When examining single-utterance ratings, the weaker assumption that $\mathbb{E}\left[Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right) \leq y\right] = h_T(g_T(\boldsymbol{t})) + h_A(g_A(\boldsymbol{a}))$ will generally suffice. As with Assumption 3, we require stronger assumptions when analyzing the paired-profile forced-choice design of Experiment 1.

"maintaining the strongest military the world has ever known"—have no discernible effect on voter evaluations. Respondents reading the utterance transcripts had no statistically significant preference for either phrasing ($p = 0.754$), though slightly more selected the earlier variant as being consistent with an inspiring leader (difference in choice probability of 4 percentage points). In contrast, respondents exposed to the audio recordings were able to hear the dynamicism and emphasis in Obama's earlier speech. As a result, they were 40 percentage points more likely to select it as the more inspirational variant, compared to the later, listless recording. In a χ^2 test of equal proportions, we reject **H3** at $p = 0.018$.

All in all, despite the subtle variation in vocal delivery utilized in this experiment, we find strong evidence that speech shapes voter evaluations. Aggregating across catchphrases, we estimate that the average magnitude of vocal style effects is an 11.4-percentage-point (p.p.) change in choice probability. (Here, we define the audio effect as the deviation of audio choice probability from $\frac{1}{2}$ when wording is identical, or deviation from the text choice proportion otherwise.) Substantively speaking, it does not appear that the visual component of speech strengthens these effects (11.1 p.p. difference relative to text). Audio effect estimates are smallest for "consistent with a knowledgeable leader" (9.7 p.p.), which may be a more difficult concept to gauge in a short utterance; they are largest for "consistent with a strong leader" (12.5 p.p.).

To account for multiple testing across a large number of voter evaluation metrics, as well as the nesting of these evaluations within catchphrases, we adopt the hierarchical procedure of Peterson et al. (2016). This approach uses on a combination of (1) the Simes method (Simes, 1986) for testing the intersection null, that choice probabilities on any evaluation metric are unaffected by vocal style within a catchphrase and (2) the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) for controlling the false discovery rate across catchphrases. After applying this procedure, we find vocal style effects are significant at the 0.05 level for catchphrases spanning a "fair shot" at social mobility for hard workers, "offshoring" of American jobs, America's resolve in the face of "terror," real "change" taking time, economic "opportunity," rejection of "top-down" economics putting Americans back to "work," and broken "promises" to save Medicare. Complete transcripts for these and

other catchphrases, identified by their abbreviated names (quoted above), are provided in Tables 3–7.

6   Experiment 2: Voice Actor Treatments

While Experiment 1 demonstrates that naturalistic variation in vocal delivery affects how voters respond to candidates, it is constrained in two ways. First, it is constructed exclusively from practiced campaign speeches delivered by candidates competing for the presidency. However, if indeed candidates are selected in part due to their ability to effectively communicate with prospective voters, both President Obama and Senator Romney ought to be especially well-practiced and competent speakers, given the stakes of the campaign and their relatively extensive electoral success. The range of rhetorical skill within less experienced candidates, especially those running for down-ballot offices, is likely much wider than that displayed by Obama and Romney, making our test a rather conservative one. In addition, for most pairs, we are unable to hold text completely fixed.[11] As our framework in Section 4 establishes, these textual differences complicate interpretation.

With these considerations in mind, we design a second experiment in which we hire 10 actors to record themselves reading a series of scripts in varied fashion. We then further computationally manipulated these recordings to create a total of 960 audio recordings, which serve as the basis for the audio conjoint experiment that we now describe.

6.1   Designing an Actor-Assisted Experiment

To identify the effects of different components of campaign speech delivery, we create our own audio treatments in order to carefully control elements of $g_A(\boldsymbol{a})$, the experimental manipulations, beyond what is possible with naturalistic treatments. To do so, we first selected six scripts from actual political speeches, insuring that the topics of these scripts vary in substance and partisanship. Appendix Section D provides the complete scripts and indicates the speeches from which they are drawn. Two are selected from the 2012 campaign catchphrases identified in our first experiment, two are statements made by former President

---

[11]See Appendix Section B for the complete text of Experiment 1.

Donald Trump, one is from a speech by former Secretary of Education Betsy DeVos, and the last is from former President Obama's 2009 address to the U.N. on climate change. We use a variety of scripts to avoid drawing inferences that are overly reliant on unique interactions between a vocal characteristic and a particular topic.

We then hired 10 actors—five women and five men—to read and record each script four times: (1) in a monotonous voice with a slow rate of speech; (2) in a monotonous voice with a fast rate of speech; (3) in a modulated voice with a slow rate of speech; and (4) in a modulated voice with a high rate of speed. That is, actors record all combinations of low and high values on modulation and rate. After doing so, we obtain 240 audio recordings (10 actors × 6 scripts × 4 versions). We use these recordings, pooling over the six scripts, to estimate the effect of modulation and rate of speech on voter appraisals of hypothetical candidates.

We manipulate these two components of speech, modulation and speech rate, because they are among the simplest ways to differentiate skilled and practiced speakers from their untrained counterparts. Skilled orators rarely deliver a rapid, monotonous campaign speech—an observation that is anecdotally supported by our interactions with professional voice actors, who consistently balked at our request that they deliver a monotonous, hurried speech and insisted that it would not sound convincing.

We then computationally manipulate these actor-provided recordings, shifting average pitch and average loudness. In contrast with modulation and loudness, which cannot be reasonably manipulated in an automated fashion without sounding unnatural, loudness and pitch are arguably easier to manipulate with audio editing software than by actors. It is difficult to naturally shift loudness or pitch by a constant fixed factor, but trivial to do so computationally.[12]

Importantly, it is not the case that actor-controlled manipulations—rate and modulation—are independent of and do not influence pitch and loudness. Rather, the actor-controlled manipulations represent a type of multidimensional variation in $g_A(\boldsymbol{a})$, the summarized audio

---

[12]For loudness, this is equivalent to simply "turning up the volume." For pitch, the algorithm proceeds by simply changing the timescale and sampling rate of the audio. We refer interested readers to Dolson (1986) for further detail.

charactistics that describe an utterance, that correspond broadly to speech skill. In contrast, our computationally-manipulated conditions represent mean shifts in features commonly used to study non-textual components of human speech. Appendix Section E considers this distinction in greater detail.

In sum, then, our experiment consists of four fully-crossed binary conditions (fast/slow rate, low/high modulation, low/high pitch, low/high volume), for a total of 16 unique vocal manipulations. In combination with six scripts and 10 actors, we obtain 960 unique values of $\boldsymbol{a}$. Table 1 presents each of these experimental manipulations.

Feature	Condition	Manipulator
Topic	(1) Budget	Researcher
	(2) Climate	
	(3) Education	
	(4) Military	
	(5) Nationalism	
	(6) Social Policy	
Pitch	(1) High	Researcher
	(2) Low	
Loudness	(1) Loud	Researcher
	(2) Soft	
Rate	(1) Fast	Actor
	(2) Slow	
Variation	(1) Modulated	Actor
	(2) Monotonous	

Table 1: Conjoint Design

After creating these recordings, we fielded an experiment on Mechanical Turk. Each subject heard six recordings—one for each script—drawn randomly from the set of recordings created from that script. After listening to an audio recording, the respondents evaluated the speaker on their competence, enthusiasm, inspiration, passion, persuasion and trustworthiness. Finally, respondents indicated on a scale from 0 to 100 how likely they were to vote for the candidate in an election. In the notation of Section 4, these are $Y_{ik}(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing)$. We

account for the textual contribution to respondent evaluations by only comparing recordings from the same script, t, allowing us to hold fixed the textual information used by respondents, $g_T(t)$. Our quantities of interest relate to average marginal component effect (AMCE, Hainmueller et al., 2014)—either for manipulations targeting a single element of $g_A(a)$, as in our edited recordings, or in multidimensional manipulations that shape multiple elements simultaneously, as in our actor encouragements. In each case, we present estimates that marginalize over all other uniformly randomized treatments (the uniform AMCE, De la Cuesta et al., 2022). We randomized both the assignment of treatment as well as the order of the thematic script presented. This design allows us to manipulate vocal cues directly, which has two benefits. First, we gain insight into which vocal mechanics impact voter perceptions. Next, we can observe the effects of highly varying speech in the presented audio unlike the previous experiment where natural variation between audio was minimal.

6.2   Evaluation by Speech Feature

Our results indicate that how a candidate communicates has substantial effect on voter perception. First, in Figure 4, we plot average willingness to vote for each of the voice actors—a decision that was based only on a brief audio recording. Note here, unlike many of the contrasts in Section 5, the text is held exactly constant since actors read the same scripts. This figure pools over our primary treatments of interest—the effect of speech rate, pitch, volume and modulation—but demonstrates that voice alone, as determined by actor identity, has a strong effect on expressed support. Each actor, anonymously labeled A–J, expressed the same policy positions and manipulated their speech similarly, yet some received considerably more support than others based only on the character of their voice. And while we did not explicitly highlight actor gender, on average, subjects showed significantly more support for men speakers compared to women.

Next, Figure 5 pools speakers and estimates the effect of variation in speech delivery: how speech rate, pitch, volume, and modulation change the way a speaker is perceived on a series of positive characteristics. We report estimates separately for men and women actors and document significant gender heterogeneity. Vocal modulation and rate of speech have
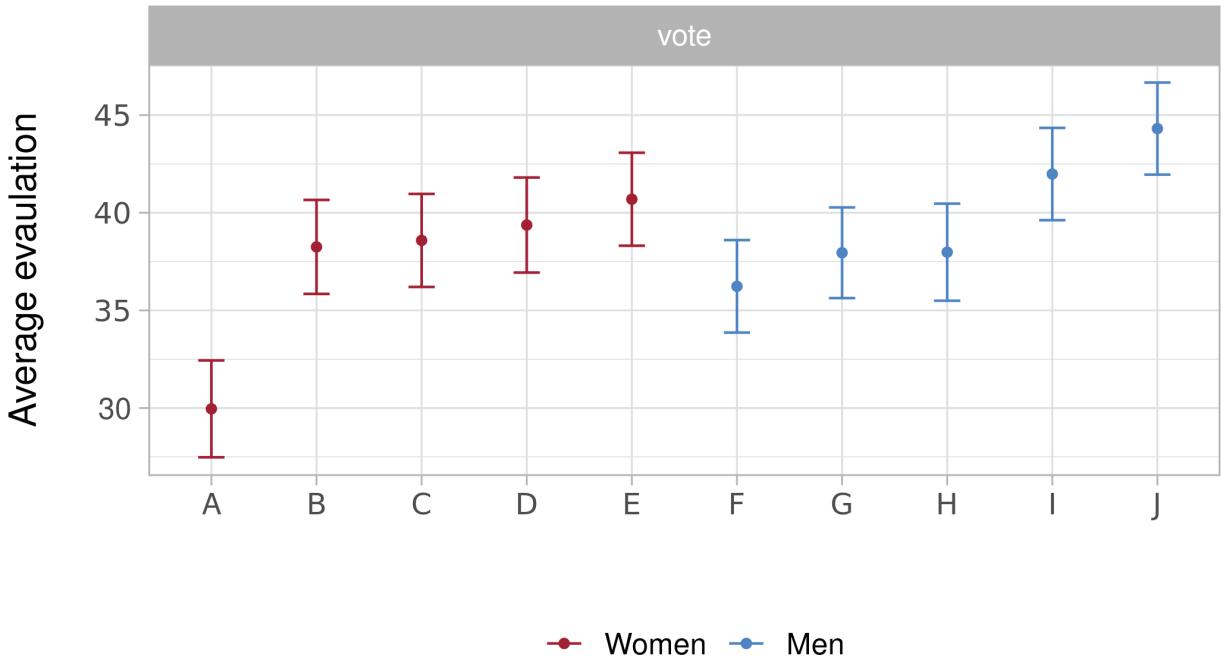
Figure 4: Average expressed willingness to vote for each actor, based only on hearing their recorded speech. Demonstrates that holding content fixed, there is sizeable variation in voter preference. Estimated from a regression with fixed effects for script and indicators for treatment condition. Table 11 presents the results of this regression.

consistently positive effects on positive evaluations of the speaker. Louder speech volumes have a small effect on perceived passion, enthusiasm and persuasion. Pitch is perceived differently than the rest of the evaluative categories. Having a higher pitched speaking voice is associated with a more negative evaluation or no effect. When examining vocal modulation, which primarily manifests in the use of heightened pitch for emphasis, respondents consistently reward women for vocal dynamicism more than they do for men. Men are also punished more than women for having a higher pitched voice, consistent with research on gender stereotypes.

In addition to having respondents evaluate speakers' positive characteristics, we also ask them how willing they are to vote for a person based on the audio of their voice. In Figure 6, we report the effect of vocal manipulations on this outcome. Interestingly, average pitch and volume appears to have relatively little effect, but variation in both—manipulated through an actor encouragement to modulate voice—has a sizeable effect not only on how subjects

perceive candidates, but also on their willingness to vote for the candidate.
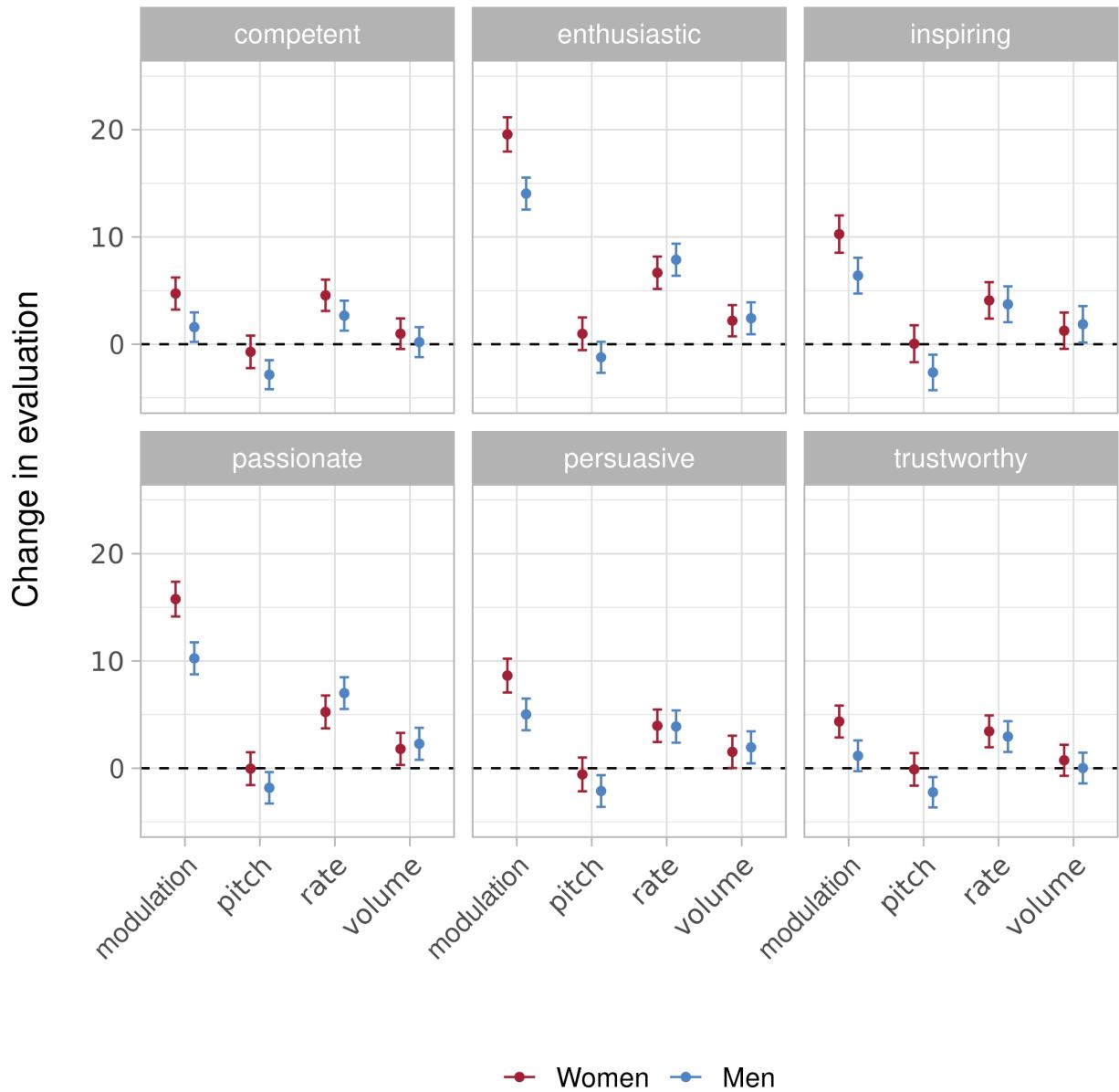
Figure 5: Effect of speech features on evaluations of the respective characteristic by speaker gender. Appendix Section G.1 presents these results in tables.
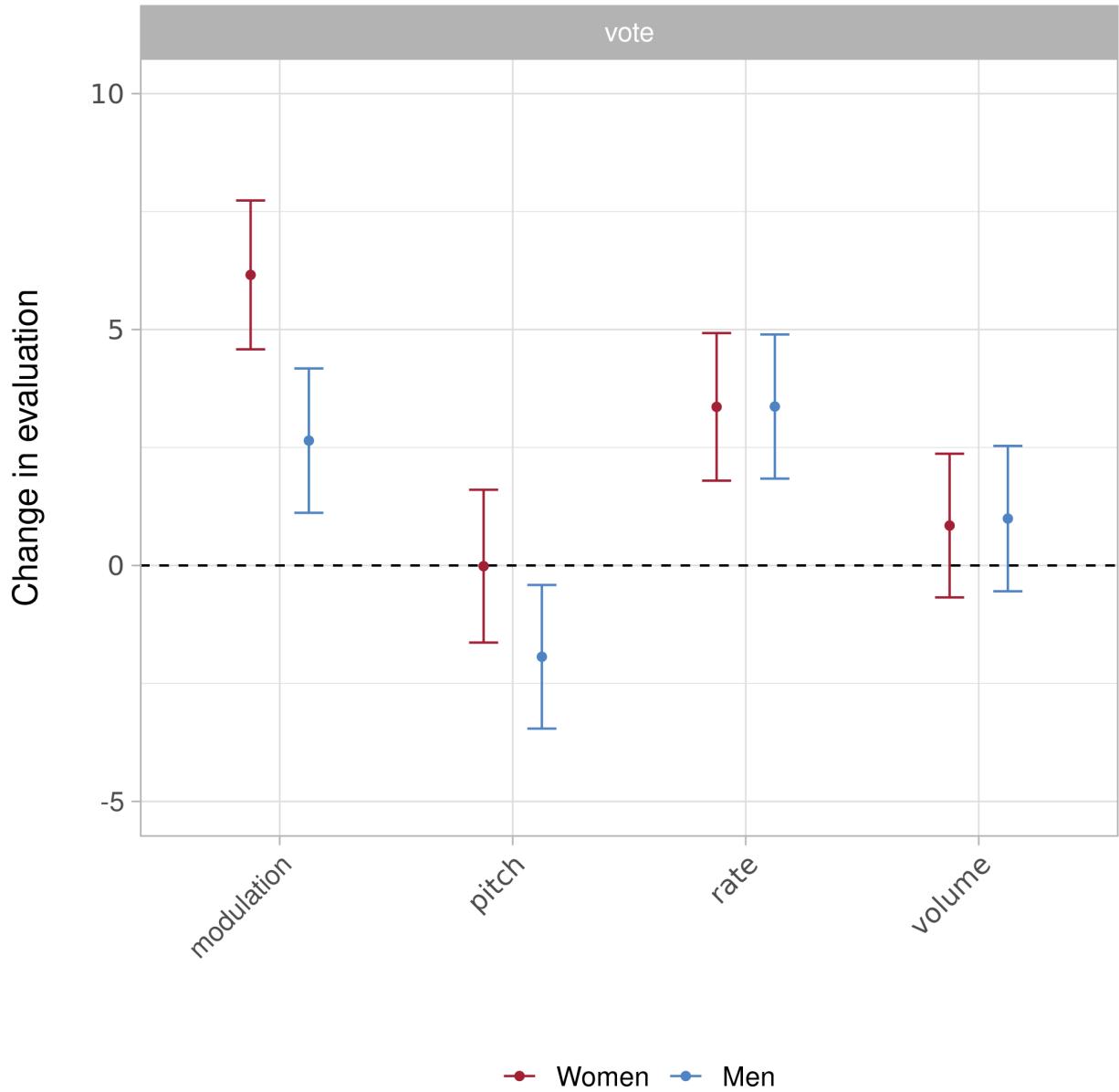
Figure 6: Effect of speech features on expressed willingness to vote for voice actor. Modulation and speech rate have relatively large effects. Appendix Section G.1 presents these results in tables.

7 Discussion and Conclusion

In this paper, we present the first corpus of audiovisual campaign recordings and present a descriptive analysis of how vocal style varies across candidates and topics. We develop a new, broadly applicable framework for drawing causal inferences about non-textual channels

of speech communication, which we used in two experiments to test the effect of non-textual communication on candidate assessment. We find strong evidence that vocal style shapes voter evaluations of candidate attributes and their willingness to vote for candidates.

As reviewed throughout this manuscript, prior work demonstrates that average vocal pitch influences voter perceptions. To our knowledge, ours is the first study to demonstrate that other features of non-textual communication—most strikingly, features related to oratory and rhetorical skill (e.g., speaking monotonously)—may have relatively larger effects on voter impressions. Moreover, we find evidence that the benefit of skillful communication is larger for women than for men, but that this relatively larger effect is due to a greater penalty imposed on women candidates at baseline. In other words, if women candidates do not communicate in a rhetorically skillful manner, they are punished more than their men counterparts. However, we note that the we are only able to draw limited inferences about these gendered effects. Specifically, our study relies on ten speakers. We hope these results lay the groundwork for future research that extends these tests to a larger number of unique speakers, to mitigate concerns that the differences we observe are due to idiosyncratic differences between the relatively small number of men and women speakers in our sample.

Our causal framework implies additional areas for future research. While the framework that we develop allows for learning dynamics that shape a voter's evaluation gradually across the course of an election, we assume for the sake of analytic tractability that this learning is negligible in the narrow timescale of our experiments. For the same reasons that causal inference in time series is complicated, incorporating these temporal dynamics requires careful thinking about the causal structure of opinion formation, particularly with respect to the potential for post-treatment bias. An important direction for future work is to extend approaches such as Blackwell and Glynn (2018) to the context studied in this article. Second, our focus is primarily on the importance of non-textual cues, and on the effects of specific audio features. Experimental designs utilizing similar visual manipulations are a promising avenue for future research. Finally, our observation that voters value different aspects of politicians' vocal styles depending on gender is exploratory in nature. Future research should focus on this relationship more explicitly.

All of these extensions suggest directions for building on our substantive results, which suggest that candidates vary how they communicate with voters and that this variation shapes perceptions of and support for the candidate—even holding fixed the actual policy content of speech. This result highlights the potential of new methods for analyzing speech audio, and also opens up a new area of study in communication.

Works Cited

Albaugh, Quinn, Julie Sevenans, Stuart Soroka, and Peter John Loewen. "The automated coding of policy agendas: A dictionary-based approach". *6th Annual Comparative Agendas Conference, Antwerp, Belgium*, 2013.

Amir, Amidhood, Yonatan Aumann, Gad M. Landau, Moshe Lewenstein, and Noa Lewenstein. "Pattern Matching with Swaps". *Proceedings of the 38th IEEE Symposium on Foundations of Computer Science*, 1997, pp. 144–153.

Banse, Rainer and Klaus R. Scherer. "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, 1996, p. 614.

Bänziger, Tanja and Klaus R. Scherer. "The role of intonation in emotional expressions". *Speech communication*, vol. 46, no. 3, 2005, pp. 252–267.

Barari, Soubhik, Christopher Lucas, and Kevin Munger. "Political deepfakes are as credible as other fake media and (sometimes) real media". *OSF Preprints*, 2021.

Baum, Matthew A. "Going private: Public opinion, presidential rhetoric, and the domestic politics of audience costs in US foreign policy crises". *Journal of Conflict Resolution*, vol. 48, no. 5, 2004, pp. 603–631.

Benjamini, Yoav and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, 1995, pp. 289–300.

Benoit, William L. "The functional theory of political campaign communication". *The Oxford Handbook of Political Communication*, 2017, p. 195.

Blackwell, Matthew and Adam N. Glynn. "How to make causal inferences with time-series cross-sectional data under selection on observables". *American Political Science Review*, vol. 112, no. 4, 2018, pp. 1067–1082.

Bligh, Michelle, Jennifer Merolla, Jean Reith Schroedel, and Randall Gonzalez. "Finding her voice: Hillary Clinton's rhetoric in the 2008 presidential campaign". *Women's Studies*, vol. 39, no. 8, 2010, pp. 823–850.

Boussalis, Constantine, Travis G Coan, Mirya R. Holman, and Stefan Müller. "Gender, candidate emotional expression, and voter reactions during televised debates". *American Political Science Review*, vol. 115, no. 4, 2021, pp. 1242–1257.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. "Peer effects in networks: A survey". *Annual Review of Economics*, vol. 12, 2020, pp. 603–629.

Canes-Wrone, Brandice. "The president's legislative influence from public appeals". *American Journal of Political Science*, 2001, pp. 313–329.

Carlson, David and Jacob M. Montgomery. "A pairwise comparison framework for fast, flexible, and reliable human coding of political texts". *American Political Science Review*, vol. 111, no. 4, 2017, pp. 835–843.

Carney, Dana R., Judith A. Hall, and Lavonia Smith LeBeau. "Beliefs about the nonverbal expression of social power". *Journal of Nonverbal Behavior*, vol. 29, no. 2, 2005, pp. 105–123.

Cohen, Jeffrey E. "Presidential rhetoric and the public agenda". *American journal of political science*, 1995, pp. 87–107.

Conway III, Lucian Gideon, Laura Janelle Gornick, Chelsea Burfeind, Paul Mandella, Andrea Kuenzli, Shannon C. Houck, and Deven Theresa Fullerton. "Does complex or simple rhetoric win elections? An integrative complexity analysis of US presidential campaigns". *Political Psychology*, vol. 33, no. 5, 2012, pp. 599–618.

De la Cuesta, Brandon, Naoki Egami, and Kosuke Imai. "Improving the external validity of conjoint analysis: the essential role of profile distribution". *Political Analysis*, vol. 30, no. 1, 2022, pp. 19–45.

Degani, Marta. *Framing the rhetoric of a leader: an analysis of Obama's election campaign speeches*. Springer, 2015.

Dietrich, Bryce J, Matthew Hayes, and Diana Z O'brien. "Pitch perfect: Vocal pitch and the emotional intensity of congressional speech". *American Political Science Review*, vol. 113, no. 4, 2019, pp. 941–962.

Dietrich, Bryce J., Ryan D. Enos, and Maya Sen. "Emotional arousal predicts voting on the US supreme court". *Political Analysis*, vol. 27, no. 2, 2019, pp. 237–243.

Dolson, Mark. "The phase vocoder: A tutorial". *Computer Music Journal*, vol. 10, no. 4, 1986, pp. 14–27.

Eckles, Dean, René F. Kizilcec, and Eytan Bakshy. "Estimating peer effects in networks with peer encouragement designs". *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, 2016, pp. 7316–7322.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. "How to make causal inferences using texts". *arXiv preprint arXiv:1802.02163*, 2018.

Fleishman, Jeffrey. "Eloquence and literary power make President Obama one of the nation's great orators". *Los Angeles Times. https://www. latimes. com/entertainment/movies/la-ca-obama-eloquent-speeches-20170111-story. html [Accessed 28 July 2020]*, 2017.

Fong, Christian and Justin Grimmer. "Discovery of treatments from text corpora". *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1600–1609.

Franz, Michael M., Erika Franklin Fowler, and Travis N. Ridout. "Loose cannons or loyal foot soldiers? Toward a more complex theory of interest group advertising strategies". *American Journal of Political Science*, vol. 60, no. 3, 2016, pp. 738–751.

Fridkin, Kim L. and Patrick Kenney. "Variability in citizens' reactions to different types of negative campaigns". *American Journal of Political Science*, vol. 55, no. 2, 2011, pp. 307–325.

Fridkin, Kim L. and Patrick J. Kenney. "The role of candidate traits in campaigns". *The Journal of Politics*, vol. 73, no. 1, 2011, pp. 61–73.

Fridkin, Kim L., Patrick J. Kenney, Sarah Allen Gershon, Karen Shafer, and Gina Serignese Woodall. "Capturing the power of a campaign event: The 2004 presidential debate in Tempe". *The Journal of Politics*, vol. 69, no. 3, 2007, pp. 770–785.

Gobl, Christer and Ailbhe Nı Chasaide. "The role of voice quality in communicating emotion, mood and attitude". *Speech communication*, vol. 40, no. 1, 2003, pp. 189–212.

Gregory, Stanford W. and Timothy J. Gallagher. "Spectral analysis of candidates nonverbal communication predicts national debate outcomes". *American Sociological Association meetings, Chicago, IL*, 1999.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments". *Political analysis*, vol. 22, no. 1, 2014, pp. 1–30.

Hamming, Richard. W. "Error detecting and error correcting codes". *Bell System Technical Journal*, vol. 29, no. 2, 1950, pp. 147–160.

Hassanieh, Haitham, Piotr Indyk, Dina Katabi, and Eric Price. "Simple and Practical Algorithm for Sparse Fourier Transform". *ACM-SIAM Symposium on Discrete Algorithms*, 2012.

Johnstone, Tom and Klaus R. Scherer. "Vocal communication of emotion". *Handbook of emotions*, vol. 2, 2000, pp. 220–235.

Kalkhoff, Will, Shane R. Thye, and Stanford W. Gregory Jr. "Nonverbal Vocal Adaptation and Audience Perceptions of Dominance and Prestige". *Social Psychology Quarterly*, vol. 80, no. 4, 2017, pp. 342–354.

Kececioglu, John and David Sankoff. "Exact and approximation algorithms for the inversion distance between two permutations". *Algorithmica*, vol. 13, 1995, pp. 180–210.

Khazan, Olga. "Would you really like Hillary more if she sounded different". *The Atlantic*, vol. 1, 2016.

Klofstad, Casey A. "Candidate voice pitch influences election outcomes". *Political psychology*, vol. 37, no. 5, 2016, pp. 725–738.

Klofstad, Casey A. "Looks and sounds like a winner: Perceptions of competence in candidates' faces and voices influences vote choice". *Journal of experimental political science*, vol. 4, no. 3, 2017, pp. 229–240.

Klofstad, Casey A., Rindy C. Anderson, and Susan Peters. "Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women". *Proceedings of the Royal Society B: Biological Sciences*, vol. 279, no. 1738, 2012, pp. 2698–2704.

Knox, Dean and Christopher Lucas. "A dynamic model of speech for the social sciences". *American Political Science Review*, vol. 115, no. 2, 2021, pp. 649–666.

Moore, Charles. *Margaret Thatcher: From Grantham to the Falklands.* Vintage, 2013.

Navarro, Gonzalo. "A Guided Tour to Approximate String Matching". *ACM Computing Surveys*, vol. 33, no. 1, 2001, pp. 31–88.

Needleman, Saul B. and Christian D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequences of two proteins". *Journal of Molecular Biology*, vol. 44, 1970, pp. 444–453.

Neyman, Jerzy S. "On the application of probability theory to agricultural experiments. essay on principles. section 9.(translated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480)". *Annals of Agricultural Sciences*, vol. 10, 1923, pp. 1–51.

Niebuhr, Oliver, Alexander Brem, and Silke Tegtmeier. "Advancing research and practice in entrepreneurship through speech analysis–From descriptive rhetorical terms to phonetically informed acoustic charisma profiles". *Journal of Speech Sciences*, vol. 6, no. 1, 2017, pp. 3–26.

Novák-Tót, Eszter, Oliver Niebuhr, and Aoju Chen. "A gender bias in the acoustic-melodic features of charismatic speech?" *INTERSPEECH*, 2017, pp. 2248–2252.

Peterson, Christine B., Marina Bogomolov, Yoav Benjamini, and Chiara Sabatti. "Many phenotypes without many false discoveries: error controlling strategies for multitrait association studies". *Genetic epidemiology*, vol. 40, no. 1, 2016, pp. 45–56.

Reece, Andrew, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. "Advancing an Interdisciplinary Science of Conversation: Insights from a Large Multimodal Corpus of Human Speech". *arXiv preprint arXiv:2203.00674*, 2022.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. "A model of text for experimentation in the social sciences". *Journal of the American Statistical Association*, vol. 111, no. 515, 2016, pp. 988–1003.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. "Structural topic models for open-ended survey responses". *American journal of political science*, vol. 58, no. 4, 2014, pp. 1064–1082.

Rodriguez, Pedro L. and Arthur Spirling. "Word embeddings: What works, what doesn't, and how to tell the difference for applied research". *The Journal of Politics*, vol. 84, no. 1, 2022, pp. 000–000.

Rosenbaum, Martin. *From soapbox to soundbite: Party political campaigning in Britain since 1945*. Springer, 2016.

Rubin, Donald B. "Estimating causal effects of treatments in randomized and non-randomized studies". *Journal of Educational Psychology*, vol. 66, no. 5, 1974, pp. 688–701.

Rubin, Donald B. "Randomization analysis of experimental data: The Fisher randomization test comment". *Journal of the American statistical association*, vol. 75, no. 371, 1980, pp. 591–593.

Rule, Alix, Jean-Philippe Cointet, and Peter S Bearman. "Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014". *Proceedings of the National Academy of Sciences*, vol. 112, no. 35, 2015, pp. 10837–10844.

Run for Something Website. Run for Something Strategic Plan. 2021, Accessed: 2020-03-07.

Scherer, Klaus R. "Vocal communication of emotion: A review of research paradigms". *Speech communication*, vol. 40, no. 1, 2003, pp. 227–256.

Schroedel, Jean, Michelle Bligh, Jennifer Merolla, and Randall Gonzalez. "Charismatic rhetoric in the 2008 presidential campaign: Commonalities and differences". *Presidential Studies Quarterly*, vol. 43, no. 1, 2013, pp. 101–128.

Sides, John and Andrew Karch. "Messages that mobilize? Issue publics and the content of campaign advertising". *The Journal of Politics*, vol. 70, no. 2, 2008, pp. 466–476.

Simes, R. John. "An improved Bonferroni procedure for multiple tests of significance". *Biometrika*, vol. 73, no. 3, 1986, pp. 751–754.

Spiliotes, Constantine J. and Lynn Vavreck. "Campaign advertising: Partisan convergence or divergence?" *The Journal of Politics*, vol. 64, no. 1, 2002, pp. 249–261.

Surawski, Melissa K. and Elizabeth P. Ossoff. "The effects of physical and vocal attractiveness on impression formation of politicians". *Current Psychology*, vol. 25, no. 1, 2006, pp. 15–27.

Tichy, Walter F. "The string-to-string correction problem with block moves". *ACM Transactions on Computer Systems*, vol. 2, 4 1984, pp. 309–321.

Tigue, Cara C, Diana J Borak, Jillian JM O'Connor, Charles Schandl, and David R Feinberg. "Voice pitch influences voting behavior". *Evolution and Human Behavior*, vol. 33, no. 3, 2012, pp. 210–216.

Torres, Michelle. "Give me the full picture: Using computer vision to understand visual frames and political communication". *URL: http://qssi. psu. edu/new-faces-papers-2018/torres-computer-vision-and-politicalcommunication*, 2018.

Ukkonen, Esko. "Approximate string matching with q-grams and maximal matches". *Theoretical Computer Science*, vol. 1, 1992, pp. 191–211.

Woolley, John T. and Gerhard Peters. "The American presidency project". *Santa Barbara, CA. Available from World Wide Web: http://www. presidency. ucsb. edu/ws*, 2008.

Appendix (Online Publication Only)

Table of Contents

A Finding Approximate String Matches

In this section, we briefly describe how we defined and efficiently discovered approximate string matches for the naturalistic treatments used in Experiment 1 (see Appendix Section B for the complete text of the matches that we selected for use in the experiment).

A.1 A Computationally Amenable Metric For String Similarity

A wide range of string distances have been proposed for quantifying general and domain-specific similarity, including the classical Levenshtein edit distance, simplified variants (Hamming, 1950; Needleman and Wunsch, 1970), and numerous modifications, generalizations, and alternative approaches (Amir et al., 1997; Kececioglu and Sankoff, 1995; Tichy, 1984; Ukkonen, 1992). For a review of this extensive literature, we refer the reader to Navarro, 2001.

    We encode each string as a word-letter matrix in which the k-th row contains frequencies for each of the L letters—e.g., $L = 4$ in genomics, $L = 26$ in English. The result is a lossy representation of the original string that discards information about letter ordering within

words. This representation of the pattern is denoted $\boldsymbol{P}_{K \times L}$, and target i is $\boldsymbol{T}_{i}{}_{J_i \times L}$. An example is given in Table 2. It is worth noting that word-embedding matrices may be substituted for word-letter matrices with no further modification of the algorithm proposed below.

Table 2: **Word-letter matrix.** Excerpted words from a President Barack Obama's campaign speech during the 2012 presidential election are represented using their letter counts. Word-letter matrix representations are used for approximate string alignment in `ffgrep`.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
we've					2																	1	1			
doubled		1		2	1							1			1						1					
the					1			1												1						
amount	1												1	1	1					1	1					
of						1									1											
renewable	1	1			3							1		1				1					1			
energy					2		1							1				1							1	
that	1							1												2						
we					1																		1			
generate	1				3		1							1				1		1						

The similarity between two K-word sequences, $\boldsymbol{P}$ and $\boldsymbol{Q}$, is then operationalized as

$$\mathcal{S}(\boldsymbol{P}, \boldsymbol{Q}) = \frac{\sum_{k=1}^{K} \sum_{\ell=1}^{L} \tilde{p}_{k,\ell} \ \tilde{q}_{k,\ell}}{||\tilde{\boldsymbol{P}}||_F \ ||\tilde{\boldsymbol{Q}}||_F} \tag{1}$$

where $\tilde{\boldsymbol{A}} = [a_{k\ell} - \bar{a}_\ell]$ indicates the column-demeaned transformation of $\boldsymbol{A}$, $\tilde{a}_{k,\ell}$ is the (k, ℓ)-th element of $\tilde{\boldsymbol{A}}$, and $||\boldsymbol{A}||_F = \sqrt{\sum_k \sum_\ell a_{k,\ell}^2}$ is the Frobenius norm.

In intuitive terms, $||\tilde{\boldsymbol{P}}||_F^2$ is proportional to the pattern's total variance, or the sum of letter-specific variances, and the numerator is proportional to $\sum_{\ell=1}^{L} \text{Cov}(P_\ell, Q_\ell)$, where P_ℓ is the sequence of counts for letter ℓ. Thus, when $L = 1$, Equation 1 yields the correlation coefficient. For lack of imagination, we refer to $1 - \mathcal{S}(\boldsymbol{P}, \boldsymbol{Q})$ as the string correlation distance. $\mathcal{S}(\cdot, \cdot)$ is symmetric, bounded in $[-1, 1]$, and has the property $\mathcal{S}(\boldsymbol{P}, \boldsymbol{P}) = 1$.

A.2   The Algorithm

Approximate string search involves examining all target documents i and candidate offsets j within each document. Figure 7 illustrates how this sequence can be obtained by sweeping a pattern over a target document. At each position, the similarity measure is computed,

producing the alignment sequence $\left[\mathcal{S}(\boldsymbol{P}, \boldsymbol{T}_{i,1:K}), \ldots, \mathcal{S}(\boldsymbol{P}, \boldsymbol{T}_{i,(J_i - K+1):J_i}) \right]$. A "hit," or high-quality alignment, is a position in the target document that produces a spike in this similarity sequence. In this section, we show how this apparently intensive task can be reformulated using highly efficient rolling sums and Fourier transforms. We begin by examining the elements of Equation 1.
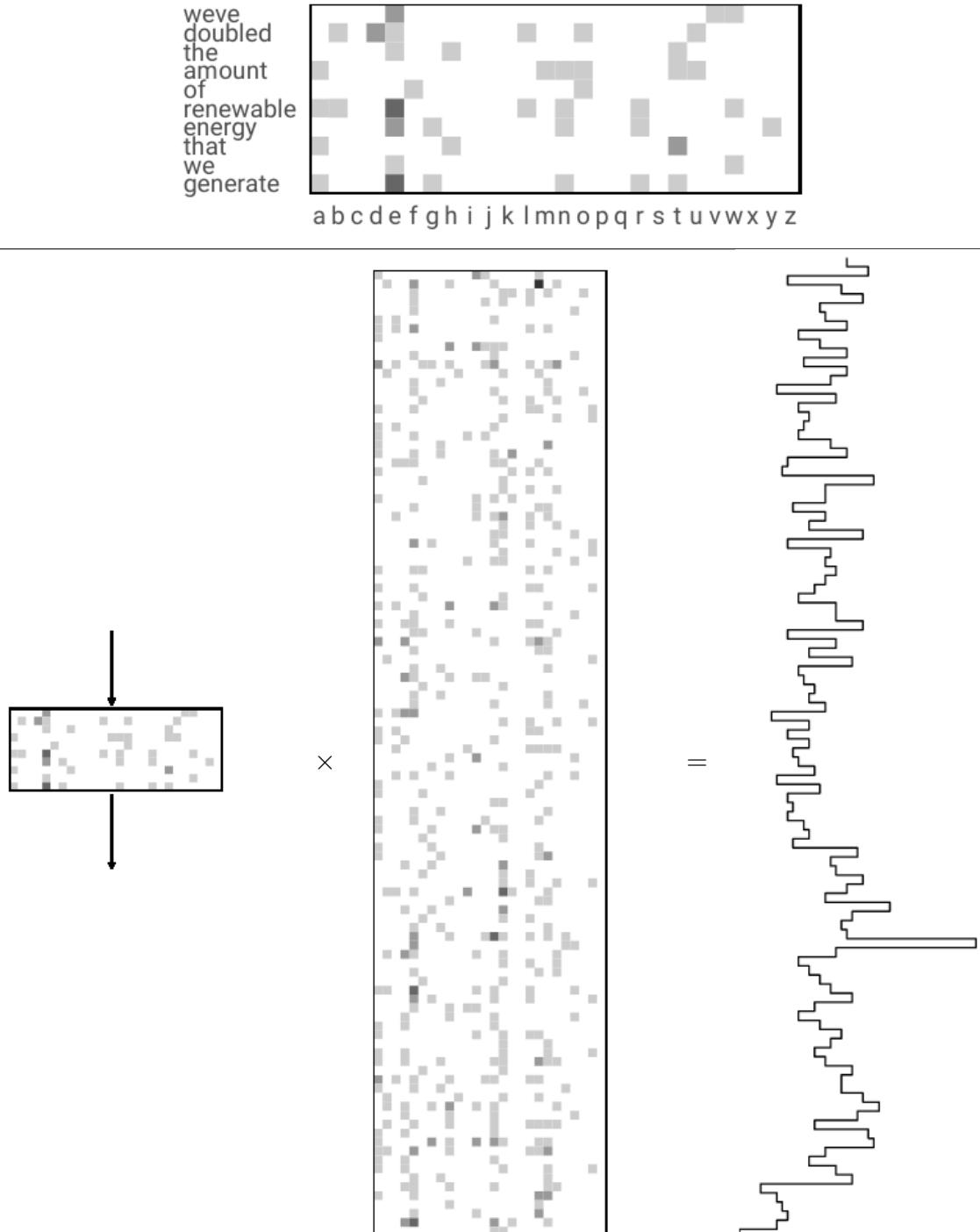
First, observe that $||\tilde{\boldsymbol{T}}_{i,1:K}||_F^2$ is the grand sum of a row subset of $[\tilde{t}_{i,j}^2]$. Corresponding values must be computed at every offset in document i to produce the sequence $\left[||\tilde{\boldsymbol{T}}_{i,1:K}||_F, \ldots, ||\tilde{\boldsymbol{T}}_{i,(J_i-K+1):J_i}||_F \right]$ which is simply a rolling windowed sum on $[\tilde{t}_{i,j}^2]\mathbf{1}$. Computation of $||\tilde{\boldsymbol{P}}||_F$ is even more straightforward.

Next, we observe that the numerator, $\sum_{k=1}^K \sum_{\ell=1}^L \tilde{p}_{k,\ell} \ \tilde{t}_{i,j+k-1,\ell}$, can be rewritten as $\sum_{k=1}^K \sum_{\ell=1}^L p_{k,\ell} \ t_{i,j+k-1,\ell} - \sum_{k=1}^K \sum_{\ell=1}^L \bar{p}_\ell \ \bar{t}_{i,j,\ell}$, where $\bar{p}_\ell$ is the mean of the pattern's ℓ-th column and $\bar{t}_{i,j,\ell}$ is the mean count of letter ℓ in the K words starting at offset j in target i. The latter term can be simultaneously evaluated for all offsets as follows: Compute the rolling column means of $\boldsymbol{T}_i$, forming $\bar{\boldsymbol{T}}_i \underset{J_i \times L}{=} [\bar{t}_{i,j,\ell}]$, then take its matrix product with the vector $[\bar{p}_\ell]$.

Finally, we are left with the term $\sum_{k=1}^K \sum_{\ell=1}^L p_{k,\ell} \ t_{i,j+k-1,\ell}$. Consider the contribution of a single letter, $x_{i,j,\ell} = \sum_{k=1}^K p_{k,\ell} \ t_{i,j+k-1,\ell}$. Evaluating this expression at every possible offset in the target, from $j = 1$ to J_i, is computationally demanding. However, the resulting vector, $[x_{i,1,\ell}, \ldots, x_{i,J_i,\ell}]$, is the convolution $P_\ell * T_{i,\ell}$. It is well-known that the Fourier convolution theorem offers a drastically more efficient approach for solving such problems. Briefly, the theorem states that $P_\ell * T_{i,\ell} = \mathcal{F}^{-1}\big(\mathcal{F}(P_\ell) \odot \mathcal{F}(T_{i,\ell})\big)$, where $\mathcal{F}$ is the Fourier transform, $\mathcal{F}^{-1}$ is the inverse transform, and $\odot$ denotes the elementwise product. Thus, $\sum_{\ell=1}^L \mathcal{F}^{-1}\big(\mathcal{F}(P_\ell) \odot \mathcal{F}(T_{i,\ell})\big)$ completes the rolling similarity score. By linearity of the Fourier transform, this can be rewritten $\mathcal{F}^{-1}\left(\sum_{\ell=1}^L \mathcal{F}(P_\ell) \odot \mathcal{F}(T_{i,\ell})\right)$, reducing complexity of the inverse step by an additional factor of L. Moreover, because the goal of approximate string matching is to identify sharp peaks in the similarity sequence, a sparse Fourier transform Hassanieh et al., 2012 in the inverse step has the potential to reduce computation time further. We do not explore sparsity-based optimizations here.

To identify approximate alignments, the resulting similarity sequence is thresholded. Among other steps, we zero-pad the pattern to a convenient length, then use the overlap-save method to cut targets into smaller batches of the same length. Target batches are also zero-padded to avoid circular convolution. After computing the Fourier transforms of the pattern and each batch, the target batch spectra are cached to accelerate subsequent searches against the same targets.

Figure 7: **Convolution of text sequences.** The top panel depicts a word-letter matrix, $\boldsymbol{P}$, for a single pattern: "we've doubled the amount of renewable energy that we generate," a quote from an Obama rally in Madison, WI. The bottom-left panel illustrates how this pattern is swept over a target document, $\boldsymbol{T}_i$, an earlier speech in West Palm Beach, FL (bottom middle). At offset j, the elementwise product with $\boldsymbol{T}_{i,(j-K+1):j_i}$ is taken and summed. This is repeated from $j = 1$ to target length J_i, and the sequence of resulting sums—the convolution—is plotted on the bottom right. Appropriate scaling yields the desired sequence of correlation similarities. The peak successfully identifies the previous usage of a similar phrase, "we've doubled our use of renewable energy like wind and..." from an earlier rally in West Palm Beach.

B   Text From Experiment 1

As described in Section 5, Experiment 1 relies on pairs of approximately matched text scripts. The table below displays the text of these scripts.

Topic	Variant A	Variant B
Tax Cuts	They want to spend 5 trillion dollars on new tax cuts, including a 25% tax cut for every millionare in the country.	Then they want to add another 5 trillion dollars in tax cuts on top of that, including a 25% tax cut for every millionaire in the country.
Fair Shot	We do believe in a country where hard work pays off, where responsibility is rewarded, where everyone gets a fair shot, and everybody is doing their fair share, and everybody plays by the same rules.	The promise that if you work hard, it will pay off. The promise that if you act responsibly, you will be rewarded. That everybody in this country gets a fair shot, and everybody gets a fair share, and everybody plays by the same rules.
Medicare	Now I've already strengthened medicare. We've already added years to the life of medicare by getting rid of taxpayer subsidies to insurance companies that weren't making people any healthier and in fact were making things more expensive for everybody.	I have strengthened medicare. We've added years to the life of medicare. We did it by getting rid of taxpayer subsidies to insurance companies that weren't making people healthier.
Energy	We can help big factories and small businesses double their exports and create a million new manufacturing jobs over the next four years. You can make that happen. I want to control more of our own energy. You know after 30 years of inaction, we raised fuel standards so after the middle of the next decade your cars and trucks will be going twice as far on a gallon of gas.	We can create a million new manufacturing jobs in the next four years, you can make that happen. Second part of our plan, let's control our own energy. You know, after 30 years of inaction, we raised fuel standards so that by the middle of the next decade your cars and trucks will go twice as far on the same gallon of gas.
Offshore	No company should have to look for workers in China because they couldn't find any with the right skills here in the United States.	No company should have to look for a worker someplace else because they couldn't find the right skills for workers here in the United States.

Table 3: Phrase versions A and B for each script.

Topic	Variant A	Variant B
Bailout	And after all we've been through, does anybody really think that somehow rolling back regulations on Wall Street that we put in place to make sure we don't have another taxpayer funded bailout, that somehow that's going to be good for the small businesswoman?	I don't think rolling back regulations on Wall Street so that we don't have another taxpayer funded bailout is a smart idea.
Terror	No act of terror will go unpunished, it will not dim the light of the values that we proudly present to the rest of the world. No act of violence shakes the resolve of the United States of America.	No act of terror will dim the light of the values that we proudly shine on the rest of the world, and no act of violence will shake the resolve of the United States of America.
Military	There are still threats in the world, and we've got to remain vigilant. That's why we have to be relentless in pursuing those who attacked us this week. That's also why so long as I'm still commander in chief, we will sustain the strongest military the world has ever known.	We still face threats in this world, and we've got to remain vigilant. But that's why we will be relentless in our pursuit of those who attacked us yesterday. But that's also why, so long as I'm commander in chief, we will sustain the strongest military the world has ever known.
College	And right now as I said because of the actions we already took, millions of young people are paying less for college because we finally took on that system that was wasting taxpayer dollars, gave it directly to students.	And we've already been working on this so millions of students are right now paying less for college because we took on a system that was wasting billions of dollars in taxpayer money to banks and lenders, we said, let's give it directly to students.
Change	From the day we began this campaign we've always said that real change takes time. It takes more than one year or one term or even one president. It takes more than one party. It certainly can't happen if you're willing to write off half the nation before you even take office.	And from the day we began this campaign, we've always said that change takes more than one term or even one president. And it certainly takes more than one party. It can't happen if you write off half the nation before you even take office.
Plurality	In 2008, 47% of the country didn't vote for me. But on the night of the election I said to those Americans, I may not have won your vote, but I hear your voices, I need your help, I'll be your president too.	In 2008, 47% of the country didn't vote for me. But on the night of the election I said to all those Americans, I may not have won your vote, but I hear your voices, I need your help, and I will be your president.

Table 4: Phrase versions A and B for each script.

Topic	Variant A	Variant B
Opportunity	We grow our economy not from the top down, but from the middle out. We don't believe that anybody's entitled to success in this country, but we do believe in something called opportunity.	Our economy does not grow from the top down, it grows from the middle out. That's how it grows. We don't believe that anybody's entitled to success in this country but we do believe in opportunity.
Students	We finally took on a system that was wasting billions of dollars on banks and lenders. We said, let's cut out the middle man, and let's give the money directly to students.	We took a system that was wasting tens of billions of dollars on banks and lenders. We said, let's cut out the middle man, give the money directly to the students.
Can't Afford	We can't afford to go down that road again. We can't afford another round of budget busting tax cuts for the wealthy. We can't afford to gut our investments in education or clean energy or research and technology. We can't afford to roll back regulations on Wall Street.	We can't afford to go down that road again. We can't afford another round of budget busting tax cuts for the wealthy. We can't afford to gut our investments in education or clean energy or research or technology. We can't afford to roll back regulations on Wall Street.
Top-Down	I have seen too much pain, seen too much struggle to let this country get hit with another round of top-down economics. One of the main reasons we had this crisis was because big banks on Wall Street were allowed to make big bets with other people's money.	I have seen too much pain and too much struggle to let this country go with another round of top-down economics. One of the main reasons we had this crisis was because we had big banks on Wall Street making bets with other people's money.
Deficit	But look, we've gotta do something about it. So what I've said - look - I've already worked with Republicans and Democrats to cut a trillion dollars in spending. I'm ready to do more.	Yes, we're gonna need to cut our deficit by 4 trillion dollars over the next 10 years. And I've already worked with Republicans and Democrats to cut a trillion dollars in spending. I'm ready to do more.
Economy	Unemployment is falling, manufacturing is coming back, our assembly lines are humming again. We've got a long way to go, but Florida we've come too far to turn back now.	Unemployment has fallen to its lowest levels since I took office. Home values and home sales are rising. Our assembly lines are humming again. We've got a long we to go Iowa but we've come too far to turn back now.
Math	And it turns out, his math and their math was just as bad back then as it is now.	Turns out, their math was just as bad back then as it is today.

Table 5: Phrase versions A and B for each script.

Topic	Variant A	Variant B
Renewables	Today, there are thousands of workers building long-lasting batteries, solar technology, and wind turbines, all across the country. Jobs that weren't there four years ago.	Today, there are thousands of workers building long-lasting batteries, and wind turbines, and solar panels, all across the country. Jobs that weren't there four years ago.
Work	Let's put Americans back to work doing the work that needs to be done.	Let's put Americans back to work doing the work that needs to be done.
Wealthy	I intend to do more. And I'll work with both parties to streamline agencies and get rid of programs that don't work. But if we're serious about the deficit, we've also go to ask the wealthiest Americans to go back to the tax rate they paid when Bill Clinton was in office.	I intend to do more. We can streamline agencies, we can get rid of programs that aren't working. But if we're serious about the deficit, we also have to ask the wealthiest Americans to go back to the tax rates they paid when Bill Clinton was in office.
Apologize	We'll stop the days of apologizing for success at home, and never again will we apologize for America abroad.	I will not apologize for success here, and I will never apologize for America abroad.
Rights	That document, the Declaration of Independence, said that we were endowed by our creator with our rights. Not the state, not the king, but our creator. And among them are life, liberty, and the pursuit of happiness.	The founders of this nation, when they said we had our rights, they did not say they came from the king or the government, they said they came from god. And among them were life and liberty and the pursuit of happiness.
Hymn	I love that stanza in own of our national hymns, America the Beautiful. 'Oh beautiful, for heroes proved, in liberating strife, who more than self their country loved, and mercy more than life.'	I love those words in one of our national hymns. 'Oh beautiful, for heroes proved, in liberating strife, who more than self their country loved, and mercy more than life.'
Better Days	My conviction that betters days are ahead is not based on promises and rhetoric, but on solid plans and proven results, and an unshakebale faith in the American spirit.	My conviction that better days are ahead is not based on promises and hollow rhetoric, but on solid plans and proven results, and an unshakeable faith in the American people and the American spirit.
Same Course	The same course we have been on will not lead to a better destination. The same path means 20 trillion in debt, it means crippling unemployment continuing. It means stagnant take-home pay and depressed home values, and a devastated military.	The same course we've been on will not lead to a better destination, Mr. President. The same path means 20 trillion dollars in debt, it means crippling unemployment, stagnant take-home pay, depressed home values, and a devastated military.

Table 6: Phrase versions A and B for each script.

Topic	Variant A	Variant B
Divide	He has not met on the economy, or on the budget, or on jobs, with either the Republican leader of the House or the Senate since July. Instead of bridging the divide, he's made it wider.	He has not met on the economy, or on the budget, or on jobs, with either the Republican leader of the House or the Senate since July. So instead of bridging the divide, he's made it wider.
Promised	He promised that he would propose a plan to save Social Security and Medicare from insolvency. He didn't. Rather he raided 716 billion dollars from medicare to pay for his vaunted Obamacare.	He promised that he'd propose a plan to save Social Security and Medicare from insolvency. And rather he raided 716 billion dollars from medicare for his vaunted Obamacare plan.
Both Sides	I'll meet with them regularly. I'll endeavor to find those good men and women on both sides of the aisle, who care more about the country than about politics.	I'm going to meet regularly with their leaders. I'll endeavor to find those good men and women on both sides of the aisle, who care more about the country than about politics.

Table 7: Phrase versions A and B for each script.

C Display of Experiment 1

In this section, we provide screenshots of the survey pages presenting the text, audio, and video conditions, respectively.

Figure 8: Display of the text condition in Experiment 1.

Statement A	**Statement B**
▶ ● ———— 0:00 / 0:14 🔊 ——●	▶ ● ———— 0:00 / 0:14 🔊 ——●

Which statement makes you feel more angry?

○ Statement A

○ Statement B

Which statement makes you feel more afraid?

○ Statement A

○ Statement B

Figure 9: Display of the audio condition in Experiment 1.

Statement A	**Statement B**

Which statement makes you feel more angry?

○ Statement A

○ Statement B

Figure 10: Display of the video condition in Experiment 1.

D Text From Experiment 2

We hired 10 professional voice actors to perform 6 scripts in 4 different manners (see Table 2 in text). The table below displays the text of each script.

Topic	Text	Source
Budget	"Yes, we're gonna need to cut our deficit by 4 trillion dollars over the next 10 years. And I've already worked with Republicans and Democrats to cut a trillion dollars in spending. I'm ready to do more."	(Text from experiment 1)
Climate	"No nation, however large or small, wealthy or poor, can escape the impact of climate change. The security and stability of each nation and all peoples – our prosperity, our health, our safety – are in jeopardy. And the time we have to reverse this tide is running out."	(Text from former President Obama's 2009 address to the U.N. on climate change)
Education	"Charter schools are here to stay. We're now seeing the first generation of charter students raising children of their own. They know the difference educational choice made in their lives, and now as parents they want the same options for their children."	(Text from Betsy DeVos's 2017 speech to the National Charter Schools Conference)
Military	"My fellow Americans, a short time ago, I ordered the United States Armed Forces to launch precision strikes on targets associated with the chemical weapons capabilities of Syrian dictator Bashar al-Assad. A combined operation with the armed forces of France and the United Kingdom is now underway. We thank them both."	(Text from April 13, 2018 form President Trump address on airstrikes in Syria)
Nationalism	"No act of terror will dim the light of the values that we proudly shine on the rest of the world, and no act of violence will shake the resolve of the United States of America."	(Text from experiment 1)
Social Policy	"I am also proud to be the first president to include in my budget a plan for nationwide paid family leave — so that every new parent has the chance to bond with their newborn child."	(Text from 2019 State of the Union)

Table 8: Voice actors read four versions of each of these scripts.

E   How Actor-Controlled Manipulations Affect Volume and Pitch

As discussed, experiment 2 contains four experimental manipulations: volume, rate, pitch, and modulation. To implement these manipulations, 10 actors recorded 4 versions of 6 scripts. These four versions were readings of each script but: [1] spoken slowly (low rate) and in a monotonous voice (low modulation), [2] spoken slowly (low rate) and in a modulated voice (high modulation), [3] spoken quickly (high rate) and in a monotonous voice (low modulation), [2] spoken quickly (high rate) and in a modulated voice (high modulation). For these manipulations, we relied on actors because computational manipulations of rate of speech and modulation do not sound naturalistic. In total, this resulted in 240 recordings (10 actors * 6 scripts * 4 versions).

Using these actor-controlled recordings, we further computationally manipulated the volume and pitch of each each recording, resulting in 960 recordings in total (240 * high/low volume * high/low pitch). In contrast with rate and modulation, volume and pitch are better manipulated through computational interventions, for the following reasons. Volume is trivially easy to adjust digitally, whereas increasing spoken volume into a microphone can sound unnatural (shouting or whispering, on either end of the continuum). Pitch is similar. It is difficult for an actor to increase the overall pitch of speech in a constant way, but it's trivially easy to increase a segment of speech by several semitones.

Figure 11 plots the difference between the high and low versions of each of these four manipulations: the two actor-controlled manipulations (rate and modulation) and the two researcher-controlled manipulations (volume and pitch). For each of these manipulations, we plot the difference between the high and low versions in each of 5 summary features: average loudness, loudness variance, average pitch, pitch variance, and rate of speech.

First, note the two researcher-controlled manipulations, pitch and volume. Predictably, each only affect features related to the manipulation. For example, the difference on rate of speech between the high and low versions of these recordings is precisely zero. The reason for this is straightforward: the high and low versions are equivalent except with the pitch and volume raised/lowered. Similarly, computationally manipulating the volume obviously changed the volume, but had no effect on pitch, whereas computationally manipulating the pitch shifted only the pitch but not the volume.

Next, note the actor-controlled manipulations from which these recordings are constructed (rate and modulation). Predictably, when human actors speak faster/slower or in a modulated/monotous voice, they naturally vary pitch and volume. This is a feature of our design: *by relying on actors to construct these manipulations, we capture realistic variation in speech that cannot be convincingly manipulated computationally.*

Importantly, it is not possible to conduct these manipulations in any other way. For example, it is not possible to increase the modulation in speech without also shifting the average pitch. When a speaker modulates, they rarely drop their voice to very low pitches, but rather raise pitch to emphasize certain points and phrases. Doing so results in an overall upward shift in the mean, but it also highlights why a computational manipulation is not possible: modulated speech uses pitch and loudness to emphasize certain words in a phrase in order to heighten semantic meaning. Simply increasing the overall variance

of pitch would not appropriately pair the pitch increases to the terms that substantively ought to be emphasized in the relevant piece of text. For example, a candidate reading our "Nationalism" script (Table 8), which begins "No act of terror will dim the light of the values we proudly shine on the rest of the world..." A naturally modulated reading would likely increase pitch and loudness when reading the word "No", in order to emphasize the negation implied by the sentence. Trained actors, like those in our sample, can manipulate their speech in such ways with ease. A computational manipulation, however, would require tremendous sophistication in order to realistically approximate this, and there is ultimately no reason to do so when we can instead rely on professional voice actors.

However, as a result, estimates of the effect of speech modulation and speech rate should not be thought of as completely independent of speech features like loudness and pitch. Rather, they are complex manipulations, involving every component of spoken speech, from pitch contours to pronunciation. In contrast, the computational manipulations that we cross with these actor-controlled manipulations capture the effect of mean shifts on variables commonly used to summarize the sound of political speech. In sum, our results indicate that human evaluation of speech is considerably more complex than simply the mean shifts in easily measured features: our human-manipulated treatment conditions in general are considerably more effective than simply shifting the mean. This highlights the importance of subtler ways for summarizing speech, compared to simply summarizing it according to averages and variances, and potentially highlights the importance of using human coders rather than low dimensional summaries like the mean.
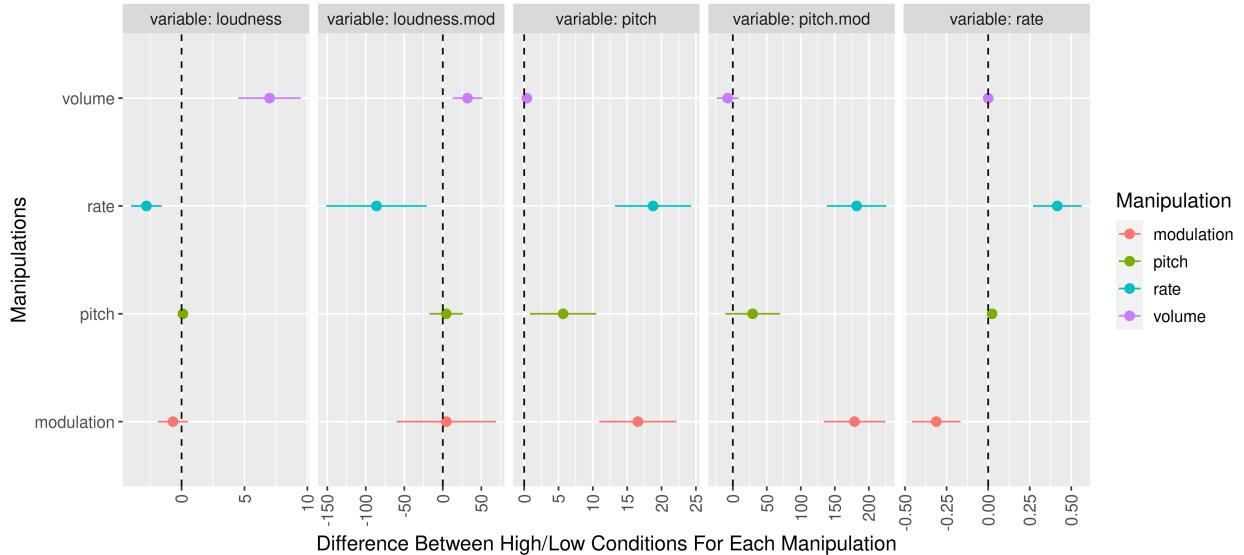
Figure 11: Comparison of manipulations used in Experiment 2 across five summary features. Of the four manipulations, two were controlled by actors recording different versions of each script (rate and modulation), while the other two were implemented by computationally manipulating all actor-produced recordings. Note that the computationally-implemented manipulations (volume and pitch) only affect features related to those manipulations (e.g., neither have any effect on the rate of speech, but the pitch manipulation affects pitch-related features and the loudness manipulation affects loudness-related features). In contrast, the actor-controlled manipulations affected other features. Section E discusses this in greater detail.

F   Supplementary Figures
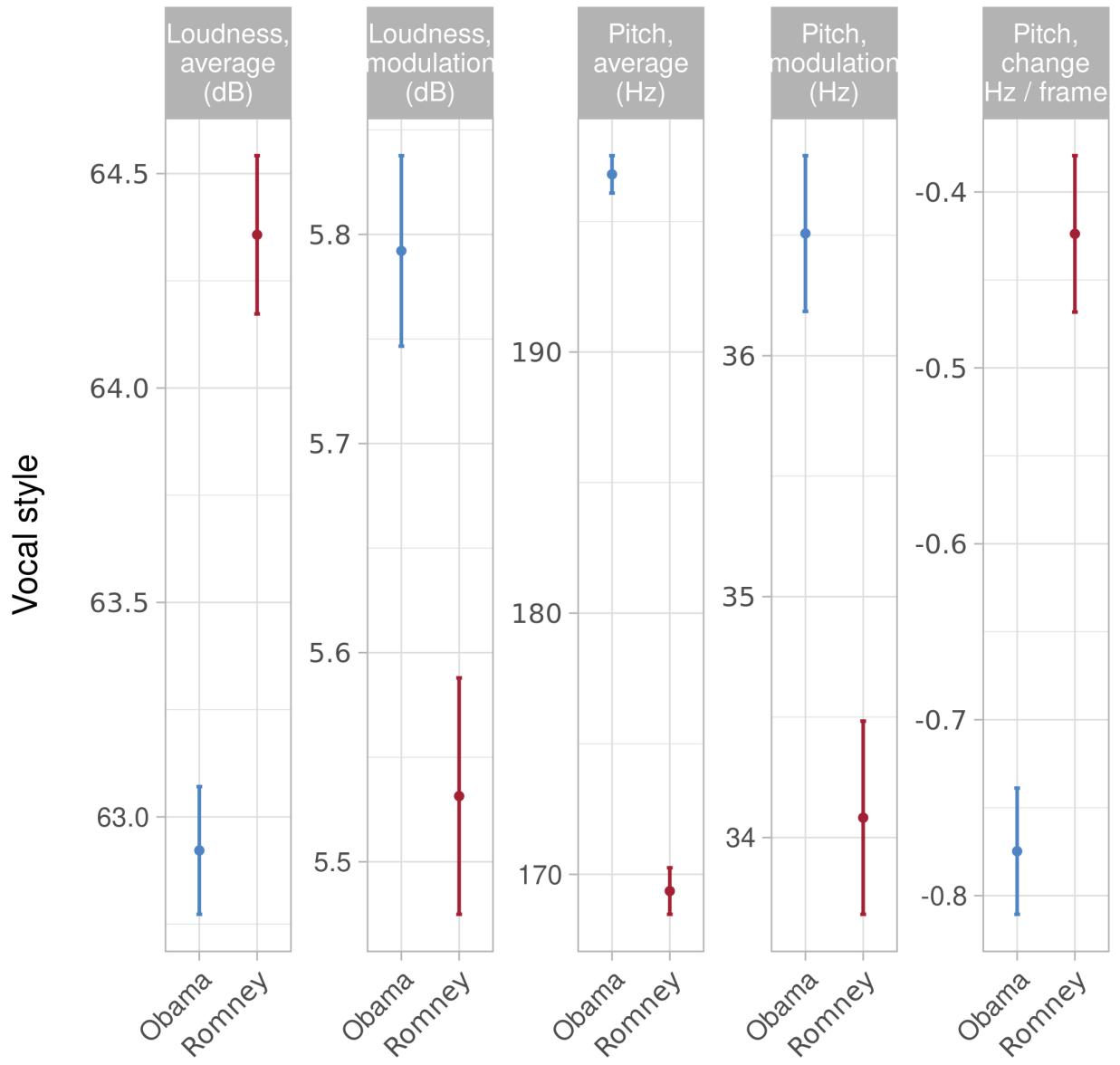
Figure 12: Comparison of campaign speech by Obama and Romney on common speech audio features. On average, Obama displays considerably more variation in loudness and pitch, consistent with popular accounts of Obama being a talented public speaker (Fleishman, 2017).
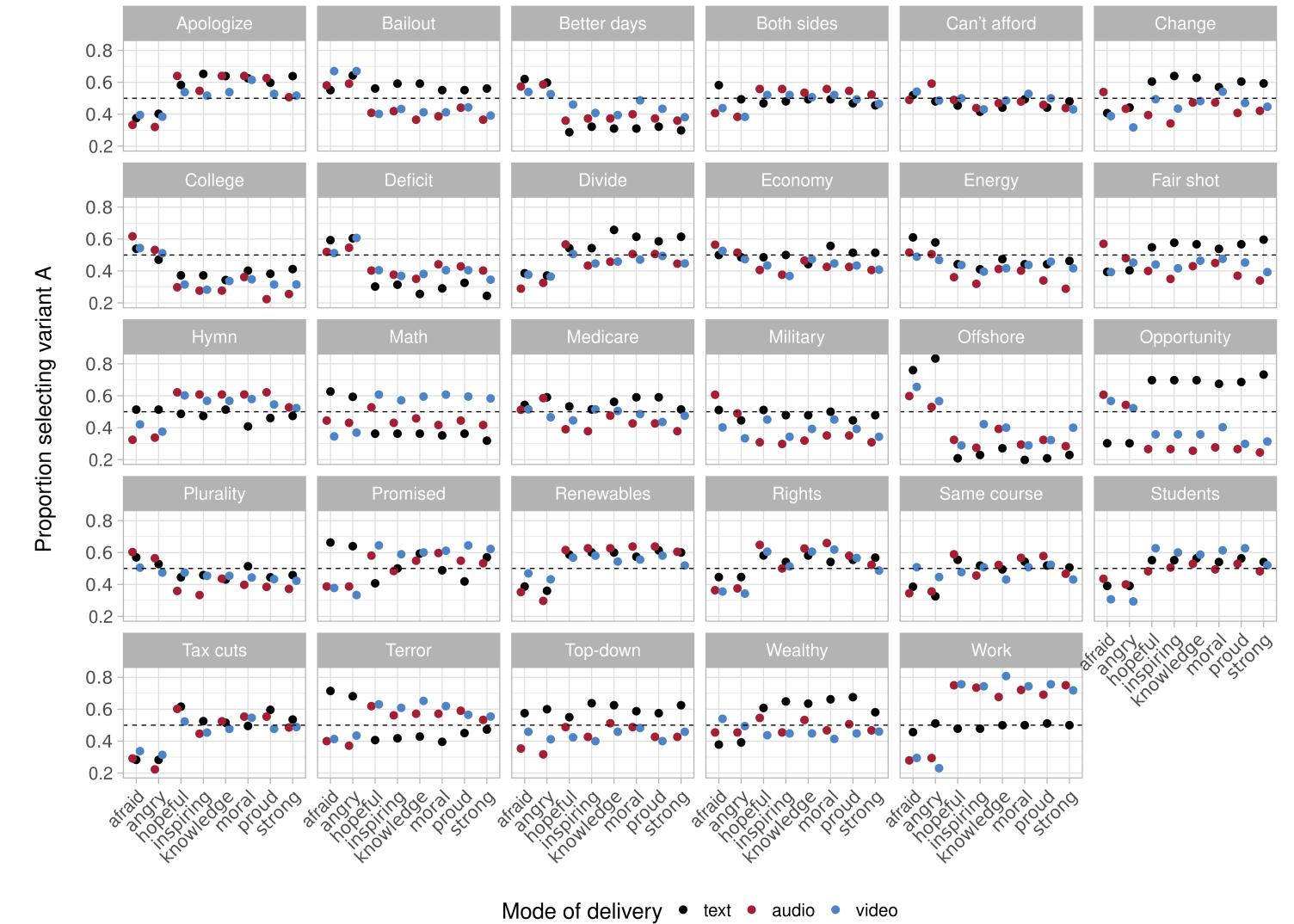
Figure 13: Each panel plots the proportion of subjects selecting variant A of a matched text pair over variant B. Within the pair, assignment of a variant to be A or B is arbitrary, so there are no directional expectations. Each panel in the plot shows the proportion of subjects selecting variant A over B for each eight characteristics, separately depending on whether the subject read, heard, or watched the paired variants. The panel labels denote manual labeling of the text topic of the pairs. The primary takeaway is that there is considerable variation as a result of speech mode, as the text of each variant is constant in the text, audio, and video comparisons.

G   Supplementary Tables

Variable	Loudness (avg)	Loudness (mod)	Pitch (avg)	Pitch (mod)	Pitch (change)
			Outcome		
Economy	-0.027 (0.02)	0.101 (0.02)*	0.167 (0.024)*	-0.012 (0.022)	0.118 (0.021)*
Civil rights	0.195 (0.087)*	0.104 (0.104)	0.092 (0.078)	0.228 (0.117)	-0.094 (0.056)
Healthcare	0.029 (0.035)	0.095 (0.039)*	0.081 (0.049)	0.053 (0.055)	-0.028 (0.046)
Labor	0.028 (0.05)	0.038 (0.054)	0.126 (0.05)*	-0.024 (0.066)	-0.022 (0.04)
Education	0.083 (0.029)*	0.142 (0.025)*	0.021 (0.043)	0.148 (0.057)*	-0.101 (0.038)*
Energy	-0.119 (0.057)*	-0.031 (0.062)	0.085 (0.063)	0.109 (0.062)	-0.09 (0.044)*
Transportation	0.05 (0.03)	0.089 (0.042)*	0.108 (0.035)*	-0.029 (0.041)	-0.009 (0.035)
Crime	-0.102 (0.044)*	0.073 (0.08)	0.068 (0.107)	-0.161 (0.106)	0.008 (0.067)
Social Welfare	-0.023 (0.054)	0.036 (0.065)	-0.034 (0.103)	0.091 (0.155)	0.021 (0.059)
Finance	-0.015 (0.048)	0.129 (0.051)*	0.133 (0.041)*	-0.043 (0.056)	0.096 (0.042)*
Defense	0.107 (0.052)*	-0.018 (0.044)	0.09 (0.069)	-0.023 (0.076)	0.067 (0.036)
Technology	-0.169 (0.056)*	-0.004 (0.075)	0.189 (0.178)	-0.114 (0.15)	-0.024 (0.085)
Environment	0.012 (0.087)	0.159 (0.103)	0.179 (0.083)*	0.086 (0.14)	0.113 (0.076)
Culture	0.059 (0.096)	0.143 (0.137)	0.214 (0.144)	0.234 (0.174)	0.028 (0.141)
Religion	0.318 (0.104)*	0.027 (0.116)	-0.749 (0.084)*	0.634 (0.11)*	-0.142 (0.061)*
Speech Fixed Effect	✓	✓	✓	✓	✓

Table 9:   Change in vocal style across different speech topics. This table presents the results displayed in Figure 2, but in tabular form, where each column is a separate regression on a different outcome variable, and the rows are the covariates.

	Outcome				
Variable	Loudness (avg)	Loudness (mod)	Pitch (avg)	Pitch (mod)	Pitch (change)
Economy	-0.052 (0.023)*	0.134 (0.027)*	0.216 (0.031)*	0.013 (0.047)	0.19 (0.029)*
Civil rights	-0.073 (0.034)*	-0.03 (0.051)	0.124 (0.076)	-0.141 (0.098)	0.093 (0.118)
Healthcare	-0.002 (0.028)	0.012 (0.031)	0.096 (0.042)*	-0.1 (0.056)	0.113 (0.05)*
Labor	-0.079 (0.026)*	-0.042 (0.036)	0.072 (0.049)	-0.003 (0.06)	0.151 (0.051)*
Education	-0.138 (0.034)*	-0.029 (0.046)	-0.152 (0.049)*	-0.21 (0.044)*	-0.035 (0.05)
Energy	0.038 (0.027)	0.051 (0.031)	0.245 (0.04)*	0.149 (0.049)*	0.148 (0.06)*
Transportation	-0.148 (0.048)*	-0.049 (0.047)	-0.15 (0.096)	-0.262 (0.075)*	0.051 (0.072)
Crime	-0.122 (0.071)	-0.018 (0.088)	-0.289 (0.097)*	-0.314 (0.087)*	-0.124 (0.08)
Social welfare	-0.115 (0.057)*	-0.013 (0.061)	0.234 (0.121)	-0.088 (0.112)	0.259 (0.099)*
Finance	0.095 (0.035)*	0.158 (0.039)*	0.189 (0.039)*	0.185 (0.062)*	-0.197 (0.077)*
Defense	-0.161 (0.052)*	-0.106 (0.042)*	-0.4 (0.097)*	-0.396 (0.079)*	0.066 (0.06)
Technology	-0.367 (0.053)*	-0.405 (0.084)*	-0.483 (0.06)*	-0.399 (0.099)*	0.016 (0.063)
Environment	0.008 (0.051)	0.038 (0.053)	0.174 (0.105)	-0.015 (0.097)	0.289 (0.087)*
Culture	-0.121 (0.085)	-0.118 (0.091)	-0.19 (0.119)	-0.18 (0.074)*	-0.057 (0.126)
Religion	0.068 (0.064)	0.155 (0.082)	-0.061 (0.08)	-0.049 (0.114)	0.004 (0.082)
Speech Fixed Effect	✓	✓	✓	✓	✓

Table 10: Change in vocal style across different speech topics. This table presents the results displayed in Figure 3, but in tabular form, where each column is a separate regression on a different outcome variable, and the rows are the covariates.

	Estimate	Std. Error	t value	Pr(>|t|)
Speaker A	29.9594	1.1959	25.05	0.0000
Speaker B	38.2496	1.1887	32.18	0.0000
Speaker C	38.5827	1.1950	32.29	0.0000
Speaker D	39.3700	1.1955	32.93	0.0000
Speaker E	40.6919	1.1942	34.07	0.0000
Speaker F	36.2316	1.1977	30.25	0.0000
Speaker G	37.9518	1.1984	31.67	0.0000
Speaker H	37.9805	1.2008	31.63	0.0000
Speaker I	41.9813	1.2012	34.95	0.0000
Speaker J	44.3087	1.1922	37.16	0.0000
Modulated Speech	4.4103	0.5503	8.01	0.0000
High Pitch	-0.9744	0.5503	-1.77	0.0766
High Rate	3.3685	0.5501	6.12	0.0000
High Volume	0.9289	0.5501	1.69	0.0913

Table 11: Also contains script fixed effects. The indicators for speaker are the source of Figure 4.

G.1 Tabular Representation of Figures 5 and 6

In this section, we present in tabular form estimates presented visually in plots 5 and 6. Each table reports estimates from a model regressing each outcome (competence, enthusiasm, etc) on the four treatment indicators (modulation, pitch, rate, and volume), with separate indicators for recordings by male and female actors (speakers).

Outcome: Competence

	Estimate	Std. Error	t value	Pr(>|t|)
Modulated Speech (Female)	4.7251	0.7194	6.57	0.0000
Modulated Speech (Male)	1.5912	0.7233	2.20	0.0278
High Pitch (Female)	-0.7118	0.7193	-0.99	0.3224
High Pitch (Male)	-2.8412	0.7234	-3.93	0.0001
Fast Rate (Female)	4.5602	0.7193	6.34	0.0000
Fast Rate (Male)	2.6631	0.7229	3.68	0.0002
High Volume (Female)	0.9794	0.7191	1.36	0.1732
High Volume (Male)	0.1970	0.7234	0.27	0.7854

Table 12: Also includes speaker and script fixed effects.

Outcome: Enthusiastic

	Estimate	Std. Error	t value	Pr(>|t|)
Modulated Speech (Female)	19.5617	0.7470	26.19	0.0000
Modulated Speech (Male)	14.0432	0.7511	18.70	0.0000
High Pitch (Female)	0.9728	0.7470	1.30	0.1928
High Pitch (Male)	-1.2214	0.7512	-1.63	0.1040
Fast Rate (Female)	6.6630	0.7470	8.92	0.0000
Fast Rate (Male)	7.8793	0.7507	10.50	0.0000
High Volume (Female)	2.1929	0.7468	2.94	0.0033
High Volume (Male)	2.4184	0.7512	3.22	0.0013

Table 13: Also includes speaker and script fixed effects.

Outcome: Inspiring

	Estimate	Std. Error	t value	Pr(>|t|)
Modulated Speech (Female)	10.2660	0.8549	12.01	0.0000
Modulated Speech (Male)	6.3931	0.8593	7.44	0.0000
High Pitch (Female)	0.0464	0.8550	0.05	0.9567
High Pitch (Male)	-2.6305	0.8597	-3.06	0.0022
Fast Rate (Female)	4.0855	0.8548	4.78	0.0000
Fast Rate (Male)	3.7253	0.8592	4.34	0.0000
High Volume (Female)	1.2588	0.8548	1.47	0.1409
High Volume (Male)	1.8586	0.8595	2.16	0.0306

Table 14: Also includes speaker and script fixed effects.

Outcome: Passionate

	Estimate	Std. Error	t value	Pr(>|t|)
Modulated Speech (Female)	15.7657	0.7574	20.81	0.0000
Modulated Speech (Male)	10.2431	0.7615	13.45	0.0000
High Pitch (Female)	-0.0429	0.7574	-0.06	0.9548
High Pitch (Male)	-1.8242	0.7617	-2.39	0.0166
Fast Rate (Female)	5.2483	0.7574	6.93	0.0000
Fast Rate (Male)	7.0002	0.7612	9.20	0.0000
High Volume (Female)	1.7968	0.7572	2.37	0.0177
High Volume (Male)	2.2793	0.7617	2.99	0.0028

Table 15: Also includes speaker and script fixed effects.

Outcome: Persuasive

	Estimate	Std. Error	t value	Pr(>|t|)
Modulated Speech (Female)	8.6354	0.7630	11.32	0.0000
Modulated Speech (Male)	5.0168	0.7671	6.54	0.0000
High Pitch (Female)	-0.5768	0.7629	-0.76	0.4496
High Pitch (Male)	-2.1267	0.7673	-2.77	0.0056
Fast Rate (Female)	3.9583	0.7629	5.19	0.0000
Fast Rate (Male)	3.8845	0.7668	5.07	0.0000
High Volume (Female)	1.5258	0.7627	2.00	0.0455
High Volume (Male)	1.9443	0.7673	2.53	0.0113

Table 16: Also includes speaker and script fixed effects.

Outcome: Trustworthy

	Estimate	Std. Error	t value	Pr(>|t|)
Modulated Speech (Female)	4.3560	0.7321	5.95	0.0000
Modulated Speech (Male)	1.1580	0.7360	1.57	0.1157
High Pitch (Female)	-0.1090	0.7320	-0.15	0.8816
High Pitch (Male)	-2.2383	0.7362	-3.04	0.0024
Fast Rate (Female)	3.4406	0.7320	4.70	0.0000
Fast Rate (Male)	2.9516	0.7357	4.01	0.0001
High Volume (Female)	0.7403	0.7318	1.01	0.3117
High Volume (Male)	0.0187	0.7362	0.03	0.9797

Table 17: Also includes speaker and script fixed effects.

Outcome: Willingness to vote for				
	Estimate	Std. Error	t value	Pr(>|t|)
Modulated Speech (Female)	6.1578	0.7759	7.94	0.0000
Modulated Speech (Male)	2.6456	0.7801	3.39	0.0007
High Pitch (Female)	-0.0146	0.7758	-0.02	0.9850
High Pitch (Male)	-1.9351	0.7802	-2.48	0.0131
Fast Rate (Female)	3.3608	0.7758	4.33	0.0000
Fast Rate (Male)	3.3682	0.7797	4.32	0.0000
High Volume (Female)	0.8456	0.7756	1.09	0.2756
High Volume (Male)	0.9934	0.7802	1.27	0.2030

Table 18: Also includes speaker and script fixed effects.