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

Taylor J. Damann † Dean Knox ‡ Christopher Lucas §

August 19, 2024

Abstract

Numerous disciplines hypothesize about the causal effects of speech, which influences others not only through *which* words are spoken, but also *how* they are spoken. Yet applied research focuses almost exclusively on the textual component of speech—ignoring its audiovisual components or reducing them to coarse summary statistics. We show that text-only analyses are biased, except in implausible scenarios where (a) non-textual speech elements are irrelevant to listeners, or (b) speakers' vocal style does not change with the words spoken. Even analyses including audiovisual summary measures are biased unless (c) these measures satisfy a "sufficient reduction" condition, fully capturing the non-textual mechanisms through which speech operates. To show this, we develop a formal causal framework that explicitly accounts for the unstructured and multimodal nature of speech. We present an application to U.S. presidential campaign speeches, using the framework to clarify implicit assumptions in prior work. We then demonstrate two designs that permit valid hypothesis tests and causal effect estimates under more plausible conditions: (1) a naturalistic experiment exploiting subtle variation in campaign speech "catchphrases" with near-identical wording, identified with 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.

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

*For helpful comments, we thank Taylor Carlson, Ted Enamorado, Justin Grimmer, Kosuke Imai, Jacob Montgomery, and Matthew Tyler, as well as participants in the 2020 Meeting of the Japanese Society for Quantitative Political Science, the Rice University Speaker Series, the Hot Politics Lab at the University of Amsterdam, the Political Data Science Lab at Washington University, the CIVICA Data Science Seminar, the 2021 Summer Political Methodology Meeting, and Junior Faculty Working Group at Washington University.

†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

Spoken language has long been used to persuade listeners by shaping perceptions, evoking emotion, and presenting evidence (Aristotle, c. 330 BCE). A vast literature studies the causal effects of spoken language across disciplines such as business (Allison et al., 2022; Xu et al., 2023), law (Bucholtz, 2009; Elias-Bursac, 2015), political science (Anderson and Klofstad, 2012; Boussalis et al., 2021a; Dietrich, Hayes, et al., 2019a; Osnabrügge et al., 2021; Rittmann, 2023; Klofstad, 2016; Klofstad, 2017), and social psychology (Carli, 1990; Ji et al., 2004), to name only a few. It is widely recognized that these effects do not operate solely through the textual component of human speech such as word choice and syntax. Rather, the effects of speech depend heavily on paralinguistic (e.g., auditory and visual) components including intonation, emphatic stress, hand movement, and facial expression (Bänziger and Scherer, 2005; Eaves and Leathers, 2017; Krauss et al., 1996; Wagner et al., 2014).

Puzzlingly, much applied research uses sophisticated text-analysis techniques to study communication (Grimmer et al., 2022; Rodriguez and Spirling, 2022) while concurrently ignoring paralinguistic cues or capturing them only with coarse measures, such as the average pitch of a speaker's voice (Cohen-Mohliver et al., 2023; Krahé and Papakonstantinou, 2020). In this paper, we develop a formal causal framework for studying the effects of speech that explicitly accounts for its textual, auditory, and visual components. We use this framework to reexamine a variety of common research designs, clarify the often-implicit assumptions upon which they rest, and demonstrate new designs that address their limitations.

The remainder of this paper proceeds as follows. In Section 2, we introduce an original corpus of U.S. presidential campaign speeches that will serve as a running example for the application of our proposed framework. A large body of academic work studies political campaigns by analyzing textual transcripts of speeches alone—an approach that contrasts sharply with accounts by journalists and political strategists, which frequently emphasize candidates' vocal style. We summarize this work and briefly review the literature on paralinguistic cues in persuasion, with particular attention to the auditory component of speech.

In Section 3, we formally define our proposed framework and use it to show that text-based analyses of rhetorical effectiveness will generally fail to recover the desired quantities of interest. This is because speakers modulate the non-textual aspects of their speeches in a way that depends on textual topic, and these non-textual aspects also influence listener perceptions. In other words, audiovisual speech elements operate as omitted confounders, distorting regressions that seek to explain listener responses using speech transcripts alone.

Next, we turn our attention to recent analyses of speech audio, which have drawn competing conclusions about how vocal style influences listeners. We examine and extend research designs from past work, showing how apparent contradictions in this literature can be explained by differing implicit assumptions of the designs. One common approach is to extend text-based regressions by incorporating additional auditory summary statistics, such as the speaker's average pitch, which are reductive proxies of their vocal style. Our framework clarifies that this is generally biased as well, for much the same reason that text-only regressions are biased: when speakers raise their pitch, they inevitably shift other aspects of vocal style (e.g. pitch, volume). Because these other vocal elements may also influence listeners, they behave as omitted confounders. To clarify implicit assumptions in prior work, we extend ideas in the text domain from Egami et al. (2018) to formalize the notion of a "sufficient reduction," or set of summary statistics that jointly capture all mechanisms by which speech can influence listeners, and show that many claims in the literature rest on such assumptions.

In Section 5, we argue that this sufficient-reduction condition is unlikely to be satisfied for a number of reasons, not least of which is because speech researchers continue to propose new ways of operationalizing vocal style. To make progress in spite of this challenge, we propose a design in which researchers construct pairs of utterances with matching transcripts but distinct vocal style. We illustrate this design in the political-campaign context by using real-world speeches to identify "catchphrases" that candidates repeat across multiple speeches—sometimes in more animated or emphatic tones, and sometimes with verbal stumbles or flat delivery. While these matched pairs differ on multiple vocal dimensions, meaning that researchers cannot isolate the effects of pitch as compared to speed, we show that the designs allow researchers to conduct valid null-hypothesis tests. We further show how the design

can be extended to accommodate utterance pairs with imperfectly matched transcripts, as exact matches cannot always be found when working with natural-speech corpora, using a bias-correction step under a parallel-trends-like assumption.

Finally, in Section 6, we demonstrate how these assumptions can be made more plausible through experimental designs in which elements of speech style are manipulated directly by the researcher, while also formally clarifying the limitations of such manipulations. We begin by presenting a set of manipulations in which pitch and volume of a recorded campaign speech segment are artificially shifted using audio-editing software. While this approach can be used to obtain unbiased estimates of the quantities that some prior work seems to target in regression-type analyses, the tradeoff with artificial manipulations is that they are fundamentally limited, in the sense that researchers cannot easily create the kind of vocal variation that real-world speakers tend to exhibit. To complement these artificial manipulations, we work with voice actors to naturalistically vary the modulation and speed of campaign speeches. This approach better approximates interventions that are possible in real-world speech, but with the tradeoff of creating compound treatments that simultaneously manipulate many aspects of speech in a difficult-to-control manner. We further demonstrate how artificial and naturalistic manipulations can be combined in a factorial design and clarify the causal estimands that such designs permit researchers to estimate.

In sum, we develop a formal framework that clarifies the consequences of ignoring the non-textual components of speech and demonstrate how common approaches to incorporating speech audio through auditory summary statistics yield biased estimates. As a solution, we propose designs that address these threats and clarify the assumptions on which they rest. Throughout, we demonstrate our proposed methods in the context of political campaign speech. Our results suggest substantial differences in the effects of various vocal interventions, and that the most commonly studied measure of vocal style—average pitch of a speaker's voice—does not necessarily have the largest effect on listeners. Overall, our analysis and results highlight the need for further study of the complex channels through which speech influences listener beliefs and behavior.

2 Running Example: Vocal Style in U.S. Presidential Candidates

A considerable body of interdisciplinary research has studied political campaigns, much of which 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). While suitable for certain goals (e.g., categorizing speeches according to their topic), we will demonstrate that text-based analyses do not in general recover quantities of interest related to rhetorical effectiveness and speech style. This should come as no surprise; though recent progress has made text analysis increasingly powerful and convenient, most humans intuitively recognize that effective communication and persuasion involves a broad range of non-textual cues.[1]

Conceptually, listeners' inferences from speech can be divided into two categories: time-invariant traits and time-varying status of the speaker. The former is what a listener comes to believe about 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 convey intelligence to the listener: speech characteristics such as rate, pitch, pronunciation, and disfluencies can signal the speaker's competence (Klofstad et al., 2012; Tigue et al., 2012; Surawski and Ossoff, 2006). Vocal characteristics are also the primary way through which a speaker's charisma is perceived (Niebuhr et al., 2017; Novák-Tót et al., 2017). Traits such as dominance, intelligence, and charisma change gradually, if at all, and one speech can leave lasting impressions about these qualities on a listener. The voice also offers unique insight into dynamic—or time-varying—attributes of the speaker, such as their expressed emotion (Knox and Lucas, 2021). Numerous auditory features help convey this information (Banse and Scherer, 1996; Johnstone and Scherer, 2000; Scherer, 2003; Dietrich, Hayes, et al., 2019b), including tone and intonation

[1] Due to space constraints, we only 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 speech components, returned about 101,000 results in May 2024.

patterns (Bänziger and Scherer, 2005) and qualities such as breathiness and meekness (Gobl and Chasaide, 2003). These aspects of vocal style are highly correlated, making it difficult to disentangle the effect of a single element of speech.

Throughout this paper, we demonstrate our framework for studying the causal effects of speech by collecting and analyzing a corpus of campaign speeches from the 2012 Presidential Election. Our corpus contains 100 recorded campaign-speech videos—38 of Barack Obama and 62 of Mitt Romney—that we scraped from the nonpartisan website ElectAd.[2] To help illustrate the importance of auditory information, Figure 1 shows how then-candidate Obama's voice varied over the course of his 2012 Democratic National Convention speech. Our basic unit of analysis is an utterance, or continuous segment of speech roughly analogous to a sentence. The left panels present different views of a key speech segment: four sequential utterances in which Obama criticizes his opponent's laissez-faire positions and argues that governments must defend citizenship rights: "you know what", "that's not who we are", "that's not what this country is about", and "as Americans, we believe we are endowed by our creator with certain inalienable rights."

While this textual channel can convey semantic meaning, it fails to capture the sincerity and conviction with which those words are spoken—information that is readily apparent in the auditory channel. In Panels A1–3, the horizontal axis represents time. Panel A1 depicts the waveform, or raw audio signal: the vertical axis is the air-pressure signal received by a microphone or eardrum at a particular instant in time. Panel A2 presents one time-series auditory feature that can be extracted from this raw signal: the speaker's vocal pitch, including the rising tone of "you know what?" that asks a rhetorical question of the audience.[3] Panel A3 depicts another time-series auditory feature, the speech volume or loudness, which shows a long pause as the question hangs in the air, followed by loud bursts that punctuate "that's *not* who we are!"

Analysts often reduce this rich time-series information to a handful of utterance-level

[2]We manually preprocessed videos to remove music and segments involving individuals other than Obama and Romney.

[3]Pitch is an estimated quantity that cannot be directly observed and is undefined during unvoiced speech such as sibilants and plosives. We plot it only during periods estimated to be voiced speech.

summary statistics. Common measures include the average vocal pitch and volume of an utterance; occasionally, analysts also compute the variance of these features over the course of an utterance, which can quantify the modulation of a speaker's voice. After a continuous audio recording is reduced into a sequence of utterances, these summary statistics can be represented as a second, coarser time series. In Panels B1–4, the horizontal axis represents the index of an utterance (e.g. the first sentence of a speech), and the vertical axis represents one particular utterance-level summary statistic. While some general patterns can be observed, such as the increasing modulation of volume toward the end of the speech as Obama rallies the crowd, note that these sentence-level summaries discard much of the finer-grained within-utterance information displayed in panels A1–3.

In Appendix A, we present two descriptive analyses of campaign speech. First, we characterize differences in vocal style between Obama and Romney. Obama's speech exhibits considerably greater variation in within-utterance pitch and volume, and he utilizes greater emphasis with falling tones—consistent with popular accounts that characterize him as a dynamic public speaker. Second, we show how each speaker modifies their vocal style within a campaign speech, depending on the topic of the current utterance. In Appendix Figure 11 reveals that Obama uses rhetorical flourishes, in the form of pitch and volume modulation, to emphasize his discussion of religious and economic issues. Appendix Figure 12 shows a similar emphasis on economic issues by Romney, but a considerably more subdued and monotonous voice on on technology, education and defense issues.

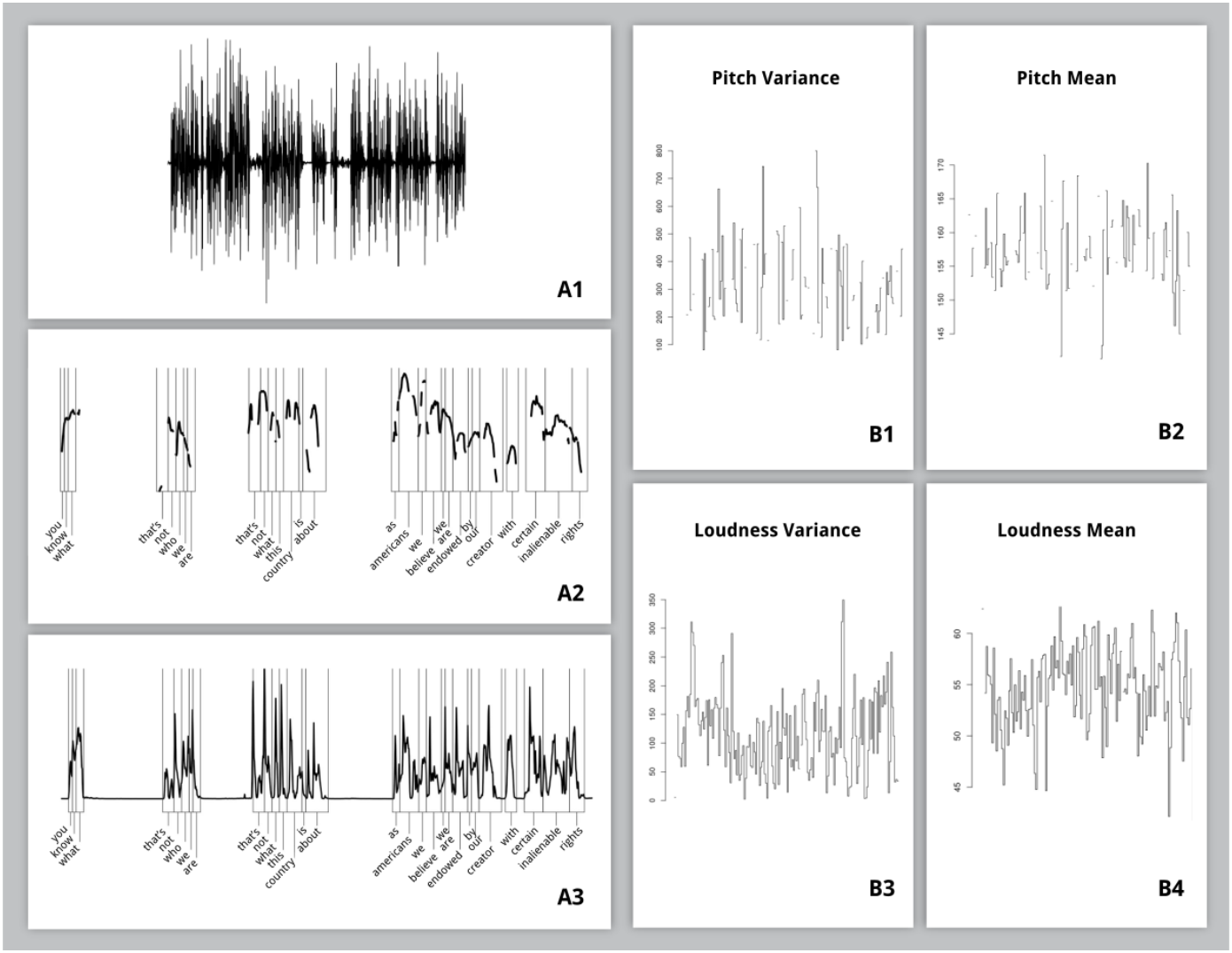


Figure 1: Auditory features of Obama's 2012 Democratic National Convention speech. Panels A1–3 depict four sequential utterances. Panel A1 shows the raw audio waveform, or time-series signal of pressure received by a microphone or eardrum. Panel A2 shows pitch, one time-series auditory feature that can be extracted from the raw waveform, along with the timestamped words of the sentence; together, the text and audio clarify that "you know what?" is a rhetorical question posed to the audience. Panel A3 shows the same words alongside the volume of the speaker's voice; this conveys the long pause as the question hangs in the air, followed by loud bursts that emphasize "that's *not* who we are!" Utterance-level summary statistics such as mean and variance of pitch and volume are computed from these fine-grained time series. Panels B1–4 show vocal variation at a coarser level, showing how utterance summary measures change over sequential utterances in the longer speech. For example, panel B3 shows how Obama starts his speech dynamically, settles down for a period, and gradually ramps up again toward the end as he rallies the crowd.

3 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. Our approach builds 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 $j \in \{1, \ldots, J\}$. We denote the j-th utterance consumed by respondent i with the triple $\boldsymbol{U}_{ij} = \{\boldsymbol{T}_{ij}, \boldsymbol{A}_{ij}, \boldsymbol{V}_{ij}\}$: the raw components of the utterance, respectively representing the transcript, audio recording, and silent video.[4] In what follows, we will use bars to denote collections such as the J utterances observed by respondent i, $\bar{\boldsymbol{U}}_i = \{\boldsymbol{U}_{i1}, \ldots, \boldsymbol{U}_{iJ}\}$; similarly, the collection of utterance transcripts will be $\bar{\boldsymbol{T}}_i = \{T_{i1}, \ldots, T_{iJ}\}$; audio recordings, $\bar{\boldsymbol{A}}_i = \{A_{i1}, \ldots, A_{iJ}\}$; and silent videos, $\bar{\boldsymbol{V}}_i = \{V_{i1}, \ldots, V_{iJ}\}$. After consuming the candidate's j-th utterance, the i-th voter forms a K-dimensional evaluation—containing different evaluations such as perceived competence and trustworthiness—indexed by $k \in \{1, \ldots, K\}$. We collect these evaluations in $\boldsymbol{Y}_{ij} = \{Y_{ij1}, \ldots, Y_{ijK}\}$.

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. To facilitate this, we propose a potential-outcome framework (Neyman, 1923; Rubin, 1974) for studying the effects of speech. Let $Y_{ijk}(\bar{\boldsymbol{u}}) = Y_{ijk}(\bar{\boldsymbol{t}}, \bar{\boldsymbol{a}}, \bar{\boldsymbol{v}})$ 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 $\bar{\boldsymbol{u}} = \{\bar{\boldsymbol{t}}, \bar{\boldsymbol{a}}, \bar{\boldsymbol{v}}\}$—respectively, the counterfactual sequences of transcripts, audio recordings, and silent videos.

One immediate takeaway from this formalization is that the common practice of textual regressions—i.e. seeking to explain respondents' evaluations using some textual attributes of the speech to which respondents were exposed—will generally produce biased estimates when

[4] Throughout, we use braces for ordered sets.

using real-world utterances. For example, if analysts seek to understand whether certain ways of formulating arguments or framing questions are effective at changing listeners' minds, they will generally fail to recover the true quantity of interest, even when respondents are randomly assigned to utterances. The reason is twofold. First, as Section 2 illustrates, the vocal style of real-world speech is often correlated with its topic or other textual attributes, i.e. $\boldsymbol{T} \not\perp\!\!\!\perp \boldsymbol{A}$. And second, as we show below in Experiments 1–2, this vocal style has its own effects on listener perceptions, so that $\boldsymbol{Y} \not\perp\!\!\!\perp \boldsymbol{A}$. Thus, if listeners are exposed to audiovisual speech recordings, these omitted variables will confound inferences even if researchers are only interested in textual effects.

4 Assumptions and Experimental Designs for Studying Effects of Speech

Conceptualizing the effects of speech, a highly unstructured stimulus, is a challenging task. Much of the existing literature sidesteps this challenge by simply describing modeling procedures, such as the regression equations that were used, rather than articulating and justifying the assumptions under which these approaches would yield a defensible answer. In this section, we formalize causal assumptions implicit in much prior work and propose experimental designs that allow more relaxed assumptions.

4.1 Candidate Assumptions for Studying Effects of Speech

One critical, unstated assumption in speech research is 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 an unstructured or high-dimensional sequence of words, $\boldsymbol{t}$, into a low-dimensional representation, $g_T(\boldsymbol{t})$, such as the presence or absence of 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 as well. 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), which also studies the Supreme Court, employs a $g_A(\boldsymbol{a})$ reduction that classifies justice utterances into domain-relevant categories—"skeptical" or "neutral" questioning. These sufficient reductions are estimated with supervised hidden Markov models, then used to study the flow of conversation in judicial deliberations. In this paper, for illustrative purposes, we quantify the vocal characteristics of each candidate utterance with a multidimensional $g_A(\boldsymbol{a})$ that includes the mean and variance of pitch and volume, along with speech speed (measured in words per second).[5] In principle, analysts can employ visual reductions, $g_V(\boldsymbol{v})$, to represent elements of visual style such as facial expressions and head movements (as in Torres, 2018; Boussalis et al., 2021b; 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, for compactness, we will use $\bar{g}_X()$ to denote the repeated application of the sufficient reduction function to multiple utterances, so that $\bar{g}_X(\bar{\boldsymbol{X}}_i) = \{g_X(\boldsymbol{X}_{i1}), \ldots, g_X(\boldsymbol{X}_{iJ})\}$. We are now ready to formally state the assumption.

[5] However, we are generally skeptical that these characteristics alone are sufficient to capture all auditory persuasive mechanisms in political speech, and we do not advocate for the use of Assumption 1 except to clarify implicit assumptions in prior work.

Assumption 1 (Sufficiency of reduced representation)**.** $Y_{ijk}(\bar{\boldsymbol{u}}) = Y_{ijk}(\bar{\boldsymbol{u}}')$ *for* $\bar{\boldsymbol{u}} = \{\bar{\boldsymbol{t}}, \bar{\boldsymbol{a}}, \bar{\boldsymbol{v}}\}$ *and* $\bar{\boldsymbol{u}}' = \{\bar{\boldsymbol{t}}, \bar{\boldsymbol{a}}, \bar{\boldsymbol{v}}\}$ *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 means that apart from the sufficient reductions, all other elements of the text, audio, and video are irrelevant in the sense that they would not lead any respondent to evaluate any candidate differently. The assumption allows us to rewrite $Y_{ijk}(\bar{\boldsymbol{u}})$ in terms of the sufficient reductions for each utterance, $Y_{ijk}\left(\bar{g}_T(\bar{\boldsymbol{t}}), \bar{g}_A(\bar{\boldsymbol{a}}), \bar{g}_V(\bar{\boldsymbol{v}})\right)$, which is notationally convenient. However, as a reviewer pointed out, this is a strong assumption that is analogous to a sharp null. In many contexts, it can be relaxed to an assumption about distributional equality, $Y_{ijk}(\bar{\boldsymbol{u}}) \stackrel{d}{=} Y_{ijk}(\bar{\boldsymbol{u}}')$ if $\bar{\boldsymbol{u}}$ and $\bar{\boldsymbol{u}}'$ have the same sufficient reductions, a point that we explore further below.[6]

This formulation is without loss of generality because text, audio, and visual reduction functions are allowed to be arbitrarily complex. One way to guarantee that the assumption will hold is by considering a set of "reductions" that does not change anything at all, like the identity function $g_T(\boldsymbol{t}) = \boldsymbol{t}$. In this case, the assumption states only that a speech will receive the same evaluation as an identical copy of itself. This means analysts effectively assume that even the slightest deviation—a stray "uh," the slightest pause, or a miscolored pixel—can produce entirely different potential evaluations. 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 more restrictive assumptions about sufficient reductions are used, 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

[6] Note that our formulation is stronger than the sufficiency assumption of Egami et al. (2018), which (adapted to our context) would be the equality-of-expectations assumption that $\mathbb{E}[Y_{ijk}(\bar{\boldsymbol{u}})] = \mathbb{E}[Y_{ijk}(\bar{\boldsymbol{u}}')]$ if $\boldsymbol{u}$ and $\boldsymbol{u}'$ have the same reductions. This is because in paired-comparison tasks, where respondents choose whether $\boldsymbol{u}$ or $\boldsymbol{u}'$ is a more appealing speech, equality of expectations alone does not guarantee that $\Pr[Y_{ijk}(\bar{\boldsymbol{u}}) > Y_{ijk}(\bar{\boldsymbol{u}}')]$ is the same as $\Pr[Y_{ijk}(\bar{\boldsymbol{u}}) < Y_{ijk}(\bar{\boldsymbol{u}}')]$. However, an equality-of-distributions assumption is sufficient to resolve this issue.

[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 $g_V(\bar{\boldsymbol{v}}) = g_V(\bar{\boldsymbol{v}}')$.

effect on opinion formation. Similarly, when $g_A(\cdot)$ measures only the average pitch, analysts assume that a monotonous 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).

Importantly, this formulation makes clear that violations of Assumption 1 will generally lead to bias. Suppose that listeners are influenced by vocal style in ways that are captured in a true sufficient reduction that includes the average pitch and volume of a speaker's voice as well as the variance or modulation of those auditory features. In this case, the common practice of explaining evaluations using average pitch, excluding the other auditory variables, will lead to omitted variable bias as well. This is because most speakers cannot raise their pitch without simultaneously changing their voice in a host of other ways—and as we show in Experiment 2, many of these other elements of vocal style have their own effects on listeners.

Next, 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)**.** $Y_{ijk}\left(\bar{\boldsymbol{t}}, \bar{\boldsymbol{a}}, \bar{\boldsymbol{v}}\right) = Y_{ijk}\left(\bar{\boldsymbol{t}}', \bar{\boldsymbol{a}}', \bar{\boldsymbol{v}}'\right)$ *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.[8] With this simplifying assumption, we can eliminate past and future utterances from our potential outcomes, dropping the j subscript to obtain the simplified notation $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right)$. 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 $Y_{it}\left(g_T(\boldsymbol{t}), \varnothing, \varnothing\right)$ 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)**.**
$Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right) = Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), g_V(\boldsymbol{v}')\right)$ *for all* $\boldsymbol{a}$, $\boldsymbol{a}'$, $\boldsymbol{v}$, $\boldsymbol{v}'$.

As with Assumption 1, in many settings, this can be weakened to require only equality of expectations or distributions.

4.2 Experimental Designs for Studying Effects of Speech

With these preliminaries, we are now ready to describe the design of experiments that will be reported 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,

[8] In general, we suggest that researchers should carefully evaluate the observable implications of this assumption by randomizing stimuli ordering and testing for ordering effects—i.e., whether earlier and later stimuli tend to score differently—wherever possible. While the paired-utterance forced-choice design of Experiment 1 does not permit such a test, we are able to do so in the single-utterance rating design of Experiment 2. When testing for order effects here, regardless of whether we regress stimuli scores on presentation order or on a binary indicator for presentation in the second half of the experiment, we find no evidence that Assumption 2 is violated.

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 $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right)$. 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 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 candidate's typical range of speaking styles. 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 in isolation from others. 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 utilize a single-utterance rating design in which respondents are presented 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-World Campaign Speech

We now design a naturalistic experiment that leverages variation in how candidates deliver campaign catchphrases in the corpus described in Section 2. We first use a new computational text-analysis technique to segment and cluster speech transcripts into 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 evaluate 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 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 string comparisons. Accomplishing this task in an efficient manner required the development of a new text-matching algorithm, which we details in Appendix Section B. 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. 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 only once, ensuring that no individual was repeatedly exposed to a particular catchphrase. After dropping subjects who failed an audio-based attention check or had duplicate IP addresses, 773 voters participated in the first experiment. On average, more than 250 voters evaluated each phrase. Appendix Section D displays screenshots of the survey interface. Appendix Figure 14 shows the choice proportions for each experimental condition and recording pair, demonstrating substantial variation in choice patterns.[9]

5.2 Null Hypothesis Testing in the Naturalistic Experiment

In Section 5.2.1, we first develop the basic logic of the experimental design in the simple case when two utterances have perfectly matched text. This setting demonstrates the clarifying value of the notation previously introduced in Section 3. In Section 5.2.2, we then provide an alternative design that extends the approach by accounting for the slight variations in phrasing that appear in other identified catchphrase pairs.

5.2.1 Paired Utterance Design with Exact Text Matching

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 emphatically ($\boldsymbol{u}' = \{\boldsymbol{t}', \boldsymbol{a}', \boldsymbol{v}'\}$). In this utterance pair, the two transcripts are a perfect textual 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 $\boldsymbol{v} \neq \boldsymbol{v}'$.

Can these paired utterances be used to identify the auditory and visual elements that are most compelling to voters, allowing candidates to modify their speech style and attract more votes? Unfortunately, the high-dimensional nature of speech makes estimation of these effects difficult. For example, even if analysts assume that key aspects of vocal style

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

can be captured with the sufficient-reduction function $g_A()$, there will usually be multiple elements of the difference vector, $g_A(\boldsymbol{a}) - g_A(\boldsymbol{a}')$, that could explain differences in voter preferences. In Experiment 1, we will therefore focus on using the matched-pair design to show that Assumption 3 is violated, by demonstrating that different non-textual content leads to different perceptions. Later, in Experiment 2, we propose an alternative factorial design that addresses these limitations.

Experiment 1 proceeds as follows. We utilize a forced-choice design in which a subset of respondents are exposed to videos of both $\boldsymbol{u}$ and $\boldsymbol{u}'$ in randomized order, then asked to select the one that scores higher. Assumption 3 implies that the two evaluations will be exactly equal, in which case $Y_{ik}(\boldsymbol{u}) = Y_{ik}(\boldsymbol{u}')$ for all respondents i and all evaluation metrics k, so that the first term of (1) below will evaluate to zero. In this case, respondents would select an utterance uniformly between the randomly ordered $\boldsymbol{u}$ and $\boldsymbol{u}'$, leading to a $\boldsymbol{u}$ choice probability of $\frac{1}{2}$. However, a reviewer noted that relaxed versions of Assumption 3 are possible as well, such as the distributional equality assumption $Y_{ik}(\boldsymbol{u}) \stackrel{d}{=} Y_{ik}(\boldsymbol{u}')$. In this case, the first term of (1) accounts for the instances where one utterance's evaluation exceeds that of its identically distributed counterpart, and the second term accounts for ties.

$$\begin{aligned} &\Pr\left[Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right) > Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), g_V(\boldsymbol{v}')\right)\right] \\ &\quad + \frac{1}{2}\Pr\left[Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right) = Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), g_V(\boldsymbol{v}')\right)\right] \;=\; \frac{1}{2} \end{aligned} \tag{1}$$

This leads to Hypothesis 1: that the video of $\boldsymbol{u}$ will be selected over the video of $\boldsymbol{u}'$ with probability $\frac{1}{2}$. 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 modified the matched-text design for a different subset of respondents. In this version, respondents were given audio recordings only, so that evaluations were based on a comparison between $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right)$

and $Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right)$. Assumption 3 then suggests

$$\Pr\left[Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right) > Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right)\right] + \frac{1}{2}\Pr\left[Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right) = Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right)\right] = \frac{1}{2}. \quad (2)$$

This leads to Hypothesis 2: that the audio of $\boldsymbol{u}$ will be selected over the audio of $\boldsymbol{u}'$ with probability $\frac{1}{2}$. We reject the null again, finding strong evidence that vocal style matters. Across every evaluation criterion, respondents exhibited a preference for one video over the other by more than 35 percentage points, and the null hypothesis was rejected at $p = 0.005$ or less for each of the $K = 8$ evaluation criteria.

5.2.2 Paired Utterance Design with Adjustment for Approximate Text Matching

While the simple tests in Section 5.2.1 allow us to decisively reject the assumption that vocal delivery is irrelevant—an assumption that is implicit in much prior work—it is rare that naturalistic speech will permit such perfectly matched experiments. Even in the campaign context, 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—providing natural variation that seems well-suited for researchers seeking to evaluate the impact of vocal delivery. However, strictly speaking, a direct comparison between the 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 next develop a "difference in differences" design that compares the pair of audio recordings, $\{\boldsymbol{t}, \boldsymbol{a}\}$ and $\{\boldsymbol{t}', \boldsymbol{a}'\}$, to the pair of utterance transcripts alone, $\boldsymbol{t}$ and $\boldsymbol{t}'$. 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 an 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)**.**
$Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right) = \alpha_{ik} + h_{ik}^T(g_T(\boldsymbol{t})) + h_{ik}^A(g_A(\boldsymbol{a})) + h_{ik}^V(g_V(\boldsymbol{v}))$*, 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 trends assumption in conventional difference-in-differences analyses. An important special case that automatically satisfies Assumption 4 is when the $\boldsymbol{t}$-to-$\boldsymbol{t}'$, $\boldsymbol{a}$-to-$\boldsymbol{a}'$, or $\boldsymbol{v}$-to-$\boldsymbol{v}'$ manipulations have constant treatment effects. Due to the complexity of candidate speech and voter evaluations, this assumption is unlikely to be generally satisfied. However, because the manipulations studied in Experiment 1 are quite 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 $\{\boldsymbol{t}, \boldsymbol{a}\}$

[10] In many settings, Assumption 4 can be weakened to an assumption about additive separability of the conditional expectation function, rather than the individual-level potential-outcome function itself. When examining single-utterance ratings, the weaker assumption that $\mathbb{E}\left[Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right)\right] = \alpha + h_T(g_T(\boldsymbol{t})) + h_A(g_A(\boldsymbol{a})) + h_V(g_V(\boldsymbol{v}))$ will generally suffice. As with Assumption 3, we require stronger assumptions when analyzing the paired-profile forced-choice design of Experiment 1.

and $\{\boldsymbol{t}', \boldsymbol{a}'\}$ can be rewritten

$$\begin{aligned}
&\Pr\left[Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), \varnothing\right) > Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right)\right] \\
&\quad + \frac{1}{2}\Pr\left(Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}), g_V(\boldsymbol{v})\right) = Y_{ik}\left(g_T(\boldsymbol{t}), g_A(\boldsymbol{a}'), \varnothing\right)\right) \\
&= \Pr\left[h_{ik}^T(g_T(\boldsymbol{t})) + h_{ik}^A(g_A(\boldsymbol{a})) > h_{ik}^T(g_T(\boldsymbol{t}')) + h_{ik}^A(g_A(\boldsymbol{a}')) > 0\right] \\
&\quad + \frac{1}{2}\Pr\left[h_{ik}^T(g_T(\boldsymbol{t})) + h_{ik}^A(g_A(\boldsymbol{a})) = h_{ik}^T(g_T(\boldsymbol{t}')) + h_{ik}^A(g_A(\boldsymbol{a}')) = 0\right] \qquad (3)
\end{aligned}$$

Assumption 3 then suggests Hypothesis 3: that the proportion of respondents who choose the text of $\boldsymbol{u}$ (over the text of $\boldsymbol{u}'$) should be equal to the proportion of respondents who choose the audio of $\boldsymbol{u}$ (over the audio of $\boldsymbol{u}'$).

We 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 "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 the null 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] which complicates interpretation of results if Assumption 4 does not hold.

With these considerations in mind, we design a second experiment in which we hire 10

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

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 E 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 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 $\times$ 6 scripts $\times$ 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 rate, 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 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 F 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.

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 3, 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, $\boldsymbol{t}$, allowing us to hold fixed the textual information used by respondents, $g_T(\boldsymbol{t})$. Our quantities of interest relate to average marginal component effect

[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.

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

(AMCE, Hainmueller et al., 2014)—either for manipulations targeting a single element of $g_A(\boldsymbol{a})$, 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 in campaign speech style 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 2, we plot average willingness to vote for each of the voice actors—a decision that was based only on a brief audio recording. Note that 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 male speakers, compared to female counterparts.

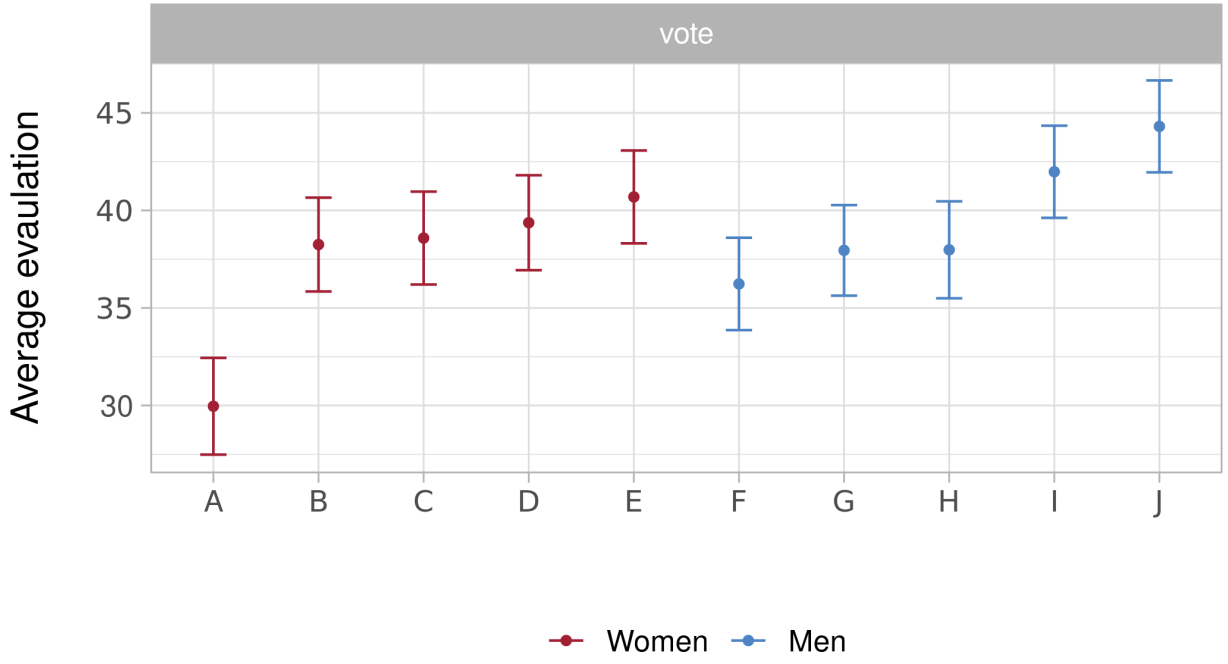


Figure 2: Average expressed willingness to vote for each actor, based only on hearing their recorded speech. Demonstrates that holding content fixed, there is sizeable variation in voter preference. Estimated from a regression with fixed effects for script and indicators for treatment condition. Table 11 presents the results of this regression.

Next, Figure 3 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 male and female voice actors, documenting significant gender heterogeneity. Vocal modulation and rate of speech have 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 gave greater rewards women who were instructed to modulate their voice, compared to men who were given the same instructions—though as Appendix Figure 10 shows, this may be driven by the fact that female actors appear to have had exhibited greater changes in speaking style in response to this prompt. Among our speakers, men are also punished more than women when voices are artifically raised in pitch, 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 4, 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.

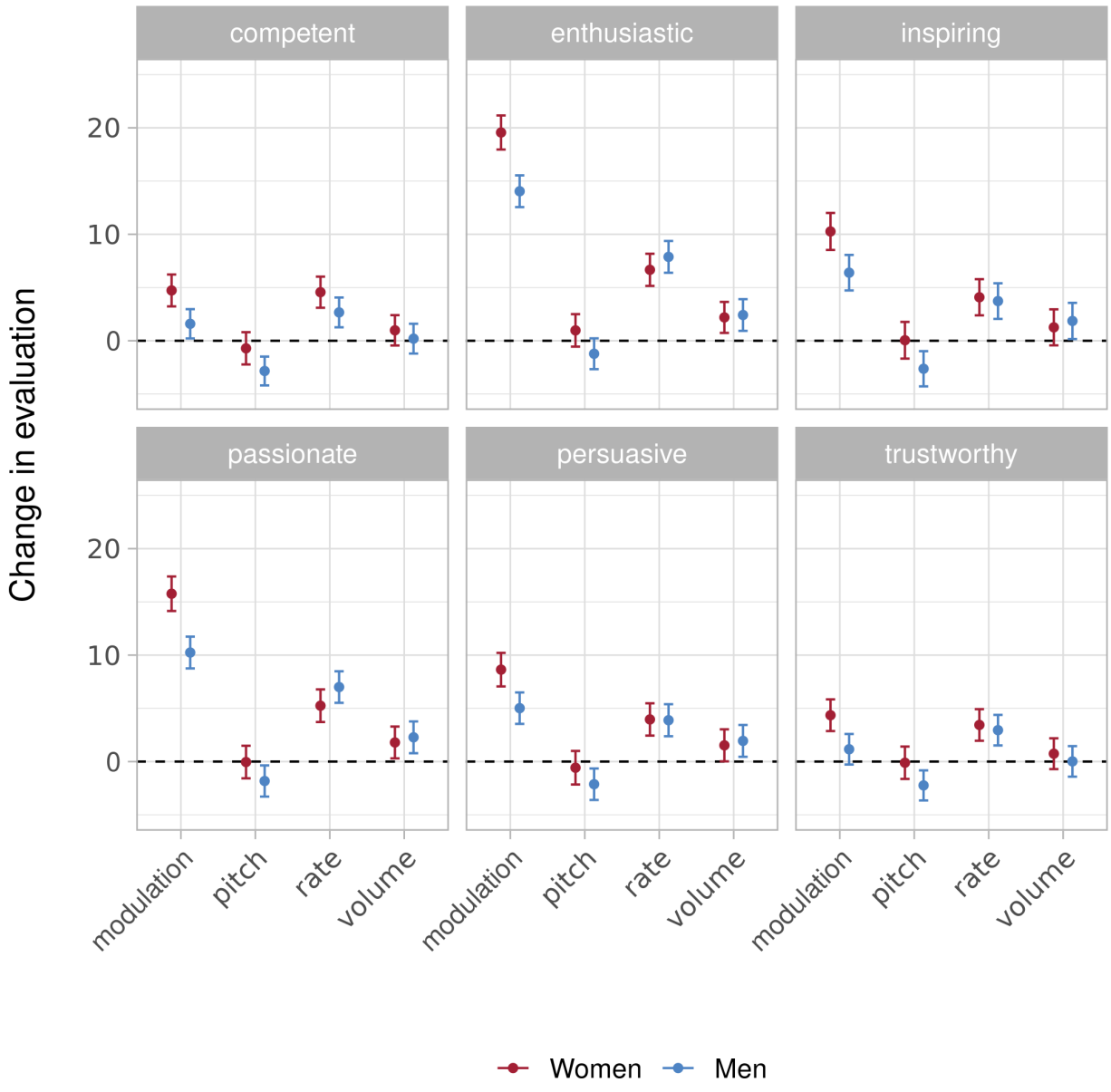


Figure 3: Effect of speech features on evaluations of the respective characteristic by speaker gender. Appendix Section H.1 presents these results in tables.

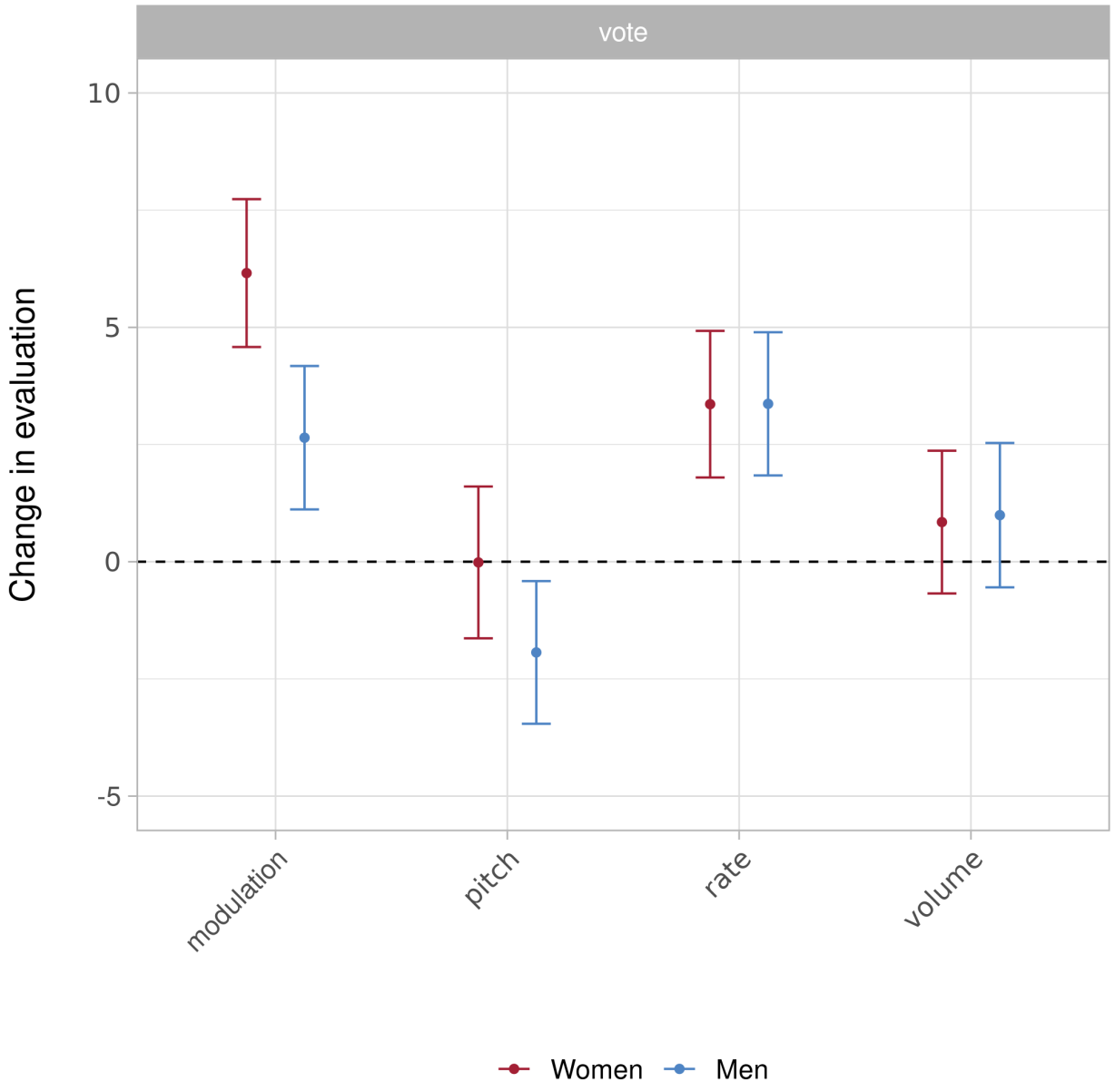


Figure 4: Effect of speech features on expressed willingness to vote for voice actor. Modulation and speech rate have relatively large effects. Appendix Section H.1 presents these results in tables.

7 Discussion and Conclusion

In this paper, We develop a general framework for drawing causal inferences about non-textual channels of speech communication, which we applied to the campaign-speech context in two experiments evaluating the role of vocal style in voter evaluation. We find strong evidence that candidate voice shapes voter perceptions of candidate attributes and their willingness to vote for candidates.

While prior work has shown 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. Beyond this, we also find suggestive evidence that the benefit of skillful communication may be larger for women than for men, perhaps driven by a greater penalty imposed on women candidates who do not communicate in a rhetorically skillful manner. However, we note that we are only able to draw limited inferences about these gendered effects, because our study relies on only ten voice actors who each implement these concepts in their own way and because gender heterogeneity was not a primary focus of this study. Future work may assess these differences more credibly identify them, either by hiring many more actors or leveraging recent developments in artificial intelligence to generate many voices without human actors (Barari et al., 2021).

Our causal framework also highlights several additional directions for future methodological research. While it allows for learning dynamics that shape a voter's evaluation gradually across the course of an election, in this paper, we assumed for simplicity that learning was 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 vocal style, particularly the effects of several

specific auditory interventions. Experimental designs evaluating a broader range of auditory effects, as well as the effect of new visual manipulations, represent another promising avenue for future research. Each of these methodological extensions suggest directions for building on our substantive analyses of campaign speech style—that is, how candidates can better communicate with voters to shape perceptions and gain support, without distorting their actual policy content of their speech.

8 Data Availability Statement

All audio and video recordings, transcriptions, survey files, and analysis code will be made available as an interactive, reproducible CodeOcean capsule upon acceptance.

9 Funding Statement

Dean Knox and Christopher Lucas gratefully acknowledge financial support through the National Science Foundation (award #2120087 through the Methodology, Measurement, and Statistics Program).

10 Conflict of Interest Statement

The authors declare no conflicts of interest.

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.

Allison, Thomas H, Benjamin J Warnick, Blakley C Davis, and Melissa S Cardon. "Can you hear me now? Engendering passion and preparedness perceptions with vocal expressions in crowdfunding pitches". *Journal of Business Venturing*, vol. 37, no. 3, 2022, p. 106193.

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.

Anderson, Rindy C and Casey A Klofstad. "Preference for leaders with masculine voices holds in the case of feminine leadership roles". *PloS one*, vol. 7, no. 12, 2012, e51216.

Aristotle. *Rhetoric*. Translated by George A. Kennedy, circa 330 BCE, Oxford UP, 1991.

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.

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.

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.

Bucholtz, Mary. "Captured on tape: Professional hearing and competing entextualizations in the criminal justice system". 2009.

Carli, Linda L. "Gender, language, and influence." *Journal of personality and social psychology*, vol. 59, no. 5, 1990, p. 941.

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-Mohliver, Aharon, Anantha Krishna Divakaruni, and Laura Fritsch. "Financial Analysts Penalize Female CEOs for Female-Type Speech". *Academy of Management Proceedings*, Academy of Management Briarcliff Manor, NY 10510, 2023, p. 18972.

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, 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.

Eaves, Michael and Dale G Leathers. *Successful nonverbal communication: Principles and applications*. Routledge, 2017.

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.

Elias-Bursac, Ellen. *Translating evidence and interpreting testimony at a war crimes tribunal: Working in a tug-of-war*. Springer, 2015.

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.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. *Text as data: A new framework for machine learning and the social sciences*. Princeton UP, 2022.

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.

Ji, Li-Jun, Zhiyong Zhang, and Richard E Nisbett. "Is it culture or is it language? Examination of language effects in cross-cultural research on categorization." *Journal of personality and social psychology*, vol. 87, no. 1, 2004, p. 57.

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.

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.

Krahé, Barbara and Lida Papakonstantinou. "Speaking like a man: Women's pitch as a cue for gender stereotyping". *Sex Roles*, vol. 82, no. 1, 2020, pp. 94–101.

Krauss, Robert M, Yihsiu Chen, and Purnima Chawla. "Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us?" *Advances in experimental social psychology*, vol. 28, Elsevier, 1996, pp. 389–450.

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.

Osnabrügge, Moritz, Sara B Hobolt, and Toni Rodon. "Playing to the gallery: Emotive rhetoric in parliaments". *American Political Science Review*, vol. 115, no. 3, 2021, pp. 885–899.

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.

Rittmann, Oliver. "Legislators' emotional engagement with women's issues: Gendered patterns of vocal pitch in the German Bundestag". *British Journal of Political Science*, 2023, pp. 1–9.

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. 101–115.

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.

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.

Wagner, Petra, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview. *Speech communication*, vol. 57, 2014, pp. 209–232.

Xu, Wei, Xiaohan Zhang, Runyu Chen, and Zhan Yang. "How do you say it matters? A multimodal analytics framework for product return prediction in live streaming e-commerce". *Decision Support Systems*, vol. 172, 2023, p. 113984.

Appendix (Online Publication Only)

Table of Contents

A Vocal Style Depends on Speech Topic

As noted in Section 2, our corpus consists of 100 video-recorded speeches obtained from ElectAd, a nonpartisan website, each corresponding to a campaign event in the 2012 U.S. presidential election. Of these, 38 are speeches by Barack Obama, and the remaining 62 are by Mitt Romney. Most occurred in the three months before election day. We removed introductions, concluding music, and other material to obtain single-speaker recordings.

We obtained timestamped transcripts of each speech using the Google Speech transcription API, which segments audio files into utterances (roughly, sentences) and provides the start and end time of each utterance. Within each utterance, we use the `communication` `R` package to compute time-series speech volume (measured in decibels, dB) and pitch (measured in Hertz, Hz). We exclude volume from analysis during silences, such as inter-word pauses, by setting it to `NA`; similarly, we exclude pitch during unvoiced speech such as sibilants and plosives, where the quantity is undefined.

We then aggregate these vocal characteristics to the level of the utterance as follows. We first compute within-utterance mean speech volume and mean vocal pitch. We then

quantify the amount of vocal modulation by taking the within-utterance standard deviation of pitch and volume. Finally, we compute the within-utterance average first derivative of pitch, which is positive for rising tones (typical of questions, e.g. "yes?") and negative for falling tones (typical of emphatic statements, e.g. "yes!"), then take its average value. To assess the baseline vocal style of each candidate, we then aggregate these utterance-level auditory features to the level of the speaker. Figure 13 shows the results, which reveal substantial differences in vocal style between Obama and Romney. In particular, Obama's speech exhibits considerably greater variation in within-utterance pitch and volume, and he utilizes greater emphasis—consistent with popular accounts that characterize him as a dynamic public speaker. (Note that cross-speaker differences in mean volume should be interpreted with caution, as it is heavily influenced by preprocessing techniques such as audio normalization that may differ across campaigns.)

Next, to assess whether speakers vary their vocal style based on speech topic, we used the Lexicoder policy-agenda dictionary (Albaugh et al., 2013), which provides a list of words associated with civil rights, crime, culture, defense, the economy, education, energy, the environment, finance, healthcare, labor, religion, social welfare, technology, and transportation. Utterances were coded as related to a topic if their stemmed transcripts contained any of the topic's keywords. Finally, for each speaker, we conducted ten linear regressions with the utterance-level datasets: one for each of the five vocal characteristic and two speakers. In this analysis, each row represents one utterance, outcomes are auditory features, and regressors consisted of binary topic indicators along with speech (campaign event) fixed effects. Standard errors were clustered at the level of the speech, or campaign event.

Figures 11 and 12 show the results, which reveal how Obama and Romney respectively varied their vocal styles depending on the topic. Points plotted in red are significant after multiple testing correction (Benjamini and Hochberg, 1995). Results show that Obama uses rhetorical flourishes to draw attention to issues of religion and the economy while speaking less emphatically when discussing national defense. Romney is similarly emphatic when discussing economic topics, as well as the environment and energy policy, and is considerably less emphatic when discussing technology, education and defense.

B 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 C for the complete text of the matches that we selected for use in the experiment).

B.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 $\underset{K \times L}{\boldsymbol{P}}$, and target i is $\underset{J_i \times L}{\boldsymbol{T}_i}$. 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

$$\mathcal{S}(\boldsymbol{P}, \boldsymbol{Q}) = \frac{\sum_{k=1}^{K} \sum_{\ell=1}^{L} \tilde{p}_{k,\ell} \; \tilde{q}_{k,\ell}}{||\tilde{\boldsymbol{P}}||_F \; ||\tilde{\boldsymbol{Q}}||_F} \tag{4}$$

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 $\sum_{\ell=1}^{L} \mathrm{Cov}\,(P_\ell, Q_\ell)$, where P_ℓ is the sequence of counts for letter ℓ. Thus, when $L = 1$, Equation 4 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 $\mathcal{S}(\boldsymbol{P}, \boldsymbol{P}) = 1$.

B.2 The Algorithm

Approximate string search involves examining all target documents i and candidate offsets j within each document. Figure 5 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 $\big[\mathcal{S}(\boldsymbol{P}, \boldsymbol{T}_{i,1:K}), \ldots, \mathcal{S}(\boldsymbol{P}, \boldsymbol{T}_{i,(J_i-K+1):J_i})\big]$. 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 4.

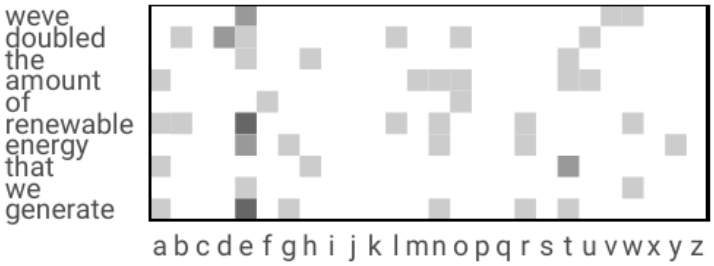
First, observe that $||\tilde{\boldsymbol{T}}_{i,1:K}||_F^2$ is the grand sum of a row subset of $[\tilde{t}_{i,j}^2]$. Corresponding values must be computed at every offset in document i to produce the sequence $\Big[||\tilde{\boldsymbol{T}}_{i,1:K}||_F, \ldots, ||\tilde{\boldsymbol{T}}_{i,(J_i-K+1):J_i}||_F$ 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.

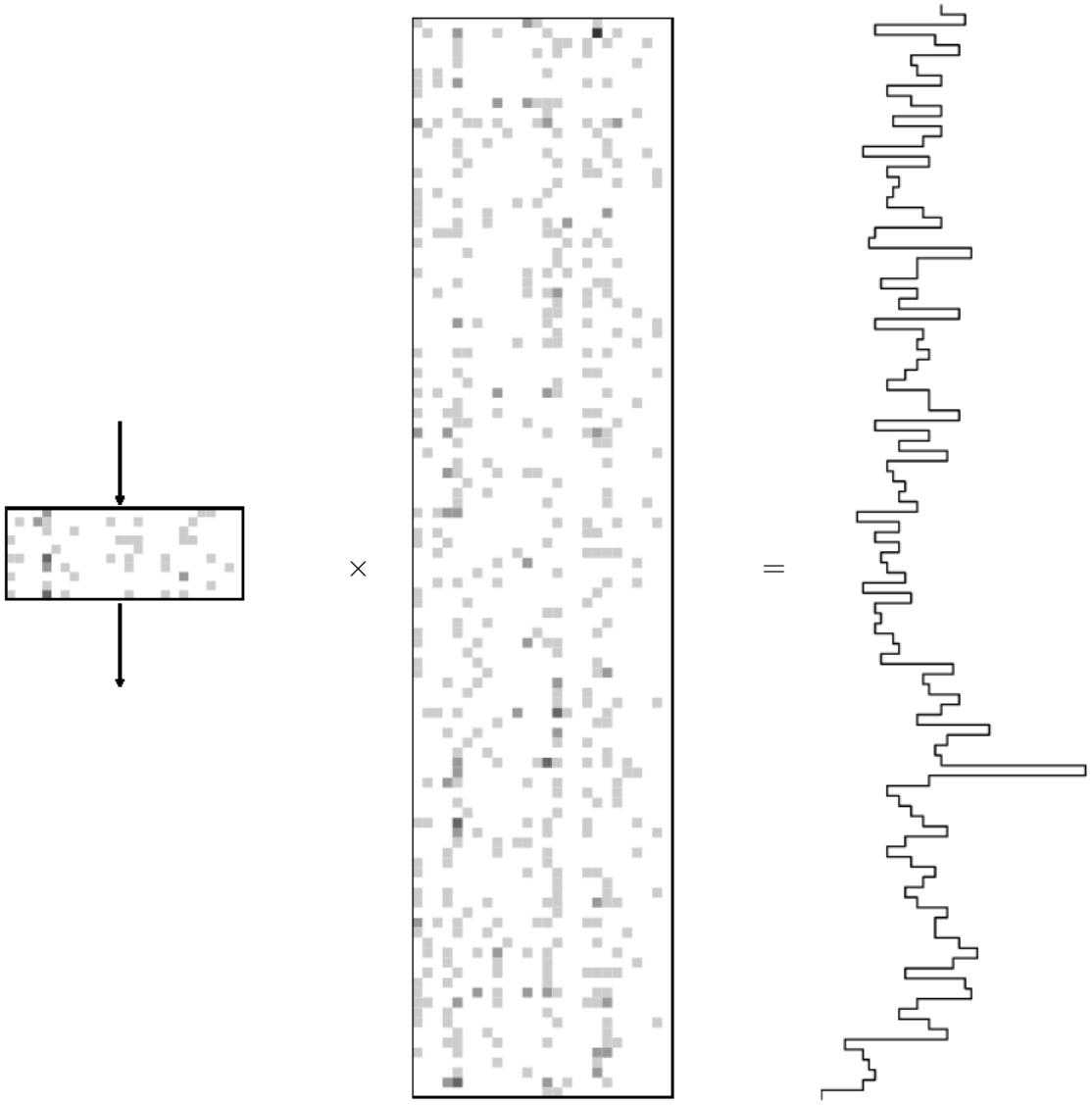
Next, we observe that the numerator, $\sum_{k=1}^{K}\sum_{\ell=1}^{L} \tilde{p}_{k,\ell}\ \tilde{t}_{i,j+k-1,\ell}$, can be rewritten as $\sum_{k=1}^{K}\sum_{\ell=1}^{L} p_{k,\ell}\ t_{i,j+k-1,\ell} - \sum_{k=1}^{K}\sum_{\ell=1}^{L} \bar{p}_{\ell}\ \bar{t}_{i,j,\ell}$, where $\bar{p}_\ell$ is the mean 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 $\underset{J_i \times L}{\bar{\boldsymbol{T}}_i} = [\bar{t}_{i,j,\ell}]$, 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 5: **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.





C 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.

D Display of Experiment 1

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

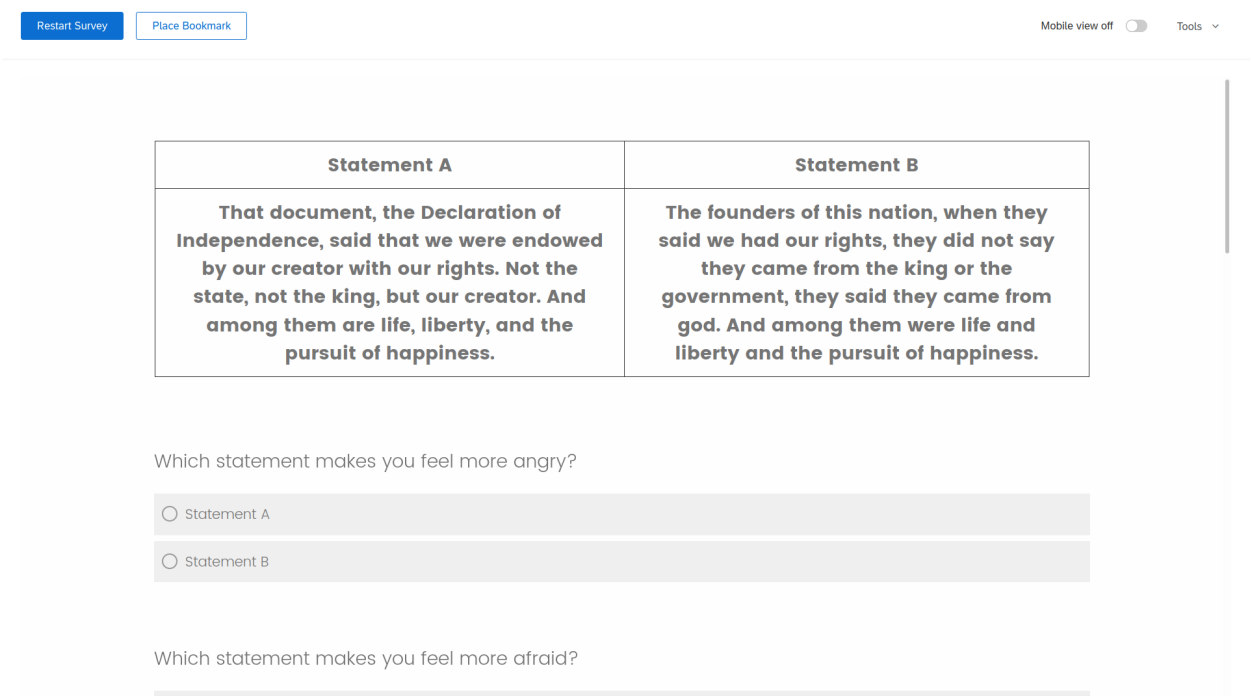


Figure 6: Display of the text condition in Experiment 1.

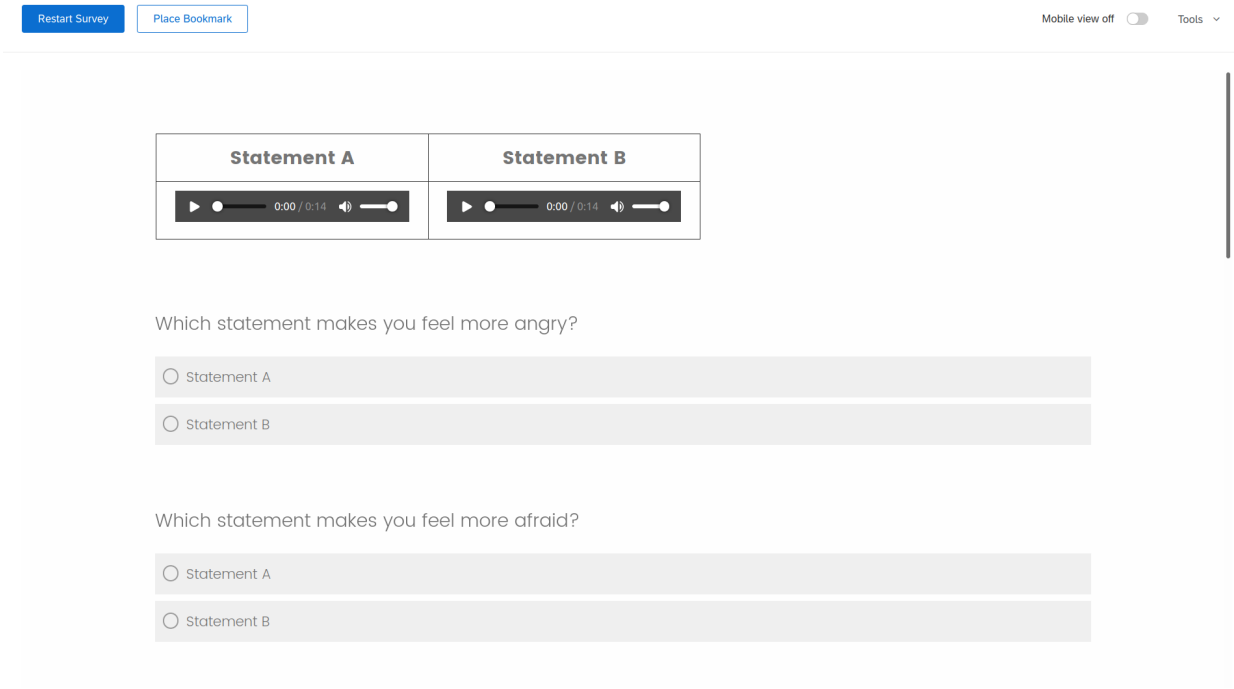


Figure 7: Display of the audio condition in Experiment 1.

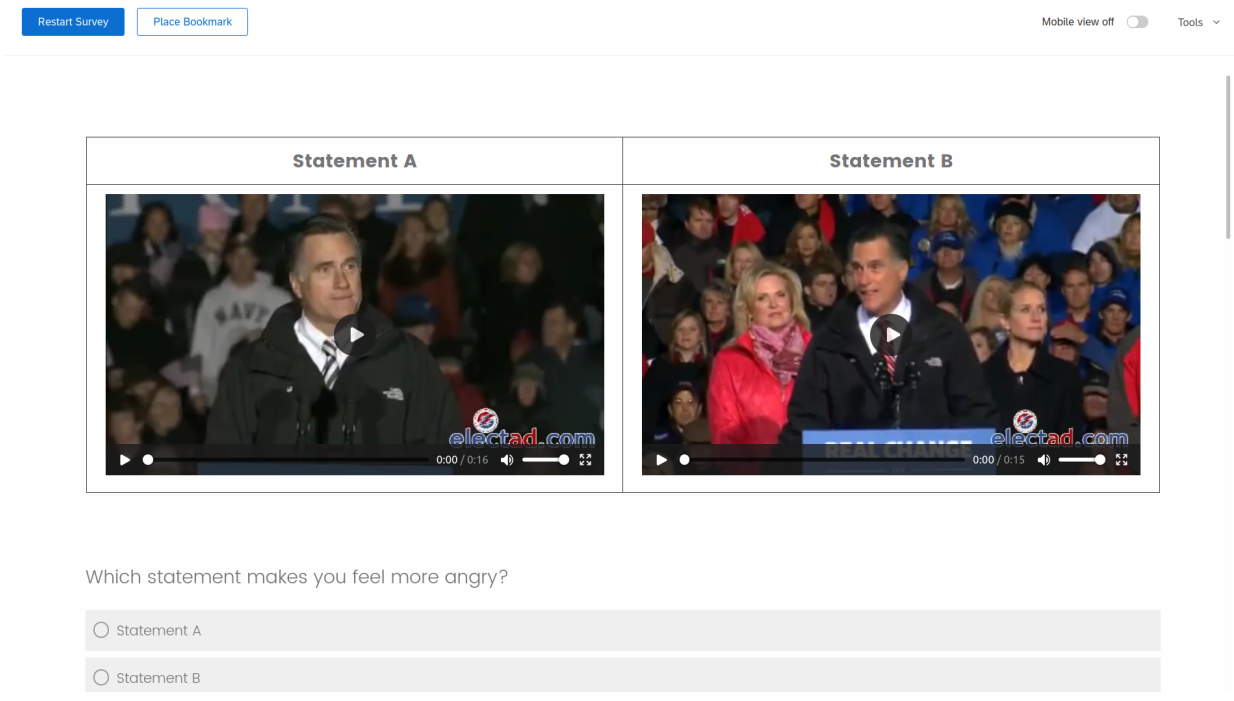


Figure 8: Display of the video condition in Experiment 1.

E 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.

F 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 $\times$ high/low volume $\times$ 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 9 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/monotonous 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.

Finally, it is possible the the differential results by gender are at least partly explained by absolute differences in the actor manipulations (speech rate and modulation. To determine if this is the case, we split the estimates visualized in Figure 9 out by the gender of the speaker. The results suggest that this may in fact drive the observed gender differences, at least in part. It is possible that when women are asked to modulate speech or speak at a faster rate, they do so differently than men, but it is more likely that these differences are the result of having relatively few speakers, and that these differences are due to idiosyncratic differences resulting from the small sample of speakers (10 total, five men and five women).

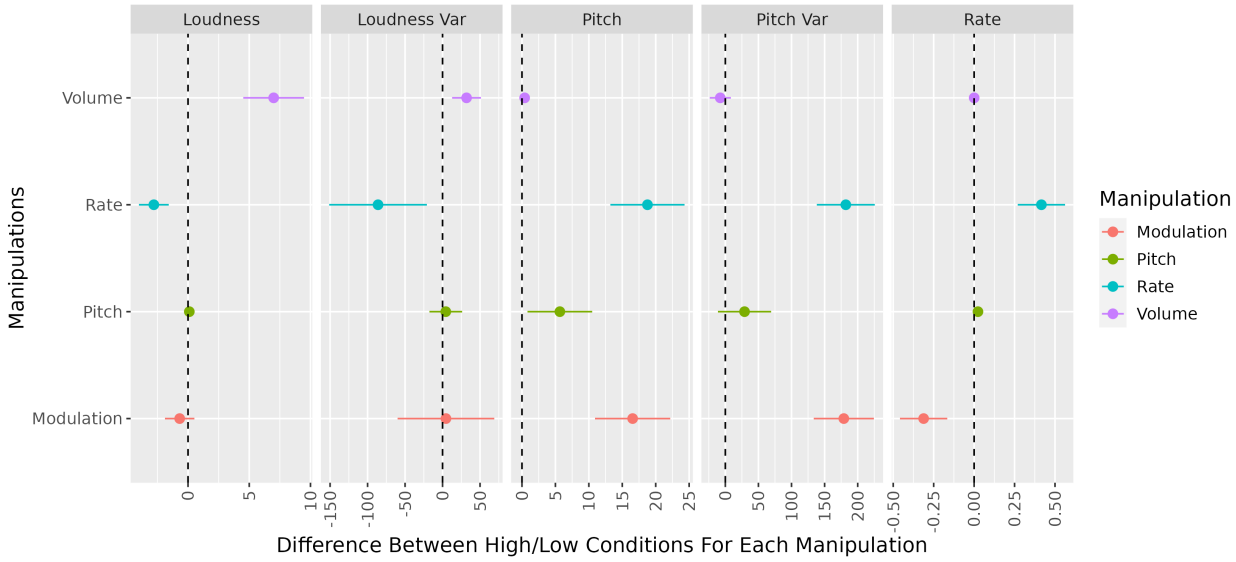


Figure 9: Comparison of manipulations used in Experiment 2 across five summary features. Of the four manipulations, two were controlled by actors recording different versions of each script (rate and modulation), while the other two were implemented by computationally manipulating all actor-produced 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 F discusses this in greater detail.

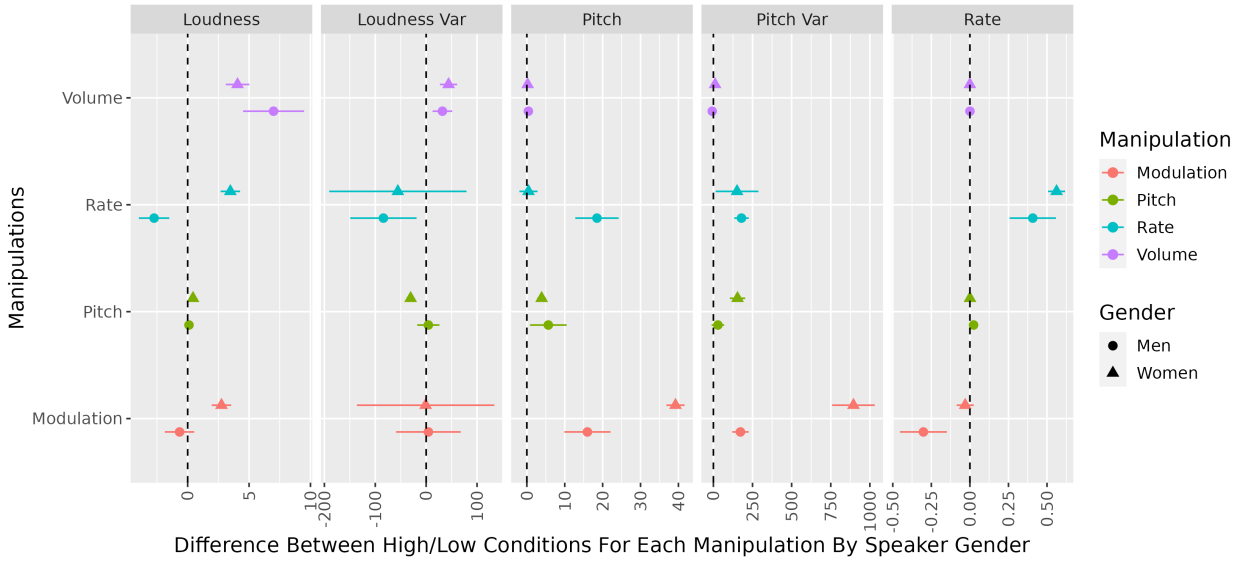


Figure 10: Comparison of manipulations used in Experiment 2 across five summary features, split by the gender of the actor. Of the four manipulations, two were controlled by actors recording different versions of each script (rate and modulation), while the other two were implemented by computationally 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 F discusses this in greater detail.

G Supplementary Figures

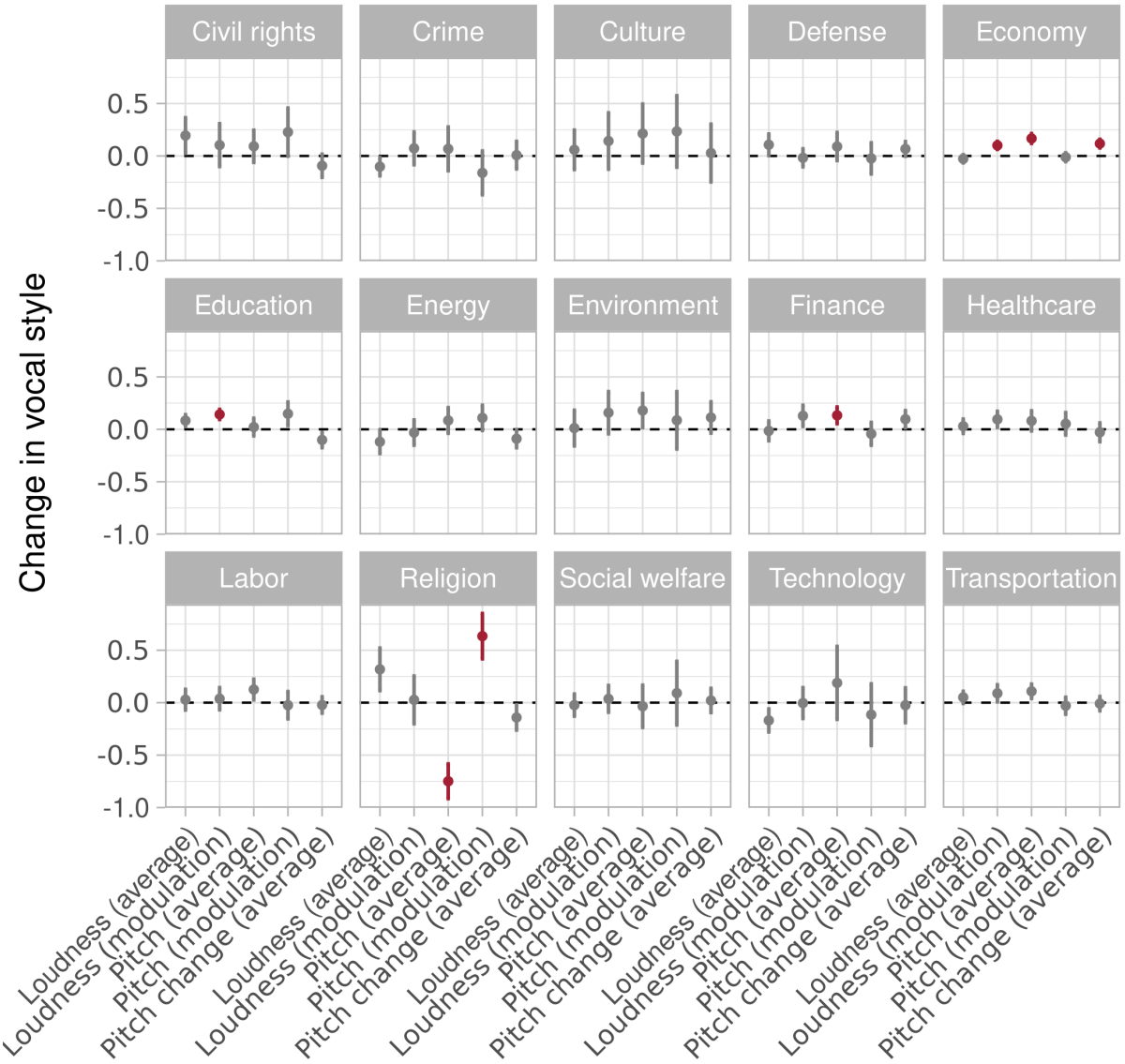


Figure 11: Change in vocal style by Obama, conditional on topic of speech. The vertical position of each point represents the average deviation from a speaker's baseline when speaking about a topic (measured in standard deviations to facilitate comparisons across elements of vocal style). Error bars represent 95% confidence intervals. Red estimates are those which remain significant after a multiple testing correction. Table 9 in the appendix presents these results in tabular form.

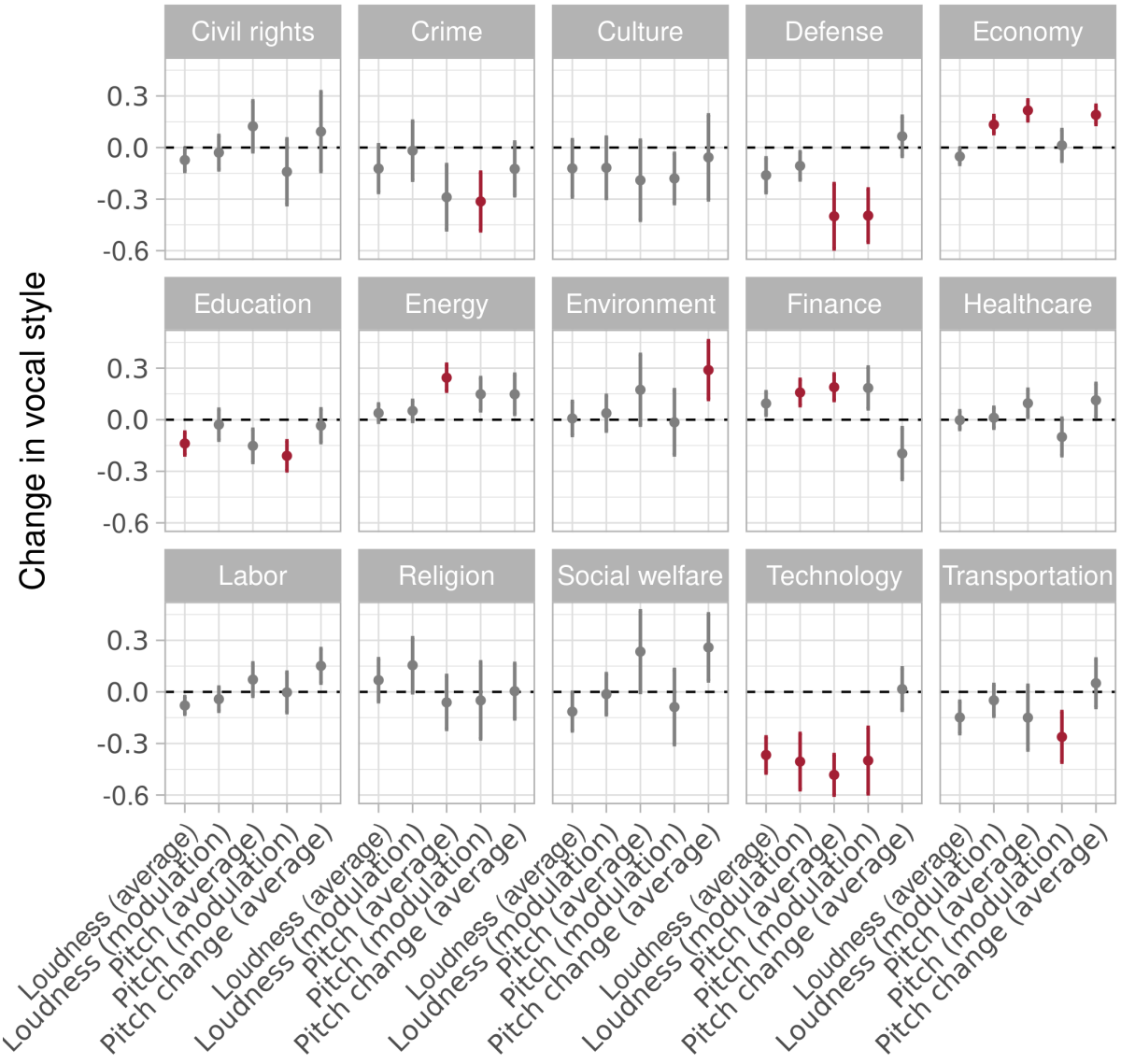


Figure 12: Change in vocal style by Romney conditional on the topic of speech. The vertical position of each point represents the average deviation from a speaker's baseline when speaking about a topic (measured in standard deviations to facilitate comparisons across elements of vocal style). Error bars represent 95% confidence intervals. Red estimates are those which remain significant after a multiple testing correction. Table 10 in the appendix presents these results in tabular form.

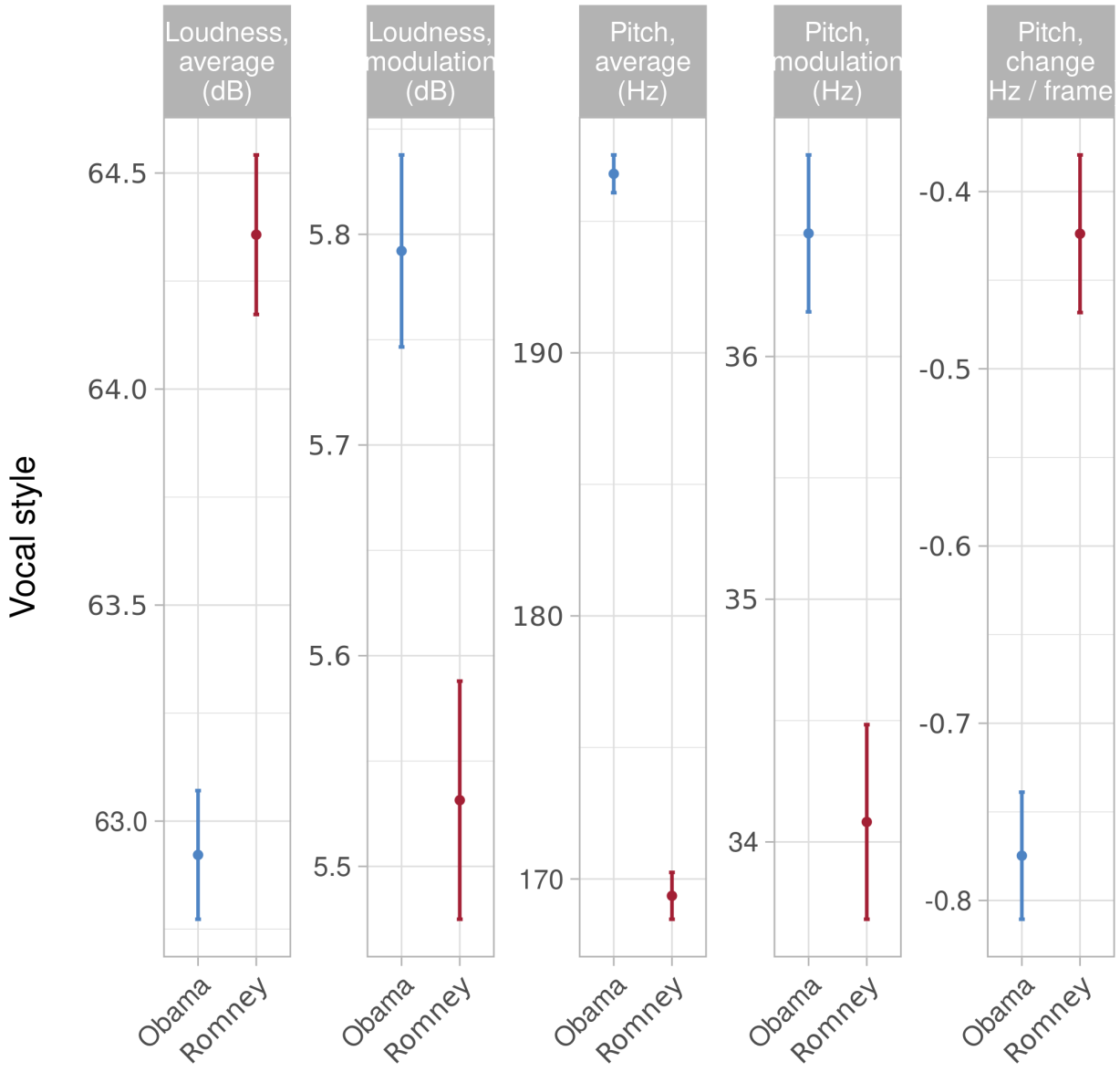


Figure 13: Comparison of campaign speech by Obama and Romney on common speech audio features. On average, Obama displays considerably more variation in loudness and pitch, consistent with popular accounts of Obama being a talented public speaker (Fleishman, 2017).

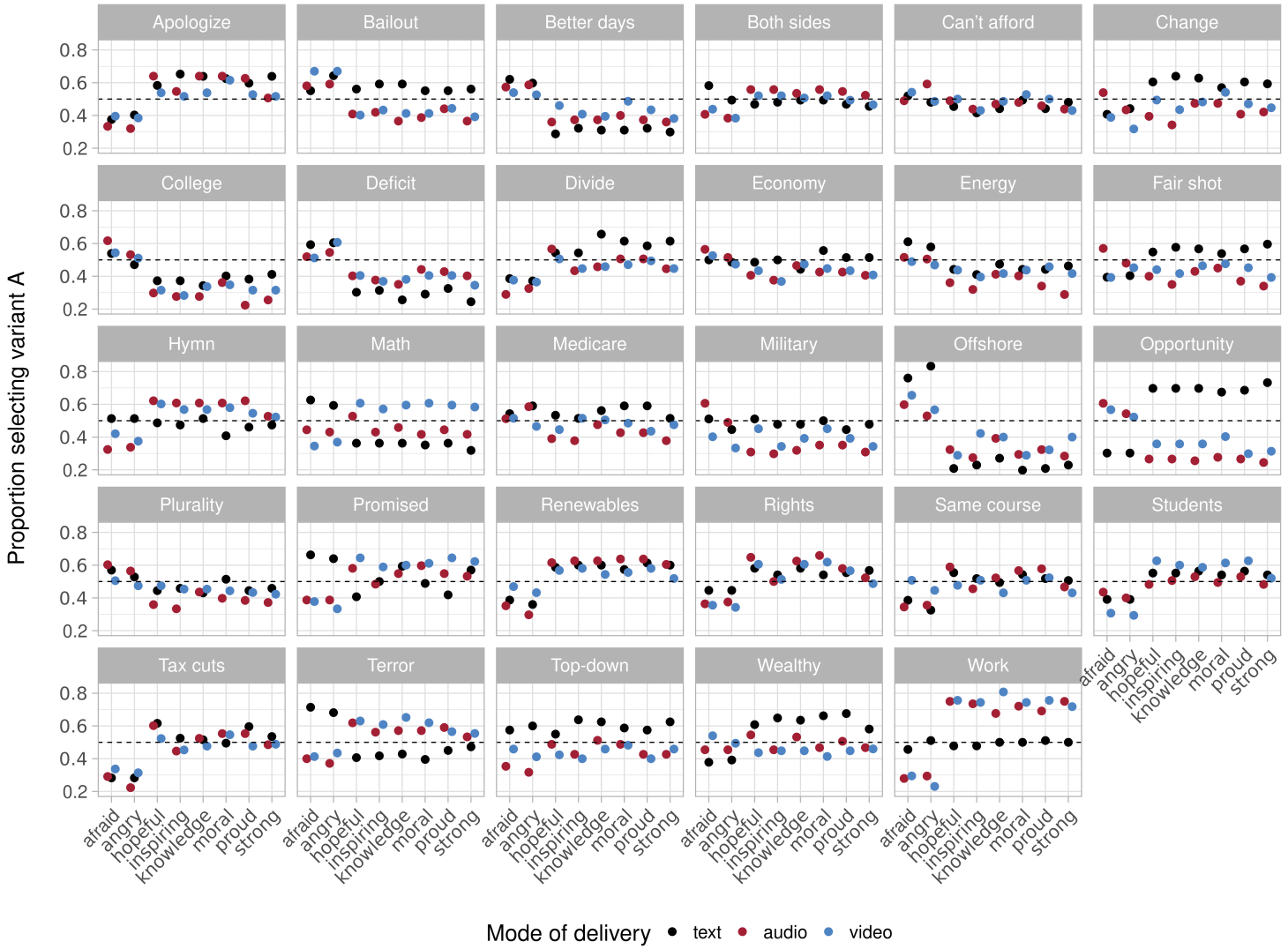


Figure 14: Each panel plots the proportion of subjects selecting variant A of a matched text pair over variant B. Within the pair, assignment of a variant to be A or B is arbitrary, so there are no directional expectations. Each panel in the plot shows the proportion of subjects selecting variant A over 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.

H Supplementary Tables

	Outcome				
Variable	Loudness (avg)	Loudness (mod)	Pitch (avg)	Pitch (mod)	Pitch (change)
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 11, 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 12, 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 2.

H.1 Tabular Representation of Figures 5 and 6

In this section, we present in tabular form estimates presented visually in plots 3 and 4. 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.