A Dynamic Model of Speech for the Social Sciences*

Dean Knox† and Christopher Lucas‡

November 22, 2019

Abstract

Speech and dialogue are the heart of politics: Nearly every political institution in the world involves verbal communication. Yet vast literatures on political communication focus almost exclusively on *what* words were spoken, entirely ignoring *how* they were delivered—auditory cues that convey emotion, signal positions, and establish reputation. We develop a model that opens this untapped information to principled statistical inquiry: the model of speech and audio structure (MASS). Our approach models political speech as a stochastic process shaped by fixed and time-varying covariates, including the history of the conversation itself. In an application to Supreme Court oral arguments, we demonstrate how vocal delivery signals crucial information—skepticism of legal arguments—that is indecipherable to text models. Results show that justices do not use questioning to strategically manipulate their peers, but rather engage in genuine fact-finding efforts. Our easy-to-use `R` package, `speech`, implements the model and many more tools for audio analysis.

*For excellent research assistance, we thank Taylor Damann. For helpful comments, we thank Justin de Benedictis-Kessner, Josh Boston, Bryce Dietrich, JB Duck-Mayr, Seth Hill, Luke Keele, Gary King, Connor Huff, In Song Kim, Adeline Lo, Jacob Montgomery, Jonathan Mummolo, David Romney, Jake Shapiro, Brandon Stewart, Dustin Tingley, Michelle Torres, Ariel White, Teppei Yamamoto, and Xiang Zhou, as well as participants at the Harvard Applied Statistics Workshop, the International Methods Colloquium, the Texas Political Methodology Conference, and the Washington University in St. Louis Political Data Science Lab. Dean Knox acknowledges financial support from the National Science Foundation (Graduate Research Fellowship under Grant No. 1122374) and the Microsoft Research Computational Social Science postdoctoral researcher program. Christopher Lucas gratefully acknowledges his Dean's support for this project.

†Assistant Professor, Princeton University, Fisher Hall, Princeton, NJ 08544; http://www.dcknox.com/

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

What we say can sometimes... be challenging for the sake of eliciting a response, sometimes it can be genuine doubt about what the position of a person might be, sometimes we're talking to each other and we're raising points through the questions that we want our colleagues to consider with us... there's lots of reasons for what we're doing, but none of them are ever perfectly understood... among the people who are listening.

Sotomayer (2019)

1 Introduction

Speech and dialogue are at the heart of politics. Candidates appeal to voters through rhetoric, legislators contest policy during floor debates, courts probe legal argumentation with questioning, and citizens discuss all of these with friends and family. Indeed, nearly every political institution in the world, from small-town council meetings to the United Nations General Assembly, involves verbal communication. Yet quantitative political science has only just begun to analyze political speech—and when it does, analysts have done so with the tools of text analysis. While fruitful, these methods entirely ignore the way in which words are spoken.

This unexploited channel of communication, *vocal tone*, is a rich data stream. In Section 2, we explain how tone is used by listeners to assess speakers' types and mental states. When politicians express themselves skeptically, compassionately, or fervently, their speech conveys more than the mere words that are spoken. Among other information, tone conveys

beliefs, preferences, and the intensity with which they are held—or at least, the impressions of these things that the speaker hopes to leave with their audience. The role of voice in persuading others—be they colleagues, adversaries, or electorates—to do what they would not otherwise do has been known at least since Aristotle's *On Rhetoric*. In this paper, we introduce a new method that not only allows researchers to measure vocal tone in high-dimensional audio data, but also study how it is used in political interactions.

Given the ubiquity of verbal communication in political institutions, why are models of speech only now being introduced? In this paper, we identify and resolve three challenges to the widespread analysis of speech audio. First, speech audio recordings do not naturally arrive in a format that is easy to analyze statistically. In Section 3, we demonstrate how audio recordings of speech can be represented numerically, drawing on insights from signal processing and linguistics. Second, the theoretically interesting quantities signaled by human speech—which may range from basic emotions (e.g. anger) to abstract concepts (decisiveness), depending on substantive domain—are latent variables that are not directly observed in audio data but must be inferred. In Section 4, we address this by developing a semi-supervised model for human speech, which infers latent tones of voice and provides a principled measure of uncertainty for the estimated patterns in their use. Third, speech is inherently context-dependent—theories may suggest that candidates' vocal tone should vary depending on their position in the polls or the topic at hand—and dynamic, with back-and-forth interaction between speakers. In dialogues like political debates or media interviews, an interruption by one person may be met with a polite but firm response, whereas a different interjection could produce an angry retort. We refer to these and other temporal patterns in the choice of tone as the *flow of speech*. Interdependence is often treated as a nuisance

or ignored entirely, but we show that speech flow is a valuable source of information about human interaction. By directly modeling this phenomenon, new substantive questions about dynamic interactions are opened to principled statistical inquiry.

In Section 6, we demonstrate that speech tone and flow matter in substantive political science by resolving an ongoing debate in the study of the Supreme Court. Judicial scholars have long disagreed over models of justice behavior in deliberation; some argue that justices are shrewd political actors that maneuver to influence the decisions of their peers, while others hold that justices cast sincere votes and engage in genuine expression according to their respective legal philosophies. We measure and extensively validate a key auditory quantity of interest that is unmeasurable with only the text of speech: skepticism, an individual's vocal expression of their judgment about an assertion or argument. Existing theories suggest diverging observable implications for the expected flow of questioning—when one justice should express skepticism, and how other subsequent justices should respond in their own speech—which we use to construct previously infeasible empirical tests of judicial behavior. We find that estimated flow is highly consistent with a theory of genuine fact-finding, and it undermines the competing account of strategic manipulation. Finally, we introduce fast, easy-to-use software to facilitate the study of "audio as data" and conclude with general guidelines for applied researchers analyzing the many political domains in which audio data is available.

2 The Importance of Audio

It is well-known that vocal delivery plays an important role in human communication, above and beyond a speaker's choice of words. In this section, we review evidence that listeners extract considerable information from vocal delivery, drawing on numerous studies in linguistics and psychology. We then identify a wide range of literatures in political science that currently study rich speech recordings using textual methods alone, ignoring the auditory component. Taken together, these points suggest that political science is currently missing a great deal by discarding the raw source audio. We briefly note a handful of the many possible applications in which our speech model can help test existing and new theories about political deliberation.

2.1 Audio Contains Exclusive Information

Tens of thousands of studies have decisively established the importance of nontextual cues in human speech.[1] Among other well-replicated results, they have shown that respondents' perceptions—including dehumanization of those with opposing political views (Schroeder and Epley 2016)—differ dramatically when exposed to audio recordings of speech, as opposed to transcripts alone. Countless experiments in linguistics and psychology have explored specific auditory mechanisms through which a speaker's voice shapes the information gleaned by listeners on numerous dimensions. Speech audio has been shown to convey information about speakers' static type (e.g., power), as well as their time-varying characteristics (e.g., emotional state). We briefly overview this literature.

[1] A Google Scholar search for "paralinguistic" returned over 64,000 results in late 2019.

2.1.1 Signals of a Speaker's Type

Much is inferred about a speaker simply from the general sound of their voice. Humans use vocal tone to draw conclusions about competence, education, and trustworthiness, among many other attributes (Zuckerman and Driver 1989; Anderson et al. 2014). In the political domain, both observational and experimental research has shown that voters's preferences are affected by the sound of a candidate's voice (Gregory Jr and Gallagher 2002; Surawski and Ossoff 2006; Klofstad, Anderson, and Peters 2012; Tigue et al. 2012; Podesva et al. 2015; Klofstad 2016). This effect is mediated, at least in part, by the belief that certain speaking styles are associated with strong, dominant leaders who are more likely to prevail in conflict (Laustsen, Petersen, and Klofstad 2015; Klofstad, Anderson, and Nowicki 2015). These patterns appear to generalize well across a range of political roles (Anderson and Klofstad 2012), building on well-established findings in psychology that nontextual elements of human conversation shift perceptions of speakers' social power (Scherer, London, and Wolf 1973; Carney, Hall, and LeBeau 2005; Hodges-Simeon, Gaulin, and Puts 2010), credibility, persuasiveness (Apple, Streeter, and Krauss 1979; Burgoon, Birk, and Pfau 1990; Hamilton and Stewart 1993), and intelligence (Brown, Strong, and Rencher 1974; Smith et al. 1975; Schroeder and Epley 2015). These results suggest that the way in which individuals speak reveals, at a minimum, how they wish to be perceived by their audiences. For example, Touati (1993) found that a politician's pitch contour was more variable before an election and monotonic after, suggesting a pre-election desire to signal attributes like competence that are associated with this vocal delivery.

In political science, research on debates has shown the way that politicians express them-

selves matters. Hinck and Hinck (2002) argues that audiences scrutinize rhetoric for "politeness strategies" which are used to "measure a candidate's good will and judgment regarding his image, his opponent's, and the values he aspires to represent for the community." Evidence from real-time voter reactions suggests over-aggressive attacks can lead to immediate drops in ratings (McKinney, Kaid, and Robertson 2001). The instantaneous effect of utterances depends on factors like audience composition (Boydstun et al. 2014), implying politicians may benefit by presenting themselves differently depending on context. In this vein, scholars such as Benoit, Blaney, and Pier (1998) have advanced contextualized theories of rhetoric indicating when candidates should attack, defend, or self-acclaim across a campaign—not only in debates but also in speeches and advertisements; Brader (2006) provides evidence that politicians selectively target emotional appeals based on context-specific expected benefits. Still others have posited that conversation flow plays a particularly important role in voter evaluations. Discussing debates, town halls, and talk show appearances, Beck (1996) asserts that to be perceived as presidential, "candidates must project themselves as able to cope and excel in spite of conflict over issues and despite their opponents' attempts to define them," and that success "hinges on their interactional skills, not just their ability to look a certain way or say particular things." Nor does the role of rhetoric diminish after election: A broad literature on "going public" (Kernell 2006) explores how presidents increasingly make televised appeals directly to the public. However, work on political rhetoric has largely relied on qualitative methods or labor-intensive human coding due to the difficulty of empirical analysis—a gap that our method directly addresses.

2.1.2 Signals of a Speaker's Current State

Speech signals much more than "presidentialness" or other time-invariant characteristics. Temporal variation in vocal tone also indicates the mental state—such as emotions or impressions (Scherer, Koivumaki, and Rosenthal 1972; Kappas, Hess, and Scherer 1991; Scherer 1995; Banse and Scherer 1996; Johnstone and Scherer 2000)—that a speaker currently wishes to convey. Some of this variation has subconscious biological origins. Certain emotional states produce physiological effects, such as mouth dryness, accelerated breathing, muscular tension, or tremors (Ohala 1981), which have audible effects on speech. (Naturally, practiced speakers like actors and politicians can also emulate these.[2]) Vocal mannerisms also convey a speaker's current impressions of a conversation, such as their level of certainty, understanding, agreement, and belief (Manusov and Trees 2002). These vocal cues form an important part of the dynamic interplay between speakers in a conversation (Leathers 1979).

In Section 6, we build on this work by studying one particular time-varying speech tone, Supreme Court justices' expressed skepticism. Skepticism is an important signal—perhaps genuine—of disagreement with or incredulity about assertions and legal arguments in the context of oral arguments. Patterns in the flow of conversation shed light on how justices deliberate, just as patterns of matter-of-fact questioning or righteous indignation in campaign debates and impeachment hearings can help reveal the nature of interactions in these political arenas. We show that different theoretical accounts of deliberation imply diverging temporal patterns in speech flow, allowing us to construct an empirical test for competing models of

[2]Much research on emotion in speech relies on actors (Scherer 2003)—a logical impossibility if emotional expression in speech was strictly subconcious.

Supreme Court behavior.

Relatedly, speech audio can affect a listener's trust in the speaker (Zuckerman et al. 1979; Schirmer et al. 2019), specifically the listener's suspicion that the speaker is lying (Zuckerman, DePaulo, and Rosenthal 1981; Manstead, Wagner, and MacDonald 1984). To our knowledge, no work has directly tested the effect of vocal tone on perceived deception by politicians. Given the importance of credibility in conflict (Fearon 1994; Guisinger and Smith 2002) and trust in elections (Hetherington 1999; Levi and Stoker 2000), this relationship warrants future investigation. However, research has firmly established that these time-varying signals play an important role elsewhere in politics. For example, Dietrich, Enos, and Sen (2019) demonstrated that the average pitch of Supreme Court justice questions is predictive of their subsequent votes, even after controlling for text, important legal covariates, and justice ideology (Martin and Quinn 2002). Similarly, the average pitch of a legislator's floor speech has also been used to proxy for issue commitment (Dietrich, Hayes, and O'Brien 2019) and shown to predict coverage on cable news (Dietrich, Schultz, and Jaquith 2018).

2.2 Political Science Already Studies Audio

Audio recordings of speech are thus indisputably a rich source of data. But how often are these recordings available in contexts of interest to political scientists? We now demonstrate that the answer is "quite often." In fact, a wide range of research has *already* studied audio recordings of political speech—but in virtually every domain, researchers have done so by extracting transcripts, then discarding the remainder of their data.

Sections 5–6 consider one such domain, judicial speech on the Supreme Court. Here,

published research has focused almost exclusively on the text of oral arguments (Ringsmuth, Bryan, and Johnson 2013; Black, Sorenson, and Johnson 2013; Kaufman, Kraft, and Sen 2018).[3] For example, Black et al. 2011 examine how justices question parties in oral arguments, showing that text-based measures of affective questioning can signal voting decisions. (In a direct comparison, Section 5 demonstrates that a comparable audio-based measure outperforms this prior work by three times, using its own evaluation task.)

However, the Supreme Court is hardly the only context in which political scientists are already studying speech. Countless other studies have examined political debates (Hart and Jarvis 1997; Bayley 2004; Thomas, Pang, and Lee 2006; Fridkin et al. 2007; Conway III et al. 2012; Benoit 2013), campaign advertisements (Spiliotes and Vavreck 2002; Fridkin and Kenney 2011; Carlson and Montgomery 2017), campaign speech (Laver, Benoit, and Garry 2003; Bligh et al. 2010; Olson et al. 2012; Schroedel et al. 2013; Degani 2015), legislative speech (Slapin and Proksch 2008; Quinn et al. 2010; Proksch and Slapin 2012, 2015; Herzog and Benoit 2015; Lauderdale and Herzog 2016; Schwarz, Traber, and Benoit 2017), television news (Behr and Iyengar 1985; Mermin 1997; Semetko and Valkenburg 2000; Oegema and Kleinnijenhuis 2000; Sanders and Gavin 2004; Young and Soroka 2012; Rozenas and Stukal 2019), talk radio (Hofstetter et al. 1999; Sobieraj and Berry 2011; Conroy-Krutz and Moehler 2015; Ross 2016), and political addresses (Cohen 1995; Ritter and Howell 2001; Young and Perkins 2005; Rule, Cointet, and Bearman 2015).

Each of these studies has used text analysis to study a political context in which communication was *not* written and read as text, but rather was spoken and heard as audio.

[3] A notable exception is Dietrich, Enos, and Sen (2019), discussed above

Given the relative youth of text analysis methods, it is perhaps surprising how often recorded speech is analyzed in this way. The mismatch between data and methods results in the inevitable loss of nontextual information, suggesting that richer models have the potential to contribute to research in each of these domains.

3 Audio as Data

The number of papers developing and applying methods for text analysis has increased rapidly in recent years (Wilkerson and Casas 2017), and workflows for preprocessing raw text are well-developed (Grimmer and Stewart 2013; Lucas et al. 2015; Benoit et al. 2018). However, little effort has been devoted to the analysis of other components of communication—like audio—"as data." In this section, we now explain how unstructured audio recordings of human speech can similarly be preprocessed into structured data in preparation for statistical analysis.

The primary unit of analysis in speech is the *utterance*: a continuous, single-speaker segment, typically concluding with a clear pause. The length of utterances are unequal but typically on the order of 10 seconds. Within each utterance, we split the recording into successive *moments*, or extremely short windows of time.[4] In each moment, the raw audio is then summarized with auditory features that are known to convey emotion and tone of voice, drawing on established literatures in psychology, phonetics, signal processing, and computer science (Ververidis and Kotropoulos 2006; El Ayadi, Kamel, and Karray 2011).

[4]We use windows that are 25 milliseconds in length, each overlapping the latter half of the previous moment (or "frame") to preserve information occurring at cutoffs.

Researchers can easily calculate these features with a single function in our accompanying `R` package, `speech`, which also implements other useful preprocessing steps. (For instance, our package includes functionality for segmenting utterances or trimming interruptions.) We briefly describe audio featurization to provide intuition.

Within each moment, the raw audio recording consists of a *waveform*, or univariate high-frequency time-series of pressure measurements. We characterize this short recording with auditory features describing the sound perceived by listeners. Selected features are illustrated in Figure 1 by describing an audio source for which that feature is high or low. For example, some features are simple functions of the raw waveform. One feature is the "zero crossing rate" (ZCR), or simply how often the waveform crosses zero (neutral pressure). Sibilants (e.g. /s/, /z/) have a particularly high ZCR, whereas vowels have low ZCR. Other auditory features, like "energy" (loudness), help distinguished /z/ (which is "voiced", i.e. involving vocal cord vibrations) from voiceless /s/.

Others features are based on the audio *spectrum*, computed via Fourier transform of the waveform. This captures (among other things) the contribution of the baritone or soprano ranges to overall loudness. Other undepicted features that can be computed from this representation include pitch (dominant frequency, or spectral peak) and Mel-frequency cepstral coefficients (MFCCs, describing overall shape). These features provide additional emotional information. For example, English words may be emphasized with higher sustained pitch or with sharp falling pitch. (Compare "We're *citing* that case?" to "We're citing *that* case!") Pitch is also higher when speakers are vocally tense, including when speakers are emotionally aroused. Online Appendix 2 documents all features used in analysis. In general, no single auditory feature distinguishes all sounds or vocal tones; to address this challenge, we develop

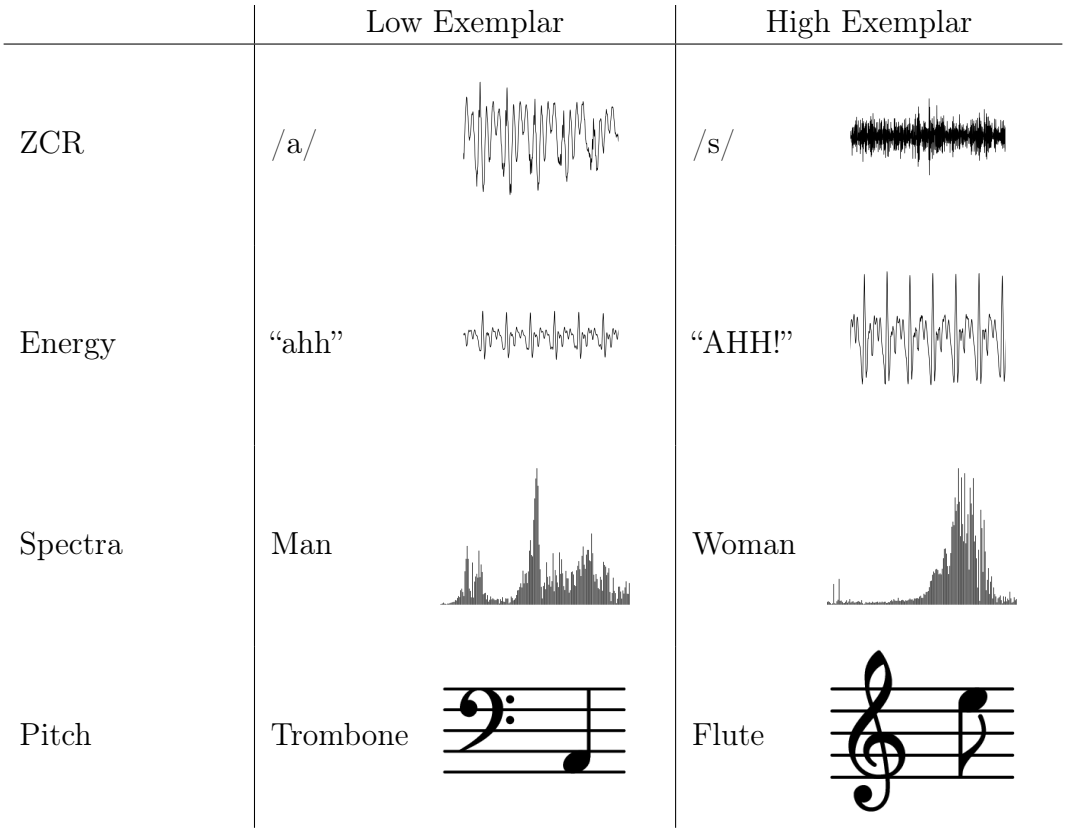

Figure 1: **Illustration of selected audio features.** The left column identifies a class of audio summaries that are used to represent the raw audio data. Subsequent columns contain textual descriptions and graphical depictions of audio sources for which the relevant feature has relatively low and high values. For example, ZCR (zero-crossing rate) has a low value for the vowel /a/ and a high value for the consonant /s/. ZCR and energy graphs depict pressure waveforms from machine-synthesized speech recordings; louder sounds are larger in amplitude and hissing sounds are higher in ZCR. Spectral graphs represent the Fourier transform of synthesized recordings; the female voices are concentrated in higher frequency ranges. Pitch is an example of a feature that can be derived from the spectral peaks.

a method that can exploit dozens or even hundreds of features in analyzing speech.

3.1 Advances Over Existing Approaches

Outside of political science, a large and interdisciplinary body of research has sought to model or classify human speech. A common approach is to collapse each utterance into a vector of descriptive statistics (e.g., mean and standard deviation of pitch), which can be

used in standard machine-learning classifiers (Dellaert, Polzin, and Waibel 1996; McGilloway et al. 2000). However, these reduced representations discard enormous amounts of auditory data. To avoid this loss of information, hidden Markov models are often used to model time-series speech. The discrete latent states in this model map well to actual human speech, which is often represented in terms of discrete phonemes, or minimal units of sound.[5]

MASS builds on these existing approaches in statistics and computer science in two main ways. First, computational constraints mean that typical HMM-based analyses are able to incorporate only a fraction of the features we incorporate. Nogueiras et al. (2001), for example, use just two features per moment, while Kwon et al. (2003) use 13 features and Mower et al. (2009) use only the MFCC coefficients. More importantly, MASS is the first to directly model the flow of speech—that is, contextual and temporal patterns in vocal tone—in terms of the meaningful structural features encoded in conversation metadata.

4 A Model of Conversation Dynamics

We now develop a generative statistical model for the audio data that arises from political speech. Section 4.1 formally defines the assumed generative model. After outlining the model, we next turn to estimation and inference in Section 4.2, then discuss practical considerations in the modeling of speech.

[5]For example, the International Phonetic Alphabet identifies 107 discrete phonemes that are grouped into broader families like "fricatives" and "plosives."

4.1 The Model

Suppose we have a conversation with U sequential utterances, each of which arises from one of M modes of speech. (To keep the exposition clear, here we consider the simplified setting in which a single conversation is analyzed. Online Appendix 1 presents the general multi-conversation model, which is essentially identical.) Let S_u denote the speech mode, or tone, of utterance u. Broadly, the model contains two levels. The "upper" level, defined in Equations 1–2, characterizes the flow of speech, or the choice of S_u. We assume that the conversation unfolds as a time-varying stochastic process in which S_u is chosen based on the conversational context at that moment, encoded in the vector $\boldsymbol{W}_u$. In the "lower" level, we then outline a generative model for utterance u's auditory characteristics, $\boldsymbol{X}_u$. Importantly, this generative model will differ depending on the selected tone of voice selected in the upper model, S_u. Equations 3–4 present the lower level more formally. The model is summarized graphically in Figure 2, and we refer readers there for a holistic view of how the various model components fit together.

We begin by modeling speech mode probabilities in each utterance as a multinomial logistic function of conversation context, $\boldsymbol{W}_u$. Note that $\boldsymbol{W}_u$ may include functions of conversation history, such as aggregate anger expressed by previous speakers over the course of an argument—that is, $\sum_{u'<u} 1(\boldsymbol{S}_{u'} = \text{anger})$—which might induce a sharp retort in utterance u.

$$\Delta_{u,m} = \exp\left(\boldsymbol{W}_u^\top \boldsymbol{\zeta}_m\right) / \sum_{m'=1}^{M} \exp\left(\boldsymbol{W}_u^\top \boldsymbol{\zeta}_{m'}\right) \tag{1}$$

$$S_u \sim \text{Cat}(\boldsymbol{\Delta}_u). \tag{2}$$

where $\boldsymbol{\Delta}_u = [\Delta_{u,1}, \ldots, \Delta_{u,M}]$ and $\boldsymbol{\zeta}_m$ is a mode-specific coefficient vector through which $\boldsymbol{W}_u$ affects the relative prevalence of mode m. The tone of utterance u, S_u, is one of the primary quantities of interest, along with the coefficients $\boldsymbol{\zeta}$ that explain why certain tones are used more in particular contexts. However, generally speaking, tone is not directly observable to the analyst; the utterance's auditory characteristics, $\boldsymbol{X}_u$, is the only available information. (As we discuss in Section 4.2, the analyst will begin estimation by obtaining a sample of utterances with human-labeled tone.)

Each $\boldsymbol{X}_u$ is a matrix describing the auditory characteristics of utterance u. In this matrix, the t-th row describes the audio at moment t in terms of D auditory features. Thus, the utterance audio is represented by a $T_u \times D$ feature matrix, where T_u is the length of the utterance; because utterances may be of differing lengths, $\boldsymbol{X}_u$ and $\boldsymbol{X}_{u'}$ may have differing numbers of rows.

To model the audio, we then assume that the m-th mode of speech is associated with its own Gaussian hidden Markov model (HMM) that produces the auditory features as follows. At moment t in utterance u, the speaker enunciates the sound $R_{u,t}$—that is, the latent state, which may represent phonemes or phoneme groups such as plosives or fricatives. In successive moments, the speaker alternates through these latent sounds according to

$$(R_{u,t} \mid S_u = m) \sim \text{Cat}(\boldsymbol{\Gamma}^m_{R_{u,t-1},*}), \tag{3}$$

with $\boldsymbol{\Gamma}^m_{k,*}$ denoting rows of the transition matrix, $[\Pr(R_{u,t} = 1|R_{u,t-1} = k), \ldots, \Pr(R_{u,t} = K|R_{u,t-1} = k)]$. By modeling the usage patterns of different sounds in this way, we approximately capture the temporal structure that plays an important role in speech. (For example, most latent

sounds are sustained for at least a few moments, and certain phonemes typically occur before the silence at the end of a word.) In turn, latent sound k is associated with its own auditory profile, which we operationalize as a multivariate Gaussian distribution with parameters $\boldsymbol{\mu}^{m,k}$ and $\boldsymbol{\Sigma}^{m,k}$. Finally, the raw audio heard at moment t of utterance u—the signal perceived by a listener—is drawn as

$$\boldsymbol{X}_{u,t} \sim \mathcal{N}(\boldsymbol{\mu}^{S_u,R_{u,t}}, \boldsymbol{\Sigma}^{S_u,R_{u,t}}), \tag{4}$$

which completes the model. Thus, each mode of speech is represented with a rich and flexible HMM that nevertheless reflects much of the known structure of human speech. It is the differences in usage patterns and sound profiles—the Gaussian HMM parameters—that enable human listeners to distinguish one tone or speaker from another.

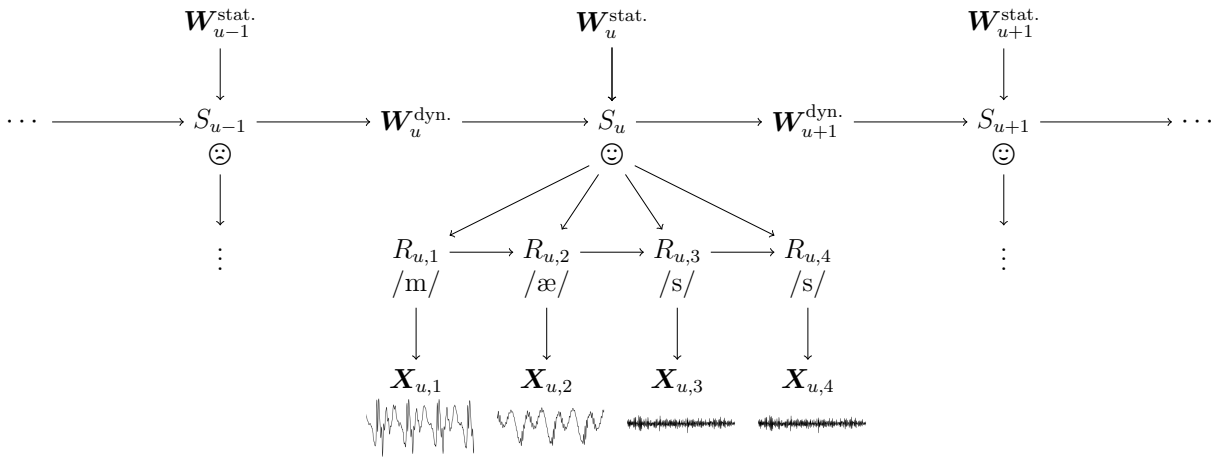

Figure 2: **Illustration of generative model.** The directed acyclic graph represents the relationships encoded in Equations 1–4. In utterance u, the speaker selects tone S_u based on "static" (i.e., externally given) time-varying covariates $\boldsymbol{W}_u^{\text{stat.}}$ as well as "dynamic" conversation history covariates $\boldsymbol{W}_u^{\text{dyn.}}$. (In the illustration, $\boldsymbol{W}_u^{\text{dyn.}}$ depends only on the prior mode of speech, but more complex dynamic covariates can be constructed.) Based on the selected tone, the speaker composes an utterance by cycling through a sequence of sounds in successive moments, $R_{u,1}, R_{u,2}, \ldots$, to form the word "mass." Each sound generates the audio perceived by a listener according to its unique profile; $\boldsymbol{X}_{u,t}$ is extracted from this audio.

4.2 Estimation

We describe a procedure for estimating the model defined in Section 4.1, incorporating elements of both unsupervised and supervised learning. The researcher begins by determining the speech modes of interest, then identifying and labeling example utterances from each class. Within this training set—which might not be a subset of the primary corpus of interest—we consider each mode of speech in turn, using a fully unsupervised approach to learn the auditory profile and cadence of that speech mode. The results are applied to the full corpus to obtain "naïve" estimates of each utterance's tone, based only on the audio features and ignoring conversational context. We then fit a model for the flow of conversation, use this to refine the "contextualized" tone estimates, and repeat in an iterative procedure. The specifics of each step are discussed below and in Online Appendix 1, and the workflow is

	Static metadata	Conversation history	Speech mode	Audio features
Primary corpus	$\boxed{\boldsymbol{W}^{\text{stat.},\mathcal{C}}}$	$\boldsymbol{W}^{\text{dyn.},\mathcal{C}}$	$\boldsymbol{S}^{\mathcal{C}}$	$\boxed{\boldsymbol{X}^{\mathcal{C}}}$
Training utterances	$\boldsymbol{W}^{\text{stat.},\mathcal{T}}$	$\boldsymbol{W}^{\text{dyn.},\mathcal{T}}$	$\boxed{\boldsymbol{S}^{\mathcal{T}}}$	$\boxed{\boldsymbol{X}^{\mathcal{T}}}$

Table 1: **Observed and Unobserved Quantities.** Data that is (un)available to the analyst are (un)boxed. Attributes of the primary corpus (training set) are indicated with $\mathcal{C}$ ($\mathcal{T}$) superscripts. Raw audio features, $\boldsymbol{X}$, are observed for all utterances. The portion of the conversational context that relates to static metadata ($\boldsymbol{W}^{\text{stat.}}$) is available for at least the primary corpus, but dynamic contextual variables that depend on the tone of prior utterances ($\boldsymbol{W}^{\text{dyn.}}$) can only be estimated. In general, the tone of each utterance ($\boldsymbol{S}$) is also unobserved, but the analyst possesses a small training set with human-labeled utterances.

outlined more formally in Algorithm 1.

Table 1 summarizes the data available for the primary corpus and training set, respectively indicated with $\mathcal{C}$ and $\mathcal{T}$. The audio characteristics of each utterance, $\boldsymbol{X}$, are observed for both the primary corpus and the training set. However, human-labeled tone of speech, $\boldsymbol{S}$, is only known for the training set. We divide the conversational context into externally given but potentially time-varying "static metadata," $\boldsymbol{W}^{\text{stat.}}$, and deterministic functions of conversation history that dynamically capture the prior tones of speech, $\boldsymbol{W}^{\text{dyn.}}$. The former is known for the primary corpus but may be unavailable for the training set, depending on how it is constructed; the latter is not directly observed for either.

Our ultimate goal is to estimate the conversation flow parameters, $\boldsymbol{\zeta}$, and the auditory parameters of each tone, which we gather in $\boldsymbol{\Theta}^m = (\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m, \boldsymbol{\Gamma}^m)$ for compactness. In what follows, we also refer to the collection of all tone parameters as $\boldsymbol{\Theta} = (\boldsymbol{\Theta}^m)_{m\in\{1,\ldots,M\}}$. Under

the model described in Equations 1–2, the likelihood can be expressed as

$$\mathcal{L}(\boldsymbol{\zeta}, \boldsymbol{\Theta} \mid \boldsymbol{X}^{\mathcal{T}}, \boldsymbol{S}^{\mathcal{T}}, \boldsymbol{X}^{\mathcal{C}}, \boldsymbol{W}^{\text{stat.},\mathcal{C}}) \;=\; f(\boldsymbol{X}^{\mathcal{C}} \mid \boldsymbol{\zeta}, \boldsymbol{\Theta}, \boldsymbol{W}^{\text{stat.},\mathcal{C}})\; f(\boldsymbol{X}^{\mathcal{T}} \mid \boldsymbol{\Theta}, \boldsymbol{S}^{\mathcal{T}}), \tag{5}$$

with one factor depending only on the primary corpus and another containing only the training set.

As a concession to computational constraints, we estimate the parameters in a stagewise fashion. The auditory parameters, $\boldsymbol{\Theta}$, are calculated by maximizing the partial likelihood, $f(\boldsymbol{X}^{\mathcal{T}} \mid \boldsymbol{\Theta}, \boldsymbol{S}^{\mathcal{T}})$, corresponding to the training factor, rather than the full likelihood in Equation 5 (Wong 1986). The full likelihood is then maximized with respect to the conversation flow parameters $\boldsymbol{\zeta}$, conditional on $\boldsymbol{\Theta}$. The factorization and a detailed discussion of stagewise estimation are presented in Online Appendix 1.1.

In Online Appendix 1.2, we detail our procedure for estimating the auditory profile and cadence for each speech mode. First, training utterances are divided according to their tone labels. Because the partial likelihood can be neatly factorized further as $f(\boldsymbol{X}^{\mathcal{T}} \mid \boldsymbol{\Theta}, \boldsymbol{S}^{\mathcal{T}}) = \prod_{m=1}^{M} \prod_{u \in \mathcal{T}} f(\boldsymbol{X}_u | \boldsymbol{\Theta}^m)^{\mathbf{1}(S_u = m)}$, $\hat{\boldsymbol{\Theta}}^m$ can be independently estimated for each speech mode with no further loss of information. For all training utterances of speech mode m, a regularized variant of the Baum-Welch algorithm, a standard estimation procedure for hidden Markov models, is used to obtain $\hat{\boldsymbol{\Theta}}^m$ for the corresponding mode. Each of the resulting M tone-specific models are then applied to each utterance u in the primary corpus to obtain the corrected emission probabilities $f(\boldsymbol{x}_u | \hat{\boldsymbol{\Theta}}^m, S_u = m)^\rho$, which represents the probability that the utterance's audio was generated by speech mode m; this captures the extent to which the audio "sounds like" the relevant training examples. Naïve tone estimates can then

be computed by combining these with the overall prevalence of each tone via Bayes' rule. The corrective factor, ρ, approximately accounts for unmodeled autocorrelation in the audio features and ensures that the naïve estimates are well-calibrated (for details, see Online Appendix 1.3). This shared correction, along with the number of latent sounds and strength of regularization, are determined by likelihood-based cross-validation (van der Laan, Dudoit, Keles, et al. 2004) in the training set.

We now briefly describe an expectation-maximization algorithm for the conversation-flow parameters, $\boldsymbol{\zeta}$, reserving derivations and other details for Online Appendix 1.4.[6] An inspection of Equation 5 shows that this procedure will depend only on $f(\boldsymbol{X}^{\mathcal{C}} \mid \boldsymbol{\zeta}, \boldsymbol{\Theta}, \boldsymbol{W}^{\text{stat.},c})$, since the remaining term does not involve $\boldsymbol{\zeta}$. We proceed by augmenting the observed data with the latent tones, $\boldsymbol{S}^{\mathcal{C}}$, and the conversation-history variables that depend on them, $\boldsymbol{W}^{\text{dyn.},\mathcal{C}}$. The augmented likelihood, $f(\boldsymbol{X}^{\mathcal{C}}, \boldsymbol{S}^{\mathcal{C}}, \boldsymbol{W}^{\text{dyn.},\mathcal{C}} \mid \boldsymbol{\zeta}, \boldsymbol{\Theta}, \boldsymbol{W}^{\text{stat.},\mathcal{C}})$, is then iteratively optimized. However, the closed-form expectation of the augmented likelihood is intractable. We therefore replace the full E step with a blockwise procedure that sweeps through the unobserved E-step variables sequentially.[7] Finally, the maximization step for $\boldsymbol{\zeta}$ reduces to a weighted multinomial logistic regression in which $\mathbf{1}(S_u = m)$ is fit on $\mathbb{E}[\boldsymbol{W}_u | S_{u-1} = m']$ for every possible m and m', with weights corresponding to the probability of that transition.

[6]This estimation procedure builds on forward-backward algorithms; interested readers are referred to Online Appendix 1.2 or standard references such as Zucchini and MacDonald (2009).

[7]The use of this alternative procedure leads to a smaller improvement of the EM objective function than a full E-step. Nevertheless, algorithms using such partial E- or M-steps ultimately converge to a local maximum, just as traditional expectation-maximization procedures do (Neal and Hinton 1998).

Finally, we observe that the unmodeled autocorrelation discussed above renders model-based inference invalid. To address this issue, we estimate the variance of parameter estimates by bootstrapping utterances in the training set, ensuring that dependence between successive moments in an utterance do not undermine our results (Online Appendix 1.5 discusses potential issues in bootstrapping). The full estimation procedure is outlined in Algorithm 1. Other quantities of interest, such as those discussed in Section 6, follow directly from the conversation-flow parameters, $\boldsymbol{\zeta}$, or the auditory parameters, $\boldsymbol{\Theta}$; inference on these quantities follows a similar bootstrap approach.

Data: Audio features $(\boldsymbol{X}^{\mathcal{C}}, \boldsymbol{X}^{\mathcal{T}})$, static metadata for primary corpus $(\boldsymbol{W}^{\text{stat.},\mathcal{C}})$

Result: Auditory parameters $\boldsymbol{\Theta}$, conversation flow parameters $\boldsymbol{\zeta}$

Procedure:

1. Define problem.
Analyst determines tones of interest and rubric for human coding.
Human-coded tone labels are obtained for training set $(\boldsymbol{S}^{\mathcal{T}})$.

2. Fit auditory parameters $(\boldsymbol{\Theta})$ by maximizing partial likelihood on training set $(\mathcal{T})$

```
for speech mode m in 1, ..., M do
    Subset to training utterances labeled as tone m.
    while not converged do
        for utterance u in T and moment t in {1, ..., T_u} do
            for sound k in 1, ..., K do
                Compute emission probability of sound (m, k) generating audio (X_{u,t}).
            end
        end
        Predict sound being pronounced at each moment (R_{u,t}).
        Update cadence (usage patterns of constituent sounds, Γ^m).
        for sound k in 1, ..., K do
            Update audio profile of sound k (μ^{m,k}, Σ^{m,k}).
        end
    end
end
```

3. Fit conversation-flow parameters $(\boldsymbol{\zeta})$ using primary corpus $(\mathcal{C})$, conditional on $\boldsymbol{\Theta}$

```
for utterance u in C do
    for speech mode m in 1, ..., M do
        Compute corrected emission probability of speech mode m generating utterance
          audio data (X_u), ignoring context.
    end
end
while not converged do
    Predict expected mode of speech for each utterance (S_u).
    Compute expected conversation context for each utterance (W_u).
    Update flow-of-speech parameters (ζ).
end
```

Algorithm 1: Stagewise estimation procedure. After defining the tones of interest and obtaining a labeled training set, the analyst conducts cross-validation to set ancillary parameters such as the number of assumed sounds in each mode of speech (not depicted). After fixing the ancillary parameters, the cadence and auditory characteristics of each speech mode are estimated from the training set by an iterative expectation-maximization procedure. These speech parameters are then fixed, and the relationship between conversation context and flow of speech is estimated from the primary corpus. In the multiple-conversation case, the utterance loop in step 3 is nested within an outer loop over conversations. Statistical inference is conducted by resampling $\mathcal{T}$ and repeating steps 2–3 within the bootstrapped training set (not depicted) to obtain bootstrap-aggregated point estimates and bootstrap variance estimates for flow-of-speech parameters and other quantities of interest.

5 A New Quantity of Interest in Judicial Behavior

In this section, we introduce an original corpus of Supreme Court oral argument audio recordings scraped from the Oyez Project (Cornell 2015)[8] and develop a new quantity of theoretical interest: *judicial skepticism.* In this section, we first describe the data. The concept of skepticism is then illustrated with a detailed case study. Finally, we extensively validate the model and compare it to related work analyzing the text of oral arguments.

5.1 Audio Data from Supreme Court Oral Arguments

We limit our analysis to the natural court that begins with the appointment of Justice Kagan and concludes with the passing of Justice Scalia, so that the composition of the Court remains constant for the entirety of the period we analyze. The Oyez data contains an accompanying textual transcript, speaker names for each utterance, and timestamps for utterance start and stop times. In addition, we inferred the target side (i.e., petitioner or respondent) of each justice's question based on the side of the most recently speaking lawyer. Additional case data was merged from the Supreme Court Database (Spaeth et al. 2014)

Using Oyez timestamps, we segmented the full-argument audio into a series of single-speaker utterances.[9] As an additional preprocessing step, we drop utterances spoken by lawyers (each of whom usually appears in only a handful of cases) and Clarence Thomas

[8]Dietrich, Enos, and Sen (2019) independently collected the same audio data and conducted an analysis of vocal pitch.

[9]Occasionally, segments are reported to have negative duration, due to errors in the original timestamps. In these cases, we drop the full "turn," or uninterrupted sequence of consecutive utterances by this speaker.

(who spoke only twice in our corpus), focusing on the behavior of the eight recurrent speakers. We also drop procedural statements, along with utterances shorter than 2.5 seconds.[10] After trimming, the resulting audio corpus contains 407 arguments and 153 hours of audio, comprising over 66,000 justice utterances and 44 million frames.

5.2 The Quantity of Interest: Judicial Skepticism

In this section and the next, we introduce and employ a new measure of substantive importance to the study of courts: *judicial skepticism*, an individual's vocal expression of their judgment about the argument at hand. Judicial skepticism is an important signal of disagreement with or incredulity about assertions and legal arguments, especially in the context of oral arguments.

To identify judicial skepticism in speech, we first randomly selected a training set of 200 utterances per justice to hand-classify as "skeptical" or "neutral" speech, allowing our assessments to reflect not only the vocal tone but also the textual content of the utterance. Thus, we define 16 modes of speech—two tones for each of the eight speaking justices.[11]

[10]We found that these extremely short utterances contained substantial amounts of crosstalk. However, they also include potentially informative interjections; future work may explore improved preprocessing techniques that do not require discarding this information.

[11]Because the transcripts attribute each utterance to a speaker, the model's decision is over whether the current statement by Anthony Kennedy was skeptical or neutral. That is, we do not conduct a joint speaker-recognition and tone-detection task. In the framework outlined in Equations 1–2, this is equivalent to introducing a covariate for the current speaker's identity, with a corresponding coefficient of $-\infty$ for the 14 speech modes that do not correspond to the current speaker.

During classification, we dropped the handful of utterances (roughly 5%) in which crosstalk or other audio anomalies occurred, or in rare instances where the speaker's identity was incorrectly recorded. The model is then estimated following Algorithm 1.

5.3 A Case Study of Judicial Skepticism

To illustrate the use of skepticism during the flow of oral arguments, we conducted a case study of *Alabama Legislative Black Caucus v. Alabama*, a racial gerrymandering case heard by the Supreme Court in 2014 which considered the legality of Alabama's 2012 redistricting efforts. The study is described in depth in Online Appendix Section 3; we briefly summarize it here and demonstrate the application of MASS to this case in Figure 3.

As background, the case arose when the Republican-led legislature redrew electoral districts in the face of declining urban population. In doing so, the legislature sought to pack black voters into a small number of already heavily Democratic districts. The Alabama Legislative Black Caucus argued that this practice violated the Voting Rights Act (VRA), a position ultimately supported by the Court's decision, whereas defenders of the new map argued that Section 5 of the VRA in fact *forced* the legislature to draw black-dominated districts.

Figure 3 depicts two exchanges in which justices spar over this legal claim about Section 5—that a state must hold or increase the numerical percentage of black voters in a district to maintain minorities' "ability to elect their preferred candidates." In the first exchange, beginning with Justice Scalia's question, "Well, I thought the Section 5 obligation...," Justice Scalia advocates this conservative view when questioning the liberal advocate. Median

Justice Kennedy, perhaps influenced by this line of questioning, pursues it further and transitions into skepticism on the technical point of whether this reading of the VRA constitutes a "one-way ratchet" on minority percentages. Justice Breyer then come to the defense of the liberal side, asking the friendly rhetorical question—ostensibly to the advocate—of whether Justice Kennedy's question was addressed by precedent. Later in the oral argument, these roles reverse. The figure depicts a subsequent exchange in which Justice Kennedy initiates a line of questioning, Justice Breyer attacks by saying to the conservative advocate "I don't know what the defense is *possibly* going to be," and Justice Scalia comes to the rescue with a softball question.

These exchanges illustrate the sort of political interactions modeled by MASS. Panel 3.A depicts how skeptical and neutral speech are deployed throughout discussion, and Panels 3.B.1–3 each highlight a justice's choice of tone (e.g., the decision to switch from neutrality to skepticism, which we model using covariates such as conversation history or the ideology of the side currently being questioned). Panels 3.C.1–2 examine two utterances in depth, showing a subset of the auditory features that MASS relies on to infer speech tone. Each tone is modeled as a sequence of discrete sounds, like "vowel" or "silence;" their usage is shown in Panels 3.D.1–2, and their auditory content is in Panel 3.E.

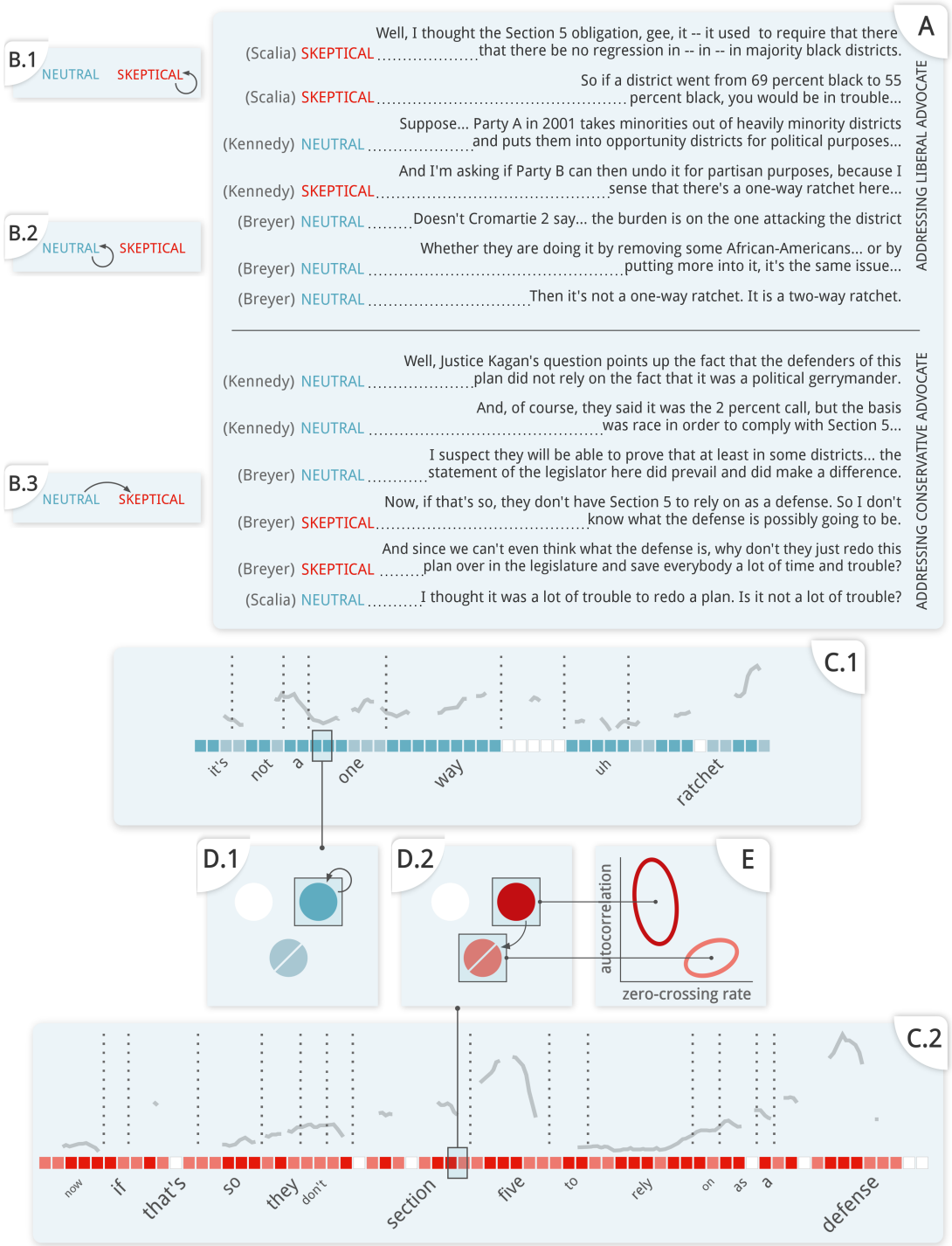

Figure 3: **An illustrative example.** Panel A contains an excerpt from Alabama Legislative Black Caucus v. Alabama, where Justices Scalia, Kennedy, and Breyer utilize neutral and skeptical tones in questioning. Call-outs highlight successive utterance-pairs in which the speaker shifted from one mode to another (B.3), and continued in the same tone of voice (B.1 and B.2). Panels C.1 and C.2 illustrate the use of loudness (text size) and pitch (contours) in a single utterance: in the neutral mode of speech (C.1), speech varies less in pitch and loudness when compared to skeptical speech (C.2). On the basis of these and other features, MASS learns to categorize sounds into vowels (dark squares), consonants (light), and pauses (white). Call-outs D.1 and D.2 respectively identify sequential frames in which a "neutral" vowel is sustained (transition from the dark blue sound back to itself, indicating repeat) and the dark red "skeptical" vowel transitions to the light red consonant. Panel E shows the differing auditory characteristics of the "skeptical" vowel and consonant, which are perceived by the listener.

5.4 Validating the Model

We conduct extensive validation of our model-based measure of judicial skepticism, confirming that MASS does in fact successfully estimate the quantity of interest. Due to space constraints, we summarize these efforts here; results are described in more detail in Online Appendix 4.

First, we demonstrate that MASS recovers a measure that has high facial validity. Online Appendix 4.1 presents utterances from the highest and lowest decile of model-predicted skepticism. Those characterized by the model as skeptical include gentle mockery and doubtful questions, whereas model-predicted neutral utterances are factual statements and straightforward legal analysis.

Second, we examine content validity in Online Appendices 4.2–4.3. Our model detects skepticism on the basis of physiologically and linguistically meaningful auditory features. In a model of Justice Kennedy's expressed skepticism, compared to neutral questioning, we find that his speech is louder, contains more pitch modulation, and is characterized by higher vocal tension. We caution that MASS does not take textual signals of skepticism into account, an important limitation on content validity. (Joint models of audio and text remain an important direction for future work.) However, we demonstrate that in the case of judicial speech, there are extremely few textual signals that distinguish skepticism from typical questioning.

Third, in Online Appendix 4.4, we estimate a lower bound on the out-of-sample performance of MASS, using cross-validation results from the lower stage of the model, corresponding to Equations 3–4 and excluding Equations 1–2. (True out-of-sample performance

of the full model is difficult to evaluate because of dependencies introduced when modeling context and conversation flow. In Online Appendix 4.6, we conduct such a test using speaker labels, which are available for the full corpus.) We find that out-of-sample accuracy of the lower-level auditory classifier is 68%, versus the 52% that would be obtained by randomly permuting labels; numerous additional performance metrics are reported in the appendix.[12]

Fourth, we compare the performance of MASS to (1) human coders and (2) text-based classifiers. For the human comparison, we recruited native English speakers on a crowdworking site and evaluated their ability to recover ground-truth labels in legal argumentation. We found that when combining responses by majority vote, non-expert listeners were able to detect 70% of judicial skepticism, outperforming MASS by a small margin. The performance of individual coders was lower, at 65%, suggesting that with a relatively small amount of domain-specific data, our model performs approximately as well as humans with a lifetime of experience in parsing non-domain-specific speech. For the textual comparison, we applied an elastic net to utterance word counts. The resulting trained text models were entirely degenerate, predicting the more common label in virtually every case.

Finally, we probe the predictive validity of our measure with a comparison to Black et al. (2011), described in the next subsection.

[12]Performance varies considerably depending on the auditory label of interest. As the comparison to non-expert human coders shows, detecting judicial skepticism is a particularly difficult task. In Online Appendix 4.6 we report results for a much simpler task, predicting speaker identity in out-of-sample utterances. Here, MASS achieves accuracies of 97% (outperforming the best competing model by 12 percentage points). Applications that involve less complex tones, like expressed anger, should expect performance between these extremes.

5.5 Comparison to an existing measure

We conduct yet another validity test by contrasting our model with the approach of Black et al. (2011), who use a measure based on directed pleasant and unpleasant words—operationalized with the Dictionary of Affect in Language DAL Whissell 2009—to predict justice voting. We replicate and contrast results with those from a comparable measure of directed skepticism.[13] Specifics are discussed in Online Appendix 4.5. We find that a one-standard-deviation increase in directed unpleasant (pleasant) words is associated with a 2.9 percentage-point decrease (no difference) in the probability that a justice votes against a side. In comparison, a one-standard-deviation increase in directed skepticism is associated with an 8 percentage-point decrease in vote probability, nearly three times as large. Moreover, Figure 4 shows that unlike text-based results, these patterns are robust to the inclusion of both justice and case fixed effects.

Why is speech tone so much more predictive of voting patterns than the use of affective words? One reason may be that DAL uses a cross-domain measure of word pleasantness, ignoring the fact that words often take on different meanings in legal contexts. For example, the 15 most common "unpleasant" words in our corpus include "argument" and "trial," whereas the most common "pleasant" words include "read" and "justice." We return to this point in our recommendations for applied researchers, in Section 7. However, as we show in other text-audio comparisons, a more likely explanation is that word choice is a noisy, high-dimensional, and difficult-to-measure signal of expressed emotion, whereas auditory tone is

[13]Our measure of directed skepticism is based on lower-level HMMs alone, since the complete model incorporates voting as a covariate.

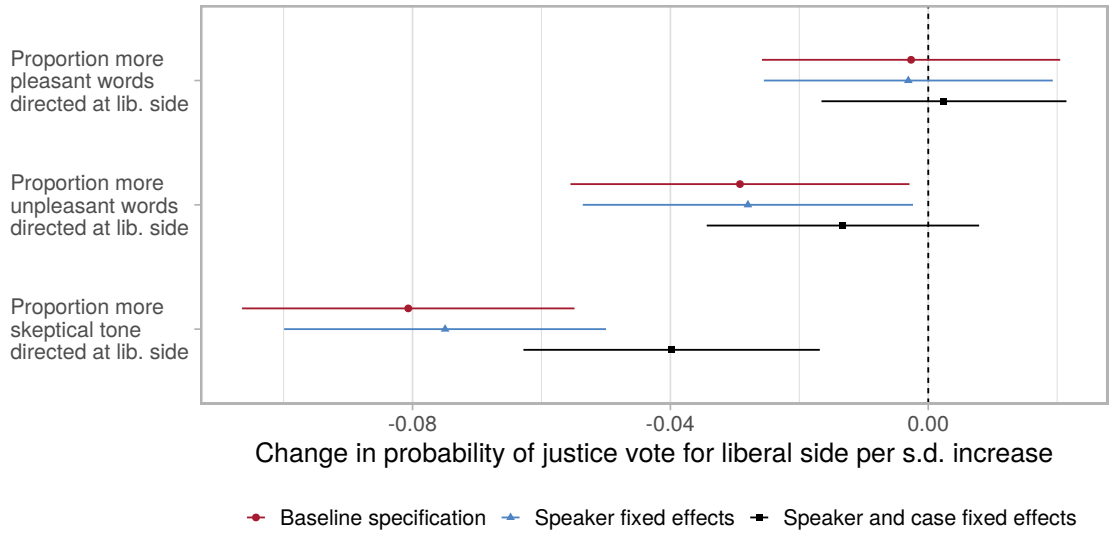

Figure 4: **Predicting justice votes with directed skepticism and directed affective language.** Horizontal errorbars represent point estimates and 95% confidence intervals from regressions of justice votes on directed pleasant words, directed unpleasant words, and our audio-based directed skepticism. Red circles correspond to a specification with no additional controls; blue triangles (black squares) report results with speaker fixed effects only (speaker and case fixed effects).

relatively structured and consistent.

However, we note that MASS exploits only the auditory channel of speech. While we show in Online Appendix 4.4 that this provides a clearer signal of skepticism than text *in general*, there are nonetheless cases when expressions of disbelief are spoken flatly. In one example, Justice Sotomayor states matter-of-factly, "Counsel, I... I... having read some of those cases that you've cited that you claim weakened or eliminated the burden-of-proof standard, most of them didn't quite eliminate it." The utterance is clearly skeptical, yet based on its vocal delivery alone, our model predicts that it is 90% likely to be neutral speech. This counterexample highlights limitations in the use of any communication channel in isolation, suggesting that joint models of text and tone are a necessary direction for future work.

Next, we demonstrate the model with an application to speech in Supreme Court oral

arguments, then conclude.

6 Testing Theories of Supreme Court Deliberation

While some scholars believe that oral arguments are inconsequential in Supreme Court decision-making (Rohde and Spaeth 1976; Segal and Spaeth 1993), others argue that they play a crucial role in shaping the Court's ultimate position on a case (Wasby, D'Amato, and Metrailer 1976; Shapiro 1984; McGuire 1995; Johnson 2001; Johnson, Wahlbeck, and Spriggs 2006; Epstein, Landes, and Posner 2010; Black, Sorenson, and Johnson 2013). The justices themselves, however, are virtually unanimous on this point. Justice Powell stated, "the fact is, as every judge knows... the oral argument... does contribute significantly to the development of precedents" (Stern, Gressman, and Shapiro 1993). Johnson and Goldman (2009) document numerous other direct quotes about the importance of oral arguments, including Justices Harlan, Hughes, Jackson, Brennan, White, Rehnquist, and others.

But there is little empirical evidence about *how* arguments matter. Courts scholars have advanced various accounts of the deliberative process, which can be roughly grouped into two broad theories. For concreteness, we discuss these theories and their associated mechanisms in the context of *Rapanos v. United States*, a narrowly decided environmental case about the Army Corps of Engineer's right to regulate pollution in wetlands, drawing extensively on legal analysis by Savage (2009).[14]

The first theory holds that justices are shrewd political actors who maneuver to influence the decisions of their peers in pursuit of a desired case outcome (Shullman 2004; Epstein,

[14]The full argument is available at https://www.oyez.org/cases/2005/04-1034.

Landes, and Posner 2010, 2013; Iaryczower and Shum 2012; Iaryczower, Shi, and Shum 2018). However, they are constrained by strong judicial norms against back-room discussions, which foreclose the possibility of private communication. In this account, oral arguments represent an opportunity for justices to *strategically signal* to their colleagues, with lawyers and their legal arguments serving merely as convenient foils. Some justices say as much, noting "We're always trying to persuade each other. But persuading each other is complicated... [one way] is to identify what you think of as the difficult problem in a case, and then pose a question that will reveal the difficulty in some way" (Justice Breyer (1998)). Justice Sotomayer (2019) concurs, saying "sometimes we're talking to each other, and we're raising points through the questions that we want our colleagues to consider with us." These attempts at persuasion appear to be on display in *Rapanos v. United States*. From the very start of arguments, battle lines were drawn over the precise definition of a watershed, which determined the Corps' jurisdiction. Justice Roberts, an well-known opponent of what he views as the agency's regulatory overreach, sought to undermine its position: "To *me* it... it suggests that even the *Corps* recognized that at *some* point you've got to say stop, because *logically* any drop of water *anywhere* is going to have *some* sort of connection through drainage."

Conversely, the liberal wing of the court attacked the Pacific Legal Foundation's (PLF, a conservative property-rights advocacy group) position that the federal government should not have jurisdiction over local pollution unless it can prove that a polluter's waste reached a navigable waterway. Justice Souter—whose support for greater environmental protection was previously made clear in *SWANCC v. United States*—ridicules this position, asking "You mean... in *every case*, then... a scientist would have to analyze the *molecules* and.... trace it to a specific discharge... ?"

On its face, it seems plausible that these justices sought to sway Justice Kennedy on this pivotal issue. The point was of critical importance to Justice Kennedy, more so than any other justice.[15]

But were those questions intentionally deployed to shape the subsequent flow of conversation—and ultimately, voting? Or would Justices Roberts and Souter have taken the same stance even if the outcome of the case were not in question? A second, largely incompatible conception of the decision-making process considers justices as neutral arbiters, each casting a *sincere vote* according to rules determined by their respective legal philosophies (Johnson 2001; Black et al. 2011; Black, Sorenson, and Johnson 2013). In this latter account, oral arguments primarily operate as opportunity for fact-finding, rather than persuasion. Justice Douglas summarized this position best, stating "The purpose of a hearing is that the Court may learn what it does not know" (Galloway 1989). (Justice Thomas (2013) famously goes a step further, saying "I think it's unnecessary to deciding cases to ask that many questions, and I don't think it's helpful.") And while justices may reveal their predispositions with a display of doubt, this theory suggests that it is merely an honest response to compelling or uncompelling legal reasoning, rather than an effort to manipulate. For example, Justice Scalia was unable to contain his skepticism after Justice Souter's attack, exclaiming, "Well, I... couldn't you simply *assume* that anything that is discharged into a tributary, ultimately,

[15]At one point, he clarifies "I think what the Court is asking you... is how to define 'significant nexus,"' to a navigable waterway. This point of contention is mentioned 29 times in Kennedy's separate concurrence—compared to 31 times in the opinion, Roberts' concurrence, and both dissents combined. Moreover, 19 of these mentions were in reference to Kennedy's position.

goes where the tributary goes? ... You *really* think it has to trace the molecules?" His outburst, which undermined the position of the PLF (his preferred side), suggested an genuine response to a difficult-to-believe position rather than an attempt at persuasion.

These competing accounts are challenging to disentangle even under the best of circumstances. This difficulty has been compounded by widespread reliance on a narrowly limited representation of judicial speech: textual transcripts alone. Here, we demonstrate that the discarded audio channel contains information of enormous value to social scientists—and that by modeling the tone it conveys, MASS not only opens the door to new research questions but can also shed new light on existing puzzles. Specifically, we use MASS to analyze the structural determinants of expressed skepticism in oral arguments. Using justices' ideological leanings, their ultimate vote, and a measure of case contentiousness, we test the observable implications of two commonly espoused but conflicting narratives of the Supreme Court decisional process: that justices are highly strategic actors jockeying for influence, on the one hand (Johnson, Wahlbeck, and Spriggs 2006), or alternatively that they are neutral arbiters who respond genuinely to compelling legal arguments (Johnson, Wahlbeck, and Spriggs 2006).

Differing theoretical accounts of deliberation suggest very different patterns in the usage of this tone. A model of *genuine voting* implies when justices communicate their skepticism (to the extent that they make it known at all), it is largely as a natural reaction to poor argumentation. In other words, we should observe generically higher rates of skepticism when justices question lawyers for the side that they find less persuasive. This leads to an observable implication of the genuine-voting theoretical account: when a justice votes against a side, we should observe that this behavior is associated with increased skeptical

questioning of the corresponding lawyers (Black et al. 2011; Black, Sorenson, and Johnson 2013).

A *strategic-signaling* model of deliberation, on the other hand, must account for the fact that many—indeed, nearly half—of all cases are decided unanimously. When all justices agree, no strategy is necessary. There is little to gain from posturing, including acted skepticism. To the extent that experienced justices are able to identify uncontroversial cases from pre-argument legal briefs and lower court decisions, this suggests a key observable implication of the strategic-signaling account: justices should exhibit greater skepticism toward their non-preferred side *especially* in contentious cases. That is, in cases that are ultimately decided by a 5-4 margin, we should see justices use markedly more skepticism toward the side they vote against. Forward-looking justices should similarly reduce skepticism toward their own side to avoid damaging its chances in close calls. (We note that persuading justices to change their vote is not the only potential incentive in a strategic model of judicial questioning. For instance, justices may wish to shift the opinion's content even when expecting a unanimous decision, though this does not undermine the general logic of our test.) To further adjudicate between these competing explanations, we turn to a dynamic test of oral arguments, in which the implications of each theory are even cleaner.

In general, justices who are ideologically close will exhibit greater similarity in preferences and judicial perspectives, relative to those who are far apart. When justice i finds a line of legal reasoning to be objectionable (as manifested in an expression of skepticism) it is likely that their ideological neighbor j will find it objectionable as well. The two narratives then diverge in their predictions for j's response. A *genuine* reaction would be to acknowledge the flaw in reasoning, perhaps following up with further skeptical probing regardless of j's

affinity for the lawyer under attack. In contrast, if i is ideologically distant from j, then i's skepticism should not provoke much of a response from j due to the relative lack of shared hot-button issues. The *strategic* account, on the other hand, implies a very different flow of questioning. Suppose that j dislikes the current lawyer. If j were a savvy justice, they should be on the lookout for weaknesses in the opposing side's arguments, seizing the chance to dogpile when an opportunity presents itself. Ideological distance from i—the preceding critic—should not restrain the shrewd operator much, if at all. Indeed, a left-right combination may be a particularly effective blow against the current lawyer.

The strategic narrative suggests a very different sequence of events when j's *preferred* side comes under attack, however. When ideological neighbor i expresses skepticism, j has an incentive to smooth things over—despite j's ideological inclination to agree with i's points. Thus, the extent to which ideological proximity colors j's response to prior skepticism is a useful point of differentiation. Specifically, we discretize ideology into "left" (Justices Breyer, Ginsburg, Kagan, and Sotomayor) and "right" (Justices Alito, Roberts, and Scalia), setting Kennedy aside given his unique position at the median. We then test whether a justice agrees with their *usual* allies—that is, expresses skepticism together—even when that skepticism is directed against their preferred side. If so, this suggests a genuine response; if not, it suggests that justices may be strategically pulling their punches to protect case-specific interests.

The observable implications described above utilize post-argument proxies for justice preferences (vote) and case divisiveness (margin of victory). A natural concern is that a justice's ultimate vote may be influenced by the course of questioning as well. In this case, persuasion may offer an alternative explanation for patterns in observed skepticism. If strategic justices are always successful in persuading their colleagues, the genuine-voting

and strategic-signaling accounts become observationally similar in many cases. While we find this level of persuasiveness to be implausible in light of extensive qualitative work on the courts (Ringsmuth, Bryan, and Johnson 2013; Wolfson 2001), there is considerable room for future work to improve on our analysis. These improvements may include more rigorous formal modeling of strategic interaction in the deliberative process; collection of better pre-argument proxies for justice predisposition and case controversiality (e.g. circuit court votes); or analysis of natural experiments (e.g. exploiting justice retirements).

Finally, note that in interpreting model parameters and testing theories, we formulate all hypotheses in the following matter: Conditional on justice i speaking, is it more likely that they do so skeptically in one conversation context, as opposed to another?[16] The competing narratives described above suggest several observable implications for the structural determinants of justice tone—for example, how the side currently being questioned or the tone of the previous questioner translate into expressed skepticism.

Figure 5 presents results from two MASS specifications. First, we model transition probabilities (i.e., the probability that the next utterance is of justice-tone m) as $\exp\left(\boldsymbol{W}_u^\top \boldsymbol{\zeta}_m\right) / \sum_{m'=1}^{M} \exp\left(\boldsymbol{W}_u^\top \boldsymbol{\zeta}_{m'}\right)$, where the conversation context $\boldsymbol{W}$ includes the eventual case margin, the ultimate vote of the justice in question, an interaction, and a justice-tone intercept. We then average across justices according to their speech frequency to obtain an average effect. The results show that justices use skepticism against their non-preferred side—either the petitioner or respondent, depending on whom they go on to vote against—at a significantly higher rate, as expected.

[16]This formulation allows us to partial out any shifts in speaker frequencies, which besides being difficult to theorize are also relatively uninteresting.

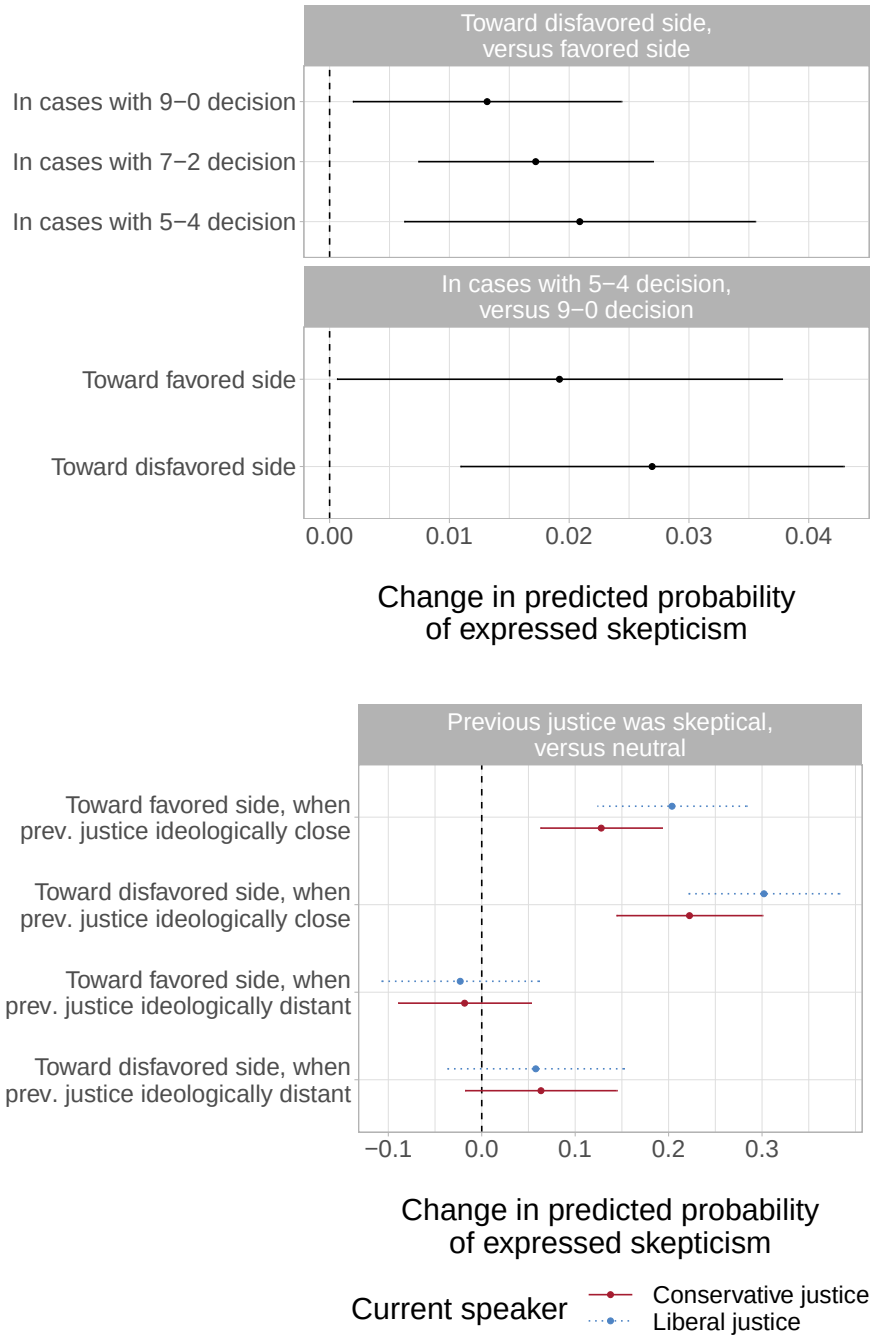

Figure 5: **Simulated quantities of interest.** Each panel manipulates a single variable from a control value (second element of panel title) to a treatment value (first). Points (error bars) represent estimated changes (95% bootstrap confidence intervals) in skepticism. We average over all other nuisance covariates (e.g., the identity of the next speaker) of the scenario-specific change in outcome, weighting by the empirical frequencies of these covariates. The top panel shows that justices deploy skepticism more often toward their non-preferred side. The second panel compares close votes to unanimous decisions, demonstrating that justices express more skepticism across the board in the former. However, justices do not attempt to target a particular side in close votes; rather, they simply ask more skeptical questions across the board. Finally, the bottom panel shows that justices mirror the tones of their ideological neighbors, who share similar legal philosophies, even when those neighbors are opposed to the justice's case-specific voting interests.

Contrary to theoretical predictions under strategic signaling, however, we find no indication that the gap between petitioner- and respondent-targeted skepticism depends on the margin of the case decision. That is, the difference in differences is far from significant. We do find that skepticism is generically higher *across the board* in close cases. These results are consistent with the account that Supreme Court justices are engaged in genuine fact-finding, not a strategic attempt to manipulate their colleagues, and that they particularly seek to get answers right when the stakes are high.[17]

To probe further, we now turn to the dynamics of argumentation. In an expanded specification, we now incorporate additional binary indicators for whether the preceding speaker belonged to the liberal or conservative wing of the Court, as well as interactions between skepticism in the preceding utterance, ideology of the previous speaker, and vote. As described above, the *strategic* model of judicial signaling implies that after a peer—any peer—criticizes a justice's preferred side, the justice should withhold comment or defuse tensions with neutral commentary. By the same token, the savvy justice should follow up with a coup de grâce after a colleague finds fault in the disfavored side's reasoning. We find no evidence that this is true. Rather, our results are highly consistent with a model of *genuine expression* in which justices concede to the criticisms of ideologically proximate peers, regardless of their case-specific interests. That is, after a liberal (conservative) justice casts doubt on a lawyer's argument, other liberals (conservatives) on the Court are likely to follow suit even if that criticism undermines their favored side.

[17]As we note in detail above, we cannot entirely rule out the alternative explanation that some of these results are due to persuasion over the course of oral arguments.

7 Practical Guidelines for Applied Audio Research

Having demonstrated the application of MASS to Supreme Court oral arguments, we now briefly discuss general principles for future speech analysis before concluding. First, we note that in the context of speech audio, substantively interesting quantities are almost always unobserved—that is, they are latent variables, not ex-ante specifiable functions of audio features. Researchers should seek to measure these as directly as possible. In practice, this almost always implies some sort of supervised approach with labeled data.

Second, precisely because the substantive variables of interest are not observable, they must be inferred with uncertainty. MASS provides estimates of the relationship between covariates of interest and these latent variables that reflect this uncertainty. In speech analysis and machine learning more generally, approaches that use predicted labels while ignoring this fact will dramatically overstate their confidence, leading to spurious results. These should be avoided.

Third, the importance of validation cannot be overstated. In text analysis, it is now widely appreciated that analysts must validate their measurements and results extensively to ensure that they capture the theoretical quantity of interest. This principle is arguably more critical with audio data, where features are typically less interpretable than textual "terms". Thus, researchers are encouraged to think carefully about validity in its various forms. We demonstrate several approachs to validation in Online Appendix 4.4.

Finally, as a practical matter, researchers may wonder how much training data is needed to conduct an analysis of the sort demonstrated here. In general, the amount of labeled data increases with the subtlety of the labels. In practice, this is likely to range from perhaps ten

minutes of labeled audio per speech mode for highly distinct categories (Online Appendix 4.6) to perhaps thirty minutes per mode for subtle labels like skepticism.

8 Concluding Remarks

While a great deal of progress has been made in analyzing *what* was said in these audio recordings, *how* these words were spoken is often equally or even more important. MASS provides a principled solution for inferring vocal tone. And by modeling patterns in its usage, MASS directly tests questions about speech dynamics like inter-speaker interaction—phenomena that are impossible to study with traditional approaches.

We view this paper as another step toward a better model of political speech, and surely not the final stride. Numerous related areas of inquiry remain open. We briefly note three here.

First, we note that MASS exploits only the auditory channel of speech. While we show in Online Appendix 4 that this provides a clearer signal of skepticism than text *in general*, there are nonetheless cases when expressions of disbelief are spoken flatly. Joint models of text and tone are a necessary direction for future work.

Second, our application to Supreme Court oral arguments is among the first attempts at supervised classification using audio data within political science. As such, there is no methodological consensus on best practices, for example on preprocessing or labeling training data. While the previous section briefly outlines our suggestions for research with audio data, more work is needed to better understand how to annotate speech audio.

Finally, while there are clear advantages to supervised models for measurement in political

science, the widespread use of unsupervised models for discovery in speech text suggests that similar approaches may also be useful in audio analysis. MASS can in principle be extended to an unsupervised setup, and we view this as an interesting direction for future work.

References

Anderson, Rindy C, and Casey A Klofstad. 2012. "Preference for leaders with masculine voices holds in the case of feminine leadership roles". *PloS one* 7 (12): e51216.

Anderson, Rindy C, et al. 2014. "Vocal fry may undermine the success of young women in the labor market". *PloS one* 9 (5): e97506.

Apple, William, Lynn A Streeter, and Robert M Krauss. 1979. "Effects of pitch and speech rate on personal attributions." *Journal of personality and social psychology* 37 (5): 715.

Banse, Rainer, and Klaus R Scherer. 1996. "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology* 70 (3): 614.

Bayley, Paul. 2004. *Cross-cultural perspectives on parliamentary discourse*. Vol. 10. John Benjamins Publishing.

Beck, Christina. 1996. ""I've got some points I'd like to make here": The achievement of social face through turn management during the 1992 vice presidential debate". *Political Communication* 13 (2): 165–180.

Behr, Roy, and Shanto Iyengar. 1985. "Television news, real-world cues, and changes in the public agenda". *Public Opinion Quarterly* 49 (1): 38–57.

Benoit, Kenneth, et al. 2018. "quanteda: An R package for the quantitative analysis of textual data". *Journal of Open Source Software* 3 (30): 774.

Benoit, William L. 2013. *Political election debates: Informing voters about policy and character*. Lexington Books.

Benoit, William, Joseph Blaney, and P.M. Pier. 1998. *Campaign '96: A Functional Analysis of Acclaiming, Attacking, and Defending*. Westport, CT: Greenwood.

Black, Ryan C, Maron W Sorenson, and Timothy R Johnson. 2013. "Toward an actor-based measure of Supreme Court case salience: Information-seeking and engagement during oral arguments". *Political Research Quarterly* 66 (4): 804–818.

Black, Ryan, et al. 2011. "Emotions, oral arguments, and Supreme Court decision making". *Journal of Politics* 73 (2): 572–581.

Bligh, Michelle, et al. 2010. "Finding her voice: Hillary Clinton's rhetoric in the 2008 presidential campaign". *Women's Studies* 39 (8): 823–850.

Boydstun, Amber, et al. 2014. "Real-Time Reactions to a 2012 Presidential Debate: A Method for Understanding Which Messages Matter". *Public Opinion Quarterly* 78 (S1): 330–343.

Brader, Ted. 2006. *Campaigning for Hearts and Minds: How Emotional Appeals in Political Ads Work*. Chicago, IL: University of Chicago Press.

Breyer, Stephen. 1998. "The Work of the Supreme Court". *Bulletin of the American Academy of Arts and Sciences* 52 (1): 47–58.

Brown, Bruce L, William J Strong, and Alvin C Rencher. 1974. "Fifty-four voices from two: the effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech". *The Journal of the Acoustical Society of America* 55 (2): 313–318.

Burgoon, Judee K, Thomas Birk, and Michael Pfau. 1990. "Nonverbal behaviors, persuasion, and credibility". *Human communication research* 17 (1): 140–169.

Carlson, David, and Jacob Montgomery. 2017. "A pairwise comparison framework for fast, flexible, and reliable human coding of political texts". *American Political Science Review* 111 (4): 835–843.

Carney, Dana R, Judith A Hall, and Lavonia Smith LeBeau. 2005. "Beliefs about the nonverbal expression of social power". *Journal of Nonverbal Behavior* 29 (2): 105–123.

Cohen, Jeffrey. 1995. "Presidential Rhetoric and the Public Agenda". *American Journal of Political Science* 39 (1): 87–107.

Conroy-Krutz, Jeffrey, and Devra C Moehler. 2015. "Moderation from bias: A field experiment on partisan media in a new democracy". *The Journal of Politics* 77 (2): 575–587.

Conway III, Lucian Gideon, et al. 2012. "Does complex or simple rhetoric win elections? An integrative complexity analysis of US presidential campaigns". *Political Psychology* 33 (5): 599–618.

Cornell, Legal Information Institute. 2015. *Oyez Project*. Accessed August 2015. `oyez.org`.

Degani, Marta. 2015. *Framing the Rhetoric of a Leader: An Analysis of Obama's Election Campaign Speeches*. Springer.

Dellaert, Frank, Thomas Polzin, and Alex Waibel. 1996. "Recognizing emotion in speech". In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 3:1970–1973. IEEE.

Dietrich, Bryce J, Matthew Hayes, and Diana Z O'Brien. 2019. "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech". *American Political Science Review*: 1–22.

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

Dietrich, Bryce, Dan Schultz, and Tracey Jaquith. 2018. "This Floor Speech Will Be Televised: Understanding the Factors that Influence When Floor Speeches Appear on Cable Television". *Working Paper*.

El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. 2011. "Survey on speech emotion recognition: features, classification schemes, and databases". *Pattern Recognition* 44:572–587.

Epstein, Lee, William Landes, and Richard Posner. 2010. "Inferring the winning party in the Supreme Court from the pattern of questioning at oral argument". *Journal of Legal Studies* 39 (2): 433–467.

— . 2013. *The behavior of federal judges: a theoretical and empirical study of rational choice*. Harvard University Press.

Fearon, James D. 1994. "Domestic political audiences and the escalation of international disputes". *American political science review* 88 (3): 577–592.

Fridkin, Kim L, and Patrick Kenney. 2011. "Variability in citizens' reactions to different types of negative campaigns". *American Journal of Political Science* 55 (2): 307–325.

Fridkin, Kim, et al. 2007. "Capturing the power of a campaign event: The 2004 presidential debate in Tempe". *Journal of Politics* 69 (3): 770–785.

Galloway, Russell. 1989. "Oral Argument in the Court". *Trial* 25 (1): 78–84.

Gregory Jr, Stanford W, and Timothy J Gallagher. 2002. "Spectral analysis of candidates' nonverbal vocal communication: Predicting US presidential election outcomes". *Social Psychology Quarterly*: 298–308.

Grimmer, Justin, and Brandon Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts". *Political Analysis* 21 (3): 267–297.

Guisinger, Alexandra, and Alastair Smith. 2002. "Honest threats: The interaction of reputation and political institutions in international crises". *Journal of Conflict Resolution* 46 (2): 175–200.

Hamilton, Mark A, and Becky L Stewart. 1993. "Extending an information processing model of language intensity effects". *Communication quarterly* 41 (2): 231–246.

Hart, Roderick P, and Sharon E Jarvis. 1997. "Political debate: Forms, styles, and media". *American Behavioral Scientist* 40 (8): 1095–1122.

Herzog, Alexander, and Kenneth Benoit. 2015. "The most unkindest cuts: speaker selection and expressed government dissent during economic crisis". *The Journal of Politics* 77 (4): 1157–1175.

Hetherington, Marc J. 1999. "The effect of political trust on the presidential vote, 1968–96". *American Political Science Review* 93 (2): 311–326.

Hinck, Edward, and Shelly Hinck. 2002. "Politeness Strategies in the 1992 Vice Presidential and Presidential Debates". *Argumentation and Advocacy* 38 (4): 234–250.

Hodges-Simeon, Carolyn R, Steven JC Gaulin, and David A Puts. 2010. "Different vocal parameters predict perceptions of dominance and attractiveness". *Human Nature* 21 (4): 406–427.

Hofstetter, C Richard, et al. 1999. "Information, misinformation, and political talk radio". *Political Research Quarterly* 52 (2): 353–369.

Iaryczower, Matias, Xiaoxia Shi, and Matthew Shum. 2018. "Can Words Get in the Way? The Effect of Deliberation in Collective Decision Making". *Journal of Political Economy* 126 (2): 688–734. doi:10.1086/696228.

Iaryczower, Matias, and Matthew Shum. 2012. "The value of information in the court: Get it right, keep it tight". *American Economic Review* 102 (1): 202–37.

Johnson, Timothy. 2001. "Information, oral arguments, and Supreme Court decision making". *American Politics Research* 29 (4): 331–351.

Johnson, Timothy R., and Jerry Goldman. 2009. "The Role of Oral Arguments in the Supreme Court". In *A Good Quarrel*, ed. by Timothy R. Johnson and Jerry Goldman, 1–10. Ann Arbor, MI: University of Michigan Press.

Johnson, Timothy, Paul Wahlbeck, and James Spriggs. 2006. "The influence of oral arguments on the U.S. Supreme Court". *American Political Science Review* 100 (01): 99–113.

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

Kappas, Arvid, Ursula Hess, and Klaus R Scherer. 1991. "Voice and emotion". *Fundamentals of nonverbal behavior* 200.

Kaufman, Aaron, Peter Kraft, and Maya Sen. 2018. "Improving Supreme Court Forecasting Using Boosted Decision Trees". *URL: j. mp/sctfore*.

Kernell, Samuel. 2006. *Going public: New strategies of presidential leadership*. Washington, DC: CQ Press.

Klofstad, Casey A. 2016. "Candidate voice pitch influences election outcomes". *Political Psychology* 37 (5): 725–738.

Klofstad, Casey A, Rindy C Anderson, and Stephen Nowicki. 2015. "Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices". *PloS one* 10 (8): e0133779.

Klofstad, Casey A, Rindy C Anderson, and Susan Peters. 2012. "Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women". *Proceedings of the Royal Society B: Biological Sciences* 279 (1738): 2698–2704.

Kwon, Oh-Wook, et al. 2003. "Emotion recognition by speech signals". In *Eighth European Conference on Speech Communication and Technology*.

Lauderdale, Benjamin, and Alexander Herzog. 2016. "Measuring political positions from legislative speech". *Political Analysis* 24 (3): 374–394.

Laustsen, Lasse, Michael Bang Petersen, and Casey A Klofstad. 2015. "Vote choice, ideology, and social dominance orientation influence preferences for lower pitched voices in political candidates". *Evolutionary Psychology* 13 (3): 1474704915600576.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting policy positions from political texts using words as data". *American political science review* 97 (2): 311–331.

Leathers, Dale G. 1979. "The impact of multichannel message inconsistency on verbal and nonverbal decoding behaviors". *Communications Monographs* 46 (2): 88–100.

Levi, Margaret, and Laura Stoker. 2000. "Political trust and trustworthiness". *Annual review of political science* 3 (1): 475–507.

Lucas, Christopher, et al. 2015. "Computer-assisted text analysis for comparative politics". *Political Analysis* 23 (2): 254–277.

Manstead, Anthony SR, Hugh L Wagner, and Christopher J MacDonald. 1984. "Face, body, and speech as channels of communication in the detection of deception". *Basic and Applied Social Psychology* 5 (4): 317–332.

Manusov, Valerie, and April R Trees. 2002. ""Are you kidding me?": The role of nonverbal cues in the verbal accounting process". *Journal of Communication* 52 (3): 640–656.

Martin, Andrew, and Kevin Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999". *Political Analysis* 10 (2): 134–153. doi:10.1093/pan/10.2.134.

McGilloway, Sinéad, et al. 2000. "Approaching automatic recognition of emotion from voice: a rough benchmark". In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

McGuire, Kevin. 1995. "Repeat players in the Supreme Court: The role of experienced lawyers in litigation success". *Journal of Politics* 57 (1): 187–196.

McKinney, Mitchell, Lynda Kaid, and Terry Robertson. 2001. "The Front-Runner, Contenders, and Also-Rans: Effects of Watching a 2000 Republican Primary Debate". *American Behavioral Scientist* 44 (12): 2232–2251.

Mermin, Jonathan. 1997. "Television news and American intervention in Somalia: The myth of a media-driven foreign policy". *Political science quarterly* 112 (3): 385–403.

Mower, Emily, et al. 2009. "Interpreting Ambiguous Emotional Expressions". In *Proceedings ACII Special Session: Recognition of Non-Prototypical Emotion From Speech - The Final Frontier?*, 662–669.

Neal, Radford, and Geoffrey Hinton. 1998. "A view of the EM algorithm that justifies incremental, sparse, and other variants". In *Learning in Graphical Models*, 355–368. Springer.

Nogueiras, Albino, et al. 2001. "Speech emotion recognition using hidden Markov models". In *Seventh European Conference on Speech Communication and Technology*.

Oegema, Dirk, and Jan Kleinnijenhuis. 2000. "Personalization in political television news: A 13-wave survey study to assess effects of text and footage". *Communications* 25 (1): 43–60.

Ohala, John J. 1981. "The listener as a source of sound change". *Papers from the parasession on language behavior. Chicago: Chicago Linguistic Association.*

Olson, Jeremiah, et al. 2012. "The Teleprompter Presidency: Comparing O bama's Campaign and Governing Rhetoric". *Social Science Quarterly* 93 (5): 1402–1423.

Podesva, Robert J, et al. 2015. "Constraints on the social meaning of released/t: A production and perception study of US politicians". *Language Variation and Change* 27 (1): 59–87.

Proksch, Sven-Oliver, and Jonathan Slapin. 2012. "Institutional Foundations of Legislative Speech" [**inlang**en]. *American Journal of Political Science* 56 (3): 520–537. ISSN: 00925853, visited on 11/01/2015.

— . 2015. *The Politics of Parliamentary Debate: Parties, Rebels and Representation.* New York: Cambridge University Press.

Quinn, Kevin, et al. 2010. "How to analyze political attention with minimal assumptions and costs". *American Journal of Political Science* 54 (1): 209–228.

Ringsmuth, Eve M, Amanda C Bryan, and Timothy R Johnson. 2013. "Voting fluidity and oral argument on the US Supreme Court". *Political research quarterly* 66 (2): 429–440.

Ritter, Kurt, and Buddy Howell. 2001. "Ending the 2000 presidential election: Gore's concession speech and Bush's victory speech". *American Behavioral Scientist* 44 (12): 2314–2330.

Rohde, David W, and Harold J Spaeth. 1976. *Supreme Court decision making.* WH Freeman.

Ross, Scott. 2016. "Encouraging rebel demobilization by radio in uganda and the dr congo: The case of "come home" messaging". *African Studies Review* 59 (1): 33–55.

Rozenas, Arturas, and Denis Stukal. 2019. "How autocrats manipulate economic news: Evidence from Russia's state-controlled television". *The Journal of Politics* 81 (3): 000–000.

Rule, Alix, Jean-Philippe Cointet, and Peter S Bearman. 2015. "Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014". *Proceedings of the National Academy of Sciences* 112 (35): 10837–10844.

Sanders, David, and Neil Gavin. 2004. "Television news, economic perceptions and political preferences in Britain, 1997–2001". *The Journal of Politics* 66 (4): 1245–1266.

Savage, David G. 2009. "Rapanos v. United States: Wading into the Wetlands". In *A Good Quarrel*, ed. by Timothy R. Johnson and Jerry Goldman, 125–144. Ann Arbor, MI: University of Michigan Press.

Scherer, Klaus R. 1995. "Expression of emotion in voice and music". *Journal of voice* 9 (3): 235–248.

— . 2003. "Vocal communication of emotion: A review of research paradigms". *Speech communication* 40 (1-2): 227–256.

Scherer, Klaus R, Judy Koivumaki, and Robert Rosenthal. 1972. "Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech". *Journal of Psycholinguistic Research* 1 (3): 269–285.

Scherer, Klaus R, Harvey London, and Jared J Wolf. 1973. "The voice of confidence: Paralinguistic cues and audience evaluation". *Journal of Research in Personality* 7 (1): 31–44.

Schirmer, Annett, et al. 2019. "Angry, old, male–and trustworthy? How expressive and person voice characteristics shape listener trust". *PloS one* 14 (1): e0210555.

Schroedel, Jean, et al. 2013. "Charismatic rhetoric in the 2008 presidential campaign: Commonalities and differences". *Presidential Studies Quarterly* 43 (1): 101–128.

Schroeder, Juliana, and Nicholas Epley. 2016. "Mistaking minds and machines: How speech affects dehumanization and anthropomorphism." *Journal of Experimental Psychology: General* 145 (11): 1427.

— . 2015. "The sound of intellect: Speech reveals a thoughtful mind, increasing a job candidate's appeal". *Psychological science* 26 (6): 877–891.

Schwarz, Daniel, Denise Traber, and Kenneth Benoit. 2017. "Estimating intra-party preferences: comparing speeches to votes". *Political Science Research and Methods* 5 (2): 379–396.

Segal, Jeffrey Allan, and Harold J Spaeth. 1993. *The Supreme Court and the attitudinal model*. Vol. 65. Cambridge University Press New York.

Semetko, Holli, and Patti Valkenburg. 2000. "Framing European politics: A content analysis of press and television news". *Journal of communication* 50 (2): 93–109.

Shapiro, Stephen. 1984. "Oral Argument in the Supreme Court of the United States". *Catholic University Law Review* 33 (3): 529–554.

Shullman, Sarah Levien. 2004. "The illusion of devil's advocacy: How the justices of the supreme court foreshadow their decisions during oral argument". *Journal of Appellate Practice and Process* 6:271.

Slapin, Jonathan B, and Sven-Oliver Proksch. 2008. "A scaling model for estimating time-series party positions from texts". *American Journal of Political Science* 52 (3): 705–722.

Smith, Bruce L, et al. 1975. "Effects of speech rate on personality perception". *Language and speech* 18 (2): 145–152.

Sobieraj, Sarah, and Jeffrey Berry. 2011. "From incivility to outrage: Political discourse in blogs, talk radio, and cable news". *Political Communication* 28 (1): 19–41.

Sotomayer, Sonia. 2019. *Life as a Supreme Court Justice*. Interview with Trevor Noah.

Spaeth, Harold, et al. 2014. *Supreme Court Database Code Book*.

Spiliotes, Constantine J, and Lynn Vavreck. 2002. "Campaign advertising: Partisan convergence or divergence?" *The Journal of Politics* 64 (1): 249–261.

Stern, Robert L., Eugene Gressman, and Stephen M. Shapiro. 1993. *Supreme Court Practice: For Practice in the Supreme Court of the United States.* Washington, DC: Bureau of National Affairs.

Surawski, Melissa K, and Elizabeth P Ossoff. 2006. "The effects of physical and vocal attractiveness on impression formation of politicians". *Current Psychology* 25 (1): 15–27.

Thomas, Clarence. 2013. *Lecture at Harvard Law School*. Accessed April 2019. `youtube.com/watch?v=heQjKdHu1P4`.

Thomas, Matt, Bo Pang, and Lillian Lee. 2006. "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts". In *Proceedings of the 2006*

conference on empirical methods in natural language processing, 327–335. Association for Computational Linguistics.

Tigue, Cara C, et al. 2012. "Voice pitch influences voting behavior". *Evolution and Human Behavior* 33 (3): 210–216.

Touati, Paul. 1993. "Prosodic aspects of political rhetoric". In *ESCA workshop on prosody*.

van der Laan, Mark J, Sandrine Dudoit, Sunduz Keles, et al. 2004. "Asymptotic optimality of likelihood-based cross-validation". *Statistical Applications in Genetics and Molecular Biology* 3 (1): 1036.

Ververidis, imitrios, and Constantine Kotropoulos. 2006. "Emotional speech recognition: Resources, features, and methods". *Speech Communication* 48:1162–1181.

Wasby, Stephen, Anthony D'Amato, and Rosemary Metrailer. 1976. "The functions of oral argument in the US Supreme Court". *Quarterly Journal of Speech* 62 (4): 410–422.

Whissell, Cynthia. 2009. "Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Language". *Psychological Reports* 105 (2).

Wilkerson, John, and Andreu Casas. 2017. "Large-scale computerized text analysis in political science: Opportunities and challenges". *Annual Review of Political Science* 20:529–544.

Wolfson, Warren D. 2001. "Oral Argument: Does It Matter". *Indiana Law Review* 35:451.

Wong, Wing Hung. 1986. "Theory of Partial Likelihood". *Annals of Statistics* 14, no. 1 (): 88–123. doi:10.1214/aos/1176349844. https://doi.org/10.1214/aos/1176349844.

Young, Garry, and William B Perkins. 2005. "Presidential rhetoric, the public agenda, and the end of presidential television's "golden age"". *The Journal of Politics* 67 (4): 1190–1205.

Young, Lori, and Stuart Soroka. 2012. "Affective news: The automated coding of sentiment in political texts". *Political Communication* 29 (2): 205–231.

Zucchini, Walter, and Iain MacDonald. 2009. *Hidden Markov Models for Time Series*. Boca Raton, FL: CRC Press.

Zuckerman, Miron, Bella M DePaulo, and Robert Rosenthal. 1981. "Verbal and nonverbal communication of deception". In *Advances in experimental social psychology*, 14:1–59. Elsevier.

Zuckerman, Miron, and Robert E Driver. 1989. "What sounds beautiful is good: The vocal attractiveness stereotype". *Journal of Nonverbal Behavior* 13 (2): 67–82.

Zuckerman, Miron, et al. 1979. "Facial and vocal cues of deception and honesty". *Journal of Experimental Social Psychology* 15 (4): 378–396.

Online Appendix for

"A Dynamic Model of Speech for the Social Sciences"

Intended for online publication only.

Contents

1 Estimation

In this section, we introduce our estimation procedure.

1.1 Factorization of the Likelihood

The complete-data likelihood is

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\zeta}, \boldsymbol{\Theta} \mid \boldsymbol{X}^{\mathcal{T}}, \boldsymbol{S}^{\mathcal{T}}, \boldsymbol{X}^{\mathcal{C}}, \boldsymbol{W}^{\text{stat.},\mathcal{C}}) \\
&\quad = f(\boldsymbol{X}^{\mathcal{T}}, \boldsymbol{X}^{\mathcal{C}}, \mid \boldsymbol{\zeta}, \boldsymbol{\Theta}, \boldsymbol{S}^{\mathcal{T}}, \boldsymbol{W}^{\text{stat.},\mathcal{C}}) \\
&\quad = f(\boldsymbol{X}^{\mathcal{C}} \mid \boldsymbol{\zeta}, \boldsymbol{\Theta}, \boldsymbol{S}^{\mathcal{T}}, \boldsymbol{W}^{\text{stat.},\mathcal{C}}, \boldsymbol{X}^{\mathcal{T}})\ f(\boldsymbol{X}^{\mathcal{T}} \mid \boldsymbol{\zeta}, \boldsymbol{\Theta}, \boldsymbol{S}^{\mathcal{T}}, \boldsymbol{W}^{\text{stat.},\mathcal{C}})
\end{aligned}
$$

By sufficiency

$$
= f(\boldsymbol{X}^{\mathcal{C}}, \boldsymbol{S}^{\mathcal{C}} \mid \boldsymbol{\zeta}, \boldsymbol{\Theta}, \boldsymbol{W}^{\text{stat.},\mathcal{C}})\ f(\boldsymbol{X}^{\mathcal{T}} \mid \boldsymbol{\Theta}, \boldsymbol{S}^{\mathcal{C}})
$$

which is Equation 5.

Our stagewise estimation procedure is primarily motivated by computational considerations. The partial-likelihood approach reduces computational complexity dramatically; simultaneously estimating $\boldsymbol{\zeta}$ and $\boldsymbol{\Theta}$ on the full data would require repeated passes over $\boldsymbol{X}^{\mathcal{C}}$, which is typically too large to hold in memory.

However, the stagewise approach has properties that make it attractive for other reasons as well. First, when the model is correctly specified, our approach remains unbiased with respect to the auditory parameters; in this case, the only sacrifice is in efficiency loss relative to joint maximization of the full likelihood. But in the presence of model misspecification—which almost certainly exists with complex phenomena like human speech, e.g., if true human speech contains more than M modes—the proposed approach can in fact outperform full maximum likelihood. More generally, semi-supervised approaches that exploit both labeled and unlabeled data often underperform those that only use the former (Masanori and Takeuchi 2014). Intuitively, this is because unsupervised methods rarely recover the analyst's preferred labels, and semi-supervised techniques are typically dominated by the much

larger unlabeled dataset.

Finally, we note that even with moderately sized training sets, the number of moments in $\boldsymbol{X}^{\mathcal{T}}$ will be already be several orders of magnitude larger than the number of parameters, due to the high-frequency nature of audio data, so that $\boldsymbol{\Theta}$ is already reasonably well-estimated from the training utterances alone.

1.2 Estimation of Lower-Level Auditory Parameters

To estimate the parameters of the M lower-level models, which each represent the auditory characteristics of a particular speech mode, we employ a non-sequential training set of example utterances that are assumed to be drawn from the same distribution as the primary corpus (conditional on mode). In the main text, the audio features of the training set are denoted $\boldsymbol{X}^{\mathcal{T}}$, and the corresponding tone labels are $\boldsymbol{S}^{\mathcal{T}}$. Here, we drop $\mathcal{T}$ for convenience and work exclusively within the training set.

Consider the subset with known mode $S_u = m$.[1] This group of utterances is assumed to be drawn from a single shared Gaussian HMM, the speech model for mode m. Below, we describe how lower-level parameters are estimated by standard HMM techniques. Interested

[1]In practice, because the perception of certain speech modes can be subjective (human coders may disagree or be uncertain), training set mode labels S_u may be a stochastic vector of length M, $\tilde{S}_u = [\Pr(S_u = 1), \ldots, \Pr(S_u = M)]$, rather than a M-valued categorical variable. In such cases the contribution of an utterance to the model for emotion m may be weighted by the m-th entry, e.g. corresponding to the proportion of human coders who classified the utterance as emotion m. After replacing $\mathbf{1}(S_u = m)$ with $\Pr(S_u = m)$, the procedure described in this appendix can be used without further modification.

readers are referred to Zucchini and MacDonald (2009) for further discussion.

We first write down the likelihood function for parameters of the m-th mode. For each utterance, at each moment t, the feature vector $\boldsymbol{X}_{u,t}$ could have been generated by any of the K sounds associated with emotion m, so there are K^{T_u} possible sequences of unobserved sounds by which the entire feature sequence $\boldsymbol{X}_u$ could have been generated. The u-th utterance's contribution to the observed-data likelihood is the joint probability of all observed features, found by summing over every possible sequence of sounds. This yields

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m, \boldsymbol{\Gamma}^m \mid \boldsymbol{X}, \boldsymbol{S}) \\
&\quad = \prod_{u=1}^{U} \Pr(\boldsymbol{X}_{u,1} = \boldsymbol{x}_{u,1}, \cdots, \boldsymbol{X}_{u,T_u} = \boldsymbol{x}_{u,T_u} \mid \boldsymbol{\mu}^{m,k}, \boldsymbol{\Sigma}^{m,k}, \boldsymbol{\Gamma}^m)^{\mathbf{1}(S_u = m)} \\
&\quad = \prod_{u=1}^{U} \left(\delta^{m\top} \boldsymbol{P}^m(\boldsymbol{x}_{u,1}) \left(\prod_{t=2}^{T_u} \boldsymbol{\Gamma}^m \boldsymbol{P}^m(\boldsymbol{x}_{u,t}) \right) \mathbf{1} \right)^{\mathbf{1}(S_u = m)},
\end{aligned} \tag{1}$$

where $\boldsymbol{\mu}^m = (\boldsymbol{\mu}^{m,k})_{k \in \{1,\ldots,K\}}$, $\boldsymbol{\Sigma}^m = (\boldsymbol{\Sigma}^{m,k})_{k \in \{1,\ldots,K\}}$, δ^m is a $1 \times K$ vector containing the initial distribution of sounds (assumed to be the stationary distribution, a unit row eigenvector of $\boldsymbol{\Gamma}^m$), the matrices $\boldsymbol{P}^m(\boldsymbol{x}_{u,t}) \equiv \operatorname{diag}\left(\phi_D(\boldsymbol{x}_{u,t} | \boldsymbol{\mu}^{m,k}, \boldsymbol{\Sigma}^{m,k})\right)$ are $K \times K$ diagonal matrices in which the (k,k)-th element is the (D-variate Gaussian) probability of $\boldsymbol{x}_{u,t}$ being generated by sound k, and $\mathbf{1}$ is a column vector of ones.

In practice, due to the high dimensionality of the audio features, we also regularize the $\boldsymbol{\Sigma}$ terms to ensure invertibility by adding a small positive value (which may be thought of as a prior) to its diagonal. We recommend setting this regularization parameter, along with the number of sounds, by selecting values that maximize the training set's cross-validated naïve probabilities (i.e., based on mode prevalence and emission probabilities, ignoring con-

text). This procedure asymptotically selects the closest approximation, in terms of the Kullback–Leibler divergence, to the true data-generating process among the candidate models considered (van der Laan, Dudoit, Keles, et al. 2004).

The parameters $\boldsymbol{\mu}^{m,k}$, $\boldsymbol{\Sigma}^{m,k}$, and $\boldsymbol{\Gamma}^{m}$ can in principle be found by directly maximizing this likelihood. However, given the vast number of parameters to optimize over, we estimate using the Baum-Welch algorithm for expectation-maximization with hidden Markov models. In what follows, we describe this procedure as it relates to the estimation of the lower-level audio parameters. Baum-Welch involves maximizing the complete-data likelihood of Equation 2, which differs from equation 1 in that it also incorporates the probability of the unobserved sounds.

$$
\begin{aligned}
&\prod_{u=1}^{U} \Pr(\boldsymbol{X}_{u,1} = \boldsymbol{x}_{u,1}, \cdots, \boldsymbol{X}_{u,T_u} = \boldsymbol{x}_{u,T_u}, R_{u,1} = r_{u,1}, \cdots, R_{u,T_u} = r_{u,T_u} \mid \boldsymbol{\mu}^{m,*}, \boldsymbol{\Sigma}^{m,*}, \boldsymbol{\Gamma}^{m})^{\mathbf{1}(S_u = m)} \\
&\quad = \prod_{u=1}^{U} \Bigg(\delta^{m}_{r_{u,1}} \; \phi_D(\boldsymbol{x}_{u,1} | \boldsymbol{\mu}^{m,r_{u,1}}, \boldsymbol{\Sigma}^{m,r_{u,1}}) \; \times \\
&\qquad \prod_{t=2}^{T_u} \Pr(R_{u,t} = r_{u,t} \mid R_{u,t-1} = r_{u,t-1}) \; \phi_D(\boldsymbol{X}_{u,t} | \boldsymbol{\mu}^{m,r_{u,t}}, \boldsymbol{\Sigma}^{m,r_{u,t}}) \Bigg)^{\mathbf{1}(S_u = m)} \\
&\quad = \prod_{u=1}^{U} \Bigg(\prod_{k=1}^{K} \left(\delta^{m}_{k} \; \phi_D(\boldsymbol{x}_{u,1} | \boldsymbol{\mu}^{m,k}, \boldsymbol{\Sigma}^{m,k}) \right)^{\mathbf{1}(R_{u,1} = k)} \; \times \\
&\qquad \prod_{t=2}^{T_u} \left(\prod_{k=1}^{K} \left(\prod_{k'=1}^{K} \left(\Gamma^{m}_{k,k'} \right)^{\mathbf{1}\{R_{u,t}=k', R_{u,t-1}=k'\}} \; \phi_D(\boldsymbol{X}_{u,t} | \boldsymbol{\mu}^{m,k}, \boldsymbol{\Sigma}^{m,k})^{\mathbf{1}(R_{u,t}=k)} \right) \right) \right) \Bigg)^{\mathbf{1}(S_u = m)},
\end{aligned}
\tag{2}
$$

1.2.1 E step

This procedure relies heavily on the joint probability of (i) all feature vectors up until time t and (ii) the sound at t, given in equation 3. These probabilities are efficiently calculated

for all t in a single recursive forward pass through the feature vectors.

$$\begin{aligned}
\alpha_{u,t,k} &= f(\boldsymbol{X}_{u,1} = \boldsymbol{x}_{u,1}, \cdots, \boldsymbol{X}_{u,t} = \boldsymbol{x}_{u,t}, R_{u,t} = k) \\
\boldsymbol{\alpha}_{u,t} &= [\alpha_{u,t,1}, \ldots, \alpha_{u,t,K}] \\
&= \delta_u^\top \boldsymbol{P}^m(\boldsymbol{x}_{u,1}) \left(\prod_{t'=2}^{t} \boldsymbol{\Gamma}^m \boldsymbol{P}^m(x_{u,t'}) \right)
\end{aligned} \tag{3}$$

It also relies on the conditional probability of (i) all feature vectors after t given (ii) the sound at t (equation 4). These are similarly calculated by backward recursion through the utterance.

$$\begin{aligned}
\beta_{u,t,k} &= f(\boldsymbol{X}_{u,t+1} = \boldsymbol{x}_{u,t+1}, \cdots, \boldsymbol{X}_{u,T_u} = \boldsymbol{x}_{u,T_u} \mid R_{u,t} = k) \\
\boldsymbol{\beta}_{u,t} &= [\beta_{u,t,1}, \ldots, \beta_{u,t,K}]^\top \\
&= \left(\prod_{t'=t+1}^{T_u} \boldsymbol{\Gamma}^m \boldsymbol{P}^m(\boldsymbol{x}_{u,t'}) \right) \mathbf{1}
\end{aligned} \tag{4}$$

The E step involves substituting (i) the unobserved sound labels, $\mathbf{1}(R_{u,t} = k)$, and (ii) the unobserved sound transitions, $\mathbf{1}(R_{u,t} = k', R_{u,t-1} = k)$, with their respective expected values, conditional on the observed training features $\boldsymbol{X}_u$ and the current estimates of $\boldsymbol{\Theta}^m = (\boldsymbol{\mu}^{m,k}, \boldsymbol{\Sigma}^{m,k}, \boldsymbol{\Gamma}^m)$.

For (i), combining equations 1, 3 and 4 immediately yields the expected sound label

$$\mathbb{E}\left[\mathbf{1}(R_{u,t} = k) \mid \boldsymbol{X}_u, S_u = m, \tilde{\boldsymbol{\Theta}} \right] \propto \tilde{\alpha}_{u,t,k} \; \tilde{\beta}_{u,t,k}, \tag{5}$$

where the tilde denotes the current approximation based on parameters from the previous

M step; $\alpha_{u,t,k}$ and $\beta_{u,t,k}$ are the k-th elements of $\boldsymbol{\alpha}_{u,t}$ and $\boldsymbol{\beta}_{u,t}$ respectively; and $\tilde{\mathcal{L}}_u^m$ is the u-th training utterance's contribution to $\tilde{\mathcal{L}}^m$.

For (ii), after some manipulation, the expected sound transitions can be expressed as

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}(R_{u,t} = k', R_{u,t-1} = k) \mid \boldsymbol{X}_u, S_u = m, \tilde{\boldsymbol{\Theta}}] \\
&\quad = \Pr(R_{u,t} = k', R_{u,t-1} = k, \boldsymbol{X}_u \mid \tilde{\boldsymbol{\Theta}}) \ / \ \Pr(\boldsymbol{X}_u \mid \tilde{\boldsymbol{\Theta}}) \\
&\quad = \Pr(\boldsymbol{X}_{u,1}, \cdots, \boldsymbol{X}_{u,t-1}, R_{u,t-1} = k \mid \tilde{\boldsymbol{\Theta}}) \ \Pr(R_{u,t} = k' \mid R_{u,t-1} = k, \tilde{\boldsymbol{\Theta}}) \times \\
&\qquad \Pr(\boldsymbol{X}_{u,t} \mid R_{u,t} = k') \ \Pr(\boldsymbol{X}_{u,t+1}, \cdots, \boldsymbol{X}_{u,T_u} \mid R_{u,t} = k') \ / \ \Pr(\boldsymbol{X}_u \mid \tilde{\boldsymbol{\Theta}}) \\
&\quad \propto \tilde{\alpha}_{u,t-1,k} \ \tilde{\Gamma}^m_{k,k'} \ \phi_D(\boldsymbol{x}_{u,t} | \tilde{\boldsymbol{\mu}}^{m,k}, \tilde{\boldsymbol{\Sigma}}^{m,k}) \ \beta_{u,t,k'}.
\end{aligned} \tag{6}$$

1.2.2 M Step

After substituting equations 5 and 6 into the complete-data likelihood (equation 2), the M step involves two straightforward calculations. First, the conditional maximum likelihood update of the transition matrix $\boldsymbol{\Gamma}^m$ follows from equation 6:

$$\tilde{\Gamma}^m_{k,k'} = \frac{\sum_{1=1}^{U} \mathbf{1}(S_u = m) \sum_{t=2}^{T_u} \mathbb{E}\left[\mathbf{1}(R_{u,t} = k', R_{u,t-1} = k) \mid \boldsymbol{X}_u, \tilde{\boldsymbol{\Theta}}\right]}{\sum_{1=1}^{U} \mathbf{1}(S_u = m) \sum_{t=2}^{T_u} \sum_{k'=1}^{K} \mathbb{E}\left[\mathbf{1}(R_{u,t} = k', R_{u,t-1} = k) \mid \boldsymbol{X}_u, \tilde{\boldsymbol{\Theta}}\right]} \tag{7}$$

Second, the optimal update of the k-th sound distribution parameters are found by fitting a Gaussian distribution to the feature vectors, with the weight of the t-th instant being given by the expected value of its k-th label.

$$\tilde{\Gamma}^m_{k,k'} = \frac{\sum_{u=1}^{U} \mathbf{1}(S_u = m) \sum_{t=2}^{T_u} \mathbb{E}\left[\mathbf{1}(R_{u,t} = k', R_{u,t-1} = k) \mid \boldsymbol{X}_u, \tilde{\boldsymbol{\Theta}}\right]}{\sum_{u=1}^{U} \mathbf{1}(S_u = m) \sum_{t=2}^{T_u} \sum_{k'=1}^{K} \mathbb{E}\left[\mathbf{1}(R_{u,t} = k', R_{u,t-1} = k) \mid \boldsymbol{X}_u, \tilde{\boldsymbol{\Theta}}\right]} \tag{8}$$

$$\tilde{\boldsymbol{\mu}}^{m,k} = \sum_{u=1}^{U} \mathbf{1}(S_u = m) \boldsymbol{X}_u^\top \boldsymbol{W}_u^{m,k} \tag{9}$$

$$\tilde{\boldsymbol{\Sigma}}^{m,k} = \sum_{u=1}^{U} \mathbf{1}(S_u = m) \left(\boldsymbol{X}_u^\top \mathrm{diag}\left(\boldsymbol{W}_u^{m,k}\right) \boldsymbol{X}_u\right) - \tilde{\boldsymbol{\mu}}^{m,k} \tilde{\boldsymbol{\mu}}^{m,k\ \top} \tag{10}$$

$$\text{where } \boldsymbol{W}_u^{m,k} \equiv \frac{\sum_{u=1}^{U} \mathbf{1}(S_u = m) \left[\mathbb{E}\left[\mathbf{1}(R_{u,1} = k) \mid \boldsymbol{X}_u, \boldsymbol{\Theta}\right], \cdots, \mathbb{E}\left[\mathbf{1}(R_{u,T_u} = k) \mid \boldsymbol{X}_u, \boldsymbol{\Theta}\right]\right]^\top}{\sum_{u=1}^{U} \mathbf{1}(S_u = m) \sum_{t=1}^{T_u} \mathbb{E}\left[\mathbf{1}(R_{u,t} = k) \mid \boldsymbol{X}_u, \boldsymbol{\Theta}\right]}$$

1.3 Unmodeled Autocorrelation

If the Gaussian HMM model of speech described in Equations 3–4 were correctly specified, then the tone of any new utterance could be classified with well-calibrated posterior probabilities based on its auditory characteristics (setting aside conversation context) by the simple application of Bayes' rule, $\Pr(S_u = m|\boldsymbol{X}_u, \boldsymbol{\Theta}) = \frac{\Pr(\boldsymbol{X}_u|S_u=m,\boldsymbol{\Theta}) \Pr(S_u=m)}{\sum_{m'=1}^{M} \Pr(\boldsymbol{X}_u|S_u=m',\boldsymbol{\Theta}) \Pr(S_u=m')}$, where $\Pr(\boldsymbol{X}_u|S_u = m, \boldsymbol{\Theta}) = \delta^{m\top} \boldsymbol{P}^m(\boldsymbol{x}_{u,1}) \left(\prod_{t=2}^{T_u} \boldsymbol{\Gamma}^m \boldsymbol{P}^m(\boldsymbol{x}_{u,t})\right) \mathbf{1}$ as in Appendix 1.2.

However, this speech model—like all simplified models of complex human behavior—is misspecified, with implications for its resulting predictions. In particular, our model assumes that the auditory features in successive moments are conditionally independent, given their respective sounds. This can be seen by noting that $\boldsymbol{X}_{u,1}$ and $\boldsymbol{X}_{u,2}$ are d-separated by $R_{u,1}$ and $R_{u,2}$ in Figure 2. In other words, the expected difference in audio between moment t and $t+1$ should be no greater than the difference between t and $t+10$, as long as a vowel is being spoken.

This assumption makes the model analytically tractable, much as the bag-of-words as-

sumption facilitates text analysis. Like the bag-of-words assumption, it is also clearly violated by actual human behavior. A speaker's vocal tract is physically incapable of changing much in a few milliseconds, but this autocorrelation in features goes unmodeled. Thus, the model mistakenly perceives the information content of an utterance to be T_u data points, when in fact it may be much less. The practical implication is that mode probabilities produced by the aforementioned approach will drift toward zero and one, leading to dramatic miscalibration. To address this issue, we use a corrective factor, $\left(\delta^{m\top}\boldsymbol{P}^m(\boldsymbol{x}_{u,1})\left(\prod_{t=2}^{T_u}\boldsymbol{\Gamma}^m\boldsymbol{P}^m(\boldsymbol{x}_{u,t})\right)\mathbf{1}\right)^{\rho}$. This scales back the utterance's contribution to the log likelihood multiplicatively, reducing the utterance's "effective value" to ρT_u. The corrective factor is estimated from out-of-sample data by maximizing the total log corrected probabilities of the correct class.

1.4 Estimation of Upper-Level Conversation Parameters

We now describe our procedure for estimating the conversation flow parameters by maximizing the observed-data likelihood of Equation 5 with respect to $\boldsymbol{\zeta}$, which amounts to maximizing $f(\boldsymbol{X}^{\mathcal{C}} \mid \boldsymbol{\zeta}, \boldsymbol{\Theta}, \boldsymbol{W}^{\text{stat.},\mathcal{C}})$. This is equivalent to estimating both the unobserved $\boldsymbol{S}^{\mathcal{C}}$ and parameters $\boldsymbol{\zeta}$ by maximizing the expected complete-data log likelihood. (All analysis in this subsection is of the primary corpus, so we drop the $\mathcal{C}$ indicator for compactness.) For complete generality, we also introduce a conversation index $v \in \{1, \ldots, V\}$. The number of utterances in conversation v is denoted U_v; metadata, speech modes and audio features for utterance u in conversation v are respectively $\boldsymbol{W}_{v,u}$, $S_{v,u}$ and $\boldsymbol{X}_{v,u}$.

First, the complete-data likelihood of the primary corpus is

$$
\begin{aligned}
&\ln f(\boldsymbol{X}, \boldsymbol{S} \mid \boldsymbol{\zeta}, \boldsymbol{\Theta}, \boldsymbol{W}^{\text{stat.}}) \\
&= \ln \left(\prod_{v=1}^{V} \delta_{v,S_{v,1}} f(\boldsymbol{x}_{v,1} | S_{v,1} = s_1, \boldsymbol{\Theta}) \prod_{u=2}^{U_v} \Pr(S_{v,u} = s_{v,u} | S_{v,u-1} = s_{v,u-1}) f(\boldsymbol{x}_{v,u} | S_{v,u} = s_{v,u}, \boldsymbol{\Theta}) \right) \\
&= \sum_{v=1}^{V} \sum_{m=1}^{M} \ln \delta_{v,m}^{\mathbf{1}(S_{v,1}=m)} + \sum_{v=1}^{V} \sum_{u=1}^{U_v} \sum_{m=1}^{M} \ln f(\boldsymbol{x}_{v,u} | S_{v,u} = m, \boldsymbol{\Theta}^m)^{\mathbf{1}(S_{v,1}=m)} \\
&\quad + \sum_{v=1}^{V} \sum_{u=2}^{U_v} \sum_{m=1}^{M} \sum_{m'=1}^{M} \Delta_{v,u,m,m'}^{\mathbf{1}(S_{v,u-1}=m, S_{v,u}=m')} \\
&= \sum_{v=1}^{V} \sum_{m=1}^{M} \mathbf{1}(S_{v,1} = m) \ln \delta_{v,m} + \sum_{v=1}^{V} \sum_{u=1}^{U_v} \sum_{m=1}^{M} \mathbf{1}(S_{v,1} = m) \ln f(\boldsymbol{x}_{v,u} | S_{v,u} = m, \boldsymbol{\Theta}^m) \\
&\quad + \sum_{v=1}^{V} \sum_{u=2}^{U_v} \sum_{m=1}^{M} \sum_{m'=1}^{M} \mathbf{1}(S_{v,u-1} = m, S_{v,u} = m') \ln \frac{\exp\left(\boldsymbol{W}_{v,u}(\boldsymbol{S}_{v,u'<u})^\top \boldsymbol{\zeta}_m\right)}{\sum_{m'=1}^{M} \exp\left(\boldsymbol{W}_{v,u}(\boldsymbol{S}_{v,u'<u})^\top \boldsymbol{\zeta}_{m'}\right)},
\end{aligned}
$$

where $\boldsymbol{W}_{v,u}(\boldsymbol{S}_{v,u'<u}) = [\boldsymbol{W}_{v,u}^{\text{stat.}\top}, \boldsymbol{W}_{v,u}^{\text{dyn.}}(\boldsymbol{S}_{v,u'<u})^\top]^\top$. $\boldsymbol{\delta}_v$ indicates the initial distribution of speech modes for conversation v.

Because the time-varying transition matrix, $\boldsymbol{\Delta}_{v,u}$, is a multinomial logistic function of conversation context, $\boldsymbol{W}_{v,u}$—which is itself a potentially complex function of unobserved prior speech modes—deriving the closed-form expectation of the complete-data likelihood is intractable. We therefore replace this expectation with the following blockwise procedure that sweeps through the unobserved variables sequentially.

1. The metadata $\boldsymbol{W}_{v,u}$ depends on conversation history, but the previous mode is unobserved. Therefore, for each utterance, we create a separate metadata vector for each possible prior mode. (This is computationally infeasible for longer-range summaries of conversation history e.g., aggregate anger expressed over the course of a debate, so we recommend a mean-field approximation for dynamic metadata based

on utterances older than $u-1$.) This step produces M possible metadata vectors, $\tilde{\boldsymbol{W}}_{v,u}(\tilde{\mathbb{E}}[\boldsymbol{S}_{v,u'<u-1}], S_{u-1}=1)$ through $\tilde{\boldsymbol{W}}_{v,u}(\tilde{\mathbb{E}}[\boldsymbol{S}_{v,u'<u-1}], S_{u-1}=M)$.

2. Each possible metadata vector implies a vector of probabilities for the next utterance, $\tilde{\boldsymbol{\Delta}}_m = [\tilde{\Pr}(S_u=1|S_{u-1}=m), \ldots, \tilde{\Pr}(S_u=M|S_{u-1}=m)] = \frac{\exp\left(\tilde{\boldsymbol{W}}_u(\tilde{\mathbb{E}}[\boldsymbol{S}_{v,u'<u-1}], S_{u-1}=m)^\top \tilde{\boldsymbol{\zeta}}_m\right)}{\sum_{m'=1}^{M}\exp\left(\tilde{\boldsymbol{W}}_u(\tilde{\mathbb{E}}[\boldsymbol{S}_{v,u'<u-1}], S_{u-1}=m)^\top \tilde{\boldsymbol{\zeta}}_{m'}\right)}$. Stack these into a transition matrix, $\tilde{\boldsymbol{\Delta}}$.

3. Compute $\tilde{\mathbb{E}}[\mathbf{1}(S_{v,u}=m)]$ and $\tilde{\mathbb{E}}[\mathbf{1}(S_{v,u-1}=m, S_{v,u}=m')]$, using a forward-backward algorithm that is essentially identical to Equations 5 and 6. We find that the use of the corrected emission probabilities, described in Appendix 1.3, is crucial in this step.

Again, tildes indicate the best guess for each variable at the current iteration. The maximization step for $\boldsymbol{\zeta}$ then reduces to weighted constrained multinomial logistic regression in which all possible transitions are included, weighted by $\tilde{\mathbb{E}}[\mathbf{1}(S_{v,u-1}=m, S_{v,u}=m')]$. A constraint on the mode-specific intercepts ensures that the fitted probabilities agree with the known tone proportions; this is implemented by first computing the relaxed update for $\boldsymbol{\zeta}$ in each iteration, then imposing the constraint. The estimated initial mode, $\boldsymbol{\delta}_v$ follows directly from the expected value of $[\mathbf{1}(S_{v,1}=m)]$. All in all, the use of this alternative procedure leads to a smaller improvement of the EM objective function than the full (infeasible) E-step would. Nevertheless, algorithms using such partial E- or M-steps ultimately converge to a local maximum, as does traditional expectation-maximization (Neal and Hinton 1998).

1.5 Bootstrapping

Because each bootstrapped speech-mode model's parameters only enter the upper model through how well or poorly they explain a particular utterance's observed auditory features,

the upper model is unaffected by likelihood invariance issues such as the label-switching problem. However, to the extent that some bootstrapped model runs are trapped in local modes and do not attain the global optimum, resulting upper-level confidence intervals will be wider (that is, more conservative), reflecting both true uncertainty and the additional random variation in the selected local mode. This pitfall may be addressed by standard optimization procedures such as simulated-annealing EM or running multiple chains.

2 Audio Features

Table 1 lists the primary features we calculate for each utterance. In addition, we calculate interactions between and derivatives of these primary features.

Feature (#)	**Description**
energy (1)	sound intensity, in decibels: $\log_{10}\sqrt{\overline{x_t^2}}$
ZCR (1)	zero-crossing rate of audio signal
autocorrelation (1)	$\text{Cor}(x_t, x_{t-1})$
TEO (1)	Teager energy operator: $\log_{10} \overline{x_t^2 - x_{t-1}x_{t+1}}$
F0 (2)	fundamental, or lowest, dominant frequency of speech signal (closely related to perceived pitch; tracked by two algorithms)
formants (6)	harmonic frequencies of speech signal, determined by shape of vocal tract (lowest three formants and their bandwidths)
MFCC (13)	Mel-frequency cepstral coefficients (characterizing the shape of the frequency spectrum, after transforming and binning the spectrum to approximate human perception of sound intensity)

Table 1: **Audio features extracted in each frame.** Parenthesized values indicate the number of scalars extracted per moment. We also include interactions between (*i*) energy and zero-crossing rate, and (*ii*) Teager energy operator and fundamental frequency, for a total of 27 primary features. In addition, first and second finite differences are often informative. For example, vocal jitter and shimmer are respectively described by the first differences in F0 and energy.

3 Case Study of *Alabama Legislative Black Caucus v. Alabama*

Here, we illustrate the model using excerpts from *Alabama Legislative Black Caucus v. Alabama*, a racial gerrymandering case heard by the Supreme Court in 2014.[2] While this example represents only a small portion from a single case, it demonstrates many of the conversation dynamics that motivate our model. We begin by discussing the legal question and positions of the justices, then walk through instances of information-seeking questions, skeptical attacks on the opposing side, and defensive interventions. We then show how MASS parameters map onto the primary theoretical quantities of interest.

As background, *Alabama Legislative Black Caucus v. Alabama* considered the legality of Alabama's 2012 redistricting efforts. The plan came after the 2010 census found substantial population decline in state legislative districts with a majority of black voters, necessitating the expansion of these districts' boundaries. (Reapportionment was required to comply with *Reynolds v. Sims*—yet another decision against the state of Alabama—which ruled overrepresentation of rural, predominantly white, voters in the Alabama state legislature unconstitutional under the Fourteenth Amendment's equal protection clause and the "one person, one vote" principle.) In response to these population shifts, the Republican-led legislature sought to pack black voters into a small number of already Democratic-dominated districts—for example, 14,500 people were added to State Senate District 26, of whom only 35 were non-black. Ultimately, the court ruled that the use of race as a "predominant" factor, even when only applied to a subset of districts rather than statewide, constituted

[2]The full argument is available at `https://www.oyez.org/cases/2014/13-895`, along with background on the case, the ruling, and dissents.

illegal racial gerrymandering.

In what follows, we consider legal jockeying in oral arguments over a contentious and highly consequential debate: Whether Section 5 of the Voting Rights Act (VRA), prohibiting retrogression in minorities' "ability to elect their preferred candidates," meant that Alabama had to continually maintain or increase the numerical percentage of black voters in black-dominated districts. If so, the state's consideration of race would be "narrowly tailored" to meeting its VRA obligations, and thus legal.[3] We focus in particular on questioning by Justices Breyer and Scalia, who respectively wrote the majority and dissenting opinions, as well as by Justice Kennedy, who cast the pivotal vote.

Panel 1 in Figure 1 presents a condensed transcript of one instance when this issue arose during arguments by a liberal advocate representing the Alabama Democratic Conference. Early on, Justice Scalia takes the position that the state was legally bound to maintain or increase black percentages. His stance was far from novel, as it had already been discussed extensively in briefs and lower-court decisions available to all justices. But Justice Scalia repeats the point nonetheless, questioning the liberal advocate not only skeptically,

[3]The ruling concluded that the Republican legislature "relied heavily upon a mechanically numerical view as to what counts as forbidden retrogression... And the difference between that view and the more purpose-oriented view reflected in the statute's language can matter. Imagine a majority-minority district with a 70% black population... it would seem highly unlikely that... reduc[ing] the percentage of the black population from, say, 70% to 65% would have a significant impact on the black voters' ability to elect their preferred candidate. And, for that reason, it would be difficult to explain just why a plan that uses racial criteria predominately to maintain the black population at 70% is "narrowly tailored" to achieve a "compelling state interest," namely the interest in preventing Section 5 retrogression.

but sarcastically—theatrically drawing out his words and even exclaiming "gee." The ploy appears to be effective. Justice Kennedy follows up on the topic, skeptically wondering why it was legal for Democrats to disperse black voters, but not for Republicans to concentrate them: a "one-way ratchet." Sensing a threat, Justice Breyer attempts to smooth things over with a matter-of-fact legal analysis of *Easley v. Cromartie*.[4] Again, the discussion hardly contained new information. In briefs by both the Alabama Black Legislative Caucus and the state of Alabama, 167 pages were devoted to analysis relating to *Easley v. Cromartie*. And more to the point, five of the nine justices had been serving on the Supreme Court when that very case was decided there in 2001.

In contrast, Panel 1 depicts an exchange—in the very same case—where the roles are reversed during questioning of the conservative advocate. Here, Justice Kennedy again seeks to clarify whether the legislature's consideration of race was a permissible attempt to comply with VRA obligations. Justice Breyer attacks, asserting that since Alabama's actions were indefensible under Section 5, "I don't know what the defense is *possibly* going to be." He seizes the opportunity to push a step further, suggesting that the Republican legislature has no case and should give up—prompting Justice Scalia to wade into the exchange defensively.

These excerpts provide a clear illustration of conversational flow: how one speaker's communication causes a subsequent speaker to communicate differently in response. In Justice Sotomayer's 2019 words, Justices Scalia and Breyer are "raising points through the questions that we want our colleagues to consider," then intervening in response to one

[4] *Easley v. Cromartie* ruled that the burden of proof is on the complainant, who must show "legislature could have achieved its legitimate political objectives in alternative [non-racial] ways."

another. In this section, we show how MASS can detect systemic patterns in speech patterns like these, allowing analysts to move beyond isolated anecdotes and test theories about oral argumentation using justices' expressions of skepticism.

To demonstrate how the MASS is able to do this, in Figure 1 (duplicated here for convenience) we turn to a close examination of two prototypical utterances by Justice Breyer. We first discuss the sounds of which each utterance is composed, along with their auditory profiles. Consider Justice Breyer's skeptical mode of speech—the tone in which he rhetorically exclaims "Now if *that's* so, they don't have *Section 5* to rely on as a *defense*!" He communicates through a sequence of sounds that, simplistically, we might categorize into "vowel," "consonant," and "silence."[5] In Panel 1, we show that our generative model of skeptical speech mirrors this structure: Vowels (dark red) are sustained for a few moments (horizontally arrayed cells) before Justice Breyer transitions to consonants (light red strikethrough) and eventually pauses in silence (white) between words[6]. One such transition is depicted in Figure 1.D.2. Just as a human can recognize phonemes from their auditory characteristics, our model automatically learns to distinguish vowels (based on their higher autocorrelation, as encoded in $\boldsymbol{\mu}^{\text{skeptical,vowel}}$) from consonants (high zero-crossing rate), as shown in Figure 1.

[5]We note that sound labels, like topic labels in latent Dirichlet allocation text models, are subjective descriptions of component distributions in unsupervised learning models. However, human speech is highly structured. Across a wide range of applications, we consistently find that HMMs recover states that correspond closely to theoretically motivated phoneme groups.

[6]Because each frame describes just milliseconds of audio, glottal stops and short pauses between words are an observable component of speech.

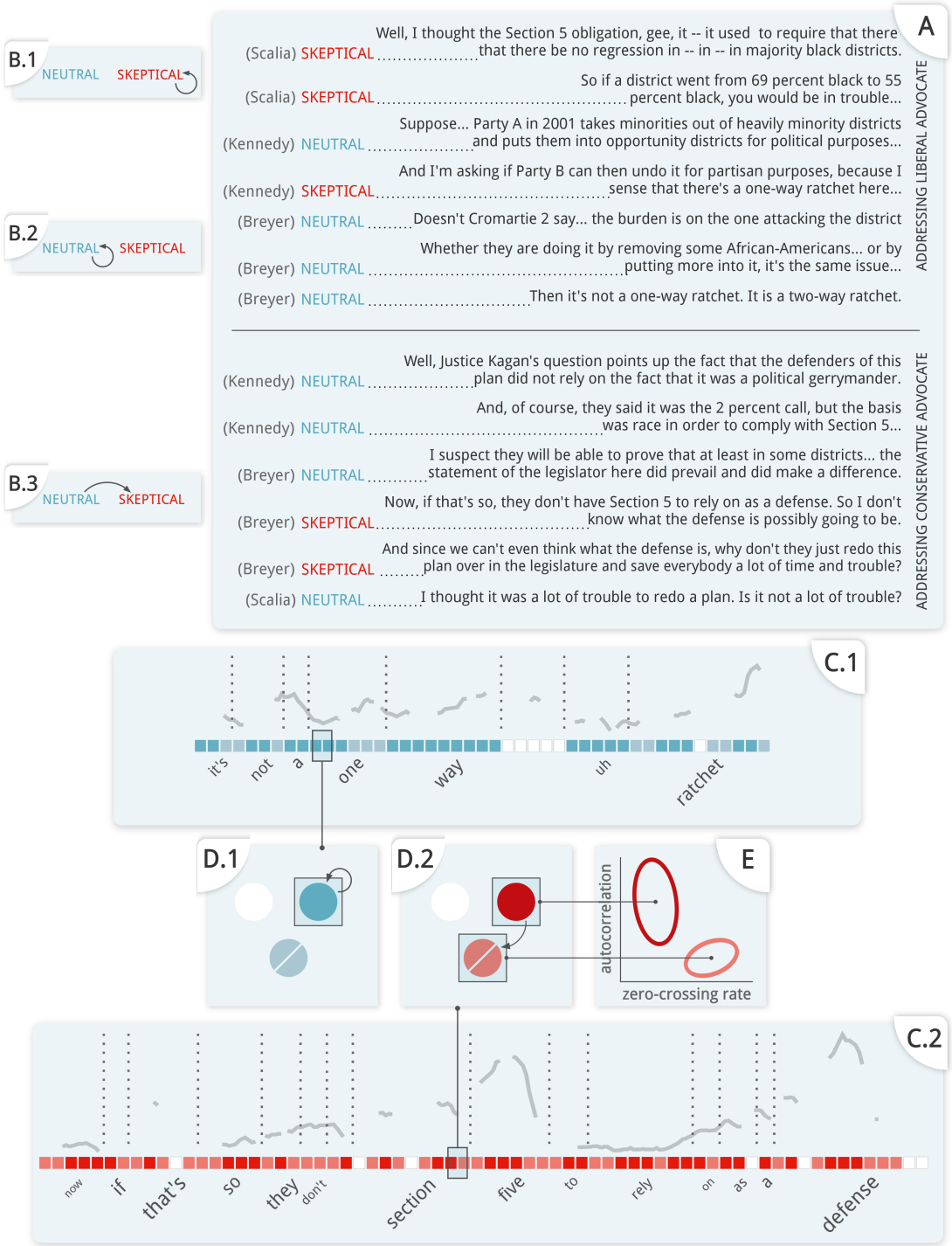

Figure 1: **An illustrative example.** Panel A contains an excerpt from Alabama Legislative Black Caucus v. Alabama, where Justices Scalia, Kennedy, and Breyer utilize neutral and skeptical tones in questioning. Call-outs highlight successive utterance-pairs in which the speaker shifted from one mode to another (B.3), and continued in the same tone of voice (B.1 and B.2). Panels C.1 and C.2 illustrate the use of loudness (text size) and pitch (contours) in a single utterance: in the neutral mode of speech (C.1), speech varies less in pitch and loudness when compared to skeptical speech (C.2). On the basis of these and other features, MASS learns to categorize sounds into vowels (dark squares), consonants (light), and pauses (white). Call-outs D.1 and D.2 respectively identify sequential frames in which a "neutral" vowel is sustained (transition from the dark blue sound back to itself, indicating repeat) and the dark red "skeptical" vowel transitions to the light red consonant. Panel E shows the differing auditory characteristics of the "skeptical" vowel and consonant, which are perceived by the listener.

Why does this matter? It is on the basis of these constituent sounds that MASS is able to discern differences between rhetorical styles. As Figure 1.A makes clear, MASS contains a parallel "neutral" model for Justice Breyer's speech alongside the "skeptical" model. While neutral speech also uses vowels and consonants, the auditory profiles of these sounds differ dramatically. Figure 1.C.2 (in which word size reflects decibel-scale energy) demonstrates Breyer's use of modulation for emphasis when advancing his argument ("if *that's* so...") and his soaring pitch when expressing incredulity and exasperation ("don't have... a *defense*!") Thus, $\boldsymbol{\Sigma}^{\text{skeptical,vowel}}$ captures higher variance in loudness and pitch when compared to neutral speech (Figure 1.C.1), where every word is delivered at near-constant volume and relatively flat pitch. Differences in average pitch—often a marker of emotional engagement—are represented in the $\boldsymbol{\mu}$ terms. Finally, shifts in cadence, like when Breyer briefly loses his train of thought before continuing "uh... ratchet", manifest in the $\boldsymbol{\Gamma}$ matrices.

These models of skeptical and neutral speech enable analysts to categorize hundreds or even thousands of hours of previously unheard speech. But learning to recognize skeptical speech is only the beginning for MASS. The most important questions in the analysis of political speech relate to its ebb and flow—when and why a speaker chooses to deploy a particular tone. After learning to distinguish tone in the lower stage (Equations 3–4), MASS moves on to model the entire Supreme Court's conversational flow by estimating the contextual determinants of speech tone (Equations 1–4). While scholars can easily listen to and compare a few short audio recordings, the amount of time required to digest an entire session's worth of argumentation—dozens of cases, each containing hundreds of utterances—rapidly grows infeasible. MASS makes it possible to identify broad patterns in the drivers of political speech, analyzing large-scale audio corpora while still incorporating human judg-

ment about tone and expressed emotion. In Section4.2, we develop a procedure for doing so; Algorithm 1 describes the steps in detail. Broadly speaking, the model learns to identify micro-level patterns, such as those described above, based on a moderately sized training set of human-provided examples. MASS then uses this knowledge to crudely categorize every utterance spoken. Finally, based on their sequence and contextual covariates, MASS identifies patterns in tone usage, then uses these patterns to iteratively refine its utterance predictions and the flow-of-speech parameters.

4 Validating the Model

4.1 Facial Validity of Predicted Skepticism

Before proceeding to more substantial results, we first demonstrate the face validity of MASS predictions in a qualitative examination of machine-generated utterance labels. Table 2 presents twenty example utterances that lie in the top decile of predicted skepticism and neutrality. Results suggest high face validity: utterances characterized by the model as skeptical include gentle mockery and doubtful questions, whereas model-predicted neutral utterances are factual statements and straightforward legal analysis.

Skeptical Speech	**Neutral Speech**
Well, I mean, you don't know; you're running away.	You would not be subject to the State suit.
You've – you've given us no – no principle the other way.	The one that has the 5-year clearly covers the situation.
So, I guess they could object on the ground that model is worthless.	But the Authorities Law does authorize the acquisition of other hospitals.
I mean, of course they would be thinking about that; that was the issue.	And the SEC apparently takes the view that this provision does cover contractors.
Next step, he goes to the grand jury or someone and says: Jones stole my horse.	But they can't do that because the statute requires a summary to be understandable and not prolix.
Counsel, it hasn't been the focus of the briefing, but you've just made it the focus here.	The ball goes back to Congress to do what it will, but it's just, in the interim, we need a solution.
Does that make any sense, given the – the class of individuals who are plaintiffs in 16(b) cases?	That sounds much more petition-like than filing a grievance pursuant to a collective bargaining agreement.
What would happen, under the reasoning of this case, what would happen to the decisions of recess-appointed judges?	Let's suppose that the district court in Washington moves expeditiously and issues a decision in mid-February.
Now, that it seems to me could include everything from a spark plug that is deficient in the airplane to a terrorist.	And the other choice is to say that "lawfully made" means it's made without contravening any provision of the Act, if the Act were applicable.
Seriously, the unions do not want to have the – they don't want to be given the status of the exclusive bargaining agent for the employees?	So leaving the language out of it, I would like you to respond to what I would call that purpose-related, fact-related argument by these particular people.

Table 2: **Transcripts of Skeptical and Neutral Utterances.** Left (right) columns contain ten transcripts of utterances in the top 10% of predicted skepticism (neutrality). While MASS is estimated solely on audio data and conversation context, its fitted values accord well with qualitative readings of the utterance text. The results suggest high face validity: Those characterized by the model as skeptical include gentle mockery and doubtful questions, whereas model-predicted neutral utterances are factual statements and straightforward legal analysis.

4.2 Textual Characteristics of Expressed Skepticism

Results from Section 4.1, which suggest that humans such as the reader (presumably) can validate model-predicted skepticism using utterance text—in extreme cases, at the least—indicate that auditory channel carries emotional information that can be detected by MASS. But they also suggest that skepticism is partially conveyed through textual channels as well. Could tone be extracted directly from the text without the need for complex audio models? To assess whether the auditory channel in fact conveys new information or is merely duplicative, we attempted to predict expressed skepticism using utterance transcripts. For each utterance, word counts were computed after stemming, stopping, and pruning words that appeared in fewer than ten utterances. A cross-validated elastic net was then applied to the utterance-term matrix, producing a maximum accuracy of 59.8%. Moreover, the textual classifier was only able to achieve this accuracy by predicting the dominant class (neutral speech, 59.4% of labeled utterances) for virtually every observation. Additional measures of classification performance, including for within-speaker classification, are reported in Appendix 4.4.

Next, to rule out the possibility that the roughly 1,600 hand-labeled utterances were too small of a training corpus, we analyze the full corpus. To do so, we treat MASS fitted probabilities of skepticism (based on audio features and conversation context) as the outcome. We then employ a post-LASSO procedure in which a cross-validated LASSO-logistic model is estimated, then an unregularized logistic regression is fit on the selected terms (Belloni, Chernozhukov, and Wei 2016).

The resulting coefficient estimates, plotted in Figure 2, demonstrate that there are ex-

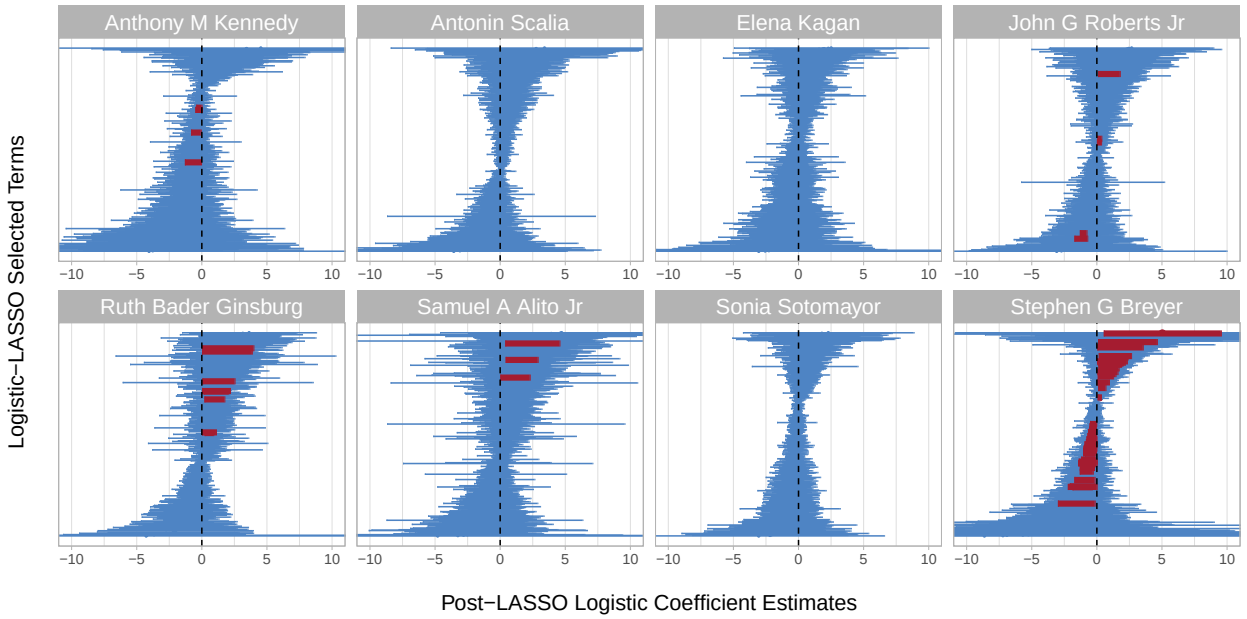

Figure 2: **Textual Signals of Justice Skepticism.** Each panel depicts a regression of MASS-predicted skepticism on word counts. Within each justice's utterances, candidate terms that may predict skepticism (selected by logistic LASSO) are arrayed on the y-axis. For each term, points and horizontal error bars depict post-LASSO logistic regression estimates and confidence intervals. Thin light blue (thick dark red) error bars reflect 95% confidence intervals that (do not) overlap zero.

traordinarily few consistent textual indicators of expressed skepticism—the vast majority are statistically indistinguishable from zero at conventional levels. In Figure 3, we arbitrarily discard speaker-terms with p-values exceeding 0.05, then investigate the remainder more closely.

For Justice Stephen Breyer, an expressive orator who is by far the most frequently speaking justice, less than 50 such terms exist. For illustrative purposes, we focus on Breyer's "ah", "block", and "lost," the three terms most heavily associated with his predicted skepticism. While these terms are not obviously associated with negative sentiments, a closer examination sheds light on Breyer's usage in his freewheeling and at times theatrical questioning:

- A sarcastic retort to an attempt to introduce new arguments, "*Ah*, now we have 'suf-

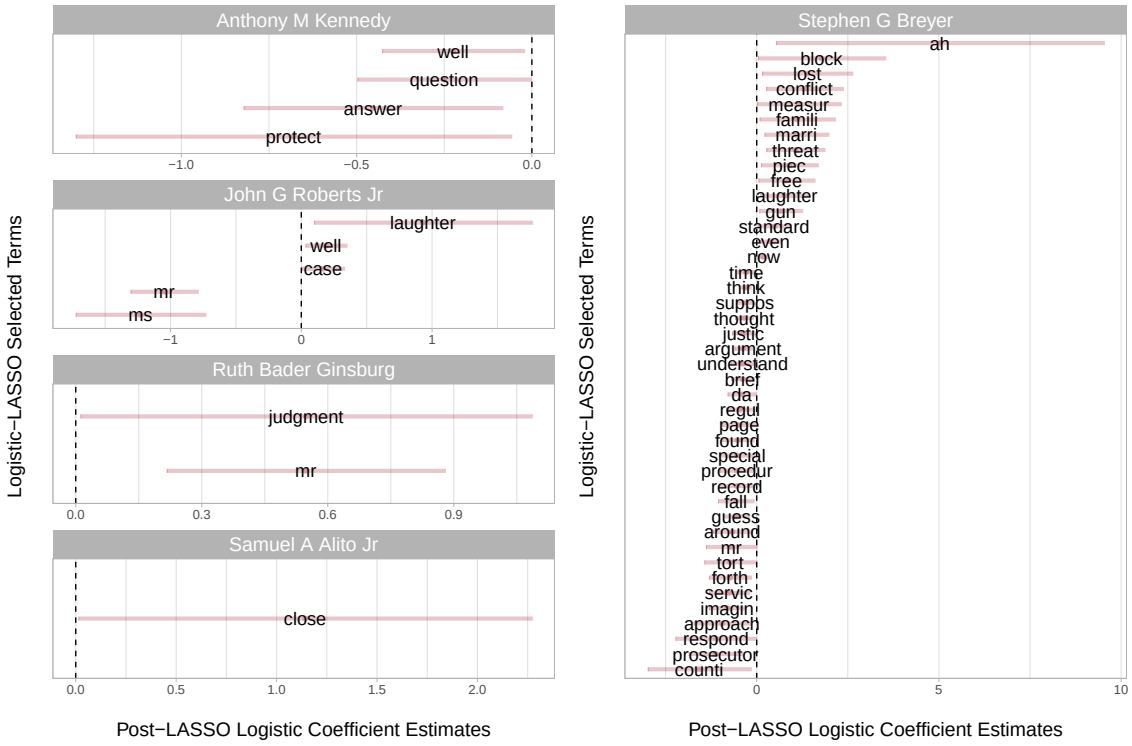

Figure 3: **Strong Textual Signals of Justice Skepticism.** Each panel depicts a regression of MASS-predicted skepticism on word counts, within a justice's utterances. Words that predict skepticism are arrayed on the y-axis. Reported terms are the subset of post-LASSO terms with post-selection logistic regression confidence intervals (error bars) that do not overlap zero. Highly specific terms (i.e., used in fewer than five cases) are not depicted.

ficiently involved;' "

- A direct legal attack, "... this is not adequate State ground that would *block* Federal habeas;"

- And a question that should strike fear into the heart of any lawyer, "... so I ask you: If it does come about, if it should come about and you *lost* this case..."

Conversely, Justice Breyer's neutral-leaning terms include technical terms ("tort," "brief," and "procedure") as well as the fairly innocuous ("guess" and "imagine"). While this particular justice's textual cues are plausible, however, his colleagues are far more difficult to read

using word frequencies alone—perhaps because they signal their position in subtle ways, or perhaps because text is just a poor indicator of expressed emotion. For all other justices, we identify fewer than five informative words through this procedure; moreover, their cumulative predictive power is virtually nonexistent.

4.3 Auditory Characteristics of Expressed Skepticism

The preceding results show that the textual channel is—at best—a noisy, idiosyncratic, or simply weak signal of a justice's expressed skepticism. What, then, distinguishes skeptical questioning from neutral speech? To demonstrate, we interpret MASS results by investigating the auditory characteristics of median justice Anthony Kennedy's speech. For Kennedy, we found that a moderately regularized speech model with $K = 3$ latent sounds minimized the total cross-validated likelihood of out-of-sample auditory features. Three well-separated sound classes can be consistently observed across model runs. We subjectively characterize these as "voiced speech" such as vowels, in which the vocal cords vibrate (high autocorrelation); "unvoiced speech," such as fricatives and sibilants, in which vocal cords are not used (moderate energy and zero-crossing rate); and "silence" (low energy). Using an alignment procedure described below, we identify the three sounds in each bootstrapped model. For illustrative purposes, we compare the auditory characteristics of voiced skeptical speech to voiced neutral speech. The top panel of Figure 4 shows that when speaking skeptically, Kennedy speaks more loudly and with higher average pitch, a consequence of tensed vocal cords. Moreover, his modulation of pitch—which rises during questions and falls sharply during emphatic statements—is markedly larger in skeptical speech, as indicated by its higher

pitch variance. We do not, however, observe similar modulation in energy: Kennedy is simply louder across the board when expressing skepticism. Finally, in the bottom panel, we contrast Justices Kennedy and Sotomayor to demonstrate that these speech dynamics are not entirely unique to individual speakers. While speaker baselines do vary—Sotomayor speaks more softly on average, and her voice is roughly six semitones higher—both communicate their skepticism by elevating pitch and raising their voices, among other auditory cues.

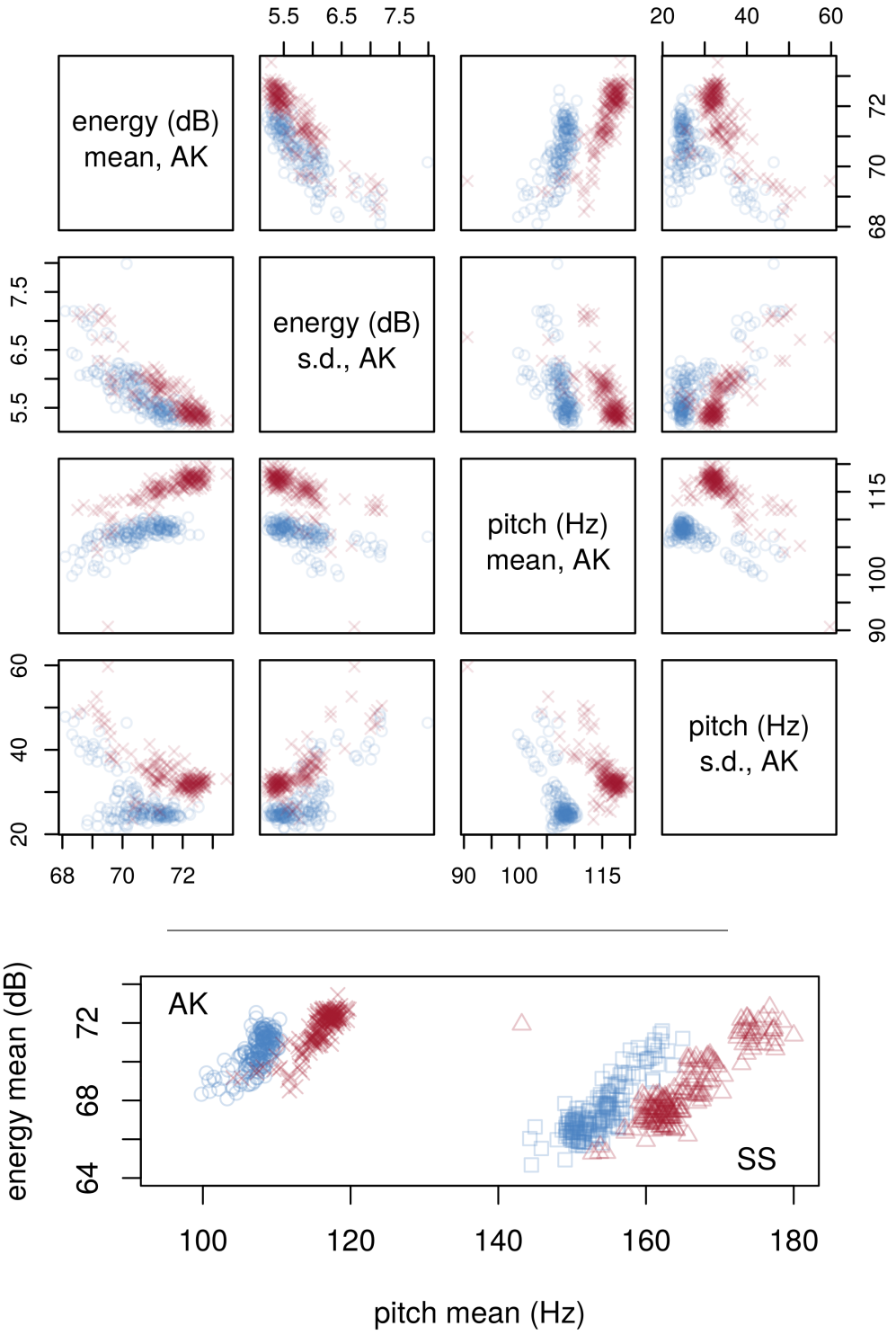

Figure 4: **Auditory characteristics of neutral and skeptical speech.** In the top panel, each dark red $\times$ (light blue $\circ$) represents a converged EM run for auditory parameters using a run-specific bootstrap draw of skeptical (neutral) training utterances for Justice Kennedy. Coordinates in a bivariate scatterplot are based on elements of $\boldsymbol{\mu}^{\text{skeptical, voiced}}$ ($\boldsymbol{\mu}^{\text{neutral, voiced}}$) and the diagonal of $\boldsymbol{\Sigma}^{\text{skeptical, voiced}}$ ($\boldsymbol{\Sigma}^{\text{neutral, voiced}}$). For example, the top right panel demonstrates that when speaking skeptically, Justice Kennedy's voice is markedly louder and exhibits more variation in pitch, relative to his neutral speech. The bottom panel compares the same parameters for Justice Sotomayor's skeptical (neutral) voiced speech, depicted with dark red $\triangle$ (light blue $\square$). While her voice is generally higher and quieter, on average, Sotomayor also communicates skepticism by elevating her pitch and speaking more loudly.

We now describe the technical details of the sound alignment procedure employed above. To identify sounds that consistently recur across the M speech modes and B trained bootstrap models, we employ an ad-hoc but effective alignment approach consisting of the following steps. First, we take the MBK separate $\boldsymbol{\mu}$ vectors, each representing the estimated average value of a sound for a particular bootstrap training set, and cluster these values using the k-means algorithm. The result of this procedure is K distinct reference points in audio-feature space, which in the main-text example corresponded to the subjective categories "voiced speech/vowel", "unvoiced speech/consonant", and "silence." In each of the MB trained models, we then determine the optimal one-to-one assignment of the K (unlabeled) sounds to the K reference categories such that the cumulative Mahalanobis distance of each sound to its assigned reference point is minimized.

This procedure produces an approximation to the far more difficult task of assigning each sound to a category while minimizing the total within-category Mahalanobis distances under the constraint of no duplicate assignments. The latter task involves optimizing over K^{MB} permutations, whereas the former consists of only MB separate K-to-K matching problems using the procedure of Hansen and Klopfer (2006).

4.4 Audio, Text, and Human Classification Performance

To validate the out-of-sample performance of the model, we treat the lower-level HMMs as auditory classifiers. (True out-of-sample performance of the full model is difficult to evaluate, because of dependencies introduced when modeling context and conversation flow.) As in the full model, bootstrap aggregation (bagging) is used to improve stability. Out-of-bag

(OOB, see e.g. Hastie, Tibshirani, and Friedman 2001, 15.3.1) performance is computed as follows. First, for labeled utterance u, we take all of the speaker's bootstrap speech models in which the utterance was out-of-bag (i.e., the roughly $\frac{1}{e}$ bootstrap resamples in which u was not drawn). For each bootstrap draw, the likelihood of utterance u is computed under the trained neutral and skeptical models, then converted to predicted tone probabilities of u. Predicted probabilities are then averaged over models. Results reflect the performance of a classifier that uses $1 - \frac{1}{e} \approx 63\%$ of the full training set. Across all speakers, we find that 68% of utterances are correctly classified ($F_1 = 0.540$). Speaker-specific results and other measures of performance are reported in Figure 5, along with measures of text classifier performance discussed in Appendix 4.2.

To assess the difficulty of the task, we contrasted the performance of supervised audio and text classifiers with that of non-expert human coders. A total of 40 native English speakers were recruited on a crowdworking site and assigned to one of eight justices (five coders per justice). Coders listened to all training utterances for their assigned justice, attempting to recover ground-truth labels. Figure 5 reports results from this evaluation in two ways. First, non-expert predictions were aggregated by majority vote, producing a set of committee predictions that were 70% correct, on average. We then disaggregated non-expert coders and found that individuals were able to recover the ground-truth label in 65% of utterances. However, individuals often disagreed in their assessment of whether a particular utterance constituted skepticism, averaging a low Cronbach's alpha of 0.50 across justices. Speaker-specific results are reported in Figure 5.

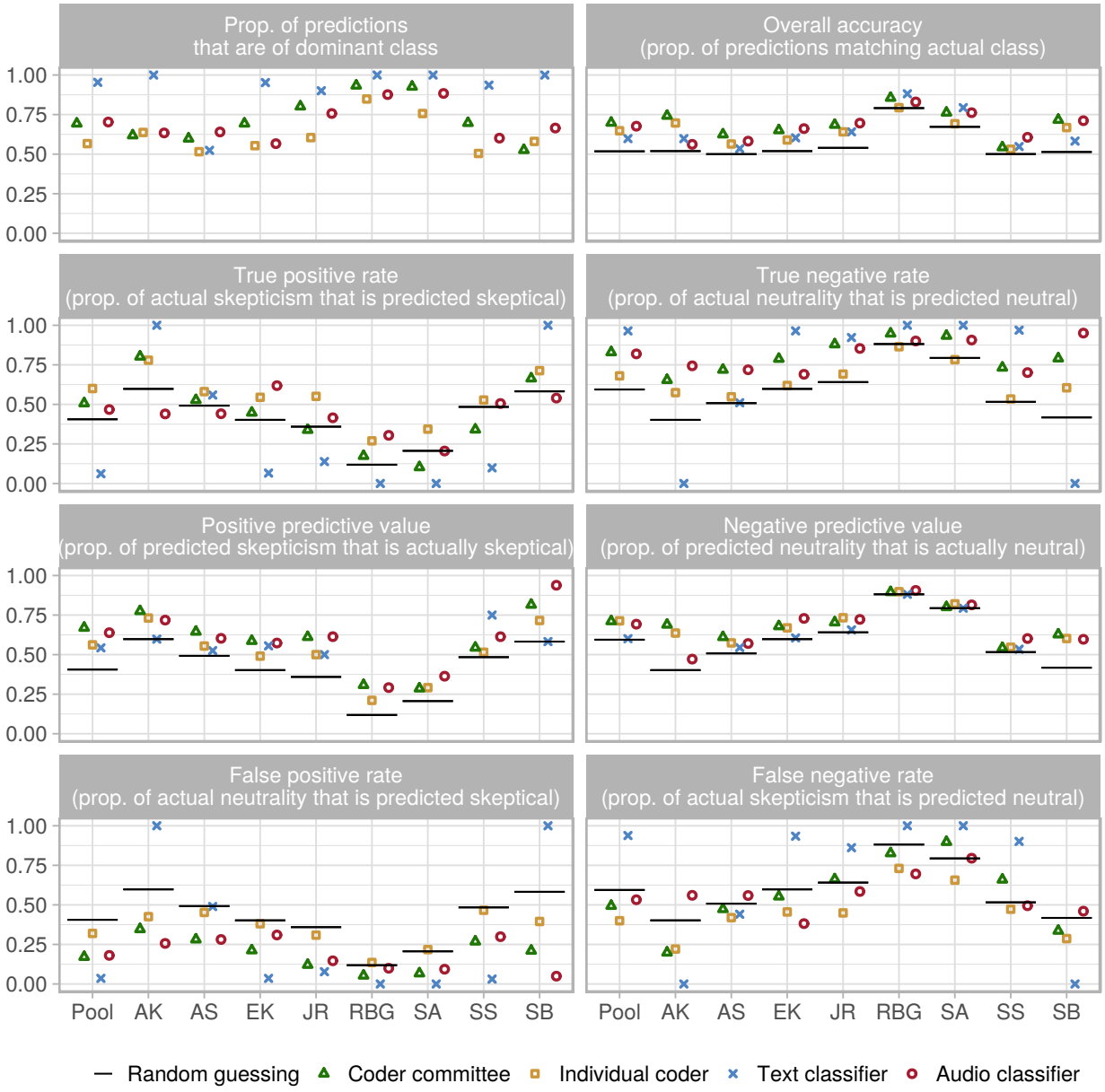

Figure 5: **Out-of-sample performance.** Red circles (blue crosses) indicate the performance of models trained on audio (text). As a point of reference, we also report the performance of individual coders (orange square) and coder predictions aggregated by majority vote (green triangle). Justices initials on the horizontal axis indicate speaker-specific results, and leftmost values indicate pooled results. For each performance measure, horizontal black lines denote the value that would be obtained by randomly guessing according to the baseline proportion of each class. The top-left panel shows that for all speakers except Justice Scalia, text classifiers almost always predict the dominant class. (In fact, speaker-specific text classifiers for Justices Kennedy and Breyer predict skepticism for every single utterance, and those for Justices Ginsburg and Alito predict neutrality for each one. As a result, predictive value for the opposite class cannot be computed for these justices in the lowest two panels.)

4.5 Comparison to Black (2011)

We compute Black et al.'s 2011 two text-based measures of justice affect as follows. Words in the top decile of "pleasantness" in the most recent Dictionary of Affect in Language (DAL) are defined as "extremely pleasant." The proportion of extremely pleasant words is defined straightforwardly as the count of pleasant words uttered by a justice toward a side, divided by the number of total directed words. Finally, we compute the difference in proportions by taking the pleasantness proportion of speech directed at the liberal side, then subtracting the conservative-directed proportion; this forms the first textual measure of directed affect. Note that under this procedure, the difference in proportions is undefined (and hence dropped) when a justice makes no utterances toward a particular side. This procedure is repeated for "extremely unpleasant" words, or words in the bottom "pleasantness" decile of DAL, to form the second textual measure. The most common pleasant and unpleasant words in Supreme Court questioning, defined in this way, are reported in Table 3. Key divergences from Black et al. (2011) are that (1) we use the most recent DAL (Whissell 2009), rather than the original (Whissell 1989), and (2) we operationalize sides in terms of Supreme Court Database (SCDB, Spaeth et al. 2014) liberal/conservative classifications (as in our main analysis) instead of petitioner/respondent. The latter coding decision makes justice fixed effects in the following analysis more informative.

Next, we compute a directly analogous measure of directed skepticism. We average predicted skepticism probabilities of utterances directed at the liberal side, then subtract the average of conservative-directed utterances. In this procedure, we use only the lower-level audio classification results, rather than the contextualized predictions from the full

model, because the full model incorporates voting as a covariate (and its predictions would therefore have have leaked information about the intended test of validity). This forms our third measure of directed affect.

Finally, we create a binary outcome of each justice's vote. This variable takes on a 1 (0) if a justice voted for the liberal (conservative) side in a case. (Observations are dropped if a justice had no recorded vote in the SCDB or sides cannot be categorized by ideology.) The voting outcome is regressed on the three directed affect covariates defined above; our expectations are that directed pleasantness textual measure will correlate positively with the voting outcome, whereas the directed unpleasantness textual measure and the directed skepticism auditory measure will correlate negatively. Figure 6 (duplicated below for ease of reference) reports coefficients on directed-affect covariates from three linear probability model specifications: (1) a "baseline" with no controls; (2) justice fixed effects, which absorb general liberal or conservative leanings; and (3) justice fixed effects and case fixed effects, which additionally absorb deficiencies in one side's legal arguments. We regard (3) as a particularly stringent test. All results are reported with standard errors clustered on case.

Across all specifications, we consistently find that the "pleasantness" textual measure is not significantly correlated with voting, thus replicating one result from Black et al. 2011. We also replicate their finding that the "unpleasantness" textual measure is negatively associated with voting, as expected, although it loses statistical significance when including case fixed effects. However, directed skepticism, as measured in the audio, is a far stronger predictor of voting patterns: a one-standard-deviation increase in this measure is associated with a change in voting that is consistently three times larger than the corresponding increase for unpleasantness, and this finding is robust across all specifications considered.

Table 3: **Common pleasant (unpleasant) words in justice speech.** Uses of the top 20 most common words in the top (bottom) decile of word pleasantness, as defined by the Dictionary of Affect in Language, in Supreme Court justice speech. The proportional contribution of each word to the measure of direct affect is computed by dividing a word's count by the total number of pleasant (unpleasant) words used.

Pleasant			Unpleasant		
Word	Count	Prop.	Words	Counts	Prop.
well	3,942	0.11	not	10,396	0.20
like	1,771	0.05	no	3,396	0.07
justice	1,659	0.05	other	3,223	0.06
us	1,120	0.03	mean	2,970	0.06
read	860	0.02	argument	1,731	0.03
talking	765	0.02	can't	1,620	0.03
money	729	0.02	problem	991	0.02
reasonable	689	0.02	over	711	0.01
clear	639	0.02	wrong	694	0.01
good	635	0.02	against	685	0.01
agree	609	0.02	trial	662	0.01
respect	570	0.02	without	650	0.01
view	558	0.01	nothing	603	0.01
yes	542	0.01	guess	578	0.01
correct	527	0.01	police	567	0.01
interest	458	0.01	number	496	0.01
sense	421	0.01	tax	443	0.01
company	347	0.01	violation	399	0.01
agreement	336	0.01	unless	362	0.01
accept	285	0.01	off	334	0.01

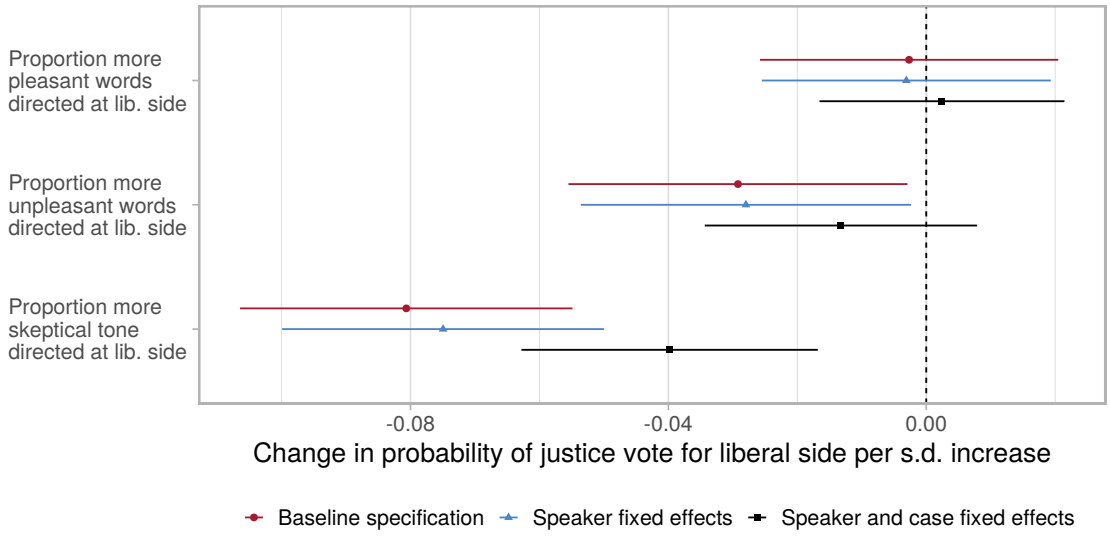

Figure 6: **Predicting justice votes with directed skepticism and directed affective language.** Horizontal errorbars represent point estimates and 95% confidence intervals from regressions of justice votes on directed pleasant words, directed unpleasant words, and our audio-based directed skepticism. Red circles correspond to a specification with no additional controls; blue triangles report results from a specification with speaker fixed effects; and black squares are from a specification with speaker and case fixed effects.

4.6 Benchmarking with Speaker Identity

Because our model incorporates temporal dependence between utterances in a conversation, a full evaluation requires a test set of multiple, completely labeled conversations. Additionally, in our primary application to skeptical speech, there is some uncertainty in the labels. Here, we benchmark against alternative models in an application where the true labels are known with certainty: identification of the speaker of each utterance, which is known for all conversations.

In this section, we first demonstrate that by explicitly modeling conversation dynamics, our hierarchical model improves on "naïve" approaches that treat each utterance individually. Specifically, the incorporation of metadata and temporal structure in the upper stage, when combined with the probabilistic predictions of the naïve lower stage, improves classification across all training set sizes and performance metrics that we examine. Next, we show that as the training set grows, model estimates converge on population parameters.

We implement the model described in Section 4 of the main text, modeling the transition probabilities (i.e., the turn-taking behavior of justices) as a multinomial logistic function of the following conversation metadata:

- Case-specific issue, indexed by i: civil rights, criminal procedure, economic activity, First Amendment rights, judicial power, or a catch-all "other" category; and
- The ideological orientation of the side of the lawyer currently arguing, indexed by j: liberal, conservative, or "unknown"; and
- A "speaker continuation" indicator for self-transitions, where the previous and current speaker are the same.

Issue and lawyer ideology variables are from Spaeth et al. (2014). The specification is

$$\Pr(S_{v,u} = m) \quad \propto \quad \exp\left(\alpha_{\mathrm{m}} + \beta \cdot 1(\mathrm{S_{v,u-1}} = \mathrm{m}) + \sum_{\mathrm{i}} \gamma^{\mathrm{issue}}_{\mathrm{m,i}} \cdot \mathrm{issue_v} + \sum_{\mathrm{j}} \gamma^{\mathrm{issue}}_{\mathrm{m,j}} \cdot \mathrm{ideology_{v,u}}\right),$$

and contains parameters respectively allowing for justice baseline frequencies of speech, justice-specific deviations based on the issue at hand or the ideology of the argument being advanced, and follow-up questions by the same justice. These factors have been shown in prior work to influence oral arguments: for example, Scalia is known to speak more frequently when First Amendment rights are under discussion, and the liberal Kagan more vigorously questions lawyers of the opposite ideological persuasion.

To examine how results improve as the training data grows, we report results for models trained with 25, 50, 100, and 200 utterances per mode. For all training set sizes, we show that contextual mode probabilities from the full model are superior in all respects to naïve mode probabilities that neglect temporal structure and metadata.

We assess performance with a variety of metrics. Using the posterior probabilities on each utterance's mode of speech $S_{v,u}$, we report average per-utterance logarithmic, quadratic, and spherical scores for each model. Because the fully labeled test set contains over 62,000 previously unseen utterances, we do not compute the miniscule confidence intervals on reported performance metrics. Training utterances are currently not excluded but represent only a small fraction of the full corpus.

While even naïve models perform well for the relatively simple task of speaker identification, we find that the upper level adds a considerable improvement. For example, across all sample sizes, the proportion of utterances misclassified by the full model falls by roughly

one quarter, relative to the lower level alone.

We also convert posterior probabilities to maximum-likelihood "hard" predictions and calculate mode-specific precision, recall, and F1 score. The prevalence-weighted average of these mode-specific performance metrics is also reported in Table 4. Note that overall and prevalence-weighted average mode accuracy equals prevalence-weighted average mode recall.

Table 4: Classification performance of lower-stage (L) model alone, versus full (F) model incorporating temporal structure and metadata, across four training set sizes and various performance metrics.

	n=25 (L)	n=25 (F)	n=50 (L)	n=50 (F)	n=100 (L)	n=100 (F)	n=200 (L)	n=200 (F)
logistic score	-0.315	-0.294	-0.278	-0.253	-0.233	-0.212	-0.211	-0.196
quadratic score	0.861	0.886	0.892	0.916	0.914	0.933	0.922	0.940
spherical score	0.917	0.934	0.935	0.951	0.949	0.962	0.954	0.965
F1 score	0.904	0.926	0.926	0.945	0.942	0.958	0.948	0.962
precision	0.912	0.933	0.929	0.947	0.943	0.959	0.950	0.963
recall	0.905	0.927	0.927	0.945	0.942	0.958	0.949	0.962

Finally, we compare our model to the best available audio classification models implemented in pyAudioAnalysis, which correctly classified speaker in 85% of out-of-sample utterances. In contrast, our model attained an accuracy of 97%. Figure 7 shows these results, with the best-performing alternative model on the top, and the results of MASS on the bottom.

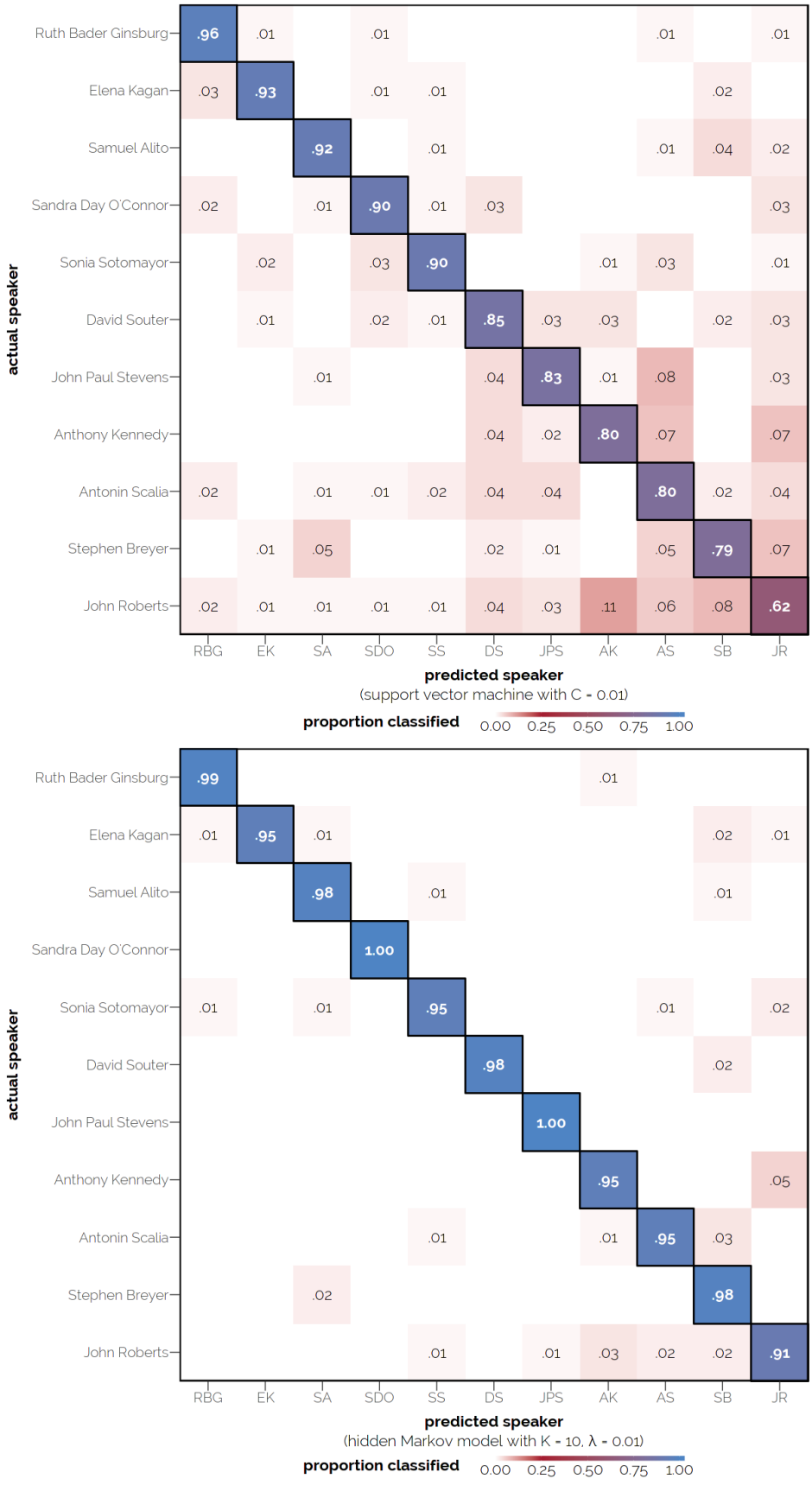

Figure 7: **Comparison to other models in a speaker identification task.**

References

Belloni, Alexandre, Victor Chernozhukov, and Ying Wei. 2016. "Post-Selection Inference for Generalized Linear Models With Many Controls". *Journal of Business and Economic Statistics* 34 (4): 606–619.

Black, Ryan, et al. 2011. "Emotions, oral arguments, and Supreme Court decision making". *Journal of Politics* 73 (2): 572–581.

Hansen, B.B., and S.O. Klopfer. 2006. "Optimal full matching and related designs via network flows". *Journal of Computational an Graphical Statistics* 15:609–627.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer.

Masanori, Kawakita, and Jun'ichi Takeuchi. 2014. "Safe semi-supervised learning based on weighted likelihood". *Neural Networks* 53:146–164.

Neal, Radford, and Geoffrey Hinton. 1998. "A view of the EM algorithm that justifies incremental, sparse, and other variants". In *Learning in Graphical Models*, 355–368. Springer.

Sotomayer, Sonia. 2019. *Life as a Supreme Court Justice*. Interview with Trevor Noah.

Spaeth, Harold, et al. 2014. *Supreme Court Database Code Book*.

van der Laan, Mark J, Sandrine Dudoit, Sunduz Keles, et al. 2004. "Asymptotic optimality of likelihood-based cross-validation". *Statistical Applications in Genetics and Molecular Biology* 3 (1): 1036.

Whissell, Cynthia. 1989. "The dictionary of affect in language". In *Emotion: theory, research, and experience*, ed. by R. Plutchik and H. Kellerman, 113–131. New York, NY: Academic Press.

— . 2009. "Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Language". *Psychological Reports* 105 (2).

Zucchini, Walter, and Iain MacDonald. 2009. *Hidden Markov Models for Time Series*. Boca Raton, FL: CRC Press.