A Dynamic Model of Speech for the Social Sciences*

Dean Knox †& Christopher Lucas ‡

April 17, 2019

Abstract

Social scientists increasingly rely on statistical models of text to resolve a wide range of questions about speech across a range of domains. However, humans communicate with more than a bag of words. Auditory cues convey important information, such as emotion, in many contexts of interest to social scientists. Nonetheless, researchers typically discard this information and work only with transcriptions of audio data. To resolve this methodological constraint, we develop the first generative model of the *sound* of political speech, the Structural Tone and Emotion Model (STEM). Our approach is grounded in human judgement of nonverbal communication, but learns and incorporates patterns of speech at scale, and is general to any class of labels that relate to the sound of speech. In addition to naturally testing theoretical predictions about individuals, STEM can also directly model the dynamic interaction between speakers. In an empirical application to speech in Supreme Court Oral Arguments, we demonstrate that the model is able to infer theoretically interesting quantities with only audio - and no text - as input. We further demonstrate that the the quantities inferred by STEM are not recoverable in the text alone, and that the out-of-sample labels inferred by STEM are valid upon human inspection. Finally, to demonstrate the substantive contributions made possible by STEM, we demonstrate that Supreme Court Justices use oral arguments as a part of a sincere fact-finding effort.

Keywords: Hidden Markov model; Signal processing; Social sciences; Latent process; Speech dynamics

*We thank Dustin Tingley for research support through the NSF-REU program; Michael May, Thomas Scanlan, Angela Su, and Shiv Sunil for excellent research assistance; and the Harvard Experiments Working Group and the MIT Department of Political Science for generously contributing funding to this project. For helpful comments, we thank Justin de Benedictis-Kessner, Bryce Dietrich, Gary King, Connor Huff, In Song Kim, Adeline Lo, Jacob Montgomery, Jonathan Mummolo, David Romney, Dustin Tingley, Michelle Torres, 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).

†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

1 Introduction

Applications of text analysis in social science often examine corpora which were first spoken, then transcribed. Though methodologically and substantively diverse, this research focuses exclusively on what people say, while entirely ignoring how those words were spoken. In this article, we develop the first generative model of political speech that explicitly represents "how" subjects express themselves: the *Structural Tone and Emotion Model* (STEM). We demonstrate that STEM is able to infer theoretically interesting quantities with only the audio - and no transcribed text - as input. And because we allow conversation metadata to structure how subjects express themselves, STEM is able to incorporate textual information when it is available, along with any additional conversational features of interest.

Substantively, we highlight the potential of our statistical model for analyzing tone of speech. However, STEM is ultimately general to *any* class of labels that relate to the "sound" of speech, whether that concerns the speaker's emotional state, gender, language, or any other imaginable category of interest to the researcher. In short, STEM allows researchers to study new quantities of interest conveyed through conversational tone. And by modeling the surrounding metadata structure, STEM naturally permits questions about how those quantities of interest measurable in the audio relate to information about speakers, the topic of speech, or other metadata available to the researcher.

A generative model of audio data is admittedly a departure from standard approaches to modeling political speech. But, political science has in fact been concerned with the study of audio data for as long as the discipline has existed, in that the study of speeches, debate, television, radio, and conversation *is* the study of audio data. And so, the question is not "should we study audio data?" (since we are already studying it), but rather, "how should we represent it?" At present, nearly all studies of political speech exclusively focus on textual representations of audio.[1]

Our approach - the Structural Tone and Emotion Model - is the first model of its kind in political science, and improves on existing approaches in statistics and computer science in three primary ways. First, existing approaches are only able to use a fraction of the features we incorporate. While we provide an overview of audio as data in Section 3, it is common to aribitrarily select a dozen features and discard the rest.[2] We provide a more principled solution through regularization that removes the need for arbitrary researcher choice with better statistical properties than alternative approaches. Second, to our knowledge, the Structural Tone and Emotion Model is the first to directly model dynamic interaction between speakers and to permit tests of hypotheses about these dynamics. For example, a researcher interested in the Supreme Court might ask, how does a hostile question from the median justice in an oral argument change the tenor of the subsequent conversation? We refer to these temporally dependent effects as changes in the *flow* of speech, and our model is the first in any field that permits the study of conversation flow. Third, while existing approaches ignore dependence and estimation uncertainty, we implement a hierarchical hidden Markov model to explicitly model these dynamics and recover estimation uncertainty through a Bayesian bootstrap.

[1]See (Dietrich et al., 2016) for a noteworthy exception.

[2]Nogueiras et al. (2001), for example, use just two features (pitch and energy, see Appendix Section A for a description) and their derivatives within each frame, while Kwon et al. (2003) use 13 total features and Mower et al. (2009) use the MFCC coefficients and their derivatives. As Section A makes clear, there are a range of additional features and researchers have not identified which are best-suited to the task El Ayadi et al. (2011*a*), as well as their interactions and derivatives. In practice, features are often selected according to preliminary results or a qualitative review of past literature. Böck et al. (2010), for example, conduct a series of experiments in order to develop prescriptions as to which features researchers ought to include and unsurprisingly generate domain-specific recommendations as opposed to a general set of rules. And at best, some sort of feature selection algorithm is used outside of the model itself, like forward selection (Ingale and Chaudhari, 2012).

We begin by expanding on the observation that political science *already* studies audio data, highlighting the many empirical analyses of speech that exclusively study the textual content of human communication. We then describe *how* audio can be preprocessed for statistical analysis. Next, we introduce our model, first positioning it within related literatures in statistics and computer science, then introducing our mathematical notation, the statistical model of human speech that we propose, and the EM algorithm that we implement for inference. Finally, we provide several empirical applications and conclude.

2 Political Science Already Studies Audio (But You Wouldn't Know It)

Politics occurs in audio. Voters are exposed to political advertisements while watching YouTube, coworkers discuss local politics around the water cooler, and viral news videos cause sudden shifts in the tenor of an election (Peters and Maheshwari, 2018). In each of these examples, political information was first *heard*. And while the textual component of speech is no doubt important, non-textual communication - the way words *sound* - is often equally, and sometimes more, important than the text of human speech.

For example, consider the historic "Daisy" ad, aired as a part of Lyndon Johnson's successful bid in the 1964 presidential election. The advertisement is arguably the most effective negative advertisement of all time (Mann, 2011). However, the text of the ad[3] contains simply a countdown from ten to one and an encouragement to vote for Johnson.

[3]The full text of the advertisement is "One, two, three, four, five, seven, six, six, eight, nine... Ten, nine, eight, seven, six, five, four, three, two, one, zero. These are the stakes: to make a world in which all of Gods children can live, are to go into the darkness. We must either love each other or we must die. Vote for President Johnson on November 3rd. The stakes are too high for you to stay home."

Nothing in the transcript suggests a cutthroat, visceral attack ad. But by juxtaposing the voice of an innocent child against that of an announcer counting down a rocket launch, the advertisement is threatening to a degree that cannot be appreciated without actually hearing the ad itself.[4] And while the "Daisy" advertisement is a particularly illustrative example, audio is an important component of political communication across a range of contexts.

For example, in Section I, we apply the Structural Tone and Emotion Model to speech in Supreme Court Oral Arguments to infer a novel measure of the sentiment expressed by justices during the argument stage. Of course, a great deal of existing work analyzes Supreme Court Oral Arguments. Johnson et al. (2006) demonstrate that the court's conclusion on a case reflects the merits of the arguments presented, and Black et al. (2011) demonstrates that a justices ultimate vote on a case is reflected in the sentiment expressed by that justice during the oral argument. And Johnson (2001) demonstrates that justice use Oral Arguments to gather information and use this information to make substantive policy choices. These cases all rely on either a manual qualitative assessment of the oral argument, or an automated analysis of strictly the text of oral arguments (ignoring the tone and tenor of speech). However, as demonstrated by Dietrich et al. (2016), the tone of speech has important information about Justice orientation toward the argument presented before them. And in Section I, we demonstrate that novel quantities of interest measured with the Structural Tone and Emotion Model can shed light on these existing questions about Justice behavior.

But, there are countless literatures which study audio data without in fact treating it as audio. To name but a few, research examining candidate debate (Abramowitz, 1978; Fridkin et al., 2007), campaign advertisements (Brader, 2005; Zhao and Chaffee, 1995; Koch, 2008;

[4] The advertisement can be viewed at youtube.com/watch?v=2cwqHB6QeUw.

Freedman et al., 2004), parliamentary debate (Proksch and Slapin, 2015; Spirling, 2016), or television and radio news (Behr and Iyengar, 1985; Sobieraj and Berry, 2011; Young and Soroka, 2012) is a study of audio data, even if the audio component of speech is ignored.

Importantly, in addition to improving on existing fields of research, the Structural Tone and Emotion Model also permits the study of new quantities of interest. For example, in our application, we study expressions of skepticism, which are not evident in the textual component of the oral argument. And in addition to permitting the analysis of new outcomes, by modeling the *dynamics* of speech, STEM allows researchers to ask and answer questions about the flow of conversation: how a line of questioning at time t affects downstream conversation. While these sorts of questions are common in theoretical models of sequential games (Dancey and Goren, 2010), they are rarely tested empirically. We suggest that this is due to the absence of a suitable empirical model, and demonstrate how STEM can answer questions of this sort.

Like text as data (Grimmer and Stewart, 2013; Lucas et al., 2015), audio requires preprocessing before it can be analyzed statistically. In the next section, we provide a straightforward introduction to audio preprocessing.

3 Audio as Data

We now explain how unstructured recordings of human speech can be represented as structured data and therefore subject to statistical analysis. The number of papers developing and applying methods for text analysis has increased rapidly in recent years (Wilkerson and Casas, 2017). However, little effort has been devoted to the analysis of other components of communication that accompany the textual component of human speech. How can the accompanying audio be similarly treated "as data"?

In its most raw form, physical sound is merely a wave of pressure transmitted through space. These waves can be recorded digitally by taking samples of the wave at constant, short intervals, each of which is a record of the strength of the original sound wave at a particular moment. We aggregate these samples to a set of representative features that are known to capture nonverbal communication, drawing on established literature in psychology, engineering, and computer science (Ververidis and Kotropoulos, 2006; El Ayadi et al., 2011*b*).

Appendix Section A.2 provides a full list of the features that we construct from the raw vector of audio samples. While these features are more abstract than a textual representation of the audio, they can still be easily understood. We divide features into those which are a function of the raw audio samples, and those which rely on first transforming the signal into its frequencies (i.e., how much treble, bass, etc, is in the samples). Features from the raw audio include the rate at which the signal crosses zero (a distinguishing feature of percusiveness), as well as features describing the loudness ("energy") of a sample. From the transformation into frequencies, we create bins across the frequency range and calculate the amount of the single in each, along with the estimated pitch and related measures that describe the shape of the vocal track, which are calculated from the maxima of this distribution over frequencies. We also include various functions of these raw features, including interactions, derivatives, and measures of dispersion.

Finally, our modeling approach can accommodate numerous types of contextual factors: From those that are constant over the entire conversation (such as a political candidate's poll numbers on the day of a speech, potentially leading to attacks on an opponent); to time-varying characteristics (the substantive topic under discussion, which necessitates a somber tone); or even variables that depend on conversation history (aggregate anger expressed by previous speakers over the course of an argument, which may make an angry retort more likely).

4 A Model of Conversation Dynamics

We begin by developing a generative statistical model for the audio data that arises from political speech. The model is illustrated with an extended example drawing on a well-known Obama speech from the 2012 presidential campaign. We first demonstrate how a skilled orator can juxtapose varying rhetorical styles within the same campaign address: A personal and intimate approach to storytelling that contrasts sharply with the fever-pitch roar of a turnout appeal. Next, we unpack the way in which these rhetorical styles manifest their distinct auditory signatures. The proposed model is shown to capture the structure of this speech well—not only the macro-level ebb and flow, which is shaped by covariates such as the issue under discussion, but also the micro-level changes in enunciation that give the rhetoric its power. After outlining the model, we next turn to estimation and inference. Finally, we discuss practical considerations in the modeling of speech. [BEEF UP?]

4.1 The Model

Suppose we have a conversation with U sequential utterances, each of which arises from one of M modes of speech. Let S_u denote the mode of utterance 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, which we represent with the vector $\boldsymbol{W}_u$. Importantly, $\boldsymbol{W}_u$ may include functions of conversation history ($\boldsymbol{S}_{u'<u}$), such as aggregate anger expressed by previous speakers over the course of an argument, which might induce a sharp retort. To keep the exposition clear, here we consider the simplified setting in which a single conversation is analyzed. In Appendix [TODO], we present the general multi-conversation model of which this is a special case.

The generative model is presented in Equations 1–4 and summarized graphically in Figure 1. We model speech mode probabilities as a multinomial logistic function of conversation context,

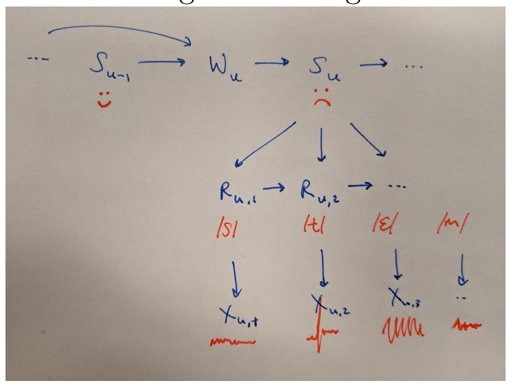
$$\Delta_m = \exp\left(\boldsymbol{W}_{\mathrm{u}}^\top \boldsymbol{\zeta}_{\mathrm{m}}\right) / \sum_{\mathrm{m}'=1}^{\mathrm{M}} \exp\left(\boldsymbol{W}_{\mathrm{u}}^\top \boldsymbol{\zeta}_{\mathrm{m}'}\right) \tag{1}$$

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

where $\boldsymbol{\Delta} = [\Delta_1, \ldots, \Delta_M]$ and $\boldsymbol{\zeta}_m$ is a mode-specific coefficient vector through which $\boldsymbol{W}_u$ affects the relative prevalence of mode m. Generally speaking, S_u is not directly observable to the analyst; the utterance's auditory characteristics, $\boldsymbol{X}_u$, is the only available information. These are matrices in which the t-th row is a D dimensional representation of the sounds perceived by a listener at a particular moment, with a total of T_u moments corresponding to the (unequal) durations of the utterances.

We assume that the m-th mode of speech is associated with its own Gaussian hidden Markov model (HMM) that produces the audio data as follows. At moment t in utterance u, the speaker enunciates the sound $R_{u,t}$, such as a plosive or fricative. In successive moments,

Figure 1: a dag



the speaker alternates through these sounds according to

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

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. In turn, sound k produces the raw audio data—the signal perceived by a listener—according to its own auditory profile, which we operationalize as a multivariate Gaussian distribution with parameters $\boldsymbol{\mu}^{m,k}$ and $\boldsymbol{\Sigma}^{m,k}$. Then

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

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.

4.2 Example

Here, we illustrate the model in the context of an excerpt from Barack Obama's final campaign address in Des Moines, Iowa, on November 6, 2012. While this example represents only a brief scene from an extended campaign, it demonstrates many of the speech dynamics that motivate our model.

In what follows, we begin with an close examination of two prototypical utterances that represent personal and inspirational speech. The selected examples are purely demonstrative, and the same approach can be used on any emotion that a candidate might seek to project, such as compassion, humor, or outrage. We first discuss the sounds from which each utterance is composed, along with their auditory profiles.

For example, consider Obama's "crowd-rousing" mode of speech—the tone in which he yells to the audience, "I've gotta turn out the vote!" He communicates through a sequence of sounds that, simplistically, we might categorize into "vowel," "consonant," and "silence."[5] In Figure TODO, we show that our generative model of crowd-rousing speech mirrors this structure: Vowels (dark red) are sustained for a few moments (horizontally arrayed cells) before transitioning to consonants (light red) and eventually pausing in silence (white) between words. 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 $\boldsymbol{\mu}^{\text{rousing,vowel}}$) from consonants (high zero-crossing rate) and silence (low volume).

[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 with theoretically motivated phoneme groups.

Why does this matter? It is on the basis of these constituent sounds that STEM is able to discern differences between rhetorical styles. As Equation TODO makes clear, STEM contains a parallel "storytelling" speech model alongside the "crowd-rousing" model. While storytelling also uses vowels and consonants, the auditory profiles of these sounds differ dramatically. Figure TODO (in which word size reflects decibel-scale energy) demonstrates Obama's use of modulation when he tells his audience that "one voice can change a room." Thus, $\mathbf{\Sigma}^{\text{story,vowel}}$ captures higher variance in loudness when compared to crowd-rousing speech, where every word is shouted at near-constant volume. Differences in pitch—often a marker of emotional engagement—are represented in the $\boldsymbol{\mu}$ terms, and shifts in cadence manifest in the $\mathbf{\Gamma}$ matrices.

The ability to quantify differences in rhetorical style presents new opportunities in the study of campaign rhetoric, such as the mechanisms by which emotional appeals sway voters. But perhaps more interestingly, models of storytelling and crowd-rousing speech enable analysts to categorize hundreds or even thousands of hours of previously unheard speech. For any new utterance u, a simple application of Bayes' theorem shows that

$$\Pr(S_u = \text{story}|\boldsymbol{X}_u, \mathbf{\Theta}) \;=\; \frac{\Pr(\boldsymbol{X}_u|S_u = \text{story}, \mathbf{\Theta})\Pr(S_u = \text{story})}{\sum_{m \in \{\text{story, rouse}\}} \Pr(\boldsymbol{X}_u|S_u = \text{m}, \mathbf{\Theta})\Pr(S_u = \text{m})},$$

where $\mathbf{\Theta}$ collects the various storytelling and crowd-rousing speech parameters, which can be learned from human-labeled examples of each mode, and $\Pr(S_u = m)$ can estimated by the proportion of each category in a randomly sampled training set. In essence, even without any knowledge about the context of an utterance, STEM can learn to distinguish human tone based on how much the utterance "sounds like" previously provided examples. Indeed, these uncontextualized probabilities play an important role in the estimation procedure described in Section TODO.

The most important questions in the analysis of political speech, however, relate to its ebb and flow—when and why a speaker chooses to deploy a particular tone. Figure TODO depicts how the two examined utterances fit into the broader context of the campaign speech. As Obama approaches the end of his rally, he launches into the story of Edith Childs, the originator of his signature "fired up, ready to go" campaign chant that animated hundreds of thousands of followers. The story begins with an intimate retelling of their meeting, building momentum before transitioning into a powerful call for turnout volunteers. Obama then repeats the narrative cycle by quieting down, drawing rallygoers in with his soft-spoken speech before ultimately finishing with a dramatic crescendo.

This isolated example, despite comprising only one minute from a yearslong political campaign, nonetheless reveals the cyclical speech patterns—the emotional appeal followed by exhortation to action—that led many observers to describe Obama as one of the most inspirational politicians in the modern era. A closer examination of the contexts in which these and other rhetorical devices are deployed may reveal deeper insights about political campaigns, just as the study of text corpora has led to insights about the usage and effectiveness of political attacks.

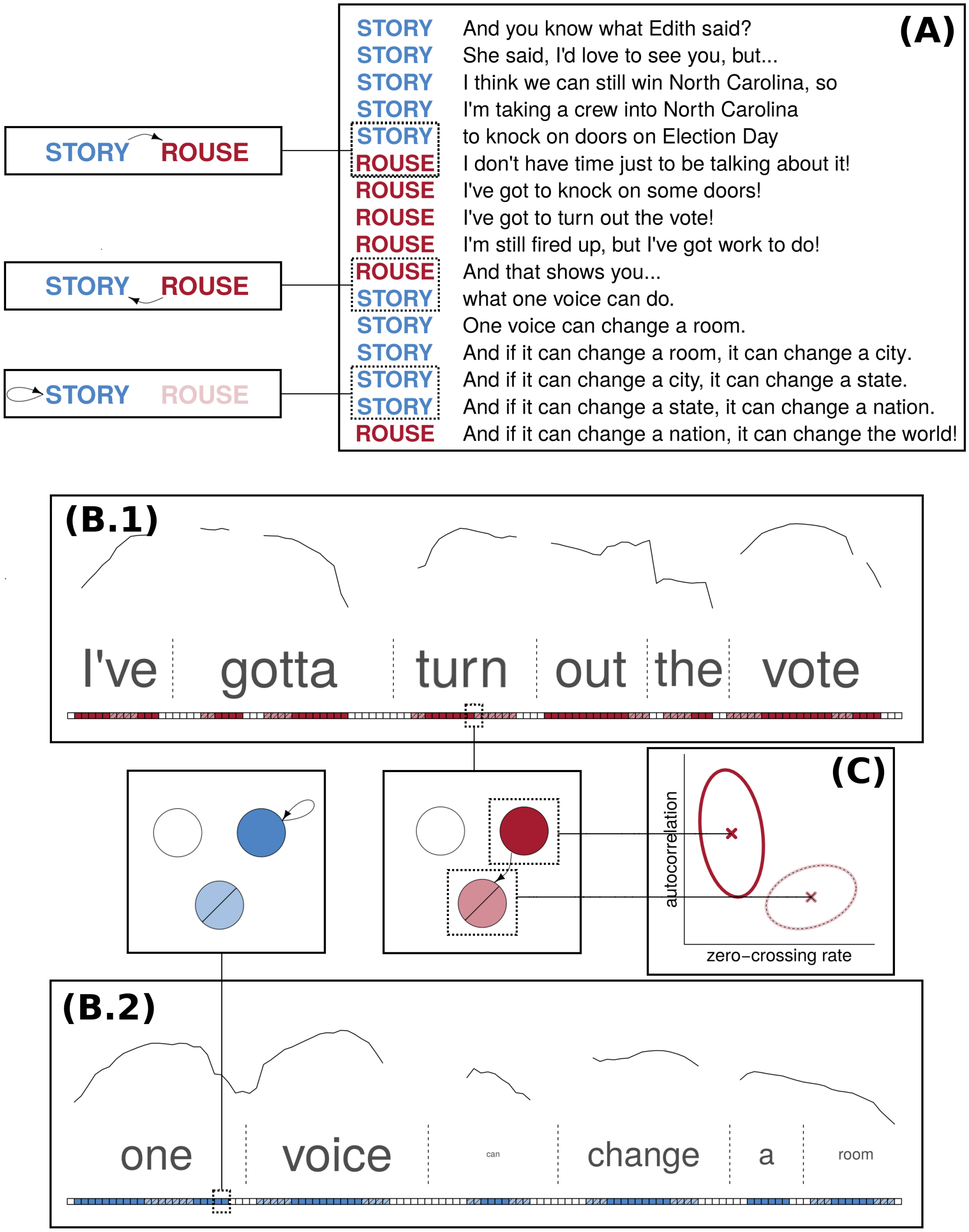


Figure 2: **An illustrative example.** Panel (A) contains an excerpt of a famous Obama campaign speech that utilized "story-telling" conversation and "crowd-rousing" shouts. Call-outs to the left of (A) highlight successive utterance-pairs in which the speaker (i) shifted from one mode to another, (ii) transitioned back, and (iii) continued in the same tone of voice. Call-out B.1 illustrates the use of loudness (text size) and pitch (contours) in a single crowd-rousing utterance: Obama maintains a high and constant volume throughout, with emphatic pitch drops on word endings. In contrast, B.2 shows the far smoother nature of his "story-telling" voice, in which pitch rises and falls gradually and the modulation of volume places emphasis on "voice" and "change." Separate call-outs identify sequential moments in which a "story-telling" vowel is sustained (transition from the dark blue sound back to itself, indicating repeat) and a shift from the dark red "crowd-rousing" vowel to the light red consonant. Panel (C) shows the differing auditory characteristics of the "crowd-rousing" vowel and consonant, which are perceived by the listener.

4.3 Estimation

Here, we describe the procedure by which we estimate the model defined in Section 4.1. At a high level, the estimation incorporates elements of unsupervised and supervised learning as follows. The researcher begins by determining the speech modes of interest, then identifying and labeling example utterances from each class. Within this training set—which is not necessarily 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 semi-supervised procedure. The specifics of each step are discussed below and in Appendix D, and the estimation is 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

	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.

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 1–4, 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; a detailed discussion is given in Appendix D.1.

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 $\boldsymbol{\Theta}$. We offer several arguments to justify this procedure. First, when the model is correctly specified, our approach remains unbiased; in this case, the only sacrifice is in efficiency loss relative to maximization of the full likelihood. However, in the presence of model misspecification—which is virtually guaranteed with complex phenomena like human speech—semi-supervised approaches that exploit both labeled and unlabeled data can counterintuitively 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. Second, 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. And third, even with moderately sized training sets, the number of audio frames 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.

In Appendix D.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 $f(\boldsymbol{X}^{\mathcal{T}} \mid \boldsymbol{\Theta}, \boldsymbol{S}^{\mathcal{T}}) = \prod_{m=1}^{M} \prod_{u \in \mathcal{T}} f(\boldsymbol{X}_u | \boldsymbol{\Theta}^m)^{\mathbf{1}(S_u = m)}$ factorizes neatly, $\hat{\boldsymbol{\Theta}}^m$ 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 $\prod_{t=1}^{T_u} f(X_{u,t} | \hat{\boldsymbol{\Theta}}^m, S_u = m)^{\rho}$, which represents the probability that $\boldsymbol{X}_u$ was generated by speech mode m and captures the extent to which the utterance 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' theorem. The corrective factor, ρ, approximately accounts for unmodeled autocorrelation in the audio features and ensures that the naïve estimates are well-calibrated; this shared correction, along with the number of latent sounds and strength of regularization, are determined by likelihood-based cross-validation (van der Laan 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 Appendix D.3. (This estimation procedure builds on Baum-Welch; readers unfamiliar with this are referred to Appendix D.2 or Zucchini and MacDonald (2009).) 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. In each step, we first compute the expected tone of each utterance, $\boldsymbol{S}$, based on its own audio profile and the conversation context, which also incorporates the audio profile of surrounding utterances. This step also fixes the conversation-history covariates, $\boldsymbol{W}^{\text{dyn.}}$. When these are linear functions of prior tone, such as average prior anger in an argument, the expectation step for $\boldsymbol{W}^{\text{dyn.},\mathcal{C}}$ follows directly by plugging in expectations for the corresponding elements of $\boldsymbol{S}$. We then update the expected transitions—in the case of the m to m' transition, this represents the current best guess for the probability that both $S_{u-1} = m$ and $S_u = m'$. 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.

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. The full estimation procedure is outlined in Algorithm 1. Other quantities of interest, such as those discussed in Section I, 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 Application

In this section, we introduce an original corpus of Supreme Court oral argument audio recordings scraped from the Oyez Project.[6] We limit our analysis to the natural court from the Kagan appointment to the death of Justice Scalia, so that the same justices are on the court for the entirety of the period we analyze. The Oyez data contains an accompanying textual transcript, speaker labels for each utterance, and accompanying timestamps for utterance start and stop times. We use these timestamps to segment the full-argument audio into a series of single-speaker utterances.[7] As an additional preprocessing step, we also drop utterances spoken by lawyers (each of whom usually appears in only a handful of cases) and Clarence Thomas (who speaks only twice in our corpus), focusing on the behavior of the eight recurrent speakers. We also drop utterances shorter than 2.5 seconds, which typically contain crosstalk or interjections.

For our primary application, we consider a tone of substantial importance in the study of courts: The expression of skepticism by a justice, indicating doubt in a lawyer's arguments. To analyze the use of skepticism, we randomly selected a training set of 200 utterances per justice to hand-classify as "skeptical" or "neutral" speech, allowing our assessments reflect not only by vocal tone but also by the textual content of the utterance. Thus, we define 16 modes of speech—two tones for each of the eight speaking justices.[8] In this process, we

[6] Dietrich et al. (2016) independently and concurrently collected the same audio data and conducted an analysis of vocal pitch.

[7] Occasionally, segments are reported to have negative duration, due to errors in the original timestamp data. In these cases, we drop the full "turn," or uninterrupted sequence of consecutive utterances in which only this speaker appeared.

[8] Because the transcripts attribute each utterance to a speaker, the model's decision is over whether (for example) the current statement by Anthony Kennedy is skeptical or neutral, rather than a joint speaker-recognition and tone-detection task. In the framework we outline 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.

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.

Four analyses of skeptical Supreme-Court speech are presented below. First, we demonstrate that STEM recovers a measure of expressed skepticism that has high facial validity, in the sense that the model's perception of skepticism (from audio and context) accord closely with human perceptions. Second, we examine the textual content of justice utterances, finding that word frequencies alone are virtually uninformative as to expressed skepticism. This reproduces a result that is well-known for other complex emotions, notably sarcasm, and is likely to generalize to other sophisticated tones in political speech such as compassion or authoritativeness. Third, we contrast the auditory content of median justice Anthony Kennedy's skeptical and neutral speech, modeled in Equations 3–4. Here we illustrate how expressed skepticism manifests in terms of loudness, vocal tension, and pitch modulation, allowing listeners to obtain a clear nontextual signal of Kennedy's projected emotional state. Finally, we analyze the structural determinants of oral-argument skepticism, modeled in Equations 1–2. 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 to influence their peers, on the one hand [CL TODO: CITE], or alternatively that they are neutral arbiters who respond genuinely to compelling legal arguments Johnson et al. (2006).

Separately, in Appendix **??**, we present an artificial validation exercise in which a "mode of speech" is defined as speech by one justice. The task is then to correctly identify the speaker of each utterance, which is known for all conversations. We demonstrate that the proposed model performs at a high level even with a limited training utterances, and that by explicitly modeling conversation dynamics, STEM improves on simpler approaches that treat each utterance individually.

5.1 Facial Validity of Predicted Skepticism

Before proceeding to more substantive results, we first demonstrate the face validity of STEM 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. The corresponding audio recordings, along with a larger selection of utterances, are available at [TODO: ONLINE CORPUS].

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 STEM is estimated solely on audio data and conversation context, its fitted values accord well with qualitative readings of the utterance text.

5.2 Textual Characteristics of Expressed Skepticism

The fact that humans can validate model-predicted skepticism using utterance text—in extreme cases, at the least—indicates that auditory channel carries emotional information that can be detected by STEM. But it also suggests 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 less than ten utterances (out of more than 50,000 total). We first sought to classify skepticism within the human-labeled utterances using a variety of approaches: Both pooling justices and considering each justice in turn to allow for speaker-specific word usage; using a wide range of supervised classifiers; and even applying pretrained neural-network models that take sentence structure into account. The results of this exercise were so poor that we do not discuss them further. 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 of over 50,000 utterances. To do so, we treat STEM 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 et al. (2016).

The resulting coefficient estimates, plotted in Figure 3, demonstrate that there are extraordinarily few consistent textual indicators of expressed skepticism—the vast majority

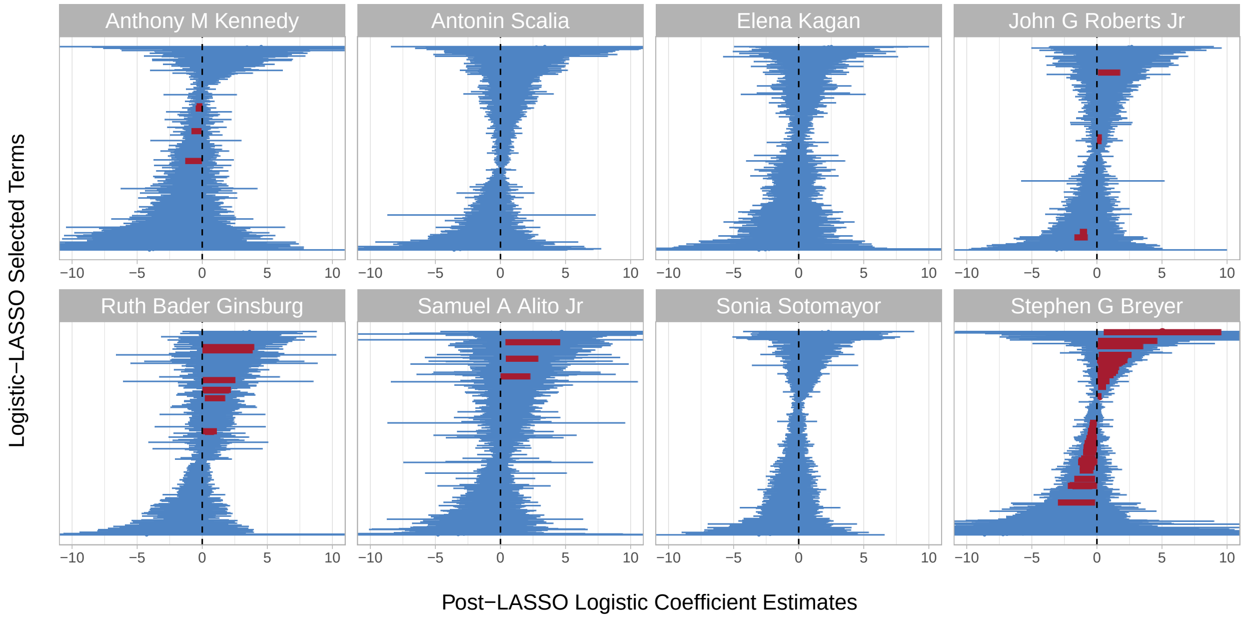


Figure 3: **Textual Signals of Justice Skepticism.** Each panel depicts a regression of STEM-predicted skepticism on word counts, within a justice's utterances. For each justice, a subset of terms are first selected by logistic LASSO; thin blue (thick red) errorbars reflect post-selection logistic regression 95% confidence intervals that (do not) overlap zero.

are statistically indistinguishable from zero at conventional levels. In Figure 4, we arbitrarily discard speaker-terms with p-values exceeding 0.05, then investigate the remainder more closely.

For 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 'sufficiently involved;' "

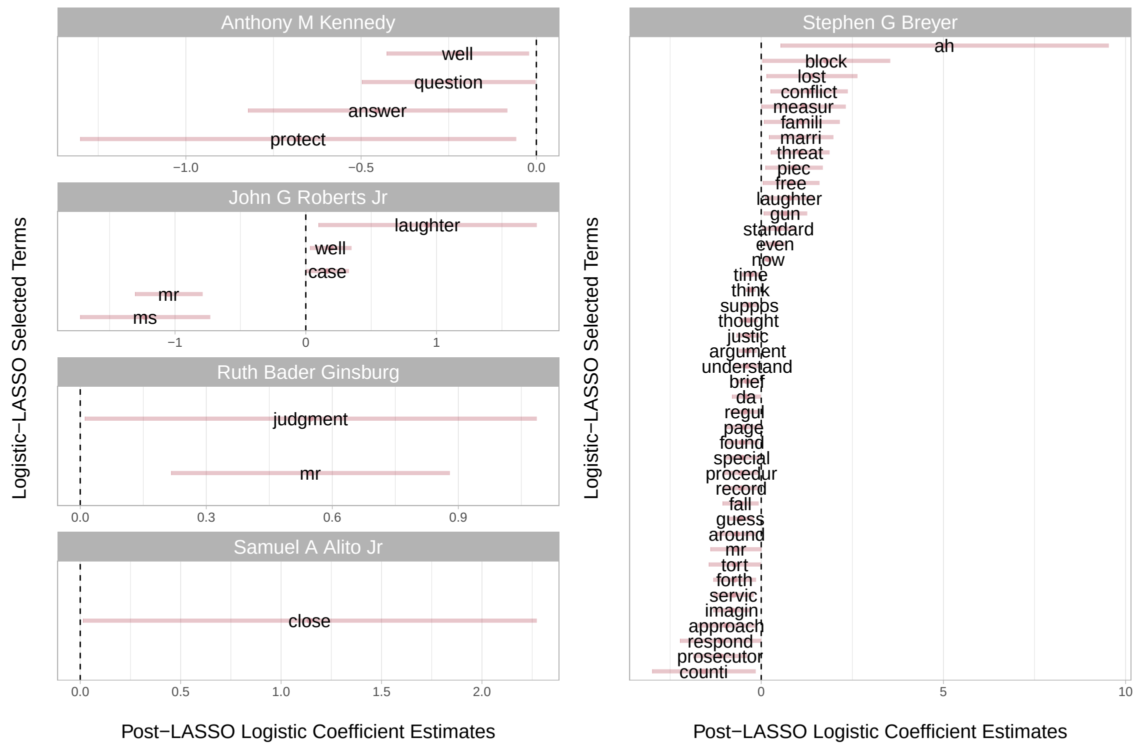


Figure 4: **Strong Textual Signals of Justice Skepticism.** Each panel depicts a regression of STEM-predicted skepticism on word counts, within a justice's utterances. Reported terms are the subset of post-LASSO terms post-selection logistic regression confidence intervals (errorbars) that do not overlap zero. Highly specific terms (i.e., used in less than five cases) are not depicted.

- 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, Breyer's neutral-leaning terms include the technical ("tort," "brief," and "procedure") and speculative ("guess", "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.

5.3 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, distingishes skeptical questioning from neutral speech? To demonstrate, we interpret STEM results by investigating the auditory characteristics of median justice Anthony Kennedy's speech. 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 (moderate energy and zero-crossing rate); and "silence" (low energy). Using an alignment procedure described in Appendix **??**, we identify the three sounds in each bootstrapped model. For illustrative purposes, we compare the auditory characteristics of voiced skeptical speech is compared to voiced neutral speech. Figure 5 demonstrates that when speaking skeptically, Kennedy speaks more loudly and with higher average pitch, a consequence of tensed vocal cords. Moreover, the modulation (variance) of pitch—which rises during questions and falls sharply during emphatic statements—is markedly larger in skeptical speech. We do not, however, observe similar modulation in energy: Kennedy is simply louder across the board when expressing skepticism.

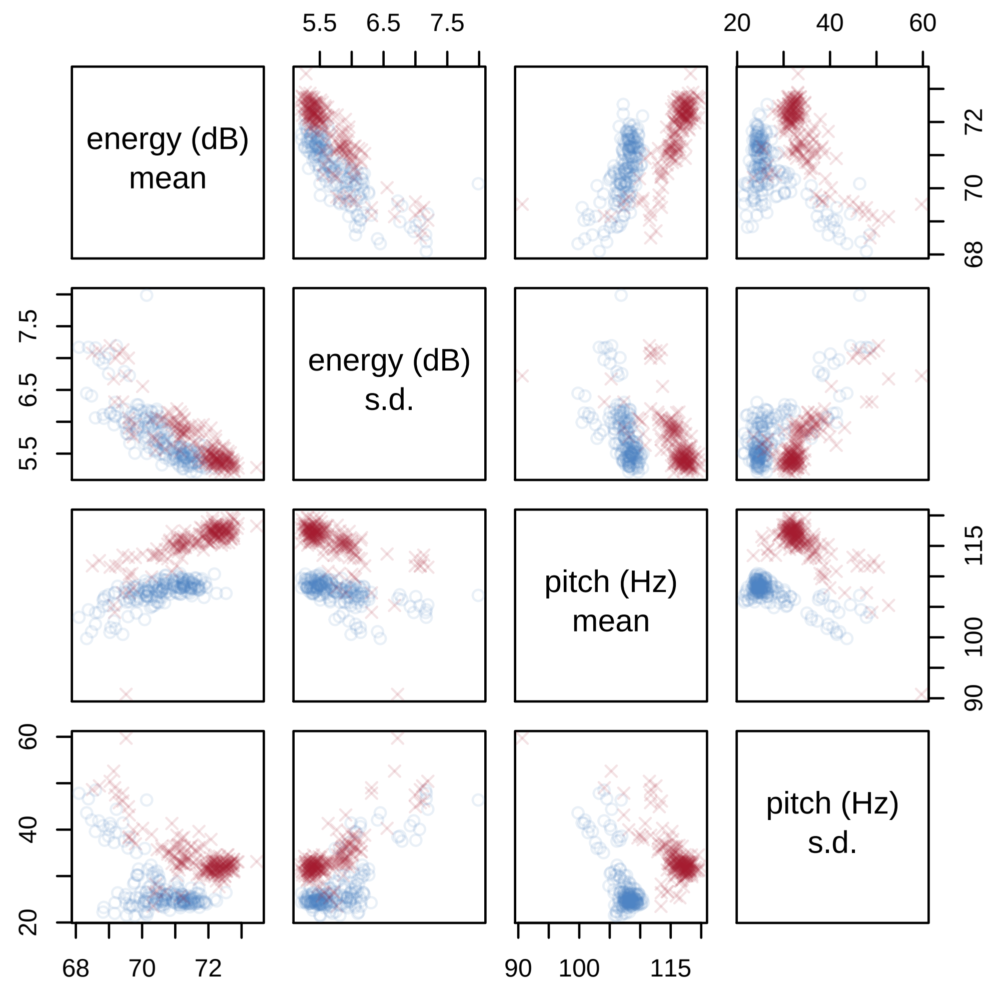


Figure 5: Each red $\times$ (blue $\circ$) represents a converged EM run for auditory parameters using a run-specific bootstrap draw of skeptical (neutral) training utterances. Coordinates in a bivariate scatterplot are based on elements of $\boldsymbol{\mu}^{\text{skeptical, voiced}}$ ($\boldsymbol{\mu}^{\text{neutral, voiced}}$) 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, Kennedy's voice is markedly louder and exhibits more variation in pitch, relative to his neutral speech.

5.4 Structural Determinants of Expressed Skepticism

When and why do members of the Supreme Court deploy skepticism in oral arguments? One common account of the deliberative process holds that justices are shrewd political actors that maneuver to influence the decisions of their peers in pursuit of a desired case outcome. However, strong judicial norms against back-room discussions foreclose the possibility of private communication. In this perspective, oral arguments represent an opportunity for justices to signal to their colleagues, with lawyers and their legal arguments serving merely as convenient foils. An alternate, largely incompatible conception of the decision-making process treats justices as neutral arbiters each voting according to rules determined by their respective legal philosophies. In this latter account, oral arguments provide an opportunity for fact-finding; while justices may tip their hand with a display of doubt, this merely reflects a genuine response to compelling or uncompelling legal reasoning.

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?[9] The competing narratives described above suggest several observable implications for the structural determinants of justice tone, or the way in which conversation context—whether justices are currently questioning the petitioner or respondent, the tone of the previous questioner, and so on—translates into expressed skepticism.

If justices react genuinely to poor argumentation, we should observe generically higher rates of skepticism when questioning lawyers for the side that they find less persuasive. A

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

strategic model of deliberation, on the other hand, must account for the fact that many—indeed, nearly half—of all cases are decided unanimously. To the extent that experienced justices are able to identify uncontroversial cases from pre-argument legal briefs and lower court decisions, this means that there would be little to gain from posturing, including acted skepticism. The strategic-signaling perspective therefore implies that justices will exhibit greater skepticism toward their non-preferred side *especially* in contentious cases. Forward-looking justices should similarly reduce skepticism toward their own side to avoid damaging its chances in close decisions. However, one practical challenge for this test is that pre-argument measures of persuasiveness and contentiousness correlate weakly with actual outcomes. As a result, we use a justice's subsequent vote and the ultimate margin of victory as the best available proxy for their legal preferences and disagreement. (We note that votes may well be influenced by the arguments; in what follows, we examine the observable implications of persuasion as well.)

A second test concerns the dynamics of oral arguments. In general, justices who are ideologically close 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. Vigorous questioning from an ideologically distant justice should not provoke much of a response from j in this view, due to the relative lack of shared hot-button issues. On the other hand, a strategic account implies a very different flow of questioning. A savvy justice 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 the preceding critic should not restrain the shrewd operator much, if at all—indeed, the left-right combination may be a particularly effective blow. When ideological neighbor i expresses skepticism toward j's preferred side, however, j has an incentive to pull punches or smooth things over, despite being predisposed 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. For operational simplicity, we discretize ideology into "left" (Breyer, Ginsburg, Kagan, and Sotomayor) and "right" (Alito, Roberts, Scalia), setting Kennedy aside given his unique position at the median. Preferences are crudely operationalized using subsequent votes, as before.

A key concern with the use of post-argument vote as a proxy for justice preferences is that it may itself be influenced by a third process in judicial deliberation. Persuasion may offer an alternative explanation for patterns observed in the way that arguments unfold. To understand its implications, we consider two stylized extreme cases. If justice i expresses skepticism and j finds the line of attack compelling, then j may both mirror that skepticism and also vote against the current side. Conversely, if j is unswayed, then they will vote for the current side and also speak neutrally. If justices are better able to persuade their ideological neighbors, then this process evidently produces predictions that are observationally equivalent to that of genuine expression. However, the proposed tests do allow us to arbitrate between strategic acting, on the one hand, and either persuasion or genuine expression, on the other.

Figure 6 presents results from two STEM specifications. First, we model transition probabilities (i.e., the probability that the next utterance is of justice-tone m) as $\exp\left(\boldsymbol{W}_{\mathrm{u}}^{\top}\boldsymbol{\zeta}_{\mathrm{m}}\right) / \sum_{\mathrm{m}'=1}^{\mathrm{M}} \exp\left(\boldsymbol{W}_{\mathrm{u}}^{\top}\boldsymbol{\zeta}_{\mathrm{m}}\right.$

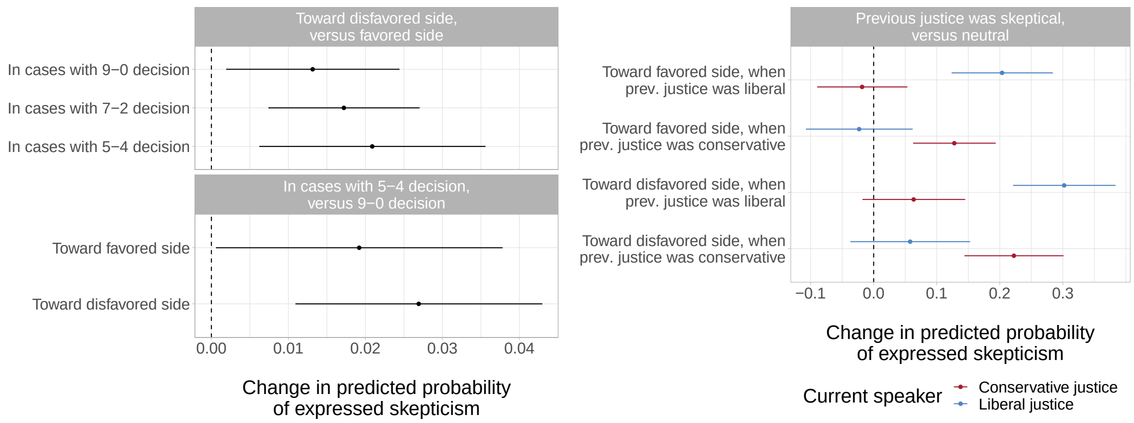


Figure 6: **Simulated quantities of interest.** Horizontal points (errorbars) represent estimated changes (95% bootstrap confidence intervals) in skepticism. Each panel manipulates a single variable from a control value (second element of panel title) to a treatment value (first). We average over all other nuisance covariates (e.g., the identity of the next speaker) of the scenario-specific change in outcome, weighting by the empirical frequencies of the nuisance covariates. The top-left panel shows that justices deploy skepticism more often toward their non-preferred side. The bottom-left 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 right panel shows that justices mirror the tones of their ideological neighbors, who share similar legal philosophies, even when those neighbors are opposed to the justices case-specific voting interests.

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. Taking the justices according to their speech frequency, we find that justices use skepticism against their non-preferred side (either the petitioner or respondent, depending on which one they go on to vote against) at a significantly higher rate. We also find no indication that the gap between petitioner- and respondent-targeted skepticism depends on the margin of the case decision. However, we do find that skepticism is generically higher *across the board* in close cases. These results suggest that Supreme Court justices are engaged in genuine fact-finding, not a strategic attempt to manipulate their colleagues, and they particularly seek to get answers right when the stakes are high.

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. (The preceding results remain virtually identical.) 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 strategic justice should follow up with a coup de grace after a peer 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 agree with the criticism 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.

Before concluding, we note that our reliance on a potentially contaminated measure of justice preferences—the post-argument vote—introduces some concerns. An alternative pathway by which one might observe the same patterns described above is if a colleague's skepticism was so persuasive that it led to a justice's concurrence (i.e., subsequent skepticism) and also to the justice voting against the current side. For persuasion to undermine the genuine-expression narrative, however, three things must hold: (1) justices must commonly begin oral arguments without a firm preference; (2) ideologically proximate peers must be more persuasive; and (3) justices must begin to signal strategically after being persuaded. While the first two are at least plausible concerns, we find the final requirement—that despite being initially indifferent, justices are sufficiently swayed to immediately initiate machinations—to be relatively unlikely. However, the structure of our tests do not permit us to clearly distiguish between genuine expression and this alternative narrative; future work by more substantively knowledgeable experts may employ superior covariates and improve on our research design using the tools developed in this paper.

References

Abramowitz, A. I. (1978), 'The impact of a presidential debate on voter rationality', *American Journal of Political Science* pp. 680–690.

Behr, R. L. and Iyengar, S. (1985), 'Television news, real-world cues, and changes in the public agenda', *Public Opinion Quarterly* **49**(1), 38–57.

Belloni, A., Chernozhukov, V. and Wei, Y. (2016), 'Post-selection inference for generalized linear models with many controls', *Journal of Business and Economic Statistics* **34**, 606–619.

Bitouk, D., Verma, R. and Nenkova, A. (2010), 'Class-level spectral features for emotion recognition', *Speech communication* **52**(7), 613–625.

Black, R. C., Treul, S. A., Johnson, T. R. and Goldman, J. (2011), 'Emotions, oral arguments, and supreme court decision making', *The Journal of Politics* **73**(2), 572–581.

Blei, D. M. and Lafferty, J. D. (2006), Dynamic topic models, *in* 'Proceedings of the 23rd international conference on Machine learning', ACM, pp. 113–120.

Böck, R., Hübner, D. and Wendemuth, A. (2010), Determining optimal signal features and parameters for hmm-based emotion classification, *in* 'MELECON 2010-2010 15th IEEE Mediterranean Electrotechnical Conference', IEEE, pp. 1586–1590.

Brader, T. (2005), 'Striking a responsive chord: How political ads motivate and persuade voters by appealing to emotions', *American Journal of Political Science* **49**(2), 388–405.

Dancey, L. and Goren, P. (2010), 'Party identification, issue attitudes, and the dynamics of political debate', *American Journal of Political Science* **54**(3), 686–699.

Dellaert, F., Polzin, T. and Waibel, A. (1996), Recognizing emotion in speech, *in* 'Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on', Vol. 3, IEEE, pp. 1970–1973.

Dietrich, B. J., Enos, R. D. and Sen, M. (2016), Emotional arousal predicts voting on the us supreme court, Technical report, Technical Report.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T. (2014), Decaf: A deep convolutional activation feature for generic visual recognition, *in* 'International conference on machine learning', pp. 647–655.

Ekman, P. (1992), 'An argument for basic emotions', *Cognition and Emotion* **6**, 169–200.

Ekman, P. (1999), Basic emotions, *in* T. Dalgleish and M. Power, eds, 'Handbook of Cognition and Emotion', Wiley, Chicester, England.

El Ayadi, M., Kamel, M. S. and Karray, F. (2011*a*), 'Survey on speech emotion recognition: Features, classification schemes, and databases', *Pattern Recognition* **44**(3), 572–587.

El Ayadi, M., Kamel, M. S. and Karray, F. (2011*b*), 'Survey on speech emotion recognition: features, classification schemes, and databases', *Pattern Recognition* **44**, 572–587.

Erhan, D., Bengio, Y., Courville, A. and Vincent, P. (2009), 'Visualizing higher-layer features of a deep network', *University of Montreal* **1341**, 3.

Freedman, P., Franz, M. and Goldstein, K. (2004), 'Campaign advertising and democratic citizenship', *American Journal of Political Science* **48**(4), 723–741.

Fridkin, K. L., Kenney, P. J., Gershon, S. A., Shafer, K. and Woodall, G. S. (2007), 'Capturing the power of a campaign event: The 2004 presidential debate in tempe', *The Journal of Politics* **69**(3), 770–785.

Gal, Y. and Ghahramani, Z. (2015), 'Bayesian convolutional neural networks with bernoulli approximate variational inference', *arXiv preprint arXiv:1506.02158* .

Gal, Y. and Ghahramani, Z. (2016*a*), Dropout as a bayesian approximation: Representing model uncertainty in deep learning, *in* 'international conference on machine learning', pp. 1050–1059.

Gal, Y. and Ghahramani, Z. (2016*b*), A theoretically grounded application of dropout in recurrent neural networks, *in* 'Advances in neural information processing systems', pp. 1019–1027.

Goplerud, M., Knox, D. and Lucas, C. (2016), 'The rhetoric of parliamentary debate', *Working Paper* .

Grimmer, J. and Stewart, B. M. (2013), 'Text as data: The promise and pitfalls of automatic content analysis methods for political texts', *Political analysis* **21**(3), 267–297.

Hall, D., Jurafsky, D. and Manning, C. D. (2008), Studying the history of ideas using topic models, *in* 'Proceedings of the conference on empirical methods in natural language processing', Association for Computational Linguistics, pp. 363–371.

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

Ingale, A. B. and Chaudhari, D. (2012), 'Speech emotion recognition', *International Journal of Soft Computing and Engineering (IJSCE)* **2**(1), 235–238.

Johnson, T. R. (2001), 'Information, oral arguments, and supreme court decision making', *American Politics Research* **29**(4), 331–351.

Johnson, T. R., Wahlbeck, P. J. and Spriggs, J. F. (2006), 'The influence of oral arguments on the us supreme court', *American Political Science Review* **100**(01), 99–113.

Karpathy, A., Johnson, J. and Fei-Fei, L. (2015), 'Visualizing and understanding recurrent networks', *arXiv preprint arXiv:1506.02078* .

Kaufman, A., Kraft, P. and Sen, M. (ND), 'Machine learning and supreme court forecasting: Improving on existing approaches'.

Kendall, A. and Gal, Y. (2017), 'What uncertainties do we need in bayesian deep learning for computer vision?', *arXiv preprint arXiv:1703.04977* .

Knox, D. and Lucas, C. (2017), 'Sam: R package for estimating emotion in audio and video', *Working Paper* .

Koch, J. W. (2008), 'Campaign advertisements' impact on voter certainty and knowledge of house candidates' ideological positions', *Political Research Quarterly* **61**(4), 609–621.

Kwon, O.-W., Chan, K., Hao, J. and Lee, T.-W. (2003), Emotion recognition by speech signals, *in* 'Eighth European Conference on Speech Communication and Technology'.

Leskovec, J., Backstrom, L. and Kleinberg, J. (2009), Meme-tracking and the dynamics of the news cycle, *in* 'Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 497–506.

Lucas, C. (2018), 'Neural networks for the social sciences', *Working Paper* .

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A. and Tingley, D. (2015), 'Computer-assisted text analysis for comparative politics', *Political Analysis* **23**(2), 254–277.

Mann, R. (2011), *Daisy petals and mushroom clouds: LBJ, Barry Goldwater, and the ad that changed American politics*, LSU Press.

Masanori, K. and Takeuchi, J. (2014), 'Safe semi-supervised learning based on weighted likelihood', *Neural Networks* **53**, 146–164.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M. and Stroeve, S. (2000), Approaching automatic recognition of emotion from voice: a rough benchmark, *in* 'ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion'.

Milner, H. V. and Tingley, D. (2015), *Sailing the water's edge: the domestic politics of American foreign policy*, Princeton University Press.

Mower, E., Metallinou, A., Lee, C.-C., Kazemzadeh, A., Busso, C., Lee, S. and Narayanan, S. (2009), Interpreting ambiguous emotional expressions, *in* 'Proceedings ACII Special Session: Recognition of Non-Prototypical Emotion From Speech - The Final Frontier?', pp. 662–669.

Neal, R. M. and Hinton, G. E. (1998), *A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants.*

Nogueiras, A., Moreno, A., Bonafonte, A. and Mariño, J. B. (2001), Speech emotion recognition using hidden markov models, *in* 'Seventh European Conference on Speech Communication and Technology'.

Panzner, M. and Cimiano, P. (2016), Comparing hidden markov models and long short term memory neural networks for learning action representations, *in* ‘International Workshop on Machine Learning, Optimization and Big Data’, Springer, pp. 94–105.

Paul, D. B. and Baker, J. M. (1992), The design for the wall street journal-based csr corpus, *in* ‘Proceedings of the workshop on Speech and Natural Language’, Association for Computational Linguistics, pp. 357–362.

Peters, J. W. and Maheshwari, S. (2018), ‘Viral videos are replacing pricey political ads. theyre cheaper, and they work.’, *The New York Times* .
URL: *https://www.nytimes.com/2018/09/10/us/politics/midterm-primaries-advertising.html*

Proksch, S.-O. and Slapin, J. B. (2015), *The politics of parliamentary debate*, Cambridge University Press.

Roberts, M. E., Stewart, B. M. and Airoldi, E. M. (2016), ‘A model of text for experimentation in the social sciences’, *Journal of the American Statistical Association* **111**(515), 988–1003.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B. and Rand, D. G. (2014), ‘Structural topic models for open-ended survey responses’, *American Journal of Political Science* **58**(4), 1064–1082.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H. and Deng, X. (2013), ‘Exploiting topic based twitter sentiment for stock prediction.’, *ACL (2)* **2013**, 24–29.

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

Socher, R., Lin, C. C., Manning, C. and Ng, A. Y. (2011), Parsing natural scenes and natural language with recursive neural networks, *in* 'Proceedings of the 28th international conference on machine learning (ICML-11)', pp. 129–136.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., Potts, C. et al. (2013), Recursive deep models for semantic compositionality over a sentiment treebank, *in* 'Proceedings of the conference on empirical methods in natural language processing (EMNLP)', Vol. 1631, Citeseer, p. 1642.

Spaeth, H., Epstein, L., Ruger, T., Whittington, K., Segal, J. and Martin, A. D. (2014), 'Supreme court database code book'.

Spirling, A. (2016), 'Democratization and linguistic complexity: The effect of franchise extension on parliamentary discourse, 1832–1915', *The Journal of Politics* **78**(1), 120–136.

Taber, C. S. and Lodge, M. (2006), 'Motivated skepticism in the evaluation of political beliefs', *American Journal of Political Science* **50**(3), 755–769.

Tvinnereim, E. and Fløttum, K. (2015), 'Explaining topic prevalence in answers to open-ended survey questions about climate change', *Nature Climate Change* **5**(8), 744–747.

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

Ververidis, i. and Kotropoulos, C. (2006), 'Emotional speech recognition: Resources, features, and methods', *Speech Communication* **48**, 1162–1181.

Wilkerson, J. and Casas, A. (2017), 'Large-scale computerized text analysis in political science: Opportunities and challenges', *Annual Review of Political Science* **20**, 529–544.

Wong, W. H. (1986), 'Theory of partial likelihood', *Annals of Statistics* **14**(1), 88–123.
URL: *https://doi.org/10.1214/aos/1176349844*

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

Zeiler, M. D. and Fergus, R. (2014), Visualizing and understanding convolutional networks, *in* 'European conference on computer vision', Springer, pp. 818–833.

Zhao, X. and Chaffee, S. H. (1995), 'Campaign advertisements versus television news as sources of political issue information', *Public Opinion Quarterly* **59**(1), 41–65.

Zucchini, W. and MacDonald, I. (2009), *Hidden Markov Models for Time Series*, CRC Press, Boca Raton, FL.

A Audio Feature Engineering

In this section, we describe the features that we use to characterize human speech, along with an overview of the mechanical process by which they are calculated. As noted in Section 1, the number of papers developing and applying methods for text analysis has increased rapidly in recent years. However, little effort has been devoted to the analysis of other data signals that often accompany text. How can the accompanying audio be similarly treated "as data"? In this section, we describe the necessary steps, beginning with a description of raw audio, then explain how that signal is processed before it may be input into a model like STEM.

A.1 The Raw Audio Signal

The human speech signal is transmitted as compression waves through air. A microphone translates air pressure into an analog electrical signal, which is then converted to sequence of signed integers by pulse code modulation. This recording process involves sampling the analog signal at a fixed sampling rate and rounding to the nearest discrete value as determined by the audio bit depth, or the number of binary digits used to encode each sample value. Higher bit depths can represent more fine-grained variation.

In order to statistically analyze audio as data, we must first format and preprocess the recordings. Recordings are typically long and composed of multiple speakers. The model presented in this paper is developed for single-speaker segments, which can be computed by calculating time stamps for words in an associated transcript, if available. If the audio corpus of interest has not been transcribed, researchers can identify unique speakers with automated methods that rely on clustering algorithms to estimate the number of speakers and when they spoke in the recording. Single-speaker speech is then cut into sentence-length *utterances*, a segment of speech in which there are no silent regions. This further stage of segmentation is accomplished within the `R` package `stem` (Knox and Lucas, 2017). For these speaker-utterances, we compute a series of *audio features*.

A.2 Raw Audio to Audio Features

We extract a wide range of features that have been used in the audio emotion-detection literature.[10] The raw audio signal is divided into overlapping 25-millisecond windows, spaced at 12.5-millisecond intervals. Some features, such as the sound intensity (measured in decibels) are extracted from the raw signal.

Next, features based on the audio frequency spectrum are extracted. The audio signal (assumed to be stationary within the short timespan of the window) is decomposed into components of various frequencies, and the power contributed by each component is estimated by discrete Fourier transform. The shape of the resulting power spectrum, particularly the location of its peaks, provides information about the shape of the speaker's vocal tract, e.g. tongue position. Some artifacts are introduced in this process, most notably by truncating the audio signal at the endpoints of the 25-millisecond frame and by the greater attenuation of high-frequency sounds as they travel through air. We ameliorate the former with a Hamming window that downweights audio samples toward the frame endpoints, and compensate for the latter using a pre-emphasis filter that boosts the higher-frequency components. Finally, we extract measures of voice quality, commonly used to diagnose pathological voice,

[10]For excellent reviews of the literature, including a more thorough discussion of these features, see (Ververidis and Kotropoulos, 2006; El Ayadi et al., 2011*b*).

based on the short-term consistency of pitch and intensity. Various interactions used in the emotion-detection literature are calculated, and the first and second finite differences of all features are also taken.

Table 3 shows the full set of features that we extract for each frame. As noted, we also include some interactions, as well as derivatives, which is possible because of the regularization step in STEM. The table divides features into those calculated directly from the raw audio, spectral features, and those measuring voice quality. Spectral features are those based on the frequency spectrum (for example, energy in the lower portion of the spectrum), while voice quality describes features that measure vocal qualities like "raspiness" and "airiness." Note as well that for some rows, we calculate many more than one feature. This is because the feature description describes a class of features, like energy in each of 12 pitch ranges, for example.

We group contiguous frames together into sentence-length *utterances*. When timestamped transcripts are available, as in our Supreme Court application in Section I, we use them to segment the audio. Otherwise, speech can be segmented using a rule-based system to pick out brief pauses in continuous speech. [11]

[11] Other classifiers can be trained to detect events of interest, such as interruptions or applause. We do so by coding a event-specific training set composed of the events of interest, as well as a few seconds before and after each instance to serve as a baseline. We then trained a linear support vector machine to classify individual audio frames as, for example, "applause" or "no applause." Framewise classifications are smoothed and thresholded to reduce false positives. This simple classifier is an effective and computationally efficient method for isolating short sounds with distinct audio profiles, such as an offstage voice. Continuous sections of speech by the same individual are thus isolated as separate segments. This allowed us to create single-speaker utterances for later analysis.

Features from raw audio samples		
energy	1 feature / frame	sound intensity, in decibels: $\log_{10}\sqrt{\overline{x_i^2}}$
ZCR	1 feature / frame	zero-crossing rate of audio signal
TEO	1 feature / frame	Teager energy operator: $\log_{10} \overline{x_i^2 - x_{i-1}x_{i+1}}$
Spectral features		
F0	2 features / frame	fundamental, or lowest, frequency of speech signal (closely related to perceived pitch; tracked by two algorithms)
formants	6 features / frame	harmonic frequencies of speech signal, determined by shape of vocal tract (lowest three formants and their bandwidths)
MFCC	12 features / frame	Mel-frequency cepstral coefficients (based on discrete Fourier transform of audio signal, transformed and pooled to approximate human perception of sound intensity in 12 pitch ranges)
Voice quality		
jitter	2 features / frame	average absolute difference in F0
shimmer	2 features / frame	average absolute difference in energy

Table 3: Audio features extracted in each frame. In addition, we include interactions between (i) energy and zero-crossing rate, and (ii) Teager energy operator and fundamental frequency. We also use the first and second finite differences of all features.

B Why not use a recurrent neural network?

Recurrent neural networks (RNNs) represent perhaps the most obvious alternative approach to time-dependent data like human speech, particularly given the increasing use of neural networks. While RNNs are not without merit, hidden Markov models are better suited to our problem for four primary reasons. First, like most applications in the social sciences, we have relatively few labeled examples, particularly in comparison to common deep learning applications to human speech.[12] Experiments comparing the performance of hidden Markov models to RNNs find that HMMs outperform neural networks where the data are limited (Panzner and Cimiano, 2016), an unsurprising result given that significant increase in the number of parameters. Second, neural networks are difficult to interpret (Lucas, 2018). And though much progress has been made in the interpretation of convolutional neural networks over the last few years (Erhan et al., 2009; Zeiler and Fergus, 2014; Donahue et al., 2014), methods for interpreting RNNs are considerably less developed (Karpathy et al., 2015). Third, the statistical foundations of deep learning are still not well-understood, though there has been some recent progress in this area (Gal and Ghahramani, 2015, 2016*a*,*b*; Kendall and Gal, 2017). Fourth and finally, we are interested not only in classifying segments of human speech, but also in analyzing the *flow* of speech - how speech of a particular tone influences the tone of subsequent speech. To our knowledge, there is no existing deep learning model that permits direct inference on statistical parameters that represent this interest.

[12] For example, the often-used Wall Street Journal speech corpus (Paul and Baker, 1992) contains 400 hours of speech, of which typically tens of hours are used as training data. By contrast, we have approximately one hour of labeled data in our application to Supreme Court Oral Arguments.

C Comparison with Text Sentiment

Given the amount of research on text and the courts, we also compare STEM to text-based sentiment analysis using the corresponding transcripts provided by Oyez. However, 100 utterances per speaker is suffciently small that it is effectively impossible to train an even remotely plausible text classifier. For example, we attempted to train an SVM on our hand-coded utterances (the same training set used in the preceeding audio benchmarks) but were unable to get even remotely plausible results. This is another argument in favor of using the audio data, as it can in fact be more informative in small samples for classification tasks like ours.

Given that we cannot effectively train a text classifier, we consider instead using a pre-trained sentiment classifier. Specificlaly, we use a state-of-the-art deep learning model, the recursive neural network (Socher et al., 2011), in which a treebank is employed to represent sentences based on their structures. Because the data in this case are too few to train our own Recursive Neural Network, we use pretrained weights provided Socher et al. (2013). Based on the the transcribed text, the neural network generates one of five possible labels for each utterance: "very negative", "negative", "neutral", "positive", and "very positive". We pool the two negative categories and treat these as predicting skepticism, because this produces the most favorable possible results for the neural network. Using this classification scheme, 78% of utterances are classified as skeptical, which leads to overall accuracy of 45% (much lower than all audio classifiers), a true positive rate of 89% (higher, because nearly all utterances were positively classified), and a true negative rate of 20% (again, much lower, because few utterances were classified negatively).

D Estimation

D.1 Factorization of the Likelihood

The full-data likelihood is as follows:

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

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.

D.2 Estimation of Lower-Level Auditory Parameters

[DK TODO: LOOK OVER AND EDIT AS NEEDED. ADD BOOTSTRAP OOS CALIBRATION/RHO.]

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. 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 $S_u = m$.[13] 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. The likelihood function for parameters of the m-th mode is then

$$\begin{aligned}
&\mathcal{L}^m(\boldsymbol{\mu}^{m,k}, \boldsymbol{\Sigma}^{m,k}, \boldsymbol{\Gamma}^m \mid \boldsymbol{X}, \boldsymbol{S}) \\
&\quad = \prod_{u=1}^{U} \Pr(\boldsymbol{X}_{u,1} = \boldsymbol{x}_{u,1}, \cdots, \boldsymbol{X}_{u,T_u} = \boldsymbol{x}_{u,T_u} \mid \boldsymbol{\mu}^{m,k}, \boldsymbol{\Sigma}^{m,k}, \boldsymbol{\Gamma}^m)^{\mathbf{1}(S_u=m)} \\
&\quad = \prod_{u=1}^{U} \left(\delta^{m\top} \boldsymbol{P}^m(\boldsymbol{x}_{u,1}) \left[\prod_{t=2}^{T_u} \boldsymbol{\Gamma}^m \boldsymbol{P}^m(\boldsymbol{x}_{u,t}) \right] \mathbf{1} \right)^{\mathbf{1}(S_u=m)}, \qquad (6)
\end{aligned}$$

where 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 Σ 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

[13] In practice, because the perception of certain speech modes can be subjective, 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.

of sounds, by selecting values that maximize the out-of-sample naïve probabilities of the training set in V-fold cross-validation. 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 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 standard procedure as it relates to the estimation of the lower-level audio parameters. Interested readers are referred to Zucchini and MacDonald (2009) for further discussion.

Baum-Welch involves maximizing the complete-data likelihood of Equation 7, which differs from equation 6 in that it also incorporates the probability of the unobserved sounds.

$$
\begin{aligned}
&\prod_{u=1}^{U} \Pr(\boldsymbol{X}_{u,1} = \boldsymbol{x}_{u,1}, \cdots, \boldsymbol{X}_{u,T_u} = \boldsymbol{x}_{u,T_u}, R_{u,1} = r_{u,1}, \cdots, R_{u,T_u} = r_{u,T_u} \mid \boldsymbol{\mu}^{m,*}, \boldsymbol{\Sigma}^{m,*}, \boldsymbol{\Gamma}^{m})^{\mathbf{1}(S_u=m)} \\
&\quad = \prod_{u=1}^{U} \Bigg(\delta_{r_{u,1}}^{m\top} \ \phi_D(\boldsymbol{x}_{u,1}; \boldsymbol{\mu}^{m,r_{u,1}}, \boldsymbol{\Sigma}^{m,r_{u,1}}) \times \\
&\qquad \prod_{t=2}^{T_u} \Pr(R_{u,t} = r_{u,t} \mid R_{u,t-1} = r_{u,t-1}) \ \phi_D(\boldsymbol{X}_{u,t}; \boldsymbol{\mu}^{m,r_{u,t}}, \boldsymbol{\Sigma}^{m,r_{u,t}}) \Bigg)^{\mathbf{1}(S_u=m)} \\
&\quad = \prod_{u=1}^{U} \Bigg(\mathbf{1}(S_u = m) \prod_{k=1}^{K} \left(\delta_k^{m\top} \ \phi_D(\boldsymbol{x}_{u,1}; \boldsymbol{\mu}^{m,k}, \boldsymbol{\Sigma}^{m,k})\right)^{\mathbf{1}(R_{u,1}=k)} \times \\
&\qquad \prod_{t=2}^{T_u} \Bigg(\prod_{k=1}^{K} \Bigg(\prod_{k'=1}^{K} \left(\Gamma_{k,k'}^{m}\right)^{\mathbf{1}\{R_{u,t}=k', R_{u,t-1}=k'\}} \ \ \phi_D(\boldsymbol{X}_{u,t}; \boldsymbol{\mu}^{m,k}, \boldsymbol{\Sigma}^{m,k})^{\mathbf{1}(R_{u,t}=k)} \Bigg) \Bigg) \Bigg)^{\mathbf{1}(S_u=m)},
\end{aligned}
\tag{7}
$$

D.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 8. These probabilities are efficiently calculated for all t in a single recursive forward pass through the feature vectors.

$$\begin{aligned} \boldsymbol{\alpha}_{u,t} &\equiv f(\boldsymbol{X}_{u,1} = \boldsymbol{x}_{u,1}, \cdots, \boldsymbol{X}_{u,t} = \boldsymbol{x}_{u,t}, R_{u,t} = k) \\ &= \delta_u^\top \boldsymbol{P}^m(\boldsymbol{x}_{u,1}) \left(\prod_{t'=2}^{t} \boldsymbol{\Gamma}^m \boldsymbol{P}^m(x_{u,t'}) \right) \end{aligned} \tag{8}$$

It also relies on the conditional probability of (i) all feature vectors after t given (ii) the sound at t (equation 9). These are similarly calculated by backward recursion through the utterance.

$$\begin{aligned} \boldsymbol{\beta}_{u,t} &\equiv f(\boldsymbol{X}_{u,t+1} = \boldsymbol{x}_{u,t+1}, \cdots, \boldsymbol{X}_{u,T_u} = \boldsymbol{x}_{u,T_u} \mid R_{u,t} = k) \\ &= \left(\prod_{t'=t+1}^{T_u} \boldsymbol{\Gamma}^m \boldsymbol{P}^m(\boldsymbol{x}_{u,t'}) \right) \mathbf{1} \end{aligned} \tag{9}$$

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 6, 8 and 9 immediately yields the expected sound label

$$\mathbb{E}\left[\mathbf{1}(R_{u,t} = k) \mid \boldsymbol{X}_u, S_u = m, \tilde{\boldsymbol{\Theta}}\right] = \tilde{\alpha}_{u,t,k} \; \tilde{\beta}_{u,t,k} \; / \; \tilde{\boldsymbol{\mathcal{L}}}_u^m, \tag{10}$$

where the tilde denotes the current approximation based on parameters from the previous M step, and $\alpha_{u,t,k}$ 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

$$
\begin{aligned}
&\mathbb{E}[\mathbf{1}(R_{u,t} = k', R_{u,t-1} = k) \mid \boldsymbol{X}_u, S_u = m, \tilde{\boldsymbol{\Theta}}] \\
&\quad = \Pr(R_{u,t} = k', R_{u,t-1} = k, \boldsymbol{X}_u \mid \tilde{\boldsymbol{\Theta}}) \ / \ \Pr(\boldsymbol{X}_u \mid \tilde{\boldsymbol{\Theta}}) \\
&\quad = \Pr(\boldsymbol{X}_{u,1}, \cdots, \boldsymbol{X}_{u,t-1}, R_{u,t-1} = k \mid \tilde{\boldsymbol{\Theta}}) \ \Pr(R_{u,t} = k' \mid R_{u,t-1} = k, \tilde{\boldsymbol{\Theta}}) \times \\
&\qquad\quad \Pr(\boldsymbol{X}_{u,t} \mid R_{u,t} = k') \ \Pr(\boldsymbol{X}_{u,t+1}, \cdots, \boldsymbol{X}_{u,T_u} \mid R_{u,t} = k') \ / \ \Pr(\boldsymbol{X}_u \mid \tilde{\boldsymbol{\Theta}}) \\
&\quad = \tilde{\alpha}_{u,t-1,k} \ \tilde{\Gamma}^m_{k,k'} \ \phi_D(\boldsymbol{x}_{u,t}; \tilde{\boldsymbol{\mu}}^{m,k}, \tilde{\boldsymbol{\Sigma}}^{m,k}) \ \beta_{u,t,k'} \ / \ \mathcal{L}_u^m.
\end{aligned}
\tag{11}
$$

implicitly conditioning on the training data throughout.

D.2.2 M Step

After substituting equations 10 and 11 into the complete-data likelihood (equation 7), the M step involves two straightforward calculations.

First, the conditional maximum likelihood update of the transition matrix $\boldsymbol{\Gamma}^m$ follows almost directly from equation 11:

$$
\tilde{\Gamma}^m_{k,k'} = \frac{\sum_{1=1}^{U} \mathbf{1}(S_u = m) \sum_{t=2}^{T_u} \mathbb{E}\left[\mathbf{1}(R_{u,t} = k', R_{u,t-1} = k) \mid \boldsymbol{X}_u, \tilde{\boldsymbol{\Theta}}\right]}{\sum_{1=1}^{U} \mathbf{1}(S_u = m) \sum_{t=2}^{T_u} \sum_{k'=1}^{K} \mathbb{E}\left[\mathbf{1}(R_{u,t} = k', R_{u,t-1} = k) \mid \boldsymbol{X}_u, \tilde{\boldsymbol{\Theta}}\right]} \tag{12}
$$

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.

$$\tilde{\Gamma}^m_{k,k'} = \frac{\sum_{u=1}^U \mathbf{1}(S_u = m) \sum_{t=2}^{T_u} \mathbb{E}\left[\mathbf{1}(R_{u,t} = k', R_{u,t-1} = k) \mid \boldsymbol{X}_u, \tilde{\boldsymbol{\Theta}}\right]}{\sum_{u=1}^U \mathbf{1}(S_u = m) \sum_{t=2}^{T_u} \sum_{k'=1}^K \mathbb{E}\left[\mathbf{1}(R_{u,t} = k', R_{u,t-1} = k) \mid \boldsymbol{X}_u, \tilde{\boldsymbol{\Theta}}\right]} \tag{13}$$

$$\tilde{\boldsymbol{\mu}}^{m,k} = \sum_{u=1}^U \mathbf{1}(S_u = m) \boldsymbol{X}_u^\top \boldsymbol{W}_u^{m,k} \tag{14}$$

$$\tilde{\boldsymbol{\Sigma}}^{m,k} = \sum_{u=1}^U \mathbf{1}(S_u = m) \left(\boldsymbol{X}_u^\top \mathrm{diag}\left(\boldsymbol{W}_u^{m,k}\right) \boldsymbol{X}_u\right) - \tilde{\boldsymbol{\mu}}^{m,k} \tilde{\boldsymbol{\mu}}^{m,k\,\top} \tag{15}$$

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

D.3 Estimation of Upper-Level Conversation Parameters

[DK: TODO ADD CONVOS, ADD INTERCEPT CONSTRAINT]

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 $f(\boldsymbol{X}^{\mathcal{C}} \mid \boldsymbol{\zeta}, \boldsymbol{\Theta}, \boldsymbol{W}^{\text{stat.},\mathcal{C}})$. It is well known that this is equivalent to jointly identifying a distribution over unobserved $\boldsymbol{S}^{\mathcal{C}}$ and parameters $\boldsymbol{\zeta}$ that together maximize the expected complete-data log likelihood,

$$
\begin{aligned}
&\ln f(\boldsymbol{X}, \boldsymbol{S} \mid \boldsymbol{\zeta}, \boldsymbol{\Theta}, \boldsymbol{W}^{\text{stat.}}) \\
&= \ln \left(\delta_{S_1} f(\boldsymbol{x}_1 | S_1 = s_1, \boldsymbol{\Theta}) \prod_{u=2}^{U} \Pr(S_u = s_u | S_{u-1} = s_{u-1}) f(\boldsymbol{x}_u | S_u = s_u, \boldsymbol{\Theta}) \right) \\
&= \sum_{m=1}^{M} \ln \delta_m^{\mathbf{1}(S_1=m)} + \sum_{u=1}^{U} \sum_{m=1}^{M} \ln f(\boldsymbol{x}_u | S_u = m, \boldsymbol{\Theta}^m)^{\mathbf{1}(S_1=m)} + \sum_{u=2}^{U} \sum_{m=1}^{M} \sum_{m'=1}^{M} \Delta_{m,m'}^{\mathbf{1}(S_{u-1}=m, S_u=m')} \\
&= \sum_{m=1}^{M} \mathbf{1}(S_1 = m) \ln \delta_m + \sum_{u=1}^{U} \sum_{m=1}^{M} \mathbf{1}(S_1 = m) \ln f(\boldsymbol{x}_u | S_u = m, \boldsymbol{\Theta}^m) \\
&\quad + \sum_{u=2}^{U} \sum_{m=1}^{M} \sum_{m'=1}^{M} \mathbf{1}(S_{u-1} = m, S_u = m') \ln \frac{\exp\left(\boldsymbol{W}_u(\boldsymbol{S}_{u'<u})^\top \boldsymbol{\zeta}_m\right)}{\sum_{m'=1}^{M} \exp\left(\boldsymbol{W}_u(\boldsymbol{S}_{u'<u})^\top \boldsymbol{\zeta}_{m'}\right)},
\end{aligned}
$$

where $\boldsymbol{W}_u(\boldsymbol{S}_{u'<u}) = [\boldsymbol{W}_u^{\text{stat.}\top}, \boldsymbol{W}_u^{\text{dyn.}}(\boldsymbol{S}_{u'<u})^\top]^\top$ and the primary-corpus indicator, $\mathcal{C}$, is omitted. Because the transition matrix, $\boldsymbol{\Delta}$, is a multinomial logistic function of conversation context, $\boldsymbol{W}_u$—which is itself a potentially complex function of 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. Update $\boldsymbol{\Delta}^{(t)} = \frac{\exp\left(\boldsymbol{W}_u^{(t-1)\top} \boldsymbol{\zeta}_m\right)}{\sum_{m'=1}^{M} \exp\left(\boldsymbol{W}_u^{(t-1)\top} \boldsymbol{\zeta}_{m'}\right)}$.

2. Compute $\mathbb{E}[\mathbf{1}(S_u = m)] = [DKTODO]$ and $\mathbb{E}[\mathbf{1}(S_{u-1} = m, S_u = m')] = [DKTODO]$, using a forward-backward algorithm paralleling Equations 10 and 11.

3. Update $\boldsymbol{W}_u^{(t)} = \boldsymbol{W}_u(\boldsymbol{S}_{u'<u}^{(t)})$.

The maximization step for $\boldsymbol{\zeta}$ then reduces to weighted multinomial logistic regression, and the estimated initial mode, $\boldsymbol{\delta}$ follows directly from the expected value of $[\mathbf{1}(S_1 = m)]$, as in Equation **??** [TODO LOWER ANALOGUE]. 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, as does the traditional expectation-maximization Neal and Hinton (1998).

E Application

E.1 Sound Alignment and Comparison

??

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).

F TODO: OLD TEXT

F.1 Existing Approaches

We assume a model of discrete speech modes, as is common in the emotion detection literature. However, in classifying political speech we depart from traditional models of so-called "basic" emotions such as anger or fear (Ekman, 1992, 1999), which are posited to be universal across cultures and often involuntarily expressed. Because such emotions are rare in political speech, of model of them is not especially useful. Instead, we argue that most actors of interest are professional communicators with a reasonable degree of practice and control over their speech. Political speakers generally employ more complex modes of speech, such as skepticism or sarcasm, in pursuit of context-specific goals such as persuasion or strategic signaling. To this end, we develop a method that can learn to distinguish between arbitrary modes of speech specified by subject-matter experts.

In Section 4, we describe how the Structural Speaker Affect Model accomplishes this task with a hierarchical hidden Markov model. Analogous extensions to models in text analysis demonstrate the utility of models for flow and structure. Within the literature on topic models, the dynamic topic model (Blei and Lafferty, 2006) and related derivations permit analysis of topical flow in text and has provided insight into the dynamics of the news cycle (Leskovec et al., 2009), stock market prediction (Si et al., 2013), and the history of ideas in a scientific field (Hall et al., 2008). The incorporation of structure into topic modeling

(Roberts et al., 2016) has been similarly influential, broadly lending insight into open-ended survey responses (Roberts et al., 2014), American foreign policy (Milner and Tingley, 2015), and public opinion about climate change (Tvinnereim and Fløttum, 2015).

A large body of research across a range of fields attempts to model human speech, nearly all of which attempts to classify short portions of speech, or *utterances*, into one of several possible labels. A common and straightforward approach is to use descriptive statistics (e.g., mean, min, max, std) of features like pitch to create a summary vector for each utterance, which is subsequently input to a classifier (Dellaert et al., 1996; McGilloway et al., 2000).[14] However, doing so necessarily discards information about the instantaneous change in the original feature (e.g., pitch).

Hidden Markov models have been widely employed in order to more directly model these transitions. Instead of summarizing each utterance with a single vector of descriptive statistics, the utterance is split into smaller *frames* in which those descriptive statistics are calculated.[15] The resulting sequence of feature vectors then summarize the utterance, rather than a single feature vector, and the transitions are typically modeled with an HMM (El Ayadi et al., 2011*a*). Additionally, HMMs permit variable-length sequences (i.e., the number of frames need not be equal), as opposed to normalizing the lengths as a preprocessing step.

[14]See Bitouk et al. (2010) for a longer review of the different approaches to feature extraction in existing research.

[15]Section 4 lays out the notation and model formally.

G Application

In this section, we introduce an original corpus of Supreme Court oral argument audio recordings scraped from the Oyez Project.[16] The corpus is used for two separate analyses in this paper. We first present a validation exercise in which we classify utterances of speech according to the identity of the speaker, then verify model predictions against known values. In the main application, we classify utterances according to their emotional characteristics.

The data for these applications are scraped from the Oyez Project.[17] We limit our analysis to the Roberts court from the Kagan appointment to the death of Justice Scalia, so that the same justices are on the court for the entirety of the period we analyze. The Oyez data contains an accompanying text transcript, as well as time stamps for utterance start and stop times and speaker labels. We use these timestamps to segment the audio into utterances in which there is a single speaker. However, occasionally, segment stop times are earlier than the stop times, due to errors in the original timestamp data. In these sections, we drop the full section of speech in which this speaker was speaking. As an additional preprocessing step, we also drop utterances spoken by lawyers (each of whom usually appears in only a handful of cases) and Clarence Thomas (who speaks only twice in our corpus). We also drop utterances shorter than 2.5 seconds, typically interjections and often containing crosstalk. To validate the remaining segments, we employ two procedures. For our main application, we randomly selected a training set of 200 utterances per Justice to code as "skeptical" or "neutral" speech, with training labels determined not only by vocal tone but also by the textual content of the utterance. In this process, we dropped the handful of utterances (5%)

[16]Dietrich et al. (2016) independently and concurrently scraped the same audio data and conducted an analysis of vocal pitch.

[17]https://www.oyez.org/

in which crosstalk or other audio anomalies occurred, or in rare instances where the speaker's identity was incorrectly recorded.

H Validation

Because our model incorporates temporal dependence between utterances in a conversation, a full evaluation requires a test set of multiple, completely labeled conversations. Because manual labeling the emotional state of entire Supreme Court oral arguments is infeasible, we first conduct an artificial validation exercise in which a "mode of speech" is defined as speech by one Supreme Court justice. The audio classification task in this validation exercise is therefore to correctly identify 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, 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

$$\Pr(S_{v,u} = m) \;\; \propto \;\; \exp\left(\mathrm{\alpha_m + \beta \cdot 1(S_{v,u-1} = m) + \sum_i \gamma_{m,i}^{issue} \cdot issue_v + \sum_j \gamma_{m,j}^{issue} \cdot ideology_{v,u}}\right),$$

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.

H.1 Predictive Performance

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, respectively defined below. Because the fully labeled test set contains over 62,000 utterances, we do not compute confidence intervals on 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.

$$\frac{1}{\sum_{v=1}^{V} U_v} \sum_{v=1}^{V} \sum_{u=1}^{U_v} \ln \Pr(S_{v,u} = s_{v,u} | \boldsymbol{X}, S^{\text{train}})$$

$$\frac{1}{\sum_{v=1}^{V} U_v} \sum_{v=1}^{V} \sum_{u=1}^{U_v} \left(2 \Pr(S_{v,u} = s_{v,u} | \boldsymbol{X}, S^{\text{train}}) - \sum_{m=1}^{M} \Pr(S_{v,u} = m | \boldsymbol{X}, S^{\text{train}})^2 \right)$$

$$\frac{1}{\sum_{v=1}^{V} U_v} \sum_{v=1}^{V} \sum_{u=1}^{U_v} \frac{\Pr(S_{v,u} = s_{v,u} | \boldsymbol{X}, S^{\text{train}})}{\sqrt{\sum_{m=1}^{M} \Pr(S_{v,u} = m | \boldsymbol{X}, S^{\text{train}})^2}}$$

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.

We find that the best available audio classification models implemented in pyAudioAnalysis correctly classify a speaker in 85% of out-of-sample utterances, whereas our model attained an accuracy of 97%.

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

H.2 Frequentist Performance

We also examine the coverage of estimated parameter confidence intervals. Population parameters are calculated by fitting the same model to the perfectly observed outcome. We opt for this naturalistic evaluation because simulated datasets are unlikely to accurately reflect performance in actual human speech corpora due to the violation of modeling assumptions. However, the conclusions about frequentist performance that can be drawn from this exercise are limited because coverage rates are poorly estimated.

We find that with a training set size of $n = 25$, four out of 57 confidence intervals fail to cover the population parameter. With $n = 50$, this number falls to two non-covering confidence intervals, and by $n = 100$ only one confidence interval (for speaker continuation) fails to cover the true parameter. The difficulty in accurately estimating the speaker continuation parameter appears to be caused by pairs of speakers, (m, m'), that are occasionally difficult to distinguish, such as Anthony Kennedy and John Roberts, that lead to utterances with large naïve posterior probability mass on the correct speaker, m, but some small mass $p_{m'}$ on m'. In this case, even if the same speaker spoke two sequential utterances, the probability of a nonexistent transition perceived by the model would be $2p_{m'}(1-p_{m'})$. We find that in practice, the bias due to misclassification in the naïve probabilities is small (leading to less than two-percentage-point difference between fitted transition probabilities and those calculated with the population parameter), diminishes as the training set grows, and is attenuating in typical scenarios of interest.

I Application

In this section, we redefine a mode of speech to correspond to a justice-emotion, e.g. skeptical speech by Antonin Scalia, for a total of 16 modes. Skepticism is a particularly interesting rhetorical category. As Johnson et al. (2006, p.99) argue, justices use oral arguments to "seek information in much the same way as members of Congress, who take advantage of information provided by interest groups and experts during committee hearings to determine their policy options or to address uncertainty over the ramifications of making a particular decision." With these intentions in mind, recent work analyzes how justice pitch when asking questions during oral arguments Dietrich et al. (2016) and the text of those questions Kaufman et al. (ND) predict that justice's vote on the respective case. We build on these results by providing the first direct classifier of a particular rhetorical mode, namely skepticism. Skepticism is especially interesting if, as Johnson et al. (2006) argue, justices use oral arguments to seek information, because skepticism is a subtle yet direct measure of the concepts and arguments that justices are willing to doubt (Taber and Lodge, 2006), which is theoretically distinct from more neutral-toned questions, in that the latter does not imply an oppositional view on the topic, whereas a question asked in a skeptical tone implies to the lawyer and the other justices that the issue at hand is not believable. Ability to measure skeptical tone, then, introduces to the literature on courts and decision-making in judicial bodies a method the permits the study of questions about when and why justices *doubt* arguments made in the courtroom, rather than simply when and why they ask questions.

The training procedure described above was implemented with a training set of the 1,600 manually coded utterances, minus the invalid segments that were dropped. We find that the use of skepticism varies widely by justice: in the training set, Sonia Sotomayor's speech was nearly evenly split between projected emotional states, whereas only 12% of the notoriously deadpan Ruth Bader Ginsburg's speech was discernably skeptical. In a cross-validation exercise, Appendix [TODO: write up], we find that imbalanced class sizes pose a severe challenge to the "flat" methods used by pyAudioAnalysis, which reduce every utterance to a vector of summary statistics. In contrast, our approach, which explicitly models the sound dynamics within each utterance, appears to be relatively unaffected.

Within each justice, we conducted 5-fold cross-validation and selected justice-specific regularization parameters and number of sounds by maximizing the total out-of-sample naïve mode probability. Overall, we found that the average accuracy of maximum-naïve-probability skepticism predictions was 72% across justices for the selected models.

We employ the following covariates:

- 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 or conservative; and

- A "speaker continuation" indicator for transitions in which the previous and current speaker are the same.
- A "speaker-mode continuation" indicator for transitions in which the previous and current speaker are the same, and the speaker's mode of speech is
- A "voted against" indicator that the justice voicing a particular mode opposed the side currently arguing
- A "skepticism" variable that candidate mode m is of skeptical projected emotion
- A "previous skepticism" variable that utterance $u-1$ was voiced with skeptical emotion

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

$$\Pr(S_{v,u} = m) \propto \exp\Big(\alpha_{\mathrm{m}} + \beta_{\mathrm{m}}^{\mathrm{mode}} \cdot 1(\mathrm{S_{v,u-1} = m}) + \beta_{\mathrm{m}}^{\mathrm{speaker}} \cdot 1(\mathrm{justice}_{\mathrm{S_{v,u-1}}} = \mathrm{justice_m}) + \sum_i \gamma_{m,i}^{\mathrm{issue}} \cdot \mathrm{issue_v} + \sum_{\mathrm{j}} \gamma_{\mathrm{m,j}}^{\mathrm{ideo}} \cdot \mathrm{ideology_{v,u}}\Big) \quad (16)$$

This specification allowing for justices to have varying baseline frequencies of both skeptical and neutral speech. It also allows each justice to have both differing volume of overall speech and differing emotional proportions (i) when questioning liberal and conservative lawyers, and (ii) while discussing cases that pertain to particular issues. Finally, it controls for justice- and justice-emotion continuation in an extremely flexible way, with one parameter for each of the four possible transitions (neutral–neutral, neutral–skeptical, skeptical–neutral, and skeptical–skeptical) that could occur if a justice spoke for two successive utterances.

Overall, we find that Kagan and Sotomayor question liberal lawyers less and Alito questions liberal lawyers more, but we find no evidence that ideological orientation alone produces greater skepticism. One possible explanation for this finding is that general ideological opposition is a crude measure of justices' preferences, and that justices take into accoun the nuances of a case. This is supported by the fact that many cases are decided unanimously, perhaps suggesting that a case-specific fixed effect is appropriate. When we introduce an additional covariate for a justice's vote on a specific case, we find that voting against a particular side are highly correlated with an increase in skeptical utterances directed toward that side, relative to neutral utterances by the same justice. However, a causal interpretation of this result depends on the assumption that justices are not persuaded during the course of the oral arguments.

Finally, we find that a justice is significantly more likely to voice skepticism in utterance u after another justice has done so in $u - 1$, but that this relationship only holds when the justice speaking at u votes against the side in question. This suggests that the piling-on of skepticism is not purely a question of low lawyer quality, but that strategic considerations may also be in play.

J Conclusion

In this paper, we introduced a new hierarchical hidden Markov model, the speaker-affect model, for classifying modes of speech using audio data. With novel data of Supreme Court oral arguments, we demonstrated that SAM consistently outperforms alternate methods of audio classification, and further showed that especially when training data are small, text classifiers are not a viable alternative for identifying modes of speech. The approach we develop has a broad range of possible substantive applications, from speech in parliamentary debates (Goplerud et al., 2016) to television news reporting on different political topics. With other interesting results on the importance of audio as data (Dietrich et al., 2016) accumulating, our approach is a useful and general solution that improves on existing approaches and broadens the set of questions open to social scientists.