

A Dynamic Model of Speech for the Social Sciences*

Dean Knox[†] and Christopher Lucas[‡]

May 2, 2019

Abstract

Auditory cues carry important information—helping politicians convey emotion, signal positions, and mark emphasis—that is lost when analysts transcribe speech recordings, then discard the original sources. We develop the first generative model of the *sound* and *flow* of political speech, the model of audio and speech structure (MASS), enabling new empirical tests of long-standing theoretical predictions about communication. Our approach models political speech as a stochastic process shaped by fixed and time-varying covariates, as well as the history of the speech or conversation itself. In an application to Supreme Court oral arguments, we demonstrate how auditory cues convey crucial information—the targeted use of skepticism—that is indecipherable to text models. Results show that justices do not use questioning to strategically signal or manipulate their peers, but rather react genuinely to information discovered in sincere fact-finding efforts. We provide a fast C++ implementation of the model in an easy-to-use R package.

Keywords: Speech dynamics; Signal processing; Conversation; Emotion; Hidden Markov model; Latent process

*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, Josh Boston, Bryce Dietrich, JB Duck-Mayr, Gary King, Connor Huff, In Song Kim, Adeline Lo, Jacob Montgomery, Jonathan Mummolo, David Romney, Jake Shapiro, 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

*Of all the talents bestowed upon men, none is so precious
as the gift of oratory. He who enjoys it wields a power
more durable than that of a great king.*

Churchill, The Scaffolding of Rhetoric

1 Introduction

Political science has always been, at least in part, the study of audio. Roosevelt’s fireside chats, Churchill’s World War II addresses to the House of Commons, the Lincoln-Douglas debates, Martin Luther King’s “I Have a Dream” speech; all pivotal moments in modern politics, each occurring in audio. Imagine analyzing these and similarly influential speeches quantitatively; text analysis would represent each no differently than if you, the reader, were the speaker. The rhetorical content would be lost entirely, and questions about the effects of that content would be completely out of reach. In this paper, we propose a new model - the model of audio and speech structure - that opens to statistical inquiry these decades-old questions.

Studies of text in social science often examine corpora which were first spoken, then transcribed. Though methodologically and substantively diverse, text analyses focus exclusively on *what* people say, while entirely ignoring the way in which those words were spoken. In this article, we develop the first generative model of political speech that explicitly represents how speakers express themselves, the model of audio and speech structure (MASS). Even without the text of speech, MASS is able to infer new quantities of interest by modeling the *sound* of speech—the auditory characteristics that makes a tone of voice recognizable—as well as the factors that shape the *flow of conversation*, or how different tones are deployed. And because

we allow conversation metadata to structure how speakers express themselves, MASS is able to incorporate textual information when it is available, along with any additional fixed or time-varying covariates.

In an application to Supreme Court oral arguments, we demonstrate how our model enables new insights into a particular political context by measuring expressions of skeptical speech on the bench. In contrast to widespread arguments that justices use oral arguments to *strategically signal* and manipulate their peers, we find that the use of skepticism is more consistent with a model of *genuine expression* in which justices engage in sincere fact-finding. But beyond this application, MASS can be used to analyze 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. By modeling the surrounding metadata structure, MASS allows researchers to assess how speech is shaped by characteristics of the speaker, the topic under discussion, or even how one individual responds to the speech of another. In short, MASS allows researchers to study new questions using new quantities of interest conveyed through speech.

A generative model of audio data is admittedly a departure from more conventional models of political communication. Here, it is useful to note that political science has in fact been concerned with the study of audio data for as long as the discipline has existed, in the sense 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 already do—but rather, “how should we represent it?” At present, virtually all studies of political speech focus exclusively on a particular representation of audio: textual content extracted by transcription. As Dietrich et al. (2016) show, discarding the audio results in the loss of

valuable information about political behavior.

While MASS builds on vast literatures in conversation analysis and audio classification, our approach is the first model of its kind in linguistics, signal processing, computer science, and the social sciences. In the study of text, the Structural Topic Model (Roberts et al., 2016, 2014) demonstrates that incorporating context—that is, explicitly modeling topic choice in terms of covariates—can result not only in improved performance, but also improved inference on the relationships of interest to social scientists. We adopt a similar approach by not only considering the auditory qualities of each utterance in isolation, but also the way in which tone is shaped by potentially time-varying covariates—how leaders project compassion after crises, for example, or whether hostility toward out-group members in a conversation provokes a similarly hostile response. We refer to these temporal effects as changes in the *flow* of speech, and our model is, as far as we are aware, the first in any field that permits the study of conversation flow. By doing so, MASS enables a range of new empirical analyses. For example, theories of persuasion focus on how logical arguments and cognitive shortcuts shape opinion formation (Festinger, 1964; McGuire, 1968; Chaiken, 1980). However, it is well-established that emotional state can affect cognition (Isen, 1984) and political information processing (Redlawsk, 2006). Yet to our knowledge, none have considered how passionate argumentation and emotional outbursts can shift the beliefs of other conversation participants, thereby altering the trajectory of political deliberation. MASS is built for precisely this kind of application, facilitating the conception and analysis of new areas of study. In negotiations, when do threats by a hawkish parties lead doves to back down, and when do they escalate tensions? During cable news interviews, do sympathetic or hostile questions shift the information content of an interview? And on the campaign trail,

do candidates adapt their tone and self-presentation based on the topic of their speech or their standing in the polls?

Each of these questions fundamentally revolve around conversation or speech flow, the process that MASS is designed to analyze. Our model of speech flow also permits MASS to draw on the sound of surrounding utterances to refine its beliefs about speech tone. Beyond innovation in the model itself, we also provide approaches for MASS model selection and regularization that remove the need for arbitrary researcher choices and allow researchers to exploit a larger range of potentially informative auditory features than past work in signal processing and computer science. Finally, we develop fast estimation procedures that permits bootstrap-based inference on model parameters and quantities of interest.

While we do not discuss our accompanying software package specifically, we implement our entire workflow in an easy-to-use R package. Our software package contains a fast C++ implementation of the model, as well as flexible functions for audio preprocessing and feature extraction, discussed in Section 3. In general, audio analysis is computationally intensive—our application to Supreme Court oral arguments, for example, analyzed over 70,000,000 rows of data with a multi-step procedure that involved batched processing, parallelization, and stagewise estimation. Our package wraps the entire procedure into a few simple functions to make its implementation broadly accessible.

In what follows, we begin by expanding on the observation that political science *already* studies audio data, reviewing the many empirical analyses that exclusively analyze the textual dimension of recorded human speech. In Section 3, we then describe how raw audio recordings can be prepared for statistical analysis, much as textual documents are preprocessed. We also review existing approaches for audio classification in signal process-

ing and computer science. Section 4 introduces our model and mathematical notation, then walks through an extended example illustrating how various MASS parameters capture Barack Obama’s rhetorical style in an inspirational campaign speech. We provide a high-level overview of the estimation procedure, reserving technical details for the Appendix. Finally, Section 5 presents a four-part application to expressed skepticism in Supreme Court oral arguments.

2 Political Science Already Studies Audio

Politics occurs in audio. Voters are exposed to political advertisements while watching YouTube [TO DO: CL CITE], coworkers discuss local politics around the water cooler [TO DO: CL CITE], and campaign speeches that go viral alter the courses of elections (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 ad’s memorable political jingle, the colleague’s partisan fervor, the candidate’s down-to-earth approachability—is often equally, and sometimes more, important than the text of human speech.

For example, consider the historic “Daisy” ad, aired by Lyndon Johnson’s successful campaign in the 1964 presidential election and described as the most effective negative advertisement of all time (Mann, 2011). The text of the ad contains a series of numbers and an encouragement to vote for Johnson.¹ Nothing in the transcript suggests a cutthroat,

¹The full text of the advertisement is “*Child*: One, two, three, four, five, seven, six, six, eight, nine... *Announcer*: Ten, nine, eight, seven, six, five, four, three, two, one, zero. *Explosion*. These are the stakes: to make a world in which all of God’s children can live, or 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.” (Italicized annotations are included for context and do not appear in transcript.)

visceral attack ad. But by juxtaposing the voice of an child innocently counting against that of an announcer counting down to a missile launch, followed by a nuclear explosion, the advertisement is threatening to a degree that cannot be appreciated without actually hearing the ad itself.² 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 5, we apply MASS 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 justices are more likely to cast “merits” votes supporting lawyers who present higher-quality arguments than the competing party, and Black et al. (2011) demonstrate that a justice’s merits vote reflects the sentiment expressed by that justice during the oral argument, making it less likely that party will prevail in the ultimate outcome. Johnson (2001) further shows that justices use oral arguments to gather information that is used to make substantive policy choices. These analyses all rely on either (1) a manual qualitative assessment of the oral argument or (2) an automated analysis of strictly the text of oral arguments, discarding the tenor of speech. However, tone carries important information about justice orientation toward the argument presented before them Dietrich et al. (2016). And in Section 5, we demonstrate that novel quantities of interest measured with MASS can shed light on these existing questions about justice behavior.

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

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

Koch, 2008; Freedman et al., 2004), parliamentary speech (Proksch and Slapin, 2015; Spiraling, 2016), or television and radio news (Behr and Iyengar, 1985; Sobieraj and Berry, 2011; Young and Soroka, 2012) inherently studies audio data, even if the audio component of speech is ignored after extracting textual transcripts.

Importantly, in addition to improving on existing fields of research, MASS also enables the study of new questions and quantities of interest. For example, in our application, we study expressions of skepticism, which are virtually impossible to extract from oral argument transcripts alone. And in addition to permitting the analysis of new outcomes, by modeling the *dynamics* of speech, MASS allows researchers to ask and answer questions about the flow of conversation, such as how a line of questioning at some point 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 Section 5 demonstrates how MASS can answer questions of this sort.

Like text (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

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 this section, we now explain how unstructured recordings of human speech can be represented as structured data, then analyzed statistically.

The primary unit of analysis in audio is the *utterance*: a continuous piece of speech by a single speaker that concludes with a clear pause. These utterances are unequal in length, but are typically on the order of ten seconds in duration. Within each utterance, we split the recording into successive *moments*, which each contain the sound in a narrow window of time; there are typically 50–100 such moments per second of speech.³ Section 4 describes this structure more formally. In each moment, we then summarize the raw audio with a set of representative features that are known to convey emotion and tone of voice, drawing on established literatures in psychology, phonetics, signal processing, and computer science (Ververidis and Kotropoulos, 2006; El Ayadi et al., 2011). Researchers can easily calculate these features with a single function in our accompanying R package, which also implements other useful preprocessing steps. For instance, our package includes functionality for trimming background noise and interruptions, along with novel approaches for splitting continuous speech into utterances.

In its most raw form, physical sound is merely a time-varying wave of pressure transmitted through space. These waves are recorded digitally by repeatedly measuring pressure at frequent intervals—tens of thousands of times per second. Thus, each *moment* that we analyze is actually a univariate time series with hundreds of such pressure samples. We then characterize each moment using a number of auditory features that describe the sound perceived by a listener. While the features we calculate are more abstract than a textual

³We use overlapping windows, each 25 milliseconds long, incrementing by 12.5 milliseconds.

representation of the audio, they can still be easily understood. We illustrate selected features in Figure 1 by describing an audio source for which that feature is high or low. For example, some features are simple functions of the raw pressure waveform. Figure 1 illustrates one such feature: the “zero crossing rate” (ZCR), which is simply the rate at which the audio wave crosses zero, or neutral pressure. Sibilance (a characteristic of /s/ or /z/, for example) has a particularly high ZCR, whereas vowels have relatively low ZCR. In general, zero crossing rate helps distinguish voiced speech (that employing the vocal cords) from unvoiced speech (that which does not use the vocal track, like /s/). However, some voiced speech nonetheless has a relatively high ZCR. For instance, /z/ is pronounced by vibrating the vocal cords while forming the mouth shape of /s/; both have a relatively high ZCR despite the fact that one is voiced. There is no single feature that distinguishes all sounds, and so we use additional features like “energy” (effectively, loudness), which can be used jointly with ZCR to distinguish the sound of /s/ and /z/.

In addition to features that are simple functions of the raw audio wave, we also calculate features based on its spectrum, which captures (among other things) the contribution of the bass or treble ranges to the overall loudness. To do so, we first decompose the raw audio using a Fourier transform. This maps the audio waveform in each moment from its original representation (pressure varying over time), into a new representation describing the energy in various frequencies. In Figure 1, this is illustrated by the “spectral density” row, which contrasts male and female speech spectra. A number of additional features can be computed from this representation: For example, the Mel-frequency cepstral coefficients (MFCCs) capture its overall shape, and pitch (the dominant frequency) can be extracted from its peaks. These features provide additional emotional information. In English, words


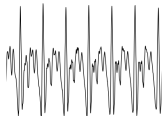
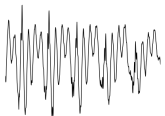

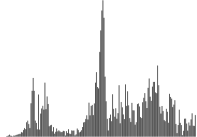
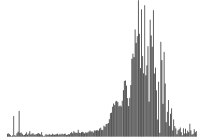


| | Low Exemplar | | High Exemplar | |
|------------------|--------------|---|---------------|---|
| Energy | “ahhh” |  | “AHHH!” |  |
| ZCR | /a/ |  | /s/ |  |
| Spectral Density | man |  | woman |  |
| Pitch | bass |  | treble |  |

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

may be emphasized with higher sustained pitch or with sharp falling pitch. (Consider “Why are you citing that case?” with “Why are you citing *that* case?”) Pitch is also higher when speakers are vocally tense, including when speakers are emotionally aroused. Additionally, we compute various functions of these raw features, including interactions and derivatives that are known to be informative. Appendix Section A.2 provides a full list of the features that we construct from the raw vector of audio samples.

In Section 4, we describe how MASS models the way in which auditory features vary with and convey a speaker’s tone of voice. Our approach further accommodates 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).

3.1 Advances Over Existing Approaches

Outside of political science, a large body of research across a range of fields attempts to model the audio of human speech. A common and straightforward approach is to use descriptive statistics (e.g., mean, minimum, maximum, standard deviation) of the audio features described above. This collapses the rich audio data (one feature vector for each moment) into a single flat summary vector per utterance, which can subsequently be input to standard machine-learning models to classify speech mode (Dellaert et al., 1996; McGilloway et al., 2000). However, these reduced representations discard enormous amounts of auditory data. To avoid this loss of information, hidden Markov models are commonly used to better model the temporal nature of speech.

MASS builds on these existing approaches in statistics and computer science in three main ways. First, typical HMM-based analyses are only able to use a fraction of the features we incorporate. It is common to arbitrarily select a dozen features and discard the rest.

Nogueiras et al. (2001), for example, use just two features⁴ 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. Second, MASS is the first to directly model the *flow* of speech—that is, the sequence of speech modes—in terms of the meaningful structural features encoded in conversation metadata.

4 A Model of Conversation Dynamics

We now develop a generative statistical model for the audio data that arises from political speech. After outlining the model in Section 4.1, we illustrate it in Section 4.2 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, crowd-rousing 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 factors 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 in Section 4.3, then discuss practical considerations in the modeling of speech.

⁴Pitch and energy; see Appendix A for a description.

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 speech 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, encoded in the vector \mathbf{W}_u . Importantly, \mathbf{W}_u may include functions of conversation history, such as aggregate anger expressed by previous speakers over the course of an argument— $\sum_{u' < u} 1(\mathbf{S}_{u'} = \text{anger})$ —which might induce a sharp retort in utterance u .

To keep the exposition clear, here we consider the simplified setting in which a single conversation is analyzed. Appendices B.2–B.4 present the general multi-conversation model, which is essentially identical. The model is defined in Equations 1–4 and summarized graphically in Figure 2. First, we model speech mode probabilities as a multinomial logistic function of conversation context,

$$\Delta_m = \exp(\mathbf{W}_u^\top \boldsymbol{\zeta}_m) / \sum_{m'=1}^M \exp(\mathbf{W}_u^\top \boldsymbol{\zeta}_{m'}) \quad (1)$$

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

where $\boldsymbol{\Delta} = [\Delta_1, \dots, \Delta_M]$ and $\boldsymbol{\zeta}_m$ is a mode-specific coefficient vector through which \mathbf{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, \mathbf{X}_u , is the only available information. Each \mathbf{X}_u is a matrix in which the t -th row contains 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 then assume that the m modes of speech are each associated with their 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, the speaker alternates through these sounds according to

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

with $\Gamma_{k,*}^m$ denoting rows of the transition matrix, $[\Pr(R_{u,t} = 1 \mid R_{u,t-1} = k), \dots, \Pr(R_{u,t} = K \mid R_{u,t-1} = k)]$.

By modeling the usage patterns of different sounds in this way, we approximately capture the temporal structure that plays an important role in speech. (For example, most sounds are sustained for at least a few moments, and certain phonemes can only occur before the silence at the end of a word.) In turn, sound k is associated with its own auditory profile, which we operationalize as a multivariate Gaussian distribution with parameters $\boldsymbol{\mu}^{m,k}$ and $\boldsymbol{\Sigma}^{m,k}$. Finally, the raw audio heard at moment t of utterance u —the signal perceived by a listener—is drawn as

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

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

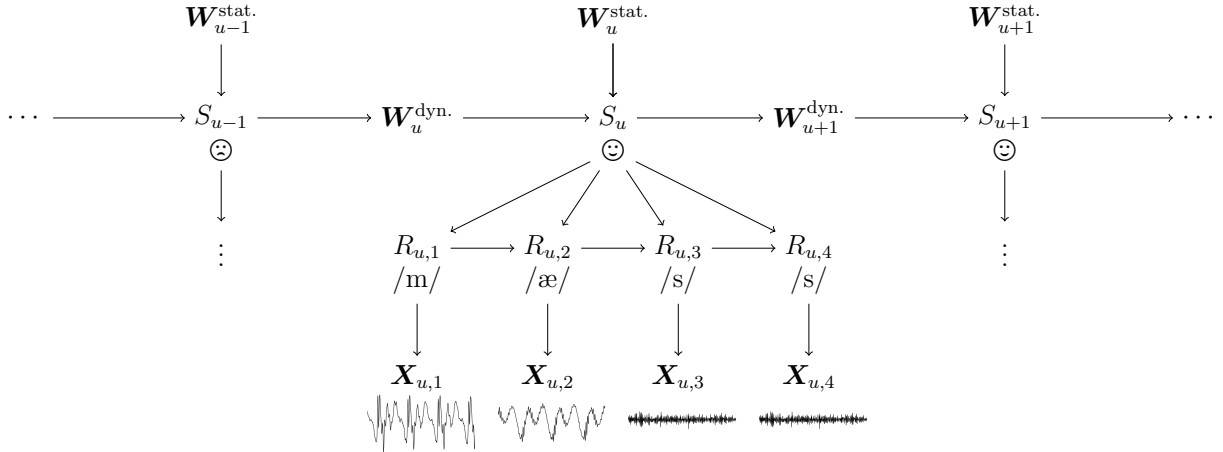


Figure 2: **Illustration of generative model.** The directed acyclic graph represents the relationships encoded in Equations 1–4. In utterance u , the speaker selects tone S_u based on “static” (i.e., externally given) time-varying covariates $\mathbf{W}_u^{\text{stat.}}$ as well as “dynamic” conversation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

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 5, 2012.⁵ While this example represents only a brief scene from an extended campaign, it demonstrates many of the speech dynamics that motivate our model. After showing how MASS parameters map onto the primary theoretical quantities of interest, we outline in detail how an analyst would go about extending the illustration to a full-scale analysis of emotional appeals in campaign rhetoric.

In what follows, we begin with an close examination of two prototypical utterances that represent what we call his “crowd-rousing” roar and his engaging “storytelling” narrative

⁵The excerpt can be viewed at [youtube.com/watch?v=pAV9Tic_bYo&t=1713](https://www.youtube.com/watch?v=pAV9Tic_bYo&t=1713), along with the full campaign speech.

voice. 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. 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.”⁶ In Figure 4.2.C.1, 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 Obama transitions to consonants (light red strikethrough) and eventually pauses in silence⁷ (white) between words. One such transition is depicted in Figure 4.2.D.2. Just as a human can recognize phonemes from their auditory characteristics, our model automatically learns to distinguish vowels (based on their higher autocorrelation, as encoded in $\mu^{\text{rousing,vowel}}$) from consonants (high zero-crossing rate), as shown in Figure 4.2.E.

⁶We 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.

⁷Because each frame describes just milliseconds of audio, the short pauses between words are an observable component of speech.

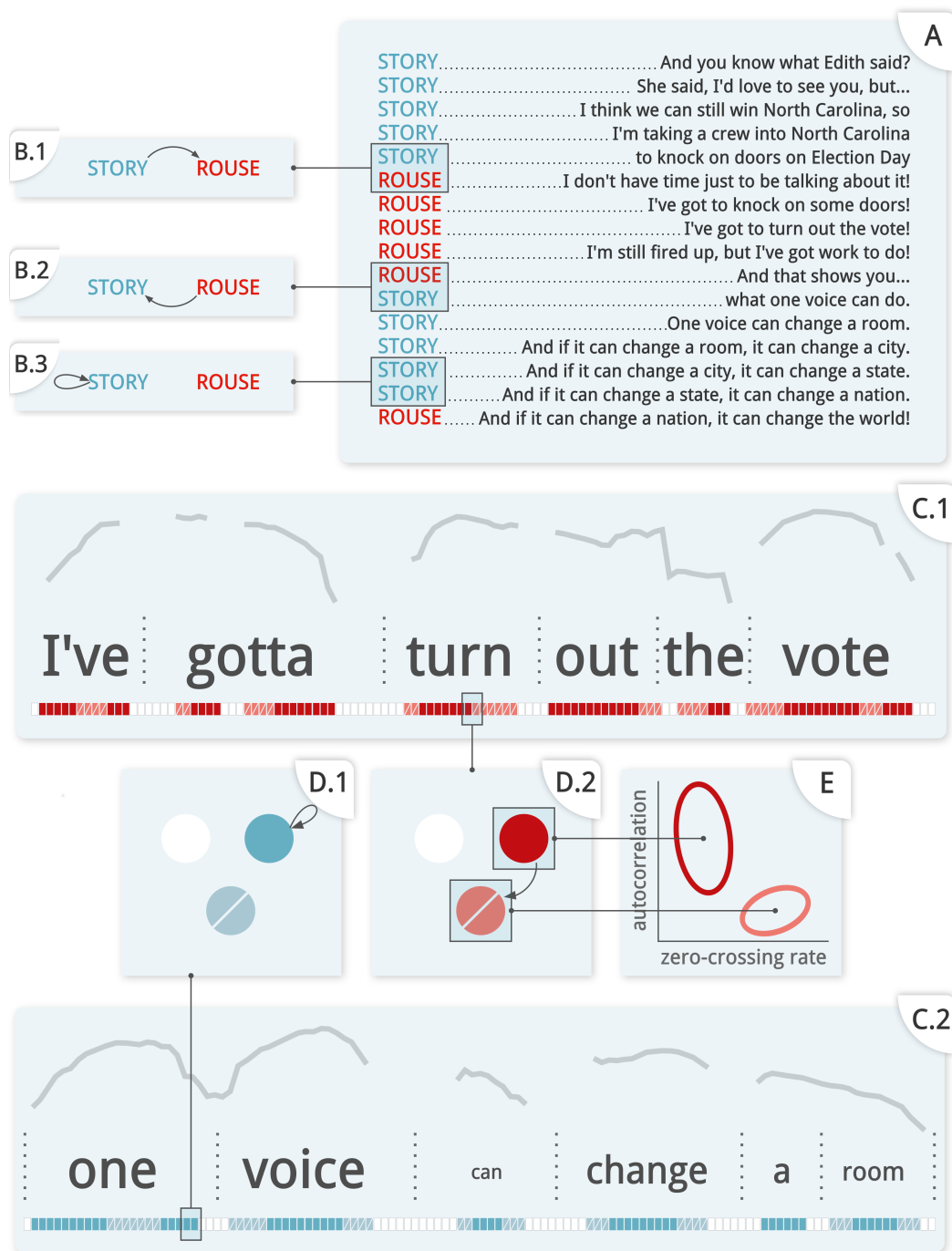


Figure 3: **An illustrative example.** Panel A contains an excerpt of Obama’s final campaign speech, utilizing “storytelling” conversation and “crowd-rousing” shouts. Call-outs highlight successive utterance-pairs in which the speaker shifted from one mode to another (B.1), transitioned back (B.2), and continued in the same tone of voice (B.3). Panel C.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, C.2 shows the far smoother nature of his “storytelling” voice, in which pitch rises and falls gradually and the modulation of volume places emphasis on “voice” and “change.” Call-outs D.1 and D.2 respectively identify sequential moments in which a “storytelling” vowel is sustained (transition from the dark blue sound back to itself, indicating repeat) and the dark red “crowd-rousing” vowel transitions to the light red consonant. Panel E shows the differing auditory characteristics of the “crowd-rousing” vowel and consonant, which are perceived by the listener.

Why does this matter? It is on the basis of these constituent sounds that MASS is able to discern differences between rhetorical styles. As Figure 4.2.A makes clear, MASS 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 4.2.C.2 (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, $\Sigma^{\text{story,vowel}}$ captures higher variance in loudness when compared to crowd-rousing speech (Figure 4.2.C.1), where every word is shouted at near-constant volume. Differences in average pitch—often a marker of emotional engagement—are represented in the μ terms, and shifts in cadence manifest in the Γ matrices.

These models of storytelling and crowd-rousing speech enable analysts to categorize hundreds or even thousands of hours of previously unheard speech. But learning to distinguish “crowd-rousing” speech from “storytelling” speech is only the beginning for MASS. The most important questions in the analysis of political speech relate to its ebb and flow—when and why a speaker chooses to deploy a particular tone. After learning to recognize tone in the lower stage (Equations 3–4), MASS moves on to estimate Obama’s flow of speech by modeling the contextual determinants of speech tone (Equations 1–2). Figure 4.2.A provides an illustration of 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 (Figure 4.2.B.1). Obama then repeats the narrative cycle by quieting down (Figure 4.2.B.2),

drawing rallygoers in with his soft-spoken speech before ultimately finishing with a dramatic crescendo.

This isolated example, despite comprising only a short excerpt 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 of the modern era. Figure 4 contrasts his dynamic performance with that of Mitt Romney’s final speech on the same night, the eve of the 2012 election.⁸ While Romney clearly seeks to inspire with “the door to a brighter future is open,” the delivery falls flat. On the highest-stakes night of the campaign season, his voice is dully monotone and weakly trails off in volume.

While scholars can easily listen to and compare a few short audio recordings, the amount of time required to digest an entire campaign’s worth of speech—hundreds of addresses, containing tens of thousands of utterances—rapidly grows infeasible. MASS makes it possible to identify broad patterns in the drivers of political speech, analyzing large-scale audio corpora while still incorporating human judgment about tone and expressed emotion. In Section 4.3, we develop a procedure for doing so; Algorithm 1 describes the steps in detail. Broadly speaking, the model learns to identify micro-level patterns, such as those described above, based on moderately sized human-provided training examples. MASS then uses its knowledge (say, of crowd-rousing speech) to crudely categorize every Obama utterance over the multi-year campaign season. Based on their sequence and contextual covariates, MASS identifies patterns in tone usage, then uses these patterns to iteratively refine its utterance

⁸The excerpt can be viewed at youtube.com/watch?v=CYhS84JbMEU&t=1530, along with the full campaign speech.



Figure 4: **Contrasting candidates.** Utterances from the conclusions of Obama’s and Romney’s election eve speeches on November 5, 2012. Both candidates sought to portray themselves as inspirational leaders. However, their approach to doing so differed tremendously. Obama’s “one voice can change a room” utilized modulation in both his pitch and volume, as the main text discusses in detail. In contrast, Romney’s “the door to a brighter future is open” consists of a monotonic drone—a single, unchanging pitch—that gradually trails off in volume. The sharp distinction illustrates the oratory skill to which Obama’s ultimate success is often attributed.

predictions and the flow-of-speech parameters.

By introducing a conversation-specific covariate for daily polling numbers, an analyst could easily incorporate additional factors into the model: e.g., studying how candidates attempt to come back from a losing position by revamping their public persona. Time-varying covariates representing the current textual topic could be used to assess how candidates modify their speech when talking about wounded veterans, as opposed to climate change or the national debt. And still other covariates might help shed light on how candidates adapt their campaign rhetoric to rural, minority, or hostile audiences. A closer examination of the contexts in which rhetorical tones and projected emotions are deployed—whether they be compassion, indignation, approachability, or fearmongering—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.

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 might not be a subset of the primary corpus of interest—we consider each mode of speech in turn, using a fully unsupervised approach to learn the auditory profile and cadence of that speech mode. The results are applied to the full corpus to obtain “naïve” estimates of each utterance’s tone, based only on the audio features and ignoring conversational context. We then fit a

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

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, \mathbf{X} , are observed for all utterances. The portion of the conversational context that relates to static metadata ($\mathbf{W}^{\text{stat.}}$) is available for at least the primary corpus, but dynamic contextual variables that depend on the tone of prior utterances ($\mathbf{W}^{\text{dyn.}}$) can only be estimated. In general, the tone of each utterance (\mathbf{S}) is also unobserved, but the analyst possesses a small training set with human-labeled utterances.

model for the flow of conversation, use this to refine the “contextualized” tone estimates, and repeat in an iterative procedure. The specifics of each step are discussed below and in Appendix B, and the workflow 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, \mathbf{X} , are observed for both the primary corpus and the training set. However, human-labeled tone of speech, \mathbf{S} , is only known for the training set. We divide the conversational context into externally given but potentially time-varying “static metadata,” $\mathbf{W}^{\text{stat.}}$, and deterministic functions of conversation history that dynamically capture the prior tones of speech, $\mathbf{W}^{\text{dyn.}}$. 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, ζ , and the auditory parameters of each tone, which we gather in $\Theta^m = (\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m, \boldsymbol{\Gamma}^m)$ for compactness. In what follows, we also refer to the collection of all tone parameters as $\Theta = (\Theta^m)_{m \in \{1, \dots, M\}}$. Under

the model described in Equations 1–4, the likelihood can be expressed as

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

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

As a concession to computational constraints, we estimate the parameters in a stagewise fashion. The auditory parameters, $\boldsymbol{\Theta}$, are calculated by maximizing the partial likelihood, $f(\mathbf{X}^{\mathcal{T}} \mid \boldsymbol{\Theta}, \mathbf{S}^{\mathcal{T}})$, corresponding to the training factor, rather than the full likelihood in Equation 5 (Wong, 1986). The full likelihood is then maximized with respect to the conversation flow parameters $\boldsymbol{\zeta}$, conditional on $\boldsymbol{\Theta}$. The factorization and a detailed discussion of stagewise estimation are presented in Appendix B.1.

In Appendix B.2, we detail our procedure for estimating the auditory profile and cadence for each speech mode. First, training utterances are divided according to their tone labels. Because the partial likelihood can be neatly factorized further as $f(\mathbf{X}^{\mathcal{T}} \mid \boldsymbol{\Theta}, \mathbf{S}^{\mathcal{T}}) = \prod_{m=1}^M \prod_{u \in \mathcal{T}} f(\mathbf{X}_u \mid \boldsymbol{\Theta}^m)^{\mathbf{1}(S_u=m)}$, $\hat{\boldsymbol{\Theta}}^m$ can be independently estimated for each speech mode with no further loss of information. For all training utterances of speech mode m , a regularized variant of the Baum-Welch algorithm, a standard estimation procedure for hidden Markov models, is used to obtain $\hat{\boldsymbol{\Theta}}^m$ for the corresponding mode. Each of the resulting M tone-specific models are then applied to each utterance u in the primary corpus to obtain the corrected emission probabilities $f(\mathbf{x}_u \mid \hat{\boldsymbol{\Theta}}^m, S_u = m)^\rho$, which represents the probability that the utterance’s audio was generated by speech mode m ; this captures the extent to which the audio “sounds like” the relevant training examples. Naïve tone estimates can

then be computed by combining these with the overall prevalence of each tone via Bayes’ rule. The corrective factor, ρ , approximately accounts for unmodeled autocorrelation in the audio features and ensures that the naïve estimates are well-calibrated (for details, see Appendix B.3). This shared correction, along with the number of latent sounds and strength of regularization, are determined by likelihood-based cross-validation (van der Laan et al., 2004) in the training set.

We now briefly describe an expectation-maximization algorithm for the conversation-flow parameters, ζ , reserving derivations and other details for Appendix B.4.⁹ An inspection of Equation 5 shows that this procedure will depend only on $f(\mathbf{X}^c \mid \zeta, \Theta, \mathbf{W}^{\text{stat.},c})$, since the remaining term does not involve ζ . We proceed by augmenting the observed data with the latent tones, \mathbf{S}^c , and the conversation-history variables that depend on them, $\mathbf{W}^{\text{dyn.},c}$. The augmented likelihood, $f(\mathbf{X}^c, \mathbf{S}^c, \mathbf{W}^{\text{dyn.},c} \mid \zeta, \Theta, \mathbf{W}^{\text{stat.},c})$, is then iteratively optimized. However, the closed-form expectation of the augmented likelihood is intractable. We therefore replace the full E step with a blockwise procedure that sweeps through the unobserved E-step variables sequentially.¹⁰ Finally, the maximization step for ζ reduces to a weighted multinomial logistic regression in which $\mathbf{1}(S_u = m)$ is fit on $\mathbb{E}[\mathbf{W}_u \mid S_{u-1} = m']$ 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

⁹This estimation procedure builds on forward-backward algorithms; interested readers are referred to Appendix B.2 or standard references such as Zucchini and MacDonald (2009).

¹⁰The use of this alternative procedure leads to a smaller improvement of the EM objective function than a full E-step. Nevertheless, algorithms using such partial E- or M-steps ultimately converge to a local maximum, just as traditional expectation-maximization procedures (Neal and Hinton, 1998).

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 5, follow directly from the conversation-flow parameters, ζ , or the auditory parameters, Θ ; inference on these quantities follows a similar bootstrap approach.

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

Result: Auditory parameters Θ , conversation flow parameters ζ

Procedure:

1. Define problem.

Analyst determines tones of interest and rubric for human coding.

Human-coded tone labels are obtained for training set ($\mathcal{S}^{\mathcal{T}}$).

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

```

for speech mode  $m$  in  $1, \dots, M$  do
    Subset to training utterances labeled as tone  $m$ .
    while not converged do
        for utterance  $u$  in  $\mathcal{T}$  and moment  $t$  in  $\{1, \dots, T_u\}$  do
            for sound  $k$  in  $1, \dots, K$  do
                | Compute emission probability of sound  $(m, k)$  generating audio ( $\mathbf{X}_{u,t}$ ).
            end
        end
        Predict sound being pronounced at each moment ( $R_{u,t}$ ).
        Update cadence (usage patterns of constituent sounds,  $\mathbf{\Gamma}^m$ ).
        for sound  $k$  in  $1, \dots, K$  do
            | Update audio profile of sound  $k$  ( $\mu^{m,k}, \Sigma^{m,k}$ ).
        end
    end
end

```

3. Fit conversation-flow parameters (ζ) using primary corpus (\mathcal{C}), conditional on Θ

```

for utterance  $u$  in  $\mathcal{C}$  do
    for speech mode  $m$  in  $1, \dots, M$  do
        | Compute corrected emission probability of speech mode  $m$  generating utterance
        | audio data ( $\mathbf{X}_u$ ), 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 ( $\mathbf{W}_u$ ).
    Update flow-of-speech parameters ( $\zeta$ ).
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 (Cornell Legal Information Institute, n.d.).¹¹ We limit our analysis to the natural court that begins with the appointment of Justice Kagan and concludes with the passing of Justice Scalia, so that the composition of the Court remains constant for the entirety of the period we analyze. The Oyez data contains an accompanying textual transcript, speaker names for each utterance, and timestamps for utterance start and stop times. In addition, we inferred the target side (i.e., petitioner or respondent) of each justice’s question based on the side of the most recently speaking lawyer. Additional case data was merged from the Supreme Court Database (Spaeth et al., 2014)

Using Oyez timestamps, we segmented the full-argument audio into a series of single-speaker utterances.¹² As an additional preprocessing step, we drop utterances spoken by lawyers (each of whom usually appears in only a handful of cases) and Clarence Thomas (who spoke only twice in our corpus), focusing on the behavior of the eight recurrent speakers. We also drop the largely formulaic introductory and concluding remarks, along with utterances shorter than 2.5 seconds.¹³ After trimming, the resulting audio corpus contains 403 arguments and 244 hours of audio, comprising nearly 107,000 justice utterances and 70 million moments.

¹¹Dietrich et al. (2016) independently collected the same audio data and conducted an analysis of vocal pitch.

¹²Occasionally, 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 by this speaker.

¹³We found that these extremely short utterances contained substantial amounts of crosstalk. However, they also include potentially informative interjections; future work may explore improved preprocessing techniques that do not require discarding this information.

In the following applications, 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 to reflect not only by vocal tone but also the textual content of the utterance.¹⁴ Thus, we define 16 modes of speech—two tones for each of the eight speaking justices.¹⁵ During classification, we dropped the handful of utterances (roughly 5%) in which crosstalk or other audio anomalies occurred, or in rare instances where the speaker’s identity was incorrectly recorded.

Four analyses of skeptical Supreme-Court speech are presented below. First, we demonstrate that MASS 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 about 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. We then demonstrate the workings of the lower-level auditory component of MASS in a third analysis, examining the auditory content of median justice Anthony Kennedy’s skeptical and neutral speech. Here we illustrate how expressed skepticism manifests in terms of loudness, vocal tension, and pitch modulation. Together, these auditory

¹⁴In the coding process, our guiding classification principle was “If I were a lawyer, would I be happy after hearing this utterance?”

¹⁵Because the transcripts attribute each utterance to a speaker, the model’s decision is over whether the current statement by Anthony Kennedy was skeptical or neutral. That is, we do not conduct a joint speaker-recognition and tone-detection task. In the framework outlined in Equations 1–2, this is equivalent to introducing a covariate for the current speaker’s identity, with a corresponding coefficient of $-\infty$ for the 14 speech modes that do not correspond to the current speaker.

characteristics allow listeners to obtain a clear nontextual signal of Kennedy’s projected emotional state; we then show how speech differs across individuals with a comparison to justice Sotomayor. Finally, we use the full model to analyze the structural determinants of oral-argument skepticism. Using justices’ ideological leanings, their ultimate vote, and a measure of case contentiousness, we test the observable implications of two commonly espoused but conflicting narratives of the Supreme Court decisional process: that justices are highly strategic actors jockeying for influence, on the one hand (Johnson et al., 2006), or alternatively that they are neutral arbiters who respond genuinely to compelling legal arguments (Johnson et al., 2006).

5.1 Facial Validity of Predicted Skepticism

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

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

Table 2: **Transcripts of Skeptical and Neutral Utterances.** Left (right) columns contain ten transcripts of utterances in the top 10% of predicted skepticism (neutrality). While MASS is estimated solely on audio data and conversation context, its fitted values accord well with qualitative readings of the utterance text.

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 MASS. 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. 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. To do so, we treat MASS fitted probabilities of skepticism (based on audio features and conversation context) as the outcome. We then employ a post-LASSO procedure in which a cross-validated LASSO-logistic model is estimated, then an unregularized logistic regression is fit on the selected terms (Belloni et al., 2016).

The resulting coefficient estimates, plotted in Figure 5, demonstrate that there are extraordinarily few consistent textual indicators of expressed skepticism—the vast majority are statistically indistinguishable from zero at conventional levels. In Figure 6, we arbitrar-

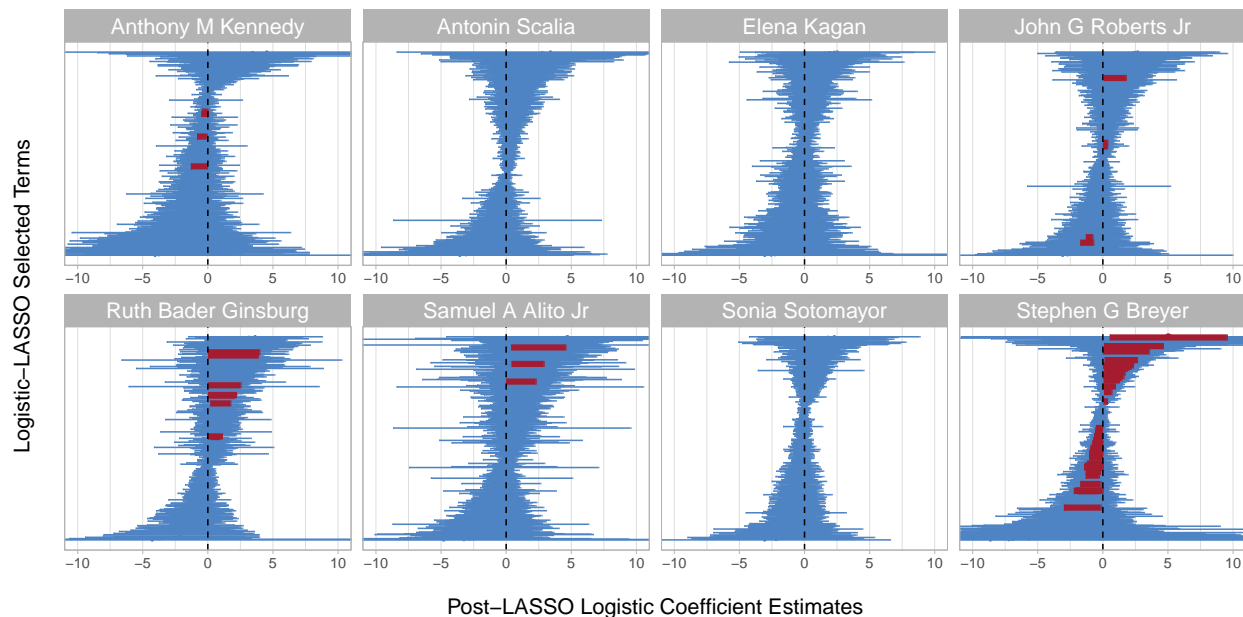


Figure 5: **Textual Signals of Justice Skepticism.** Each panel depicts a regression of MASS-predicted skepticism on word counts. Within each justice’s utterances, candidate terms that may predict skepticism (selected by logistic LASSO) are arrayed on the y -axis. For each term, points and horizontal errorbars depict post-LASSO logistic regression estimates and confidence intervals. Thin light blue (thick dark red) errorbars reflect 95% confidence intervals that (do not) overlap zero.

ily discard speaker-terms with p -values exceeding 0.05, then investigate the remainder more closely.

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

- A sarcastic retort to an attempt to introduce new arguments, “*Ah*, now we have ‘sufficiently involved;’ ”

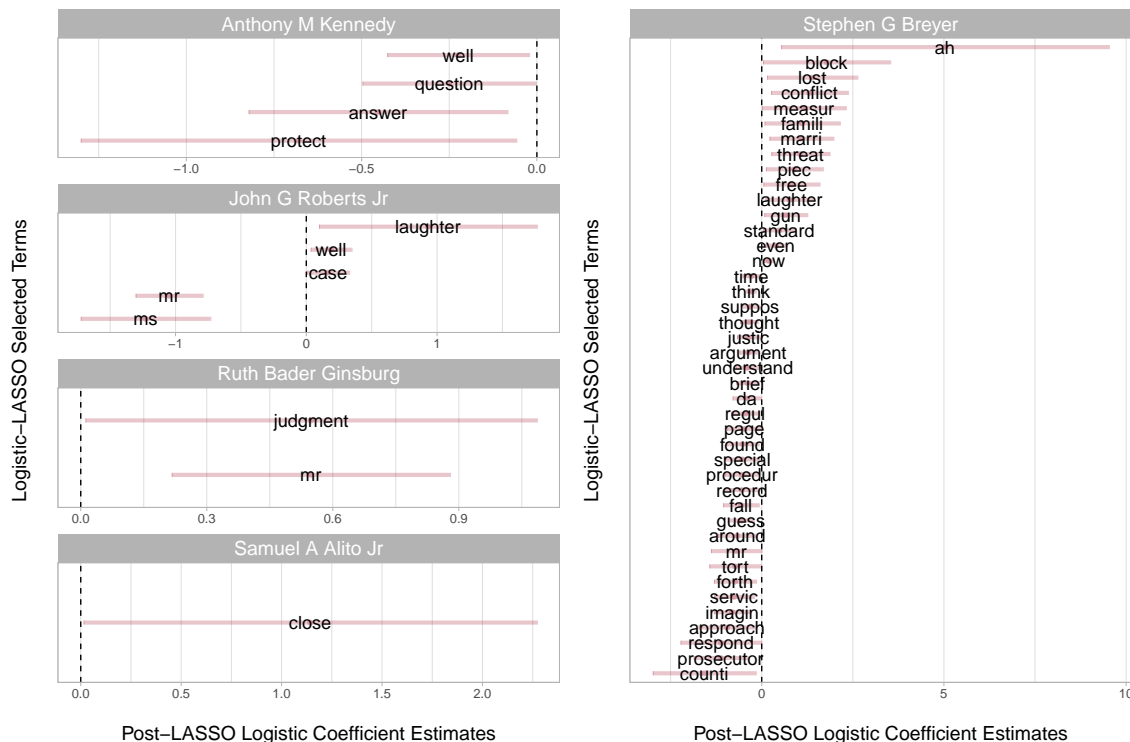


Figure 6: **Strong Textual Signals of Justice Skepticism.** Each panel depicts a regression of MASS-predicted skepticism on word counts, within a justice’s utterances. Words that predict skepticism are arrayed on the y -axis. Reported terms are the subset of post-LASSO terms with post-selection logistic regression confidence intervals (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, Justice Breyer’s neutral-leaning terms include technical terms (“tort,” “brief,” and “procedure”) as well as the fairly innocuous “guess” and “imagine”). While this particular justice’s textual cues are plausible, however, his colleagues are far more difficult to read using word frequencies alone—perhaps because they signal their position in subtle ways, or

perhaps because text is just a poor indicator of expressed emotion. For all other justices, we identify fewer than five informative words through this procedure; moreover, their cumulative predictive power is virtually nonexistent.

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, distinguishes skeptical questioning from neutral speech? To demonstrate, we interpret MASS results by investigating the auditory characteristics of median justice Anthony Kennedy’s speech. For Kennedy, we found that a moderately regularized speech model with $K = 3$ latent sounds minimized the total cross-validated likelihood of out-of-sample auditory features. Three well-separated sound classes can be consistently observed across model runs. We subjectively characterize these as “voiced speech” such as vowels, in which the vocal cords vibrate (high autocorrelation); “unvoiced speech,” such as fricatives and sibilants, in which vocal are not used (moderate energy and zero-crossing rate); and “silence” (low energy). Using an alignment procedure described in Appendix C.1, 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. The top panel of Figure 7 shows that when speaking skeptically, Kennedy speaks more loudly and with higher average pitch, a consequence of tensed vocal cords. Moreover, his modulation of pitch—which rises during questions and falls sharply during emphatic statements—is markedly larger in skeptical speech, as indicated by its higher pitch variance. We do not, however, observe similar mod-

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

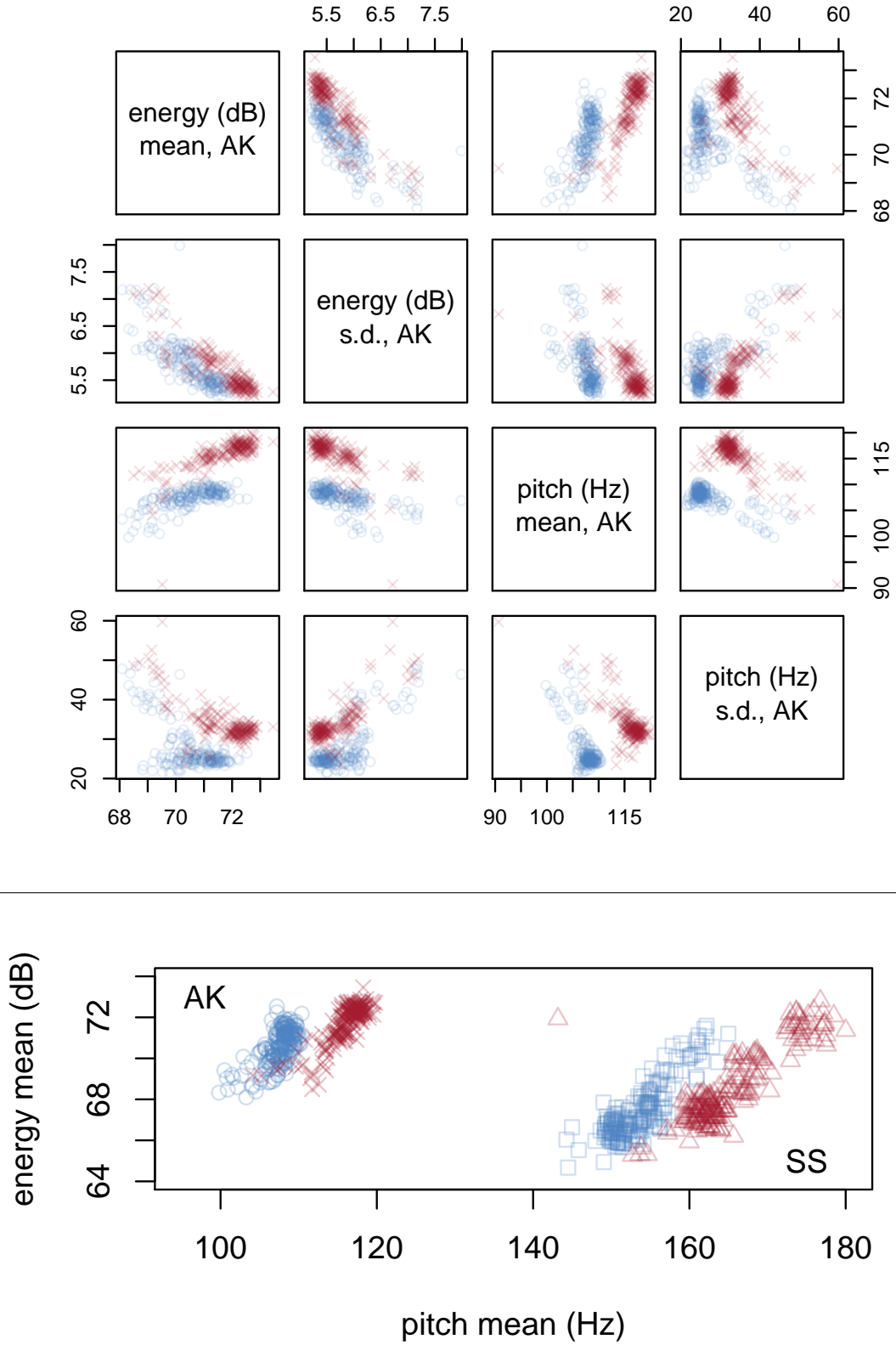


Figure 7: In the top panel, each dark red \times (light blue \circ) represents a converged EM run for auditory parameters using a run-specific bootstrap draw of skeptical (neutral) training utterances for Justice Kennedy. Coordinates in a bivariate scatterplot are based on elements of $\mu^{\text{skeptical, voiced}}$ ($\mu^{\text{neutral, voiced}}$) and the diagonal of $\Sigma^{\text{skeptical, voiced}}$ ($\Sigma^{\text{neutral, voiced}}$). For example, the top right panel demonstrates that when speaking skeptically, Justice Kennedy's voice is markedly louder and exhibits more variation in pitch, relative to his neutral speech. The bottom panel compares the same parameters for Justice Sotomayor's skeptical (neutral) voiced speech, depicted with dark red \triangle (light blue \square). While her voice is generally higher and quieter, on average, Sotomayor also communicates skepticism by elevating her pitch and speaking more loudly.

5.4 Structural Determinants of Expressed Skepticism

We now turn to a substantive question of contention in the courts literature: What purpose does questioning serve in oral arguments? The answer is subject to considerable debate (Wasby et al., 1976; Shapiro, 1984; McGuire, 1995; Johnson, 2001; Shullman, 2004; Johnson et al., 2006, 2009; Epstein et al., 2010; Wedeking, 2010; Black et al., 2011, 2013); indeed, one Supreme Court justice has even asserted that the act of questioning is neither necessary nor helpful (Thomas, 2013). While courts scholars have advanced a number of theoretical accounts of the deliberative process, these explanations can generally be grouped into two broad classes. One narrative holds that justices are shrewd political actors that maneuver to influence the decisions of their peers in pursuit of a desired case outcome (Shullman, 2004; Epstein et al., 2010; ?; Iaryczower and Shum, 2012; Iaryczower et al., 2018). However, they are constrained by strong judicial norms against back-room discussions, which foreclose the possibility of private communication. In this account, oral arguments represent an opportunity for justices to *strategically signal* to their colleagues, with lawyers and their legal arguments serving merely as convenient foils. A second, largely incompatible conception of the decision-making process considers justices as neutral arbiters, each casting a *genuine vote* according to rules determined by their respective legal philosophies (Johnson, 2001; Black et al., 2011, 2013). In this latter account, oral arguments provide an opportunity for fact-finding; while justices may reveal their predispositions with a display of doubt, this merely reflects an honest response to compelling or unconvincing legal reasoning.

These competing accounts are challenging to disentangle even under the best of circumstances. This difficulty has been compounded by widespread reliance on a narrowly limited

representation of judicial speech: textual transcripts alone. Here, we demonstrate that the discarded audio channel contains information of enormous value to social scientists—and that by modeling the tone it conveys, MASS not only opens the door to new research questions but can also shed new light on existing puzzles.

When and why justices deploy skepticism in oral arguments? Differing theoretical accounts of deliberation suggest very different patterns in the usage of this tone. A model of *genuine voting* implies when justices communicate their skepticism (to the extent that they make it known at all), it is largely as a natural reaction to poor argumentation. In other words, we should observe generically higher rates of skepticism when justices question lawyers for the side that they find less persuasive. This leads to an observable implication of the genuine-voting theoretical account: when a justice votes against a side, we should observe that this behavior is associated with increased skeptical questioning of the corresponding lawyers (Black et al., 2011, 2013).

A *strategic-signaling* model of deliberation, on the other hand, must account for the fact that many—indeed, nearly half—of all cases are decided unanimously. When all justices agree, no strategy is necessary. There is little to gain from posturing, including acted skepticism. To the extent that experienced justices are able to identify uncontroversial cases from pre-argument legal briefs and lower court decisions, this suggests a key observable implication of the strategic-signaling account: justices should exhibit greater skepticism toward their non-preferred side *especially* in contentious cases. That is, in cases that are ultimately decided by a 5-4 margin, we should see justices use markedly more skepticism toward the side they vote against. Forward-looking justices should similarly reduce skepticism toward their own side to avoid damaging its chances in close calls.

A third test concerns the dynamics of oral arguments. In general, justices who are ideologically close will exhibit greater similarity in preferences and judicial perspectives, relative to those who are far apart. When justice i finds a line of legal reasoning to be objectionable (as manifested in an expression of skepticism) it is likely that their ideological neighbor j will find it objectionable as well. The two narratives then diverge in their predictions for j 's response. A *genuine* reaction would be to acknowledge the flaw in reasoning, perhaps following up with further skeptical probing regardless of j 's affinity for the lawyer under attack. In contrast, if i is ideologically distant from j , then i 's skepticism should not provoke much of a response from j due to the relative lack of shared hot-button issues. The *strategic* account, on the other hand, implies a very different flow of questioning. Suppose that j dislikes the current lawyer. If j was a savvy justice, they should be on the lookout for weaknesses in the opposing side's arguments, seizing the chance to dogpile when an opportunity presents itself. Ideological distance from i —the preceding critic—should not restrain the shrewd operator much, if at all. Indeed, a left-right combination may be a particularly effective blow against the current lawyer.

The *strategic* narrative suggests a very different sequence of events when j 's preferred side comes under attack, however. When ideological neighbor i expresses skepticism, j has an incentive to smooth things over—despite j 's ideological inclination to agree with i 's points. Thus, the extent to which ideological proximity colors j 's response to prior skepticism is a useful point of differentiation. Specifically, we discretize ideology into “left” (Justices Breyer, Ginsburg, Kagan, and Sotomayor) and “right” (Justices Alito, Roberts, Scalia), setting Kennedy aside given his unique position at the median. We then test whether a justice agrees with their *usual* allies—that is, expresses skepticism together—even when that

skepticism is directed against their preferred side. If so, this suggests a genuine response; if not, it suggests that justices may be strategically pulling their punches to protect case-specific interests.

The observable implications described above utilize post-argument proxies for justice preferences (vote) and case divisiveness (margin of victory). A natural concern is that a justice’s ultimate vote may be influenced by the course of questioning as well. In this case, persuasion may offer an alternative explanation for patterns in observed skepticism. If strategic justices are always successful in persuading their colleagues, the genuine-voting and strategic-signaling accounts become observationally similar in many cases. While we find this level of persuasiveness to be implausible in light of extensive qualitative work on the courts (Ringsmuth et al., 2013; Wolfson, 2001), there is considerable room for future work to improve on our analysis. These improvements may include more rigorous formal modeling of strategic interaction in the deliberative process; collection of better pre-argument proxies for justice predisposition and case controversiality (e.g. circuit court votes); or natural experiments (e.g. exploiting justice retirements).

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

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

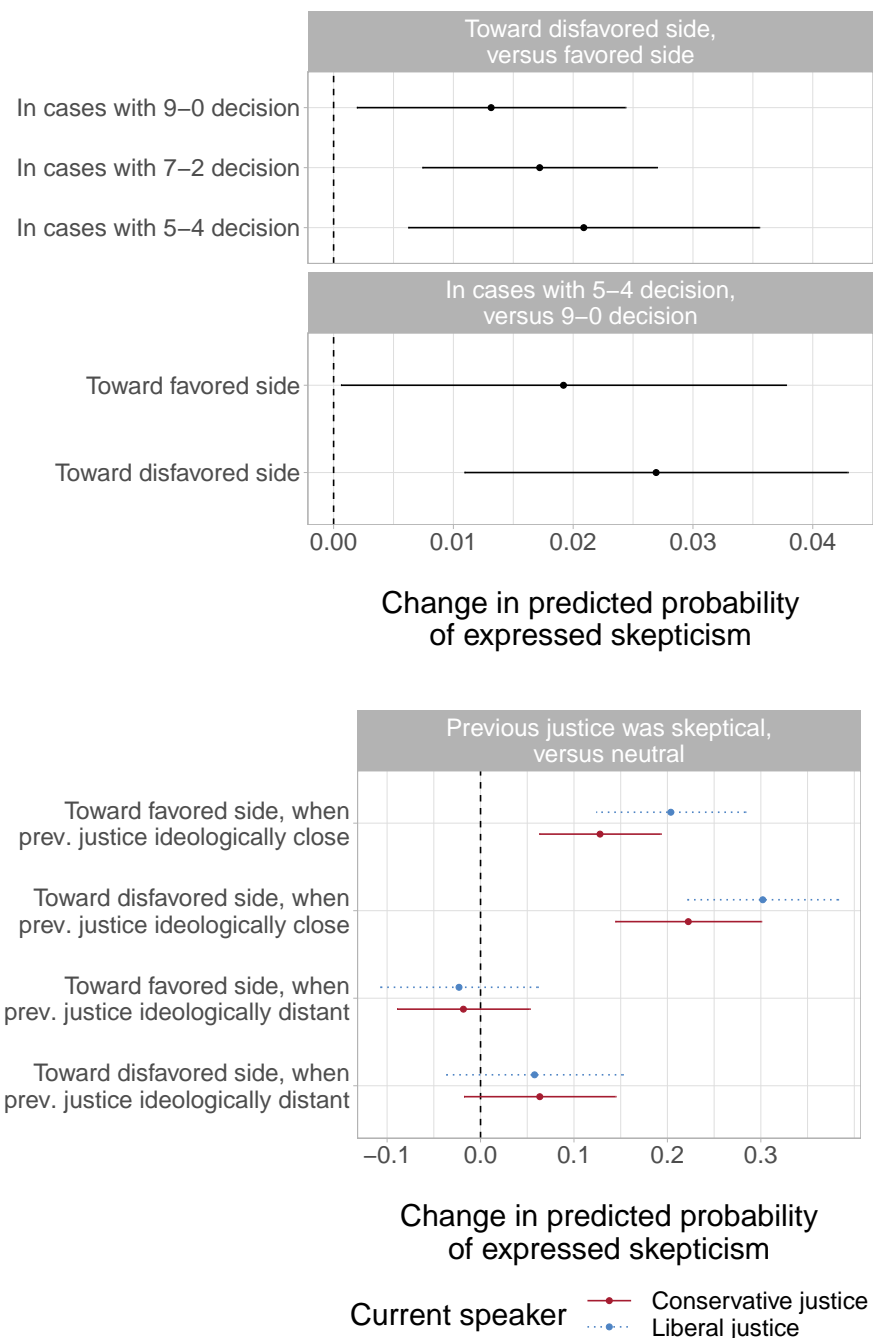


Figure 8: **Simulated quantities of interest.** Each panel manipulates a single variable from a control value (second element of panel title) to a treatment value (first). Points (errorbars) represent estimated changes (95% bootstrap confidence intervals) in skepticism. We average over all other nuisance covariates (e.g., the identity of the next speaker) of the scenario-specific change in outcome, weighting by the empirical frequencies of the nuisance covariates. The top panel shows that justices deploy skepticism more often toward their non-preferred side. The second panel compares close votes to unanimous decisions, demonstrating that justices express more skepticism across the board in the former. However, justices do not attempt to target a particular side in close votes; rather, they simply ask more skeptical questions across the board. Finally, the bottom panel shows that justices mirror the tones of their ideological neighbors, who share similar legal philosophies, even when those neighbors are opposed to the justice’s case-specific voting interests.

Figure 8 presents results from two MASS specifications. First, we model transition probabilities (i.e., the probability that the next utterance is of justice-tone m) as $\exp(\mathbf{W}_u^\top \boldsymbol{\zeta}_m) / \sum_{m'=1}^M \exp(\mathbf{W}_u^\top \boldsymbol{\zeta}_{m'})$ where the conversation context \mathbf{W} includes the eventual case margin, the ultimate vote of the justice in question, an interaction, and a justice-tone intercept. We then average across justices according to their speech frequency to obtain an average effect. The results show that justices use skepticism against their non-preferred side—either the petitioner or respondent, depending on which one they go on to vote against—at a significantly higher rate, as expected. Contrary to theoretical predictions under strategic signaling, however, we find no indication that the gap between petitioner- and respondent-targeted skepticism depends on the margin of the case decision. That is, the difference in differences is far from significant. We do find that skepticism is generically higher *across the board* in close cases. These results are consistent with the account that Supreme Court justices are engaged in genuine fact-finding, not a strategic attempt to manipulate their colleagues, and that they particularly seek to get answers right when the stakes are high. (As we note in detail above, we cannot entirely rule out the alternative explanation that some of these results are due to persuasion over the course of oral arguments.)

To probe further, we now turn to the dynamics of argumentation. In an expanded specification, we now incorporate additional binary indicators for whether the preceding speaker belonged to the liberal or conservative wing of the Court, as well as interactions between skepticism in the preceding utterance, ideology of the previous speaker, and vote. As described above, the *strategic* model of judicial signaling implies that after a peer—any peer—criticizes a justice’s preferred side, the justice should withhold comment or defuse tensions with neutral commentary. By the same token, the savvy justice should follow up

with a coup de grâce after a colleague finds fault in the disfavored side’s reasoning. We find no evidence that this is true. Rather, our results are highly consistent with a model of *genuine expression* in which justices concede to the criticisms of ideologically proximate peers, regardless of their case-specific interests. That is, after a liberal (conservative) justice casts doubt on a lawyer’s argument, other liberals (conservatives) on the Court are likely to follow suit even if that criticism undermines their favored side.

6 Concluding Remarks

Political science already studies audio. From campaign addresses to parliamentary debates, many of the field’s recurring questions deal with political speech. And while a great deal of progress has been made in analyzing *what* was said in these audio recordings, *how* these words were spoken is often equally or even more important. MASS provides a principled solution for inferring these quantities. And by modeling the flow of speech, MASS directly tests questions about speech dynamics—a phenomenon that is impossible to study with traditional text approaches. Finally, our software makes this computationally complex procedure accessible to everyone.

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.
- Black, R. C., Sorenson, M. W. and Johnson, T. R. (2013), 'Toward an actor-based measure of supreme court case salience: Information-seeking and engagement during oral arguments', *Political Research Quarterly* **66**(4), 804–818.
- Black, 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.
- 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.
- Chaiken, S. (1980), 'Heuristic versus systematic information processing and the use of source versus message cues in persuasion.', *Journal of personality and social psychology* **39**(5), 752.
- Cornell Legal Information Institute (n.d.), 'Oyez Project'. Accessed August 2015.
- URL:** *oyez.org*
- 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.
- El Ayadi, M., Kamel, M. S. and Karray, F. (2011), ‘Survey on speech emotion recognition: features, classification schemes, and databases’, *Pattern Recognition* **44**, 572–587.
- Epstein, L., Landes, W. M. and Posner, R. A. (2010), ‘Inferring the winning party in the supreme court from the pattern of questioning at oral argument’, *The Journal of Legal Studies* **39**(2), 433–467.
- Festinger, L. (1964), ‘Conflict, decision, and dissonance.’.
- 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.
- 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.
- Hansen, B. and Klopfer, S. (2006), ‘Optimal full matching and related designs via network flows’, *Journal of Computational and Graphical Statistics* **15**, 609–627.

- Iaryczower, M., Shi, X. and Shum, M. (2018), ‘Can words get in the way? the effect of deliberation in collective decision making’, *Journal of Political Economy* **126**(2), 688–734.
- Iaryczower, M. and Shum, M. (2012), ‘The value of information in the court: Get it right, keep it tight’, *American Economic Review* **102**(1), 202–37.
- Isen, A. M. (1984), ‘Toward understanding the role of affect in cognition.’.
- Johnson, T. R. (2001), ‘Information, oral arguments, and supreme court decision making’, *American Politics Research* **29**(4), 331–351.
- Johnson, T. R., Black, R. C., Goldman, J. and Treul, S. A. (2009), ‘Inquiring minds want to know: Do justices tip their hands with questions at oral argument in the us supreme court’, *Wash. UJL & Pol’y* **29**, 241.
- Johnson, T. R., Wahlbeck, P. J. and Spriggs, J. F. (2006), ‘The influence of oral arguments on the U.S. supreme court’, *American Political Science Review* **100**(01), 99–113.
- 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’.
- 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’.
- McGuire, K. T. (1995), ‘Repeat players in the supreme court: The role of experienced lawyers in litigation success’, *The Journal of Politics* **57**(1), 187–196.
- McGuire, W. J. (1968), ‘Personality and attitude change: An information-processing theory’, *Psychological foundations of attitudes* **171**, 196.
- 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’.

- Peters, J. W. and Maheshwari, S. (2018), ‘Viral videos are replacing pricey political ads. they’re 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.
- Redlawsk, D. (2006), *Feeling politics: Emotion in political information processing*, Springer.
- Ringsmuth, E. M., Bryan, A. C. and Johnson, T. R. (2013), ‘Voting fluidity and oral argument on the us supreme court’, *Political research quarterly* **66**(2), 429–440.
- Roberts, M. E., Stewart, B. M. and Airolidi, 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.
- Shapiro, S. M. (1984), ‘Oral argument in the supreme court of the united states’, *Catholic University Law Review* **33**(3), 529–554.
- Shullman, S. L. (2004), ‘The illusion of devil’s advocacy: How the justices of the supreme court foreshadow their decisions during oral argument’, *J. App. Prac. & Process* **6**, 271.

- 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.
- 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.
- Thomas, C. (2013), ‘Lecture at Harvard Law School’. Accessed April 2019.
- URL:** [youtube.com/watch?v=heQjKdHu1P4](https://www.youtube.com/watch?v=heQjKdHu1P4)
- 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.
- Wasby, S. L., D’Amato, A. A. and Metrailler, R. (1976), ‘The functions of oral argument in the us supreme court’, *Quarterly Journal of Speech* **62**(4), 410–422.
- Wedeking, J. (2010), ‘Supreme court litigants and strategic framing’, *American Journal of Political Science* **54**(3), 617–631.
- 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.
- Wolfson, W. D. (2001), ‘Oral argument: Doe it matter’, *Ind. L. Rev.* **35**, 451.

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.

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 MASS.

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 our R package. 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. For excellent reviews of the literature, including a more thorough discussion of these features, see Ververidis and Kotropoulos (2006) and El Ayadi et al. (2011). 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 com-

ponents. 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 MASS. 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 5, 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.¹⁷

¹⁷Other 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.

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

Table 3: **Audio features extracted in each frame.** Parenthesized values indicate the number of scalars extracted per moment. We also include interactions between (i) energy and zero-crossing rate, and (ii) Teager energy operator and fundamental frequency, for a total of 27 primary features. In addition, first and second finite differences are often informative. For example, vocal jitter and shimmer are respectively described by the first differences in F0 and energy.

B Estimation

B.1 Factorization of the Likelihood

The full-data likelihood is as follows:

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

By sufficiency

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

which is Equation 5.

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

However, the stagewise approach has properties that make it attractive for other reasons as well. First, when the model is correctly specified, our approach remains unbiased with respect to the auditory parameters; in this case, the only sacrifice is in efficiency loss relative to joint maximization of the full likelihood. But in the presence of model misspecification—which almost certainly exists with complex phenomena like human speech, e.g. if true human speech contains more than M modes—the proposed approach can in fact outperform

full maximum likelihood. More generally, semi-supervised approaches that exploit both labeled and unlabeled data often underperform those that only use the former (Masanori and Takeuchi, 2014). Intuitively, this is because unsupervised methods rarely recover the analyst’s preferred labels, and semi-supervised techniques are typically dominated by the much larger unlabeled dataset.

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

B.2 Estimation of Lower-Level Auditory Parameters

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

Consider the subset with known mode $S_u = m$.¹⁸ This group of utterances is assumed to be drawn from a single shared Gaussian HMM, the speech model for mode m . Below, we describe how lower-level parameters are estimated by standard HMM techniques. Interested

¹⁸In practice, because the perception of certain speech modes can be subjective (human coders may disagree or be uncertain), training set mode labels S_u may be a stochastic vector of length M , $\tilde{S}_u = [\Pr(S_u = 1), \dots, \Pr(S_u = M)]$, rather than a M -valued categorical variable. In such cases the contribution of an utterance to the model for emotion m may be weighted by the m -th entry, e.g. corresponding to the proportion of human coders who classified the utterance as emotion m . After replacing $\mathbf{1}(S_u = m)$ with $\Pr(S_u = m)$, the procedure described in this appendix can be used without further modification.

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

We first write down the likelihood function for parameters of the m -th mode. For each utterance, at each moment t , the feature vector $\mathbf{X}_{u,t}$ 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 \mathbf{X}_u could have been generated. The u -th utterance's contribution to the observed-data likelihood is the joint probability of all observed features, found by summing over every possible sequence of sounds. This yields

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

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

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

ignoring context). 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 procedure as it relates to the estimation of the lower-level audio parameters. 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.

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

B.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.

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

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.

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

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 \mathbf{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

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

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

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

B.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 from equation 11:

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

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.

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

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

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

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

B.3 Unmodeled Autocorrelation

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

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

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

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

B.4 Estimation of Upper-Level Conversation Parameters

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

conversation v are respectively $\mathbf{W}_{v,u}$, $S_{v,u}$ and $\mathbf{X}_{v,u}$.

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

where $\mathbf{W}_{v,u}(\mathbf{S}_{v,u' < u}) = [\mathbf{W}_{v,u}^{\text{stat.}\top}, \mathbf{W}_{v,u}^{\text{dyn.}}(\mathbf{S}_{v,u' < u})^\top]^\top$ and the primary-corpus indicator, \mathcal{C} , is omitted. δ_v indicates the initial distribution of speech modes for conversation v .

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

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

metadata vectors, $\tilde{\mathbf{W}}_{v,u}(\tilde{\mathbb{E}}[\mathbf{S}_{v,u' < u-1}], S_{u-1} = 1)$ through $\tilde{\mathbf{W}}_{v,u}(\tilde{\mathbb{E}}[\mathbf{S}_{v,u' < u-1}], S_{u-1} = M)$.

2. Each possible metadata vector implies a vector of probabilities for the next utterance,

$$\tilde{\Delta}_m = [\tilde{\text{Pr}}(S_u = 1|S_{u-1} = m), \dots, \tilde{\text{Pr}}(S_u = M|S_{u-1} = m)] = \frac{\exp(\tilde{\mathbf{W}}_u(\tilde{\mathbb{E}}[\mathbf{S}_{v,u' < u-1}], S_{u-1} = m)^\top \tilde{\boldsymbol{\zeta}}_m)}{\sum_{m'=1}^M \exp(\tilde{\mathbf{W}}_u(\tilde{\mathbb{E}}[\mathbf{S}_{v,u' < u-1}], S_{u-1} = m)^\top \tilde{\boldsymbol{\zeta}}_{m'})}.$$

Stack these into a transition matrix, $\tilde{\Delta}$.

3. Compute $\tilde{\mathbb{E}}[\mathbf{1}(S_{v,u} = m)]$ and $\tilde{\mathbb{E}}[\mathbf{1}(S_{v,u-1} = m, S_{v,u} = m')]$, using a forward-backward algorithm that is essentially identical to Equations 10 and 11. We find that the use of the corrected emission probabilities, described in Appendix B.3, is crucial in this step.

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

C Application

C.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 μ 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).