A Dynamic Model of Speech for the Social Sciences*

Dean Knox† & Christopher Lucas‡

March 18, 2019

Abstract

Social scientists increasingly rely on statistical models of text to resolve a wide range of questions about speech across a range of domains. However, humans communicate with more than text alone. Auditory cues convey important information, such as emotion, in many contexts of interest to social scientists. Nonetheless, researchers typically discard this information and work only with transcriptions of audio data. We develop the Structural Speaker Affect Model (SSAM), to classify auditorily distinct "modes" of speech (e.g., tone, emotion, speakers) and the transitions between them. SSAM incorporates ridge-like regularization into a nested hidden Markov model, allowing the use of high-dimensional audio features. We implement a fast estimation procedure that enables a principled approach to uncertainty based on the Bayesian bootstrap. As a validation test, we show that SSAM markedly outperforms existing audio and text approaches in both (a) identifying individual Supreme Court justices and (b) detecting human-labeled "skepticism" in their speech. We extend the analysis by examining the dynamics of expressed emotion in oral arguments.

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

*We thank Dustin Tingley for research support through the NSF-REU program; Michael May, Thomas Scanlan, Angela Su, and Shiv Sunil for excellent research assistance; and the Harvard Experiments Working Group and the MIT Department of Political Science for generously contributing funding to this project. For helpful comments, we thank Justin de Benedictis-Kessner, Bryce Dietrich, Gary King, Connor Huff, In Song Kim, Adeline Lo, Jacob Montgomery, David Romney, Dustin Tingley, Teppei Yamamoto, and Xiang Zhou, as well as participants at the Harvard Applied Statistics Workshop, the International Methods Colloquium, 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

Applications of text analysis in social science often examine corpora which were first spoken, then transcribed. Though methodologically and substantively diverse, this research focuses exclusively on what people say, while entirely ignoring how those words were spoken. In this article, we develop the first statistical model of political speech that explicitly represents "how" subjects express themselves, and demonstrate that our model is able to infer theoretically interesting quantities with only the audio - and no transcribed text - as input. This research allows researchers to study new quantities of interest conveyed through tone but not through text. And while we highlight the substantive potential of our model for analyzing tone of speech, importantly, our model is general to any class of labels that relate to the "sound" of speech, whether that concerns the speaker's emotional state, gender, language, or any other imaginable category of interest to the researcher.

Inflection and tone in human speech convey information that moderates the textual content (El Ayadi et al., 2011a), and without appropriate methods to analyze the audio signal accompanying the text transcript, researchers risk overlooking important insights into the content of political speech. But despite the frequency with which social scientists analyze speech, we are aware of no research in the social sciences that explicitly models the audio that accompanies these textual transcriptions. Nonetheless, social scientists study aspects of speech like emotion (Black et al., 2011) and rhetorical style (Sigelman and Whissell, 2002a,b), both of which are known to depend on tone of voice (Scherer and Oshinsky, 1977; Murray and Arnott, 1993; Dellaert et al., 1996). And though methods for analyzing text as data have received a great deal of attention in recent years,[1] none permit the inclusion of the accompanying audio features, even though recent work demonstrates the importance of audio features in political speech (Dietrich et al., 2016).

Our approach - the Structural Speaker Affect Model - is the first model of its kind in political science, and improves on existing approaches in statistics and computer science in two primary ways. First, existing approaches are only able to use a fraction of the features we incorporate. While we provide an overview of audio as data in Section A.2, it is common to aribitrarily select a dozen features and discard the rest.[2] We provide a more principled

[1]See, for example, Laver et al. (2003); Benoit et al. (2009); Clark and Lauderdale (2010); Hopkins and King (2010); Grimmer and Stewart (2013); Lauderdale and Clark (2014); Roberts et al. (2014); Lucas et al. (2015).

[2]Nogueiras et al. (2001), for example, use just two features (pitch and energy, see Appendix Section A

solution through regularization that removes the need for arbitrary researcher choice with better statistical properties than alternative approaches. Second, to our knowledge, the Structural Speaker Affect Model is the first to directly model dynamic interaction between speakers and to permit tests of hypotheses about these dynamics. For example, a researcher interested in the Supreme Court might ask, how does a hostile question from the median justice in an oral argument change the tenor of the subsequent conversation? We refer to these temporally dependent effects as changes in the flow of speech, and our model is the first in any field that permits the study of conversation flow. While existing approaches ignore dependence and estimation uncertainty, we implement a hierarchical hidden Markov model to explicitly model these dynamics and recover estimation uncertainty through a Bayesian bootstrap.

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

1 The Speaker-Affect Model

In this section, we introduce the structural speaker-affect model, or SSAM. SSAM is a hierarchical hidden Markov model (HHMM), meaning that each "state" in SSAM is itself

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

another hidden Markov model. Within SSAM, states are the user-defined labels, like "angry" and "neutral" or "male" and "female." Each of these states, by contrast, is modeled as an unsupervised HMM, learned during the training process. In the case of speech modes, this is useful because it permits each mode of speech to be defined by learned transitions between "sounds," which can be inferred from the user-supplied labels.

In the remainder of this section, we introduce our notation, define the model, and overview inference.

1.1 Notation

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

Our primary unit of analysis is the utterance, or a segment of continuous speech, generally bracketed by pauses. A speaker's mode of speech is assumed to be constant during an utterance. This is the quantity that we wish to measure, and it is generally unobserved unless a human coder listens to and classifies the utterance. Naturally, the mode of speech is not independent across utterances: A calm utterance is generally followed by another calm utterance. On a more granular level, each utterance is composed of an unobserved sequence of sounds, such as vowels, sibilants, and plosives. These sounds then generate a continuous stream of observed audio features.

Time-related Indices:

- Conversation index $v \in \{1, \cdots, V\}$: self-contained monologue or dialogue consisting of a sequence of utterances.

- Utterance index $u \in \{1, \cdots, U_v\}$: continuous segment of audible speech by a single speaker, preceded and followed by a period of silence or a transition between speakers.
- Time index $t \in \{1, \cdots, T_{v,u}\}$: position of an "instant" corresponding to an audio window within an utterance, advances by increments of 12.5 milliseconds.

Latent states:

- $S_{v,u} \in \{1, \cdots, M\}$: latent emotional state at for utterance u, corresponding to the emotions joy, sadness, anger, fear, surprise, disgust, and neutral. Indexed by m.
- $R_{v,u,t} \in \{1, \cdots, K\}$: latent sound at time t (e.g., sibilant, plosive). Indexed by k. Note that the same index may take on different meanings depending on the emotional state. For example, sibilants may appear in both angry and neutral speech, but exact auditory characteristics will differ by emotion, and the index corresponding to the concept of "sibilant" may not be the same for each emotion.

Features:

- $\boldsymbol{X}_{v,u,t}$: vector of D audio features at time t during utterance u of conversation v, such as sound intensity (decibels) during a brief audio window. All feature vectors in an utterance are collected in the $T_{v,u} \times D$ matrix, $\boldsymbol{X}_{v,u}$ (with $D = 27$ audio features, for example).
- $\boldsymbol{W}_{v,u}(\boldsymbol{S}_{v,u'<u}) = \left[\boldsymbol{W}_{v,u}^{\text{static}}, \boldsymbol{W}_{v,u}^{\text{dynamic}}(\boldsymbol{S}_{v,u'<u})\right]$: vector of conversation and utterance metadata, which may include functions of prior conversation history.

1.2 Model

We assume that the feature series is generated by a hierarchical hidden Markov model (HHMM) with two levels. The upper level is an HMM that generates a sequence of speech modes conditional on utterance metadata, $W_{v,u}$, and each conversation consists of one sequence of known length drawn from the upper level. The lower level that generates the observed audio features $\boldsymbol{X}_{v,u}$ conditional on the current mode of speech $S_{v,u}$.

In the upper level, speech mode probabilities are modeled as a multinomial logistic function of metadata, $\Pr(S_{v,u} = m|\boldsymbol{W}_{v,u}) \propto \exp(\mathrm{W_{v,u}}\boldsymbol{\zeta}_{\mathrm{m}})$. We note that it is more computationally demanding to estimate parameters related to longer conversation histories, because prior modes of speech are imperfectly observed. As we discuss later, when multiple values of $\boldsymbol{W}_{v,u}^{\text{dynamic}}$ are possible, each must be weighted by the total probability of speech-mode trajectories leading to that state. For simplicity, in this paper we consider the case $\boldsymbol{W}_{v,u}(\boldsymbol{S}_{v,u'<u}) = \boldsymbol{W}_{v,u}(S_{v,u-1})$, so that the upper level is a first-order HMM conditional on static metadata and mode probabilities can be collected in the $M \times M$ transition matrix $\boldsymbol{\Delta}(\boldsymbol{W}_{v,u}^{\text{static}}) = \left[\Delta_{m,m'}(\boldsymbol{W}_{v,u}^{\text{static}})\right]$. However, the model is general.

Second, given that utterance u of conversation v was spoken with emotion $S_{v,u} = m$, the sequence of sounds that comprise an utterance are assumed to be generated by the m-th emotion-specific first-order HMM. The probability of transitioning from sound k to k' is given by $\Gamma_{k,k'}^{m}$, and transition probabilities are collected in sound transition matrix $\boldsymbol{\Gamma}^{m}$.

$$(R_{v,u,t} \mid S_{v,u} = m) \sim \mathrm{Cat}(\Gamma^{\mathrm{m}}_{\mathrm{R_{v,u,t-1},*}})$$

Finally, during a particular sound, the vector of features at each point in time is assumed to be drawn from a multivariate Gaussian distribution.

$$(\boldsymbol{X}_{v,u,t} \mid S_{v,u} = m, R_{v,u,t} = k) \sim N(\boldsymbol{\mu}^{m,k}, \boldsymbol{\Sigma}^{m,k})$$

We use superscripts to index the properties of states and sounds; subscripts index the elements of a vector or matrix.

2 Estimation

2.1 Lower Level

To estimate the parameters of the M lower-level models, which each represent the auditory characteristics of a particular speech mode, a non-sequential training set of example utterances. The training set is denoted $\tilde{\boldsymbol{X}}$ and its attributes are similarly distinguished from those of the full corpus by a tilde.[3]

[3]In practice, because the perception of certain speech modes can be subjective, mode label $\tilde{S}_u$ may be a stochastic vector of length M rather than a binary indicator vector. In such cases the contribution of

Consider the subset with known mode $\tilde{S}_u = m$. For each utterance, at each time t, the feature vector $\tilde{\boldsymbol{X}}_{u,t}$ could have been generated by any of the K sounds associated with emotion m, so there are $K^{\tilde{T}_u}$ possible sequences of unobserved sounds by which the feature sequence could have been generated. The u-th utterance's contribution to the observed-data likelihood is the joint probability of all observed features, found by summing over every possible sequence of sounds. The likelihood function for parameters of the m-th mode is then

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

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

In practice, due to the high dimensionality of the audio features, we also regularize Σ to ensure invertibility by adding a small positive value (which may be thought of as a prior) to its diagonal. We recommend setting this regularization parameter, along with the number of sounds, by selecting values that maximize the out-of-sample naïve probabilities of the training set in V-fold cross-validation. This procedure possesses the "oracle" property in that it 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. In practice, given the vast number of parameters to optimize over, we estimate using the Baum–Welch algorithm for expectation–maximization with hidden Markov models. This procedure involves maximizing the complete-data likelihood, which differs from

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.

equation 1 in that it also incorporates the probability of the unobserved sounds.

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

The Baum–Welch algorithm uses the joint probability of (i) all feature vectors up until time t and (ii) the sound at t, given in equation 3. Together, these are referred to as the *forward probabilities*, because values for all t are efficiently calculated in a single recursive forward pass through the feature vectors.

$$
\begin{aligned}
\tilde{\boldsymbol{\alpha}}_{u,t} &\equiv \Pr(\tilde{\boldsymbol{X}}_{u,1} = \tilde{\boldsymbol{x}}_{u,1}, \cdots, \tilde{\boldsymbol{X}}_{u,t} = \tilde{\boldsymbol{x}}_{u,t}, \tilde{R}_{u,t} = k) \\
&= \delta_u^\top \boldsymbol{P}^m(\tilde{\boldsymbol{x}}_{u,1}) \left(\prod_{t'=2}^{t} \boldsymbol{\Gamma}^m \boldsymbol{P}^m(\tilde{x}_{u,t'}) \right)
\end{aligned}
\tag{3}
$$

The algorithm also relies on the conditional probability of (i) all feature vectors after t given (ii) the sound at t (equation 4). These are similarly called the *backward probabilities* due to their calculation by backward recursion.

$$
\begin{aligned}
\tilde{\boldsymbol{\beta}}_{u,t} &\equiv \Pr(\tilde{\boldsymbol{X}}_{u,t+1} = \tilde{\boldsymbol{x}}_{u,t+1}, \cdots, \tilde{\boldsymbol{X}}_{u,\tilde{T}_u} = \tilde{\boldsymbol{x}}_{u,\tilde{T}_u} \mid \tilde{R}_{u,t} = k) \\
&= \left(\prod_{t'=t+1}^{\tilde{T}_u} \boldsymbol{\Gamma}^m \boldsymbol{P}^m(\tilde{\boldsymbol{x}}_{u,t'}) \right) \mathbf{1}
\end{aligned}
\tag{4}
$$

2.1.1 E step

The E step involves substituting (i) the unobserved sound labels, $\mathbf{1}(\tilde{R}_{u,t} = k)$, and (ii) the unobserved sound transitions, $\mathbf{1}(\tilde{R}_{u,t} = k', \tilde{R}_{u,t-1} = k)$, with their respective expected values,

conditional on the observed training features $\tilde{\boldsymbol{X}}_u$ and the current estimates of $\boldsymbol{\mu}^{m,k}$, $\boldsymbol{\Sigma}^{m,k}$, and $\boldsymbol{\Gamma}^m$ (collectively referred to as $\boldsymbol{\Theta}$).

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

$$\mathbb{E}\left[\mathbf{1}(\tilde{R}_{u,t} = k) \mid \tilde{\boldsymbol{X}}_u, \tilde{S}_u = m, \hat{\boldsymbol{\Theta}}\right] = \hat{\tilde{\alpha}}_{u,t,k} \; \hat{\tilde{\beta}}_{u,t,k} \; / \; \hat{\mathcal{L}}_u^m, \tag{5}$$

where the hat denotes the current approximation based on parameters from the previous M step, and $\tilde{\alpha}_{u,t,k}$ and $\tilde{\beta}_{u,t,k}$ are the k-th elements of $\tilde{\boldsymbol{\alpha}}_{u,t}$ and $\tilde{\boldsymbol{\beta}}_{u,t}$ respectively, and $\hat{\mathcal{L}}_u^m$ is the u-th training utterance's contribution to $\hat{\mathcal{L}}^m$.

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

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}(\tilde{R}_{u,t} = k', \tilde{R}_{u,t-1} = k) \mid \tilde{\boldsymbol{X}}_u, \tilde{S}_u = m, \hat{\boldsymbol{\Theta}}] \\
&\quad = \Pr(\tilde{R}_{u,t} = k', \tilde{R}_{u,t-1} = k, \tilde{\boldsymbol{X}}_u \mid \hat{\boldsymbol{\Theta}}) \; / \; \Pr(\tilde{\boldsymbol{X}}_u \mid \hat{\boldsymbol{\Theta}}) \\
&\quad = \Pr(\tilde{\boldsymbol{X}}_{u,1}, \cdots, \tilde{\boldsymbol{X}}_{u,t-1}, \tilde{R}_{u,t-1} = k \mid \hat{\boldsymbol{\Theta}}) \; \Pr(\tilde{R}_{u,t} = k' \mid \tilde{R}_{u,t-1} = k, \hat{\boldsymbol{\Theta}}) \; \times \\
&\qquad \Pr(\tilde{\boldsymbol{X}}_{u,t} \mid \tilde{R}_{u,t} = k') \; \Pr(\tilde{\boldsymbol{X}}_{u,t+1}, \cdots, \tilde{\boldsymbol{X}}_{u,\tilde{T}_u} \mid \tilde{R}_{u,t} = k') \; / \; \Pr(\tilde{\boldsymbol{X}}_u \mid \hat{\boldsymbol{\Theta}}) \\
&\quad = \hat{\tilde{\alpha}}_{u,t-1,k} \; \hat{\Gamma}^m_{k,k'} \; \phi_D(\boldsymbol{x}_{u,t}; \hat{\boldsymbol{\mu}}^{m,k}, \hat{\tilde{\boldsymbol{\Sigma}}}^{m,k}) \; \tilde{\beta}_{u,t,k'} \; / \; \tilde{\mathcal{L}}_u^m.
\end{aligned} \tag{6}$$

implicitly conditioning on the training data throughout.

2.1.2 M Step

After substituting equations 5 and 6 into the complete-data likelihood (equation 2), the M step involves two straightforward calculations.

First, the conditional maximum likelihood update of the transition matrix $\boldsymbol{\Gamma}^m$ follows almost directly from equation 6:

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

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

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

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

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

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

2.1.3 Naïve Inference on Utterance Mode

The expectation–maximization procedure described in the preceding sections produces point estimates for the mode-specific HMM parameters, $\boldsymbol{\mu}^{*,k}$, $\boldsymbol{\Sigma}^{*,k}$, and $\boldsymbol{\Gamma}^{*}$. Using these parameters and the prevalence of each mode alone, the estimated posterior mode membership probabilities for each utterance in the corpus can be computed using standard mixture-model techniques.

$$\begin{aligned}
&\Pr(S_{v,u} = m | \boldsymbol{X}_{v,u}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{S}}, \boldsymbol{\Theta}) \\
&\quad = \frac{\Pr(\boldsymbol{X}_{v,u,1} = \boldsymbol{x}_{v,u,1}, \cdots, \boldsymbol{X}_{v,u,T_{v,u}} = \boldsymbol{x}_{v,u,T_{v,u}} \mid S_{v,u} = m, \boldsymbol{X}_{v,u}, \boldsymbol{\Theta}) \Pr(S_{v,u} = m | \tilde{\boldsymbol{S}})}{\sum_{m'=1}^{M} \Pr(\boldsymbol{X}_{v,u,1} = \boldsymbol{x}_{v,u,1}, \cdots, \boldsymbol{X}_{v,u,T_{v,u}} = \boldsymbol{x}_{v,u,T_{v,u}} \mid S_{v,u} = m', \boldsymbol{X}_{v,u}, \boldsymbol{\Theta}) \Pr(S_{v,u} = m' | \tilde{\boldsymbol{S}})} \\
&\quad = \frac{\left(\delta^{m\top} \boldsymbol{P}^{m}(\boldsymbol{x}_{v,u,1}) \left[\prod_{t=2}^{T_{v,u}} \boldsymbol{\Gamma}^{m} \boldsymbol{P}^{m}(\boldsymbol{x}_{v,u,t})\right] \mathbf{1}\right)}{\sum_{m'=1}^{M} \left(\delta^{m'\top} \boldsymbol{P}^{m'}(\boldsymbol{x}_{v,u,1}) \left[\prod_{t=2}^{T_{v,u}} \boldsymbol{\Gamma}^{m'} \boldsymbol{P}^{m'}(\boldsymbol{x}_{v,u,t})\right] \mathbf{1}\right)} \cdot \frac{1}{\tilde{U}} \sum_{u=1}^{\tilde{U}} \mathbf{1}(\tilde{S}_u = m)
\end{aligned}$$

Uncertainty is incorporated by integrating over the posterior of the lower-level parameters, $f(\boldsymbol{\Theta} | \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{S}})$. However, we find that in general, analytic approaches for estimating uncertainty perform extremely poorly. This is because autocorrelation in actual human speech violates the assumed conditional independence between two successive instants of the same mode and sound. To obtain more realistic measures of uncertainty, we conduct Bayesian bootstrapping of the training set. Within each reweighted bootstrap training set, the described EM algorithm is applied, the resulting lower-level parameters are used to label the full corpus, and finally bootstrap labels are averaged to produce $\Pr(S_{v,u} = m | \boldsymbol{X}_{v,u}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{S}})$.

We refer to the resulting lower-level posterior mode probabilities $\Pr(S_{v,u} = m|\boldsymbol{X}_{v,u}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{S}})$, which use only local audio characteristics and do not incorporate contextual information, as "naïve" to distinguish them from the full-model posterior, which incorporates both the (v, u)-th utterance's metadata and the audio characteristics of other utterances in conversation v.

2.2 Upper Level

In the simplest possible case, when only static metadata is used, the estimation of upper-level parameters reduces to a multinomial logistic regression of an imperfectly observed speech mode, $S_{v,u}$, on utterance metadata, $W_{v,u}$. In this case, each utterance is included in the regression M times, each with a different mode as outcome and weighted according to the naïve mode probability. When the upper-level transition function depends on static metadata and attributes of the prior utterance, so that the upper HMM is of order 1, then each utterance is duplicated M^2 times, once for each combination of possible $S_{v,u-1}$ and $S_{v,u}$ realizations, with the (m, m')-th duplicate weighted by $\Pr(S_{v,u-1} = m|\boldsymbol{X}_{v,u-1}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{S}}) \Pr(S_{v,u} = m'|\boldsymbol{X}_{v,u}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{S}})$, and assigned the value of the dynamic metadata that would be obtained if $S_{v,u-1} = m$. This approach can be easily extended to accommodate longer history dependence in the model, although computational demands grow exponentially with history length. When the conversation history incorporated into the dynamic metadata is sufficently large to make the exact approach computationally infeasible, various approximations may be used, including probabilistic sampling of conversation trajectories or mean-field approximations of dynamic metadata. The posterior of the upper-level transition function parameters, $f(\boldsymbol{\zeta}|\boldsymbol{W}, \boldsymbol{X}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{S}})$ is computed with standard Hessian-based techniques, and these parameters can be interpreted by simulation as usual.

When the upper-level transition function is known, contextualized posterior mode probabilities (i.e., incorporating metadata and audio features of the full conversation) are as

follows (implied conditioning on $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{S}}$ is omitted throughout):

$$
\begin{aligned}
&[\Pr(S_{v,u} = m|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{\zeta})] \\
&\quad \propto [\Pr(S_{v,u} = m, \boldsymbol{X}_{v,*}|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{\zeta})] \\
&\quad \propto [\Pr(\boldsymbol{X}_{v,1}, \cdots, \boldsymbol{X}_{v,u}, S_{v,u} = m|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{\zeta})] \circ [\Pr(\boldsymbol{X}_{v,u+1}, \cdots, \boldsymbol{X}_{v,U_v}|S_{v,u} = m, \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{\zeta})] \\
&\quad \propto [\Pr(S_{v,1} = m|\boldsymbol{X}_{v,1})] \prod_{u'=2}^{u} [\exp(\mathrm{W_{v,u'}(S_{v,u'-1} = m)\boldsymbol{\zeta}_{m'}})] \operatorname{diag}([\Pr(\mathrm{S_{v,u'} = m|X_{v,u'}})]) \\
&\qquad \circ \left(\prod_{u'=u+1}^{U_v} [\exp(\mathrm{W_{v,u'}(S_{v,u'-1} = m)\boldsymbol{\zeta}_{m'}})] \operatorname{diag}([\Pr(\mathrm{S_{v,u'} = m|X_{v,u'}})]) \right) \mathbf{1}
\end{aligned}
$$

where $[\Pr(S_{v,u'} = m|\boldsymbol{X}_{v,u'})]$ is an M-dimensional stochastic row vector of naïve mode probabilities; $[\exp(\mathrm{W_{v,u'}(S_{v,u'-1} = m)\boldsymbol{\zeta}_{m'}})]$ is an $M \times M$ matrix in which the (m, m')-th entry represents the probability of transitioning to mode m', given that the the previous mode was m; and $\circ$ is the elementwise product. This decomposes the contextual probabilities into their forward and backward components, then rewrites the forward/backward probabilities in terms of naïve probabilities and the contextual transition matrices.

Uncertainty due to estimation of the upper-level transition parameters, $\boldsymbol{\zeta}$, is incorporated by sampling from $f(\boldsymbol{\zeta}|\boldsymbol{W}, \boldsymbol{X}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{S}})$, calculating $\Pr(S_{v,u} = m|\boldsymbol{W}, \boldsymbol{X}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{S}}, \boldsymbol{\zeta})$ for each set of sampled parameters, and integrating out the parameters from
$\Pr(S_{v,u} = m|\boldsymbol{W}, \boldsymbol{X}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{S}}, \boldsymbol{\zeta}) f(\boldsymbol{\zeta}|\boldsymbol{W}, \boldsymbol{X}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{S}})$.

3 Data

In this section, we introduce an original corpus of Supreme Court oral argument audio recordings scraped from the Oyez Project.[4] The corpus is used for two separate analyses in this paper. We first present a validation exercise in which we classify utterances of speech according to the identity of the speaker, then verify model predictions against known values. In the main application, we classify utterances according to their emotional characteristics.

The data for these applications are scraped from the Oyez Project.[5] We limit our analysis to the Roberts court from the Kagan appointment to the death of Justice Scalia, so that

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

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

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

4 Validation

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

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

We implement the model described in Section 1, modeling the transition probabilities (i.e., the turn-taking behavior of justices) as a multinomial logistic function of the following conversation metadata:

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

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

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

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

To examine how results improve as the training data grows, we report results for models trained with 25, 50, 100, and 200 utterances per mode.

4.1 Predictive Performance

For all training set sizes, we show that contextual mode probabilities from the full model are superior in all respects to naïve mode probabilities that neglect temporal structure and metadata.

We assess performance with a variety of metrics. Using the posterior probabilities on each utterance's mode of speech $S_{v,u}$, we report average per-utterance logarithmic, quadratic, and spherical scores for each model, respectively defined below. Because the fully labeled test

set contains over 62,000 utterances, we do not compute confidence intervals on performance metrics. Training utterances are currently not excluded but represent only a small fraction of the full corpus.

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

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

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

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

We also convert posterior probabilities to maximum-likelihood "hard" predictions and calculate mode-specific precision, recall, and F1 score. The prevalence-weighted average of these mode-specific performance metrics is also reported in Table 1. Note that overall and prevalence-weighted average mode accuracy equals prevalence-weighted average mode recall.

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

4.2 Frequentist Performance

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

We find that with a training set size of $n = 25$, four out of 57 confidence intervals fail to cover the population parameter. With $n = 50$, this number falls to two non-covering confi-

Table 1: Classification performance of lower-stage (L) model alone, versus full (F) model incorporating temporal structure and metadata, across four training set sizes and various performance metrics.

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

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

5 Application

In this section, we redefine a mode of speech to correspond to a justice-emotion, e.g. skeptical speech by Antonin Scalia, for a total of 16 modes. Skepticism is a particularly interesting rhetorical category. As Johnson et al. (2006, p.99) argue, justices use oral arguments to "seek information in much the same way as members of Congress, who take advantage of

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

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

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

We employ the following covariates:

- Case-specific issue, indexed by i: civil rights, criminal procedure, economic activity, First Amendment rights, judicial power, or a catch-all "other" category; and

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

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

$$\Pr(S_{v,u} = m) \propto \exp\big(\alpha_{\mathrm{m}} + \beta^{\mathrm{mode}}_{\mathrm{m}} \cdot 1(\mathrm{S_{v,u-1}} = \mathrm{m}) + \beta^{\mathrm{speaker}}_{\mathrm{m}} \cdot 1(\mathrm{justice_{S_{v,u-1}}} = \mathrm{justice_m}) + \sum_i \gamma^{\mathrm{issue}}_{m,i} \cdot \mathrm{issue_v} + \sum_{\mathrm{j}} \gamma^{\mathrm{ideo}}_{\mathrm{m,j}} \cdot \mathrm{ideology_{v,u}}\big) \quad (11)$$

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

Overall, we find that Kagan and Sotomayor question liberal lawyers less and Alito questions liberal lawyers more, but we find no evidence that ideological orientation alone produces greater skepticism. One possible explanation for this finding is that general ideological opposition is a crude measure of justices' preferences, and that justices take into accoun the nuances of a case. This is supported by the fact that many cases are decided unanimously, perhaps suggesting that a case-specific fixed effect is appropriate. When we introduce an

additional covariate for a justice's vote on a specific case, we find that voting against a particular side are highly correlated with an increase in skeptical utterances directed toward that side, relative to neutral utterances by the same justice. However, a causal interpretation of this result depends on the assumption that justices are not persuaded during the course of the oral arguments.

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

6 Conclusion

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

References

Benoit, K., Laver, M. and Mikhaylov, S. (2009), 'Treating words as data with error: Uncertainty in text statements of policy positions', American Journal of Political Science **53**(2), 495–513.

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.

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

Clark, T. S. and Lauderdale, B. (2010), 'Locating supreme court opinions in doctrine space', American Journal of Political Science **54**(4), 871–890.

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

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

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

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

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

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

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

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

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

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

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

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

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

Hopkins, D. J. and King, G. (2010), 'A method of automated nonparametric content analysis for social science', American Journal of Political Science **54**(1), 229–247.

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

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

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

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

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

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

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'.

Lauderdale, B. E. and Clark, T. S. (2014), 'Scaling politically meaningful dimensions using texts and votes', American Journal of Political Science **58**(3), 754–771.

Laver, M., Benoit, K. and Garry, J. (2003), 'Extracting policy positions from political texts using words as data', American Political Science Review **97**(02), 311–331.

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

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

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.

Murray, I. R. and Arnott, J. L. (1993), 'Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion', The Journal of the Acoustical Society of America **93**(2), 1097–1108.

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

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

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

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.

Scherer, K. R. and Oshinsky, J. S. (1977), 'Cue utilization in emotion attribution from auditory stimuli', Motivation and emotion **1**(4), 331–346.

Sigelman, L. and Whissell, C. (2002a), '" the great communicator" and" the great talker" on the radio: Projecting presidential personas', Presidential Studies Quarterly pp. 137–146.

Sigelman, L. and Whissell, C. (2002b), 'Projecting presidential personas on the radio: An addendum on the bushes', Presidential Studies Quarterly **32**(3), 572–576.

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

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

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

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

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.

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

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 **??**, 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 SSAM.

A.1 The Raw Audio Signal

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

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

A.2 Raw Audio to Audio Features

We extract a wide range of features that have been used in the audio emotion-detection literature.[6] The raw audio signal is divided into overlapping 25-millisecond windows, spaced at 12.5-millisecond intervals. Some features, such as the sound intensity (measured in decibels) are extracted from the raw signal.

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

Table 2 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 SSAM. 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

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

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

Table 2: Audio features extracted in each frame. In addition, we include interactions between (i) energy and zero-crossing rate, and (ii) Teager energy operator and fundamental frequency. We also use the first and second finite differences of all features.

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. [7]

B Why not use a recurrent neural network?

Recurrent neural networks (RNNs) represent perhaps the most obvious alternative approach to time-dependent data like human speech, particularly given the increasing use of neural networks. While RNNs are not without merit, hidden Markov models are better suited to our problem for four primary reasons. First, like most applications in the social sciences, we have relatively few labeled examples, particularly in comparison to common deep learning applications to human speech.[8] Experiments comparing the performance of hidden Markov models to RNNs find that HMMs outperform neural networks where the data are limited (Panzner and Cimiano, 2016), an unsurprising result given that significant increase in the number of parameters. Second, neural networks are difficult to interpret (Lucas, 2018). And though much progress has been made in the interpretation of convolutional neural networks over the last few years (Erhan et al., 2009; Zeiler and Fergus, 2014; Donahue et al., 2014), methods for interpreting RNNs are considerably less developed (Karpathy et al., 2015). Third, the statistical foundations of deep learning are still not well-understood, though there has been some recent progress in this area (Gal and Ghahramani, 2015, 2016a,b; Kendall and Gal, 2017). Fourth and finally, we are interested not only in classifying segments of human speech, but also in analyzing the flow of speech - how speech of a particular tone influences

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

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

the tone of subsequent speech. To our knowledge, there is no existing deep learning model that permits direct inference on statistical parameters that represent this interest.

C Comparison with Text Sentiment

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

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