Online Appendix: A Dynamic Model of Speech for the Social Sciences

Dean Knox * Christopher Lucas †

October 27, 2020

Intended for online publication only.

Contents

*Faculty Fellow of Analytics at Wharton and Assistant Professor, The Wharton School of the University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA 19104; `dcknox.com`, dcknox@upenn.edu, `https://orcid.org/0000-0002-1945-7938`.

†Assistant Professor, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130; `christopherlucas.org`, christopher.lucas@wustl.edu, `https://orcid.org/0000-0001-8551-6935`.

1 Estimation

In this section, we introduce our estimation procedure.

1.1 Factorization of the Likelihood

The complete-data likelihood is

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

By sufficiency

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

which is Equation 5.

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

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

man speech contains more than M modes—the proposed approach can in fact outperform full maximum likelihood. More generally, semi-supervised approaches that exploit both labeled and unlabeled data often underperform those that only use the former (Masanori and Takeuchi, 2014). Intuitively, this is because unsupervised methods rarely recover the analyst's preferred labels, and semi-supervised techniques are typically dominated by the much larger unlabeled dataset.

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

1.2 Estimation of Lower-Level Auditory Parameters

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

Consider the subset with known mode $S_u = m$.[1] This group of utterances is assumed to be drawn from a single shared Gaussian HMM, the speech model for mode m. Below, we describe how lower-level parameters are estimated by standard HMM techniques. Interested readers are referred to Zucchini and MacDonald (2009) for further discussion.

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

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

where $\boldsymbol{\mu}^m = (\boldsymbol{\mu}^{m,k})_{k \in \{1,\ldots,K\}}$, $\boldsymbol{\Sigma}^m = (\boldsymbol{\Sigma}^{m,k})_{k \in \{1,\ldots,K\}}$, δ^m is a $1 \times K$ vector containing the initial

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

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

In practice, due to the high dimensionality of the audio features, we also regularize the $\mathbf{\Sigma}$ terms to ensure invertibility by adding a small positive value (which may be thought of as a prior) to its diagonal. We recommend setting this regularization parameter, along with the number of sounds, by selecting values that maximize the training set's cross-validated naïve probabilities (i.e., based on mode prevalence and emission probabilities, ignoring 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, Dudoit, and Keles, 2004).

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

unobserved sounds.

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

1.2.1 E step

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

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

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

utterance.

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

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

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

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

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

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

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

1.2.2 M Step

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

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

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

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

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

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

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

1.3 Unmodeled Autocorrelation

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

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

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

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

1.4 Estimation of Upper-Level Conversation Parameters

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

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

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

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

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

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

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

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

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

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

1.5 Bootstrapping

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

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

2 Audio Features

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

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

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

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

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

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

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

illegal racial gerrymandering.

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

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

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

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

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

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

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

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

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

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

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

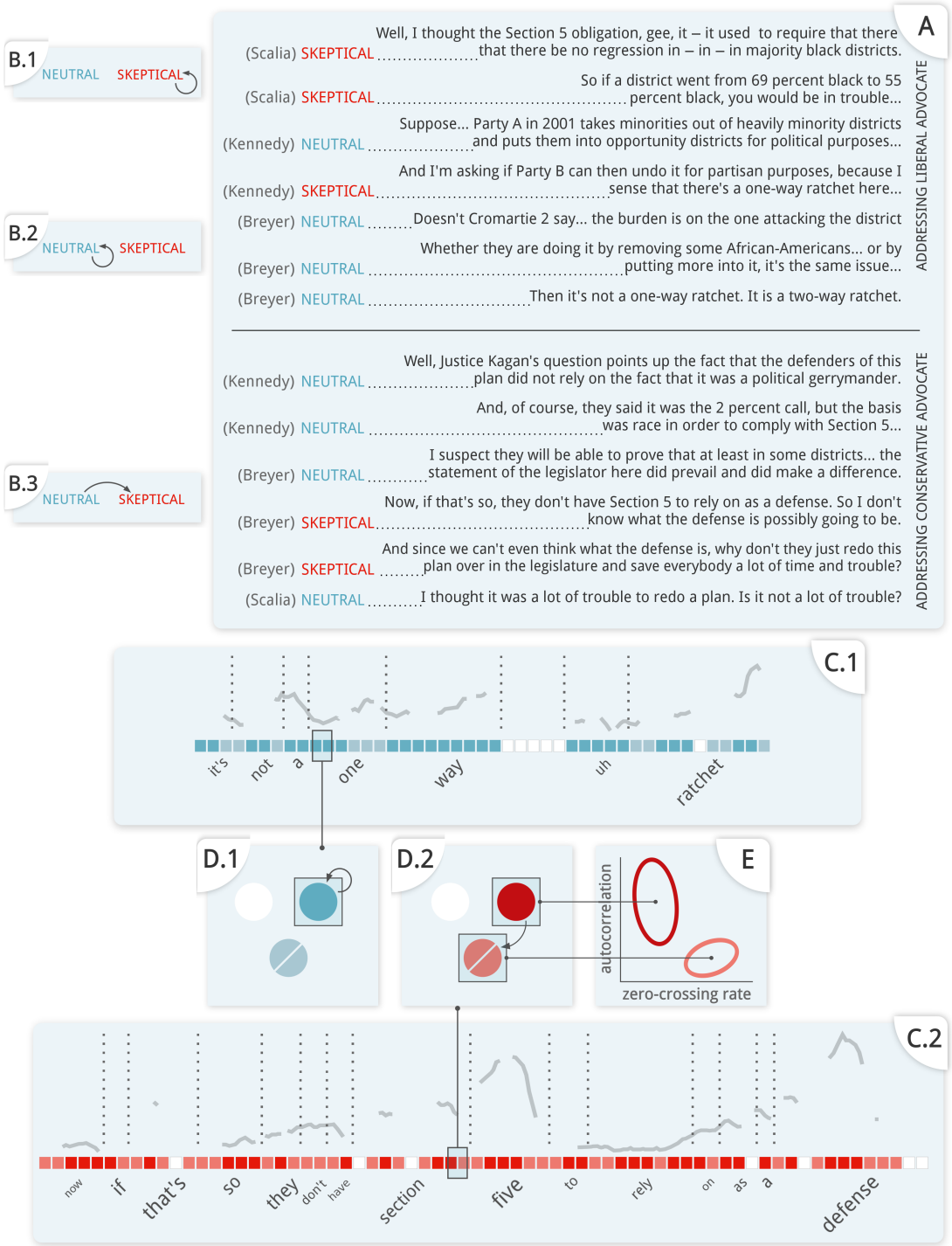

Figure 1: **An illustrative example.** Panel A contains an excerpt from Alabama Legislative Black Caucus v. Alabama, where Justices Scalia, Kennedy, and Breyer utilize neutral and skeptical tones in questioning. Call-outs highlight successive utterance-pairs in which the speaker shifted from one mode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

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

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

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

4 Validating the Model

4.1 Facial Validity of Predicted Skepticism

Before proceeding to more substantial results, we first demonstrate the face validity of MASS predictions in a qualitative examination of machine-generated utterance labels. Table 4.1 presents twenty randomly sampled 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**
And that helps women.	Mr. Frederick?
It said the rationale is unconscionable.	And because it's a regulated industry, the regulator in your view is doing one of the worst jobs in history.
You think the answer to that is clearly no.	And – and the difference between the monitoring and what happened in the past is memories are fallible, computers aren't.
Isn't it arguably in part to protect consumers?	Then if the Polynesian boat is permanently in the museum, there's a lot of objective evidence of that, it would not be a vessel.
The reason that they want to appeal is they want to win.	What about the – this as I understand it, came up originally as arbitration under the – wasn't it under the collective bargaining contract?
But the lower court said it shouldn't be weighed against the State, period.	You're talking about a very narrow range of cases, because I take it your principal position is it – it would be unusual that the defendant needs to be competent in order for the lawyer effectively to represent him on habeas.
Well, that's simply because, as we said in Allegheny Pittsburgh, the basis for considering the equal protection claim is the rights that you're given under State law.	I mean, Justice Ginsburg wrote the majority, and she said the reference to regulatory authority of a State, which is a different reference, I agree, should be read to preserve, not preempt traditional prerogative for the State.
It did command a majority of the Court, it is authoritative decision, and there are obviously different views among different judges about the extent to which they are the same or not.	If you start talking about significant effect, without those last words, " deregulatory purpose " I suddenly worry about the following: That every city in the United States depends upon towing to regulate parking within the city.
And to go back to Justice Sotomayor's question, as long as it's rational in at least some instances directly to pick out those States, at least one or two of them, then doesn't the statute survive a facial challenge?	Suppose a jurisdiction has the policy of requiring every inmate who is arrested and is going to be held in custody to disrobe and take a shower and apply medication for the prevention of the spread of lice and is observed while this is taking place from some distance by a corrections officer, let's say 10 feet away.
You made him give it to him. So what's wrong with his saying, you go give it to somebody? Now, if it's too much trouble, the judge can say he can't make you go to a lot of trouble. If it's giving it to somebody who might really do everything he wants, we'll guard against that.	Well, Buckman – Buckman was arguably a little bit different, in that there's a concern expressed in that case that requiring allowing the State suit to go forward would cause manufacturers to basically inundate the agency with proposals and warning revisions, so that there would be so many things that the agency wouldn't even be able to process them, and they would become meaningless to the consumers.

4.2 Textual Characteristics of Expressed Skepticism

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

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

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

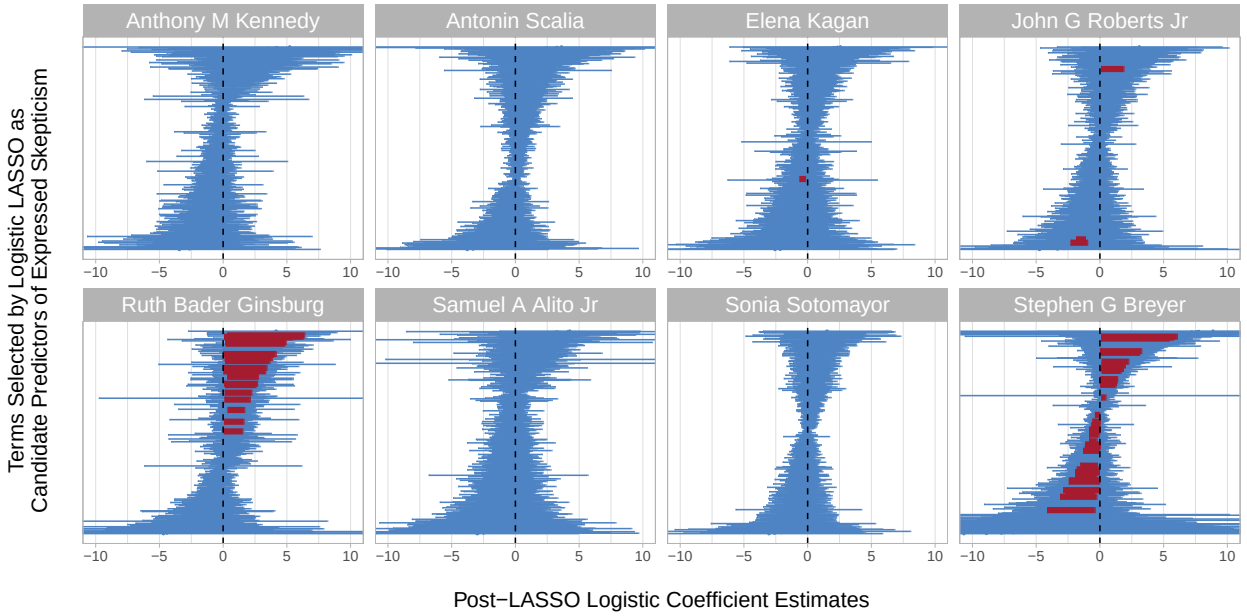

Figure 2: **Textual Signals of Justice Skepticism.** Each panel depicts a regression of MASS-predicted skepticism on word counts. Within each justice's utterances, candidate terms that may predict skepticism (selected by logistic LASSO) are arrayed on the y-axis. For each term, points and horizontal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

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

For Justice Stephen Breyer, an expressive orator who is by far the most frequently speaking justice, less than 50 such terms exist. For illustrative purposes, we focus on Breyer's "broad," "indeed," and "marry," 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:

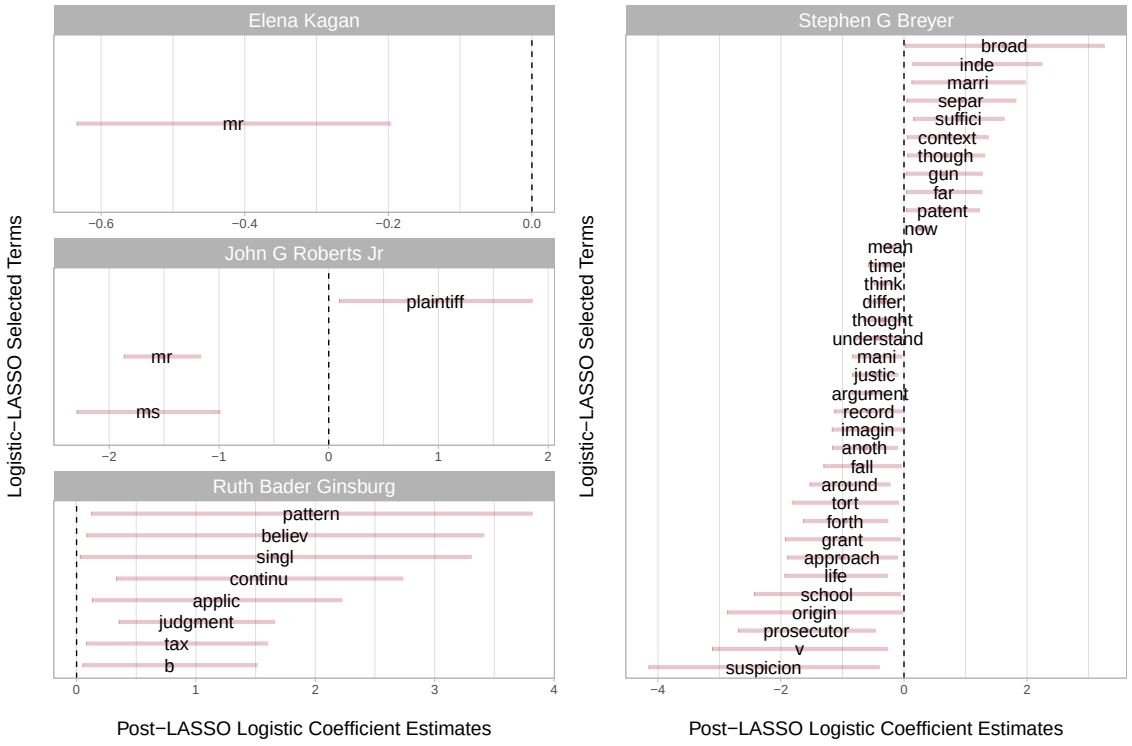

Figure 3: **Strong Textual Signals of Justice Skepticism.** Each panel depicts a regression of MASS-predicted skepticism on word counts, within a justice's utterances. Words that predict skepticism are arrayed on the y-axis. Reported terms are the subset of post-LASSO terms with post-selection logistic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

- BROAD (legal doctrine): "It's that second part, that the doctrine extends the doctrine to statutes that, while they may be clear, are far too broad, well beyond what any sensible prosecutor would even want to prosecute."

- INDEED (rhetorical device): "And indeed, you're complicating it even further for the reason that I really meant my question to be aimed at you, you know."

- MARRY (issue on which Breyer uses sarcasm heavily): "So if I marry two people in Washington D.C. and they happen to move to New York, you are saying that New York doesn't have to recognize that marriage because it doesn't comport with the marriage

of New York; is that your point?"

Conversely, Justice Breyer's neutral-leaning terms include technical terms ("prosecutor," "tort," and "argument") as well as the fairly innocuous ("thought" 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 ten informative words through this procedure; moreover, their cumulative predictive power is virtually nonexistent.

4.3 Auditory Characteristics of Expressed Skepticism

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

voiced neutral speech. The top panel of Figure 4 shows that when speaking skeptically, Kennedy speaks more loudly and with higher average pitch, a consequence of tensed vocal cords. Moreover, his modulation of pitch—which rises during questions and falls sharply during emphatic statements—is markedly larger in skeptical speech, as indicated by its higher pitch variance. We do not, however, observe similar modulation in energy: Kennedy is simply louder across the board when expressing skepticism. Finally, in the bottom panel, we contrast Justices Kennedy and Sotomayor to demonstrate that these speech dynamics are not entirely unique to individual speakers. While speaker baselines do vary—Sotomayor speaks more softly on average, and her voice is roughly six semitones higher—both communicate their skepticism by elevating pitch and raising their voices, among other auditory cues.

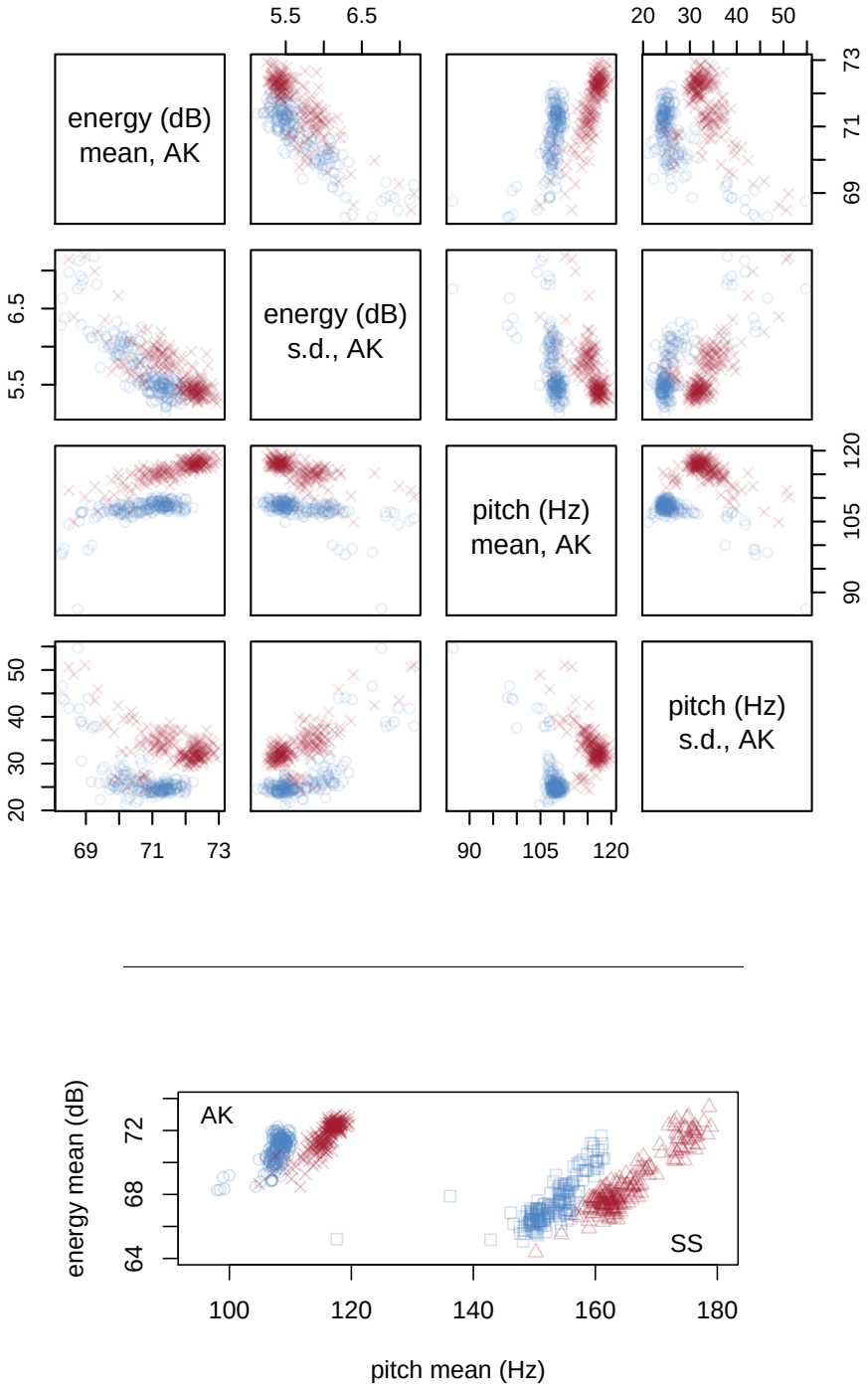

Figure 4: **Auditory characteristics of neutral and skeptical speech.** In the top panel, each dark red $\times$ (light blue $\circ$) represents a converged EM run for auditory parameters using a run-specific bootstrap draw of skeptical (neutral) training utterances for Justice Kennedy. Coordinates in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

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

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

4.4 Audio, Text, and Human Classification Performance

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

(OOB, see e.g. Hastie, Tibshirani, and Friedman, 2001, 15.3.1) performance is computed as follows. First, for labeled utterance u, we take all of the speaker's bootstrap speech models in which the utterance was out-of-bag (i.e., the roughly $\frac{1}{e}$ bootstrap resamples in which u was not drawn). For each bootstrap draw, the likelihood of utterance u is computed under the trained neutral and skeptical models, then converted to predicted tone probabilities of u. Predicted probabilities are then averaged over models. Results reflect the performance of a classifier that uses $1 - \frac{1}{e} \approx 63\%$ of the full training set. Across all speakers, we find that 68% of utterances are correctly classified ($F_1 = 0.554$). Speaker-specific results and other measures of performance are reported in Figure 5, along with measures of text classifier performance discussed in Appendix 4.2.

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

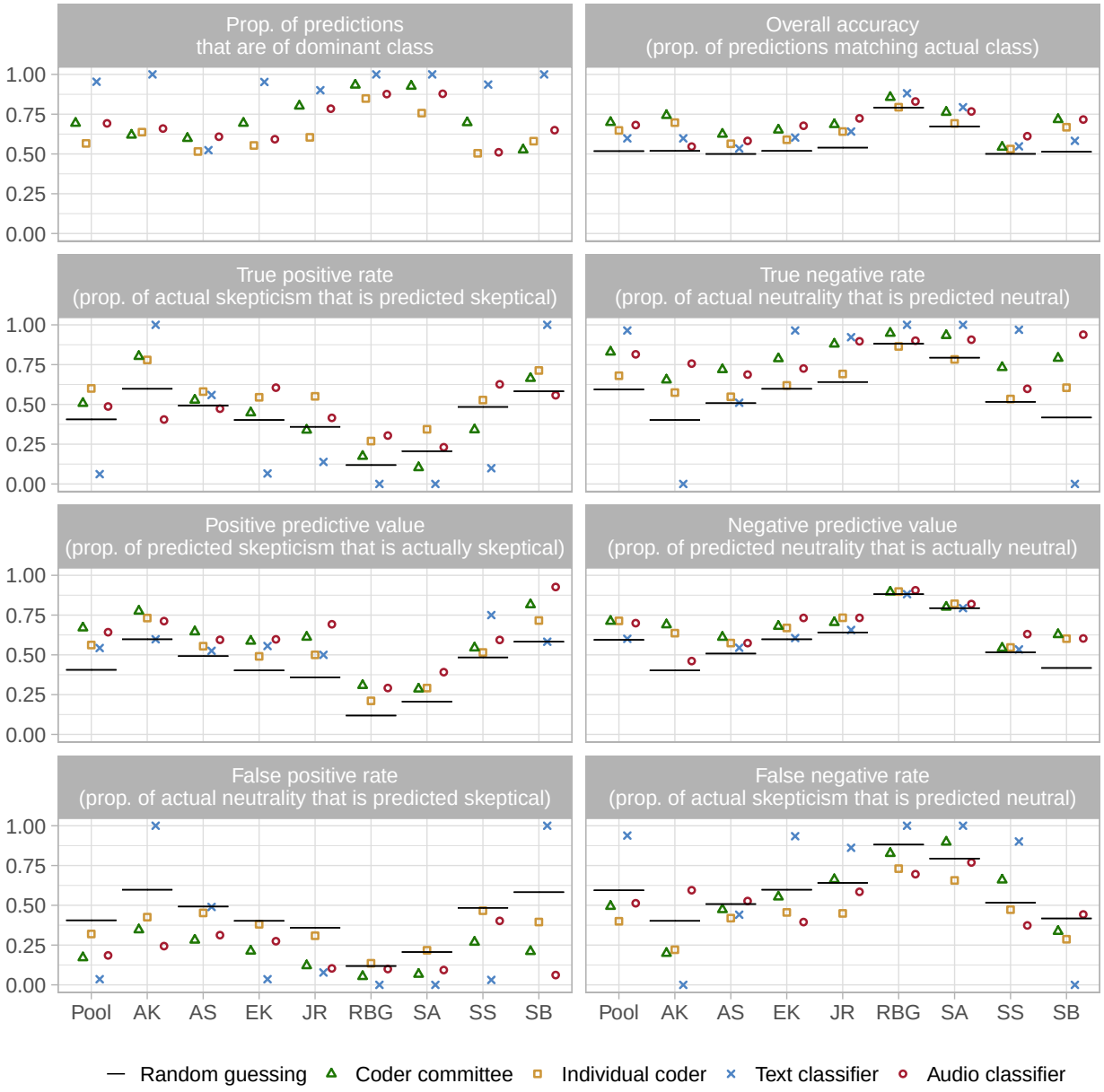

Figure 5: **Out-of-sample performance.** Red circles (blue crosses) indicate the performance of models trained on audio (text). As a point of reference, we also report the performance of individual coders (orange square) and coder predictions aggregated by majority vote (green triangle). Justices initials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

4.5 Comparison to Black (2011)

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

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

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

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

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

Table 2: **Common pleasant (unpleasant) words in justice speech.** Uses of the top 20 most common words in the top (bottom) decile of word pleasantness, as defined by the Dictionary of Affect in Language, in Supreme Court justice speech. The proportional contribution of each word to the measure of direct affect is computed by dividing a word's count by the total number of pleasant (unpleasant) words used.

Pleasant			Unpleasant		
Word	Count	Prop.	Words	Counts	Prop.
well	2912	0.11	not	7697	0.20
justice	1269	0.05	no	2630	0.07
like	1253	0.05	other	2307	0.06
us	845	0.03	mean	2144	0.06
read	648	0.02	argument	1295	0.03
talking	568	0.02	can't	1183	0.03
reasonable	559	0.02	problem	733	0.02
money	487	0.02	trial	529	0.01
agree	458	0.02	over	522	0.01
good	456	0.02	police	504	0.01
clear	451	0.02	without	504	0.01
yes	418	0.02	wrong	497	0.01
respect	387	0.01	against	473	0.01
view	385	0.01	nothing	466	0.01
correct	373	0.01	guess	406	0.01
interest	345	0.01	tax	394	0.01
sense	305	0.01	number	343	0.01
agreement	282	0.01	violation	284	0.01
company	280	0.01	unless	273	0.01
marriage	243	0.01	off	262	0.01

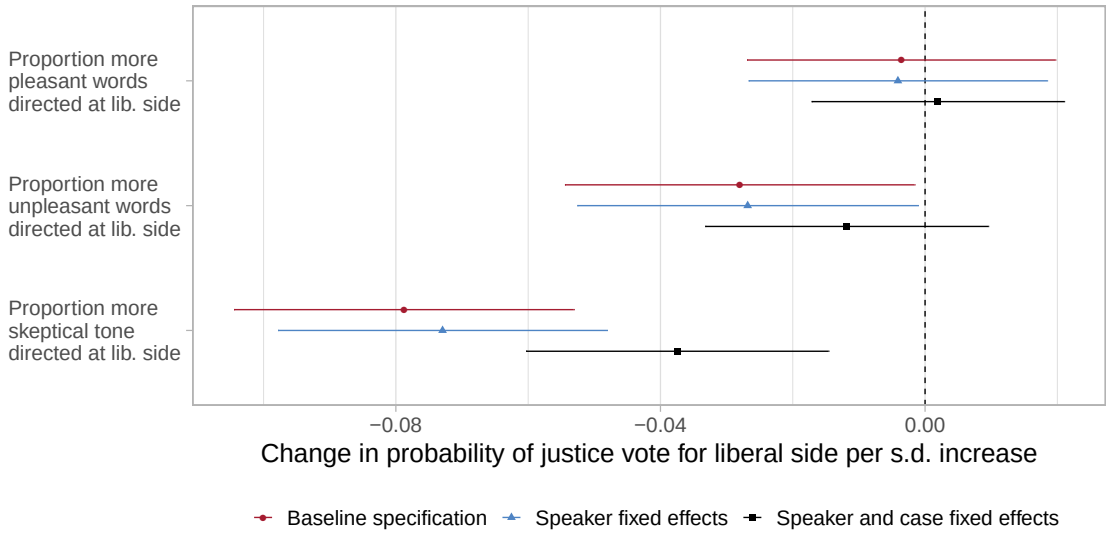

Figure 6: **Predicting justice votes with directed skepticism and directed affective language.** Horizontal errorbars represent point estimates and 95% confidence intervals from regressions of justice votes on directed pleasant words, directed unpleasant words, and our audio-based directed skepticism.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

4.6 Predicted Skepticism by Justice, Issue, and Target

In this section, we present exploratory analyses of how each justice differentially expresses skepticism depending on the legal issue and the target side's ideology. Table 3 presents issue areas in which justices appear to be strongly ideological, based on patterns of directed skepticism.

We first compute average skepticism within groups of utterances, using predicted values obtained from dynamic specification described in Section . Groups are defined by unique combinations of justice, issue area, and ideology of the target side. The latter measures are based on Supreme Court Database classifications (Spaeth et al., 2014). (Note that the specification does not include issue area or justice-issue interactions. We regard these results as suggestive, and scholars interested in issue-specific speech patterns are encouraged to model this behavior explicitly to avoid inadvertently attenuating estimates.)

Within each justice-issue, we then compare the average level of skepticism in utterances directed toward the liberal and conservative sides; results are reported for justice-issues with a substantial difference. Consistent with their known ideological predispositions, Justices Breyer and Ginsburg consistently express greater skepticism toward the conservative side, and Justice Scalia expresses greater skepticism toward the liberal side. However, the issue areas in which we observe strong ideological disparities vary by justice and appear to track the intensity of justice preferences. For example, Scalia holds strong views on the right to free speech, and this position manifests in the eight-percentage-point higher use of skepticism toward liberal advocates on First Amendment cases, relative to conservative advocates. Similarly, Justice Ginsburg is seen as a strong defender of civil rights, and uses five per-

centage points more skepticism toward conservative advocates on this issue. (Table 3 also highlights major differences in justice baselines; the notoriously stone-faced Justice Ginsburg uses relatively little discernible skepticism in general, so that this gap represents a two-thirds increase.) Finally, the table shows that Kennedy—consistent with his position as median voter—can be more skeptical of either the conservative or liberal side, depending on issue area.

Table 3: **Directed Skepticism by Justice, Issue, and Target.** Each row presents summary statistics that aggregate justice utterances on cases in a particular issue area. The final columns indicate the average predicted skepticism in speech directed toward conservative parties, liberal parties, and the difference in means. Results are shown only for justice-issues in which the absolute difference in directed skepticism exceeds five percentage points.

Justice	**Issue**	**Con.**	**Lib.**	**Diff.**
SB	Economic Activity	0.56	0.50	-0.06
SB	First Amendment	0.49	0.38	-0.11
SB	Other	0.53	0.48	-0.05
RBG	Civil Rights	0.15	0.09	-0.05
RBG	Economic Activity	0.17	0.11	-0.06
RBG	Other	0.18	0.10	-0.08
AK	Criminal Procedure	0.55	0.62	0.07
AK	First Amendment	0.60	0.50	-0.10
AS	Economic Activity	0.46	0.51	0.05
AS	First Amendment	0.47	0.55	0.08

5 `communication` `R` Package

In this section, we briefly describe our accompanying `R` package, `communication` (Duarte et al., 2020). Because the package is continually maintained and continues to be extended, researchers interested in conducting analyses with MASS will be best served by the latest package documentation. Here, we note the high-level features of our accompanying package and describe innovations over existing software.

First and most importantly, `communication` includes an efficient `C++` implementation of MASS, the model that is the primary focus of this paper. To our knowledge, there is no other structural model available for the analysis of speech dynamics.

Second, `communication` implements a number of preprocessing steps that, while not the focus of this paper, are critical for any applied research using speech data. Among many other utilities, these include input/output functions compatible with common file formats; fast extraction of auditory features that are generally understood to distinguish abstract categories of human communication; objects for corpus and metadata management; and functions for segmentation and human labeling of utterances. Notably, these tools are made available in `R` for the first time, increasingly the lingua franca of computational social scientists.

References

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

Black, Ryan, Sarah Treul, Timothy Johnson, and Jerry Goldman. 2011. "Emotions, oral arguments, and Supreme Court decision making." *Journal of Politics* 73 (2): 572–581.

Duarte, Guilherme, Alex Shmuley, Dean Knox, and Christopher Lucas. 2020. "communication: R Package for the Statistical Analysis of Human Speech." *R package version.*

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

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

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

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

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

Spaeth, Harold, Lee Epstein, Ted Ruger, Keith Whittington, Jeffrey Segal, and Andrew Martin. 2014. "Supreme Court Database Code Book." `scdb.wustl.edu`.

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

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

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

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