# Using the Audio from Telephone Surveys for Political Science Research

Bryce J. Dietrich[1,2,3*] and Jeffery J. Mondak[4]

[1]John F. Kennedy School of Government, Harvard University

[2]Department of Political Science, University of Iowa

[3]Department of Computer and Information Sciences, Northeastern University

[4]Department of Political Science, University of Illinois

[*]To whom correspondence should be addressed; E-mail: bryce_dietrich@hks.harvard.edu

**Abstract**

Despite the widespread use of telephone surveys, the audio from these surveys has yet to be used for social science research. The same can be said for in-person interviews in which answers are recorded, but no methodology has been developed to assess the way those answers are vocally delivered. In this study, we develop the first automatic emotional speech recognition (AESR) system which can effectively identify the emotional intensity associated with responses obtained from in-person interviews and telephone surveys. Using our system and the audio from two surveys, we find our audio-based measure of intensity is a statistically significant predictor of vote choice even when controls are included for party identification, ideology, age, education, income, gender and race. Ultimately, our study dramatically expands the scope of survey research and gives researchers the tools necessary to better use the audio from survey responses to answer political questions, laying an important foundation for future research.

**Early Draft and Preliminary Results:**
**Please do not circulate without permission**

# Introduction

The image of two individuals sitting silently next to one another, texting rather than talking, has gained iconic status. In a relatively short span of time, exchanges that take place via the written word have come to characterize much of interpersonal communication. People send one another texts and emails, and they post comments on platforms such as Twitter, Instagram, and Facebook. Although this increase in brief written communication brings undeniable efficiency, we all also recognize, whether implicitly or explicitly, that something is lost when we move from talking to texting. Readers of Twitter and Facebook posts often fail to capture the intended meanings of those messages because the written words alone do not convey the writer's underlying state of sarcasm, anger, or elation. Due to its tremendous capacity for expressiveness, the human voice signals the communicator's emotions in a manner the written word, even with the slap-dash aid of emojis, rarely can match.

The absence of the spoken word may deny us insights regarding a person's emotional state, but the reality is that social scientists interested in the political significance of emotion rarely have capitalized on the richness of the human voice in their research. We may want citizens to "sound off" so that their "opinions are heard," but we most often measure emotions through text – text that inherently implicates cognitive processes, and that inherently omits any meaning communicated through vocal tone and inflection. Survey respondents are asked to ponder, recall, and report what emotions they have experienced. People's feelings of happiness, fear, or annoyance then are reduced to a few self-reported numbers on five- or seven-point scales. This is how emotions almost always are measured in survey research, and how they necessarily are measured on internet surveys – surveys on which no one speaks, and no one listens.

The alternate scenario we envision, and that we seek to explore in the present paper, is that survey respondents' politically-significant emotional states can be measured validly, efficiently, and unobtrusively through analysis of those respondents' spoken words. In this study, we use as our raw material audio recordings of two surveys, one conducted in person and one by telephone,

that were administered with focus on the 2012 U.S. presidential election. The paper first reports our efforts to extract information on emotions from these recordings. Human coding of a sample of sound clips provides baseline data that we use to inform a machine-learning protocol. That machine coding then is employed to explore the acoustic properties of the remaining cases. To illustrate the utility of these exercises, the resulting data are used in assessments of the interplay between respondents' emotions and their reported likes and dislikes of Barack Obama as predictors of the 2012 vote choice.

Our central methodological innovation is an unobtrusive measure of the intensity of survey respondents' opinions. We create this measure using an automatic emotional speech recognition protocol (AESR) that can be more generally employed to analyze digital recordings of in-person and telephone survey interviews. The convergence of multiple factors related to political and psychological theories of emotion, technological innovations in speech science, and salient features of the contemporary political arena make such an inquiry particularly timely and important. As online surveys become increasingly synonymous with social science research, we demonstrate that the audio from in-person interviews and telephone surveys can yield additional insights into the underpinnings of respondent opinions.

## The Physiology of Emotion

In research on politics and emotions, information on several aspects of emotional response may be enlightening. First, and most fundamentally, it would be useful to be able to distinguish between an individual's neutral states and states of emotional activation. Second, emotional responses should be subject to some basic form of categorization, such as in terms of positive or negative valence. Third, information about the temporal aspects of emotional expressions is needed if the analyst is to differentiate among a fleeting response, an enduring state (i.e., a mood) and a chronic disposition (i.e., a trait). Fourth, if possible, it would be beneficial to identify the specific emotions activated under various circumstances and in response to various prompts. It is our contention that all of

these ends can be achieved with a focus on the physiology of emotion, and particularly the vocal expression of emotion as measured by AESR.

The plausibility of AESR in research on political behavior hinges on whether voice signals provide valid and reliable information about emotions. To understand why they do requires that we step back and consider the bases of emotional response. Our perspective emphasizes the role of biology, and especially the thesis that evolutionary forces have given rise to emotional responses that are purposive. This view draws on the foundation established by leading contemporary scholars on politics and emotion, and especially the work of George Marcus and his colleagues (e.g., Marcus 2003, 2010; Marcus, Neuman and MacKuen 2000; Neuman et al. 2007). The Marcus et al. theory of affective intelligence holds that multiple systems of emotion function to direct or manage learning, and to control attention, such as in response to threat. We add one critical point to this view, which is that the motor expression of emotion – i.e., communication via facial expressions, voice, and gestures – also is of adaptive benefit. Our particular focus in the current research is vocal expression. In part, this path is pragmatic in that we can study vocal expression, but not facial expressions or gestures, with data obtained through telephone surveys. But this attention to vocal expression also serves our interest in social communication about politics. Facial expressions and gestures signal emotions only during face-to-face exchanges, whereas vocal expression occurs during both face-to-face encounters and interactions of other types, such as telephone conversations, chats across cubicle walls and conversations among members of a carpool.[1]

Research in neuroscience establishes that the activation of emotion systems is preconscious. Indeed, it is precisely because of preconscious response that emotions serve to control attention. Two practical questions must be considered. First, given the neurological basis of emotional response, are there alternate physiological measures that might be preferable to AESR for applied research on political behavior? Second, to what extent, if any, is motor expression of emotion the involuntary consequence of antecedent neurological/physiological processes?

[1]The political significance of communication of emotions via facial expressions (e.g., Stewart, Waller and Schubert 2009; Sullivan and Masters 1988), gestures (e.g., Streeck 2008), and a combination of both (e.g., Grabe and Bucy 2009) has received occasional attention in the literature.

Aspects of social and political judgment, often with focus on emotion, have been studied in recent years using standard techniques in neuroscience. These include measures of brain function such as functional Magnetic Resonance Imaging (e.g., Greene et al. 2001; Sanfey et al. 2003; Westen et al. 2006), Event-Related Potentials (e.g., Amodio et al. 2007; Boudreau, McCubbins and Coulson 2008), an even variants of the lesion method (e.g., Knoch et al. 2006; Koenigs et al. 2007).[2] Also, several recent studies have examined emotions and politics via physiological measures of blink amplitude and skin conductance (e.g., Mutz and Reeves 2005; Oxley et al. 2008). We see value in these approaches, but we also view AESR as an important complement. A focus on vocal expression of emotions potentially offers an unobtrusive and efficient means to measure a socially-discernible manifestation of emotional response. These features–that it is unobtrusive, efficient, and that it centers on observable motor expression–make AESR a highly–desirable addition to survey-based data acquisition.

From the perspective of the survey respondent, AESR is invisible, and thus fully unobtrusive. This contrasts starkly with the techniques noted above. With fMRI, the respondent is instructed to remain extremely still, and is posed questions while encased in a confined horizontal space, and while subjected to the deafening roar of the machine's magnet. ERPs are measured with electrodes placed at multiple locations on the respondent's scalp, and the respondent is required to minimize blinking during the procedure. The traditional lesion method and its newer variants require either the availability of patients who have suffered actual damage to a particular brain area or the local disruption of brain function. Blink amplitude is measured with electrodes placed just below a person's eyes, and skin conductance is measured with sensors attached to the respondent's fingers. These techniques all instill a high level of artificiality to the data acquisition process. Further, because all of these procedures except for the traditional lesion method make use of specialized equipment, their application requires that the respondent be brought to a laboratory. Data acquisition is costly and inefficient, making these approaches infeasible at present for large N studies

[2]For discussion of neuroimaging techniques as they relate to the study of politics and emotion, see Spezio and Adolphs (2007).

such as telephone surveys.

Measures of skin conductance and blink amplitude capture physiological responses over which the individual has little or no control. But is the same true of motor expression? To some extent, people do regulate their facial expressions, their voices and their gestures. Nonetheless, the emotional cues transmitted through motor expression–and measured with techniques such as AESR and facial recognition applications–emerge primarily as the consequence of physiological changes that are beyond the individual's control, and often beyond the person's awareness. Russell (2003) equates this process to one's body temperature. Even though your body temperature is always present and you can note it whenever you want, only extreme changes become noticeable. However, regardless of the magnitude, changes exist prior to the conscious salience of words such as "hot" or "cold." Russell (2003) argues emotions work in a similar way and can affect behavior prior to conscious awareness by changing the way we process new and existing information.

Darwin (1998) first noted that vocal cues and facial expressions signal the activation of particular emotions. Darwin's focus was on the adaptive roles of these somatic changes. For instance, vocal outbursts reflective of fear tend to be loud, enabling such expressions to serve a socially-beneficial warning function. The pace of scientific research on emotion and voice accelerated starting in the 1930s (for an important early work, see Fairbanks and Pronovost (1939)). Tremendous progress has been made since then both in specifying how emotions trigger physiological changes that ultimately influence features of vocal expression and in identifying the specific prosodic qualities associated with various emotional responses. On the first of these points, the process by which emotions influence the voice's acoustical properties, the physiological changes in voice triggered by emotions are parallel to the effects that alter skin conductance. Johnstone and Scherer (2000, 222) summarize current understanding:

> The most fundamental determinants of vocal affect are the physiological changes that accompany emotions, which in turn produce changes in the functioning of the vocal production systems... (E)motions are accompanied by various adaptive responses in

the autonomic and somatic nervous systems. These responses will lead to changes in the functioning of parts of the speech production system, such as respiration, vocal fold vibration, and articulation. For example, with a highly aroused emotion such as rage, increased tension in the laryngeal musculature coupled with raised subglottal pressure will provoke a change in the production of sound at the glottis, and hence a change in voice quality.

Multiple features of vocal expression have been examined in efforts to identify the vocal signs of emotion activation and to specify the acoustical correlates of particular emotional responses.[3] One of the more important class of indicators involves fundamental frequency (i.e., pitch), which is typically designated F0. Relevant aspects of F0 for emotion recognition include the mean, range, variability, contour, and evidence of perturbation. Table 1 lists other acoustical measures which center on properties such as intensity and vocal amplitude, vocal perturbation and voice quality, and speech rate. These all can change as a consequence of an emotional reaction. For instance, an emotional response can induce physiological reactions that lead the person to speak with a higher or lower pitch, with greater or lesser intensity, and at a faster or slower rate.

Across scores of studies, convergent evidence has accumulated regarding the patterns in vocal expression associated with both the general activation of emotions and the presence of specific emotional responses. Although dozens of discrete emotions have been examined, the literature is most clear on the prosodic markers that differentiate neutral states from states of emotional activation, and on the vocal characteristics associated with a handful of basic, archetypal emotions.[4] Given the consistent use of vocal pitch as a measure of emotional arousal in the psychology and

[3]Numerous reviews discuss the voluminous research on voice and emotion, in many cases in conjunction with discussion of technological developments in research on automatic emotion recognition. Most of these reviews detail the key features of the prosodic domain, and summarize research linking those characteristics to both general emotional activation and the expression of specific emotions. Good examples include Cowie et al. (2001); Scherer (2003); Owren and Bachorowski (2007); Scherer (1986); Zeng et al. (2009).

[4]The literature is quite vast, and thus we will not dwell at length on the findings of individual studies. Banse and Scherer (1996) provide a good example of research in this area, and Simon-Thomas et al. (2009) is an interesting more recent study.

Table 1: Description of audio variables commonly used in AESR

| Variable | Description |
|---|---|
| PCM Loudness | Given that a particular change in amplitude is not perceived as a proportional change in loudness, the amplitude of a signal must be standardized. In Pulse Code Modulation (PCM) systems, perceived loudness is calculated using the ratio of the maximum amplitude and the inherent noise in the system. |
| MFCC | Speech is analyzed at the frame- and trend-level. The Mel Frequency Cepstral Coefficients (MFCC) is an example of the latter since they provide a summary of the energy distribution at specific frequencies for the entire speech signal. Ultimately, they are returned in the Mel scale which relates the perceived frequency (or pitch) to the actual measured frequency. Since the vocal tract is manipulated to change the perceived frequency (or pitch), the MFCC essentially captures the shape of the vocal tract. |
| LPCC | The Linear Predictive Coding Coefficients (LPCC) is another type of frame-level measure which provides a summary of the entire speech signal. However, instead of using a quasi-logarithmic scale similar to the MFCC, the LPCC uses the past values in a speech signal to predict the current values using a linear function. Unlike the MFCC, the LPCC is mostly used to model how a speech signal is produced rather than how it is perceived. With that said, both are trend-level measures and capture the energy distribution at specific frequencies for the entire speech signal. |
| Fundamental Frequency (F0) | The Fundamental Frequency (F0) is perceived by the human ear as pitch. It represents the frequency at which the vocal folds are opening and closing. This serves as the basis for all human speech since this quasi-periodic function resonates throughout the vocal tract ultimately producing the sound we hear. It is "fundamental" since it is often associated with the source of a speech signal (i.e., emotional activation) as compared to the vocal tract which filters the source to create specific sounds (i.e., words and phrases). |
| Jitter | The number of cycles the vocal folds make in a second is the fundamental frequency. These cycles are primarily determined by the degree of longitudinal stress placed on the vocal folds and the dimensions of the vocal folds themselves. Jitter is the variability of the fundamental frequency which ultimately captures the degree to which an individual has control over the vocal fold vibration. Rough (or hoarse) voices tend to have high jitter. |
| Shimmer | Shimmer is very similar to jitter, but instead of capturing the variability of the fundamental frequency shimmer captures the variability of amplitude. Unlike the fundamental frequency, the amplitude of a speech signal does not measure the rate of vocal fold vibration. Instead, it measures the size of the oscillations with greater amplitude implying the speech signal has more energy and will ultimately be perceived as being louder. Rough (or hoarse) voices also tend to have high shimmer which is why jitter and shimmer are often used together in the same model. |

computational linguistics literature, F0 has been used the most in the social science literature. For example, drawing from earlier work in social psychology (e.g., Gregory Jr and Gallagher 2002;

Tigue et al. 2012) Klofstad and co-authors (e.g., Klofstad and Anderson 2018; Klofstad 2017, 2016; Klofstad, Nowicki and Anderson 2016) have demonstrated pitch changes can influence vote choices. Similarly, using large-N studies of elite speech Dietrich and co-authors (e.g., Dietrich, Hayes and O'Brien 2018; Dietrich, Enos and Sen 2018; Dietrich and Juelich 2018) have used vocal pitch as an important and useful measure of emotional activation or intensity.

Noticeably lacking from this literature are the numerous other audio variables that have been used to in AESR (see Table 1). Of these, the only one that has been utilized in political science outside of F0 is the Mel Frequency Cepstral Coefficients (MFCC) which are used by Knox and Lucas (2017) to study oral argument dynamics on the Supreme Court. Numerous efforts at automatic emotional speech recognition have been conducted since the 1970s.[5] In each of these instances, a large number of audio variables are combined into a single predictive model. Unfortunately, none of the methods have been used by political scientists – especially with respect to telephone surveys. We seek to rectify this in the following pages. More specifically, our two-part objective is to develop and implement an AESR procedure in the context of the mass opinion survey, and to demonstrate that procedure's utility through a familiar application, the presidential vote choice. We now turn to the description of procedures we employed, the data we acquired, and the technology developed for the present study.

# Data

For our in-person study we used a convenience sample of 252 respondents recruited from a mid-sized county in Kentucky shortly after the 2012 Presidential Election. Respondents were recruited through advertisements both on campus (for these, university staff were specifically targeted; see Kam et al. 2007) and in the local community. All in-person interviews were conducted on the campus of small regional university between 11/11/2012 and 11/15/2012. Our phone survey was

[5]For recent examples, see Ślot et al. (2009); Xiao et al. (2010). The latter includes an interesting discussion of the possible applications of AESR in education, entertainment and business.

also conducted by a call center on the same campus and consisted of 234 respondents. All phone numbers were drawn from the county we used for the in-person interviews, and were randomly dialed. The phone interviews began on 11/11/2012 and ended on 11/30/2012. Table 2 reports demographic variables for our in-person and telephone samples and the corresponding county.

The data in Table 2 reveal that, as compared with the population of the Kentucky county as a whole, participants on our in-person survey show the greatest discrepancy on education, modest discrepancies on income and race, and are notably close matches for the state population on gender, age, and the proportion of Republicans. At the least, the sample comes much closer to representing the county population on the observed variables than would a student-based convenience sample. We find similar results for the telephone survey which diverges from the county population with respect to age, but is closer in terms of education, gender, race, and the proportion of Republicans. Our performance with respect to Democrats may be partially due to our county measure of party identification which is the proportion of the vote received by Barack Obama and Mitt Romney which does not reflect the actual number of Republicans and Democrats in the county. All other variables are derived from 2010 Census estimates.

The noise associated with audio recordings potentially poses a severe problem for AESR. To obtain high-quality recordings, it is required that audio be recorded digitally, preferably via lines that feed directly into the computer rather than running from the survey interviewer's headset to the computer. The sound quality generated by the interviewer also is important, and thus headsets must meet minimum technical specifications. Staff members from the call center had experience with digital recording of interviews and the corresponding issues regarding sound quality. Computers were equipped with the DLI Personal Logger system, which feeds audio signals directly to the computer's sound card. Interviewers also use Plantronics Supra Binaural headsets which exceed the minimum technical requirements. As a test of audio quality, the call center staff provided us with WAV files[6] from a series of interviews they recorded for other projects. Due to variation in

[6]WAV files contain uncompressed audio signals, with over 44,000 samples per second. Familiar compressed alternates (e.g., the mp3) are not adequate for AESR, which requires high-quality samples in either WAV or AIFF format.

Table 2: Respondents' Demographic Characteristics

| | Kentucky County | In-Person Interviews | Telephone Survey |
|---|---|---|---|
| *Education* | | | |
| Some High School | 0.15 | 0.01 | 0.04 |
| High School Graduate | 0.30 | 0.08 | 0.24 |
| Some College | 0.20 | 0.37 | 0.25 |
| College Degree | 0.23 | 0.26 | 0.28 |
| Post-Graduate Degree | 0.11 | 0.25 | 0.19 |
| *Income* | | | |
| Less than $24,999 | 0.28 | 0.23 | 0.17 |
| $25,000 to $49,999 | 0.27 | 0.18 | 0.29 |
| $50,000 to $74,999 | 0.18 | 0.21 | 0.25 |
| $75,000 to $99,999 | 0.12 | 0.20 | 0.15 |
| More than $100,000 | 0.14 | 0.17 | 0.13 |
| *Gender* | | | |
| Male | 0.49 | 0.47 | 0.41 |
| Female | 0.51 | 0.53 | 0.59 |
| *Age* | | | |
| Median | 32.60 | 31.50 | 62.00 |
| *Race* | | | |
| White | 0.85 | 0.74 | 0.91 |
| Black | 0.09 | 0.22 | 0.03 |
| Asian | 0.03 | 0.01 | 0.01 |
| Hispanic | 0.04 | 0.01 | 0.01 |
| Biracial | 0.02 | 0.03 | 0.02 |
| *Party Identification* | | | |
| Democrat | 0.54 | 0.42 | 0.42 |
| Republican | 0.38 | 0.33 | 0.30 |
| Independent | – | 0.25 | 0.27 |

*Note*: All county demographic variables were obtained from the 2010 Census. The number of Democrats and Republicans is the 2012 vote share for Barack Obama and Mitt Romney.

respondents' telephones, the interviews covered a range in terms of sound quality. Our initial tests determined these samples were of sufficient quality for our purposes.

# Automatic Emotional Speech Recognition

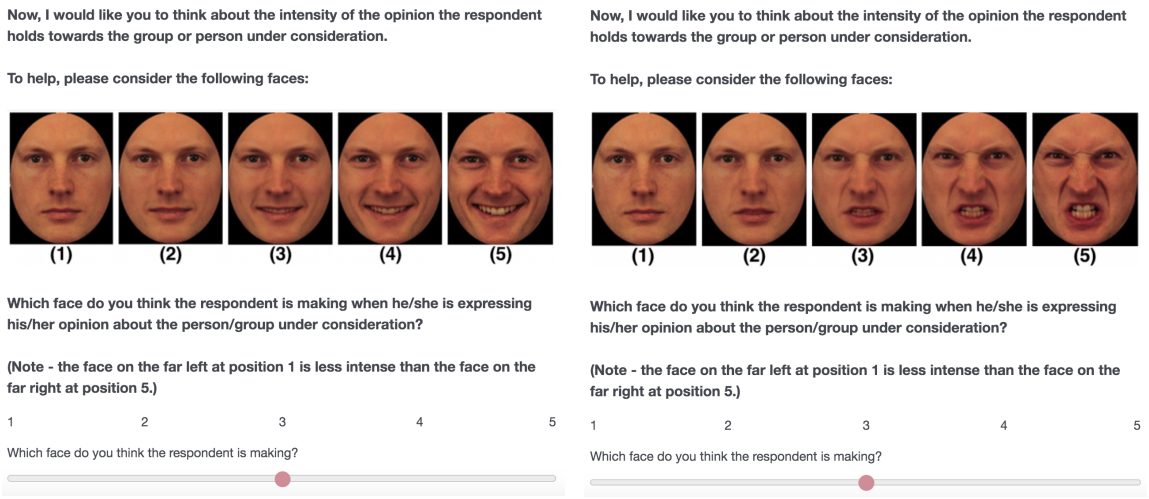
## Feature Extraction

Upon acquisition of sound samples, AESR requires samples be subjected to acoustic feature extraction and a random sample must be annotated. The audio files from the in-person interviews were manually segmented by two research assistants using Audacity (`http://audacityteam.org/`). For the phone survey, we had the interviewers record their results using an online survey instrument. Embedded within this instrument were internal timestamps which we then used to extract a single file for each survey response. To ensure all audio was from the respondents, we then manually removed extraneous dialog from the interviewer and other sounds unrelated to the interview. We did this for both the in-person and phone audio files.

Features were extracted using an open-source audio analysis and pattern recognition tool called openSMILE (`https://audeering.com/technology/opensmile/`). For this study we used the feature set from the 2010 InterSpeech challenge (`https://www.isca-speech.org/archive/interspeech_2010/index.html`), which can be found in the IS10_paraling openSMILE configuration file. The feature set includes 1,582 different permutations of 38 base-level features. These base-level features include: PCM loudness, MFCC, LPCC, fundamental frequency (F0), Jitter, and Shimmer. Each of these variables is explained in Table 1. All 38 available measures were used for the purpose of this analysis in order to preclude the need for us to impose our own a priori assumptions about which features best capture different emotional states.

Once the features were extracted, we excluded audio files in which the respondent said nothing or gave a one-word response, such as "Yes" or "No." To achieve this end, we uploaded our initial collection of 18,312 audio files to a buck on Google's cloud storage. Using the Google Cloud Speech-to-Text API, we then generated automatic transcripts. Although studies have shown such automated services yield reasonable results (e.g., Bokhove and Downey 2018; Ziman et al. 2018), we use speech-to-text simply to help filter our large collection of audio files and we do not use the

Figure 1: Audio File Annotation




*Note*: To annotate the audio files, we first asked our coders to assess whether the respondent expressed a favorable or unfavorable opinion towards the person or group described in the audio file. If they said the opinion was favorable, then they received the "happy" scale on the left. If they said the opinion was unfavorable, they received the "angry" scale on the right. The slider at the bottom could take on any value between 1 and 5 which serves as our main variable of interest.

transcription results in our final model. Once we eliminated audio files with little to no content, we had 13,523 usable files. A random sample of 3,380 was then annotated using an original scale for the intensity of the respondents' opinions which we outlined in Figure 1. Two graduate students and one undergraduate student coded the audio files. None of the students involved was aware of the purpose of the project. This human coding was done to provide training data with which to inform our subsequent automated coding procedures.

Before we annotated the full training set, we first established intercoder reliability using a random sample of 50 audio files. Using the intensity scale outlined in Figure 1, we had all three coders annotate all the sample files. Intercoder reliability was assessed using the Interclass Correlation Coefficient (ICC). Since we ultimately want to randomly assign two coders to all the files in our larger training set and then take the average, we set $k = 2$ and used a random effects model. This was done using the `ICC` function from the `psych` library in the `R` statistical software language. Ultimately, for our initial sample the ICC was 0.81 which Koo and Li (2016) describe as "good reliability."

## Supervised Machine Learning

Similar to other AESR studies, we do not make strong assumptions about which audio features best capture different levels of emotional intensity. Instead we will use several machine learning algorithms to identify the optimal configuration using (1) the Mean Absolute Error (MAE) and (2) Root Mean Squared Error (RMSE). The former is the average absolute difference between the predicted and actual score, whereas the latter is the average square root of the squared difference between the actual and predicted score. We selected these assessment metrics based on suggestions provided by Luo et al. (2016).

Before estimating any of our models, we set aside 25 percent of the annotated data (or 845 files) for testing. This left 2,535 files for training. After our training and testing sets were created, we then took the following pre-processing steps. First, we subdivided our data into male and female respondents. This was done because the vocal properties of women are distinctly different from those of men (Titze 1989). Second, we eliminated any audio variables that had close to zero variance. This was done using the `nearZeroVar` function in the `caret` library in the `R` statistical software language. Finally, we also eliminated audio variables which were highly correlated using the `findCorrelation` function in the `caret` library. In both of the preceding steps, the default settings were used and the number of audio variables incorporated into the models was around 750.

We used repeated 10-fold cross-validation to fit the training models. More specifically, we randomly assigned cases to 10 equally sized folds and used 9 of those folds to train each machine learning algorithm and the omitted fold for testing. This process was repeated for all 10 folds, meaning each fold was used exactly once in the testing data. We repeated this process using 3 different data partitions. The resulting model was tested using the 845 files left out from testing. Given that our emotional intensity scaled is a continuous variable, all machine learning algorithms use an underlying regression framework.

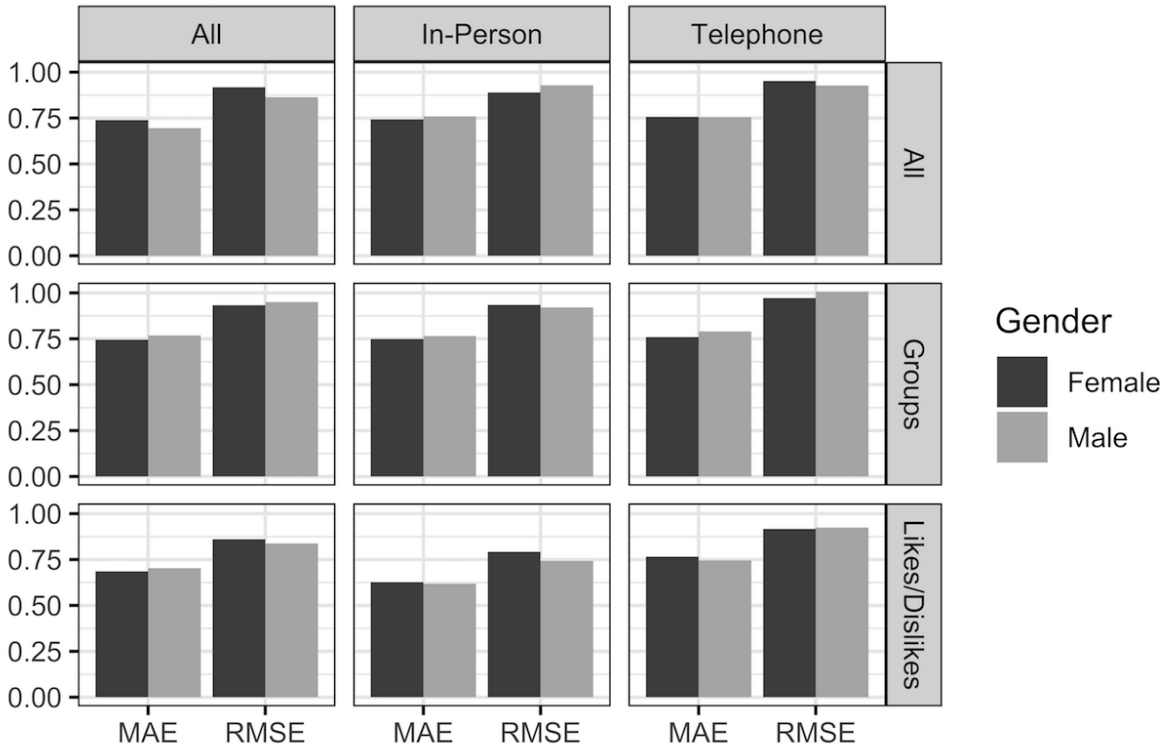
***Support Vector Machine***

Although we plan on testing a number of different algorithms, we currently only have the results for the Support Vector Machine (SVM). The SVM goal is to find the optimal separating hyperplane which maximizes the margin of the training data. The "margin" of the hyperplane is essentially the distance between the hyperplane and the closest data point, meaning no data point will lie within the margin itself. Once found, the optimal hyperplane will best predict the training data and will also best predict unobserved data, assuming the training data are representative of the types of cases that will be observed out-of-sample. In two-dimensional space (implying two covariates), the "hyperplane" is a line. In three-or-more dimensional space (implying three or more covariates), it is a more complex multidimentional object, but it functions very similarly to the line attempting to separate data points in traditional ordinary least squares.

One of the most important choices when estimating an SVM is the kernel which influences the dimensions of the estimated hyperplane. For this study, we tested three different kernels: linear, polynomial, and radial. Ultimately, we found the radial basis function (RBF) kernel performed the best. This kernel allows for non-linear combinations of the feature space while being computationally more efficient than the polynomial kernel which is why it is generally preferred over other options. As explained above, repeated cross-validation was used to select the sigma and cost parameters which determine the relative influence of a single training example and the ultimate decision function used to fit the model.

Model performance is reported in Figure 2. The black and gray bars refer to female and male respondents, respectively. Columns list the subset of data we used to estimate the SVM. Rows are the questions we included. For the "group" questions, respondents were asked to tell the first thing that came to mind when they thought of the following people and groups: the ACLU, Joe Biden, billionaires, conservatives, Democrats, homosexuals, liberals, Mitt Romney, Muslims, Barack Obama, Republicans, Paul Ryan, members of the Tea Party, and welfare recipients. In the "likes/dislikes" questions, respondents listed what they liked and disliked about Barack Obama

Figure 2: Model Performance for Various Questions and Respondent Types




*Note*: On the $x$-axis we report both the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for all models. The results of which are plotted on the $y$-axis. Black and grey bars refer to male and female respondents, respectively. Columns list the subset of data we used to estimate the SVM. Rows are the questions we included. The model used for both applications is found in the top-left (see "All" Column and "All" Row).

and Mitt Romney. Finally, on the $x$-axis we report both the RMSE and MAE for all models.

Beginning in the top-left, we find when we use all the audio files from male respondents in both the in-person and telephone surveys and all the questions the MAE is 0.70, which means on average the full model predicts our human-coded emotional score within $\pm 0.70$ units, which represents a little over half a point. For female respondents, the MAE is 0.74 using the full models which suggests we can predict the emotional intensity of male respondents slightly better than female respondents. For both male and female respondents, the best performing model is the model which only uses audio files from the in-person interviews and is restricted to candidate likes/dislikes. For men, the MAE is 0.62, which suggests our SVM can predict the emotional intensity of responses to candidate likes/dislikes statements within $\pm 0.61$ units. For women, the MAE for the same model is 0.63, which is noticeably better than the full model.

We find similar results for the RMSE. Again, we do slightly better predicting the emotional intensity of male respondents, with the full model returning a RMSE of 0.86 which is slightly lower than the 0.92 RMSE for female respondents. Similarly, we do the best when predicting emotional intensity for candidate likes/dislikes. The worst performing models are ones in which the emotional intensity of group responses is predicted and only data from telephone surveys is utilized. However, the performance decrease is slight, suggesting our SVM does a reasonable job predicting our emotional intensity scale – regardless of the data we utilize. Indeed, the model for both men and women has a MAE of 0.79 and RMSE of 1.00, which are promising results given our scale. Moreover, we expect model performance to continue to increase as we add more training data and use different supervised learning algorithms. Given that our best results for likes/dislikes, we use these responses for our first application.

# An Application: Explaining Vote Choice in the 2012 Presidential Election

## Background

What factors influence how individuals cast their ballots? This is one of the central questions in studies of American political behavior. Whether it is a general assessment of government performance (e.g., Ferejohn 1986; Ashworth 2012), evaluations of specific public policy issues (Carmines and Stimson 1980; Rabinowitz and Macdonald 1989), party preferences (Campbell et al. 1960; Bartels 2000; Fiorina 1981), or certain candidate qualities (e.g., Funk 1999; Hayes 2005), vote choices can be influenced by many things (for review, see Bartels 2010). Emotions can also play an important role with the theory of affective intelligence being the most prominent account of how emotions operate in this realm (for review, see Mutz 2007).

According to Marcus, Neuman and MacKuen (2000), when voters are anxious they are more likely to search for information, and heuristics like partisanship are less influential, which ultimately makes voters more sophisticated in the voting booth. Enthusiasm, on the other hand, leads voters to participate more, but has little effect on the quality of vote decisions. One of the main criticisms of the theory of affective intelligence is its reliance on emotion self-reports, which psychologists have been increasingly called into question (e.g., Robinson and Clore 2002). In order for self-reports to be valid, an individual has to be both willing and able to recall past emotions. In terms of the former, individuals with high social desirability bias may be reluctant to report negative emotional states (i.e., Paulhus and Reid 1991), such as anger (i.e., Welte and Russell 1993). Survey respondents also may differ more generally in what emotions they feel comfortable in reporting. For instance, research has shown that men and women differ in terms of the emotions that they are willing to express, with women being more likely to report "happiness," "sadness," "fear," and a general level of "emotionality," and men being more likely to report "anger" (Birnbaum and Chemelski 1984; Birnbaum, Nosanchuk and Croll 1980; Brody 2006). For these reasons, psychol-

ogists are increasingly relying on autonomic measures of emotion (Mauss and Robinson 2009).

The autonomic nervous system (ANS) consists of sympathetic and parasympathetic branches which are generally associated with activation and relaxation. Several scholars have suggested this dimension is associated with what Russell calls "arousal" (e.g., Russell 1980, 2003). More specifically, the circumplex model of affect posits that all affective states arise from two neurophysiological systems, one related to a pleasure-displeasure continuum (called "valence") and the other related to alertness (called "arousal" or "activation"). According to Russell (2003), at any given moment, one's emotional disposition is a single integral blend of these two dimensions. The horizontal dimension ranges from one extreme (e.g., agony) through a neutral point to its opposite extreme (e.g., ecstasy). But of particular interest is the vertical dimension, which ranges from a deactivated emotional state, such as being sleepy, to an activated emotional state, ultimately culminating in "frenetic excitement" (Russell 2003, 148).

Reticular formation (RF) is thought to regulate activation and arousal within the limbic system and thalamus (Heilman 2000; Jones 2003). When sensory stimuli are present, emotional arousal is likely relayed to the RF through the amygdalorecticular pathways (Koch and Ebert 1993; Rosen et al. 1991) which increases activity in the cerebral cortex (Heilman, Watson and Valenstein 2011; Jones 2003). This triggers changes in muscle tone and in the sweat glands (Jones 2003), both of which are associated with subjective ratings of emotional arousal (Lang et al. 1993). Increased blood flow to the vocal folds and changes in the respiratory system are why many studies have found vocal changes to be associated with emotional activation (Mauss and Robinson 2009), or what Watson and Tellegen (1985) call "engagement."

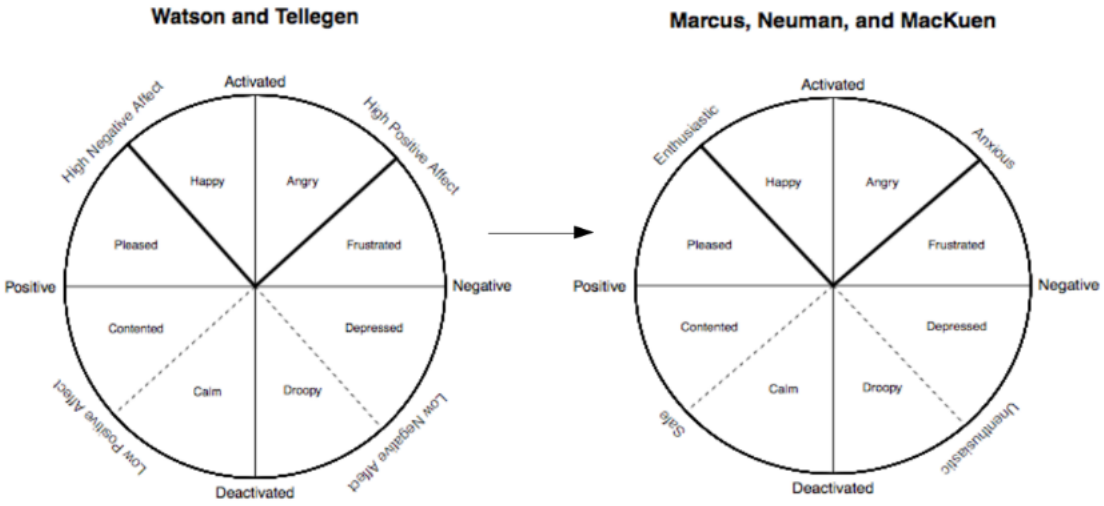
In Appendix B of *Affective Intelligence*, Marcus, Neuman and MacKuen (2000) suggest their anxiety and enthusiasm dimensions are related to Watson and Tellegen (1985)'s positive and negative affect, respectively:

> In psychology it has been common practice to label these two dimensions of emotion as "positive" and "negative." As we make clear in our exposition of the dual model

of emotional response in Chapter 3, we prefer to use the terms *enthusiasm* instead of *positive*, and *anxiety* instead of *negative*. In the present discussion of Watson's work, however, we carry over his terminology. In Tables B1-B3, two sets of labels are provided: the conventional psychological terms, *positive* and *negative*, and in parentheses the corresponding terms we have proposed, *enthusiasm* and *anxiety*. We think the latter pair of terms more precisely identifies the emotional content of the reactions associated with each factor (emphasis in original; 153).

As we show in Figure 3, positive and negative affect can be thought of as activated positive and negative emotional states (Barrett and Russell 1999; Watson et al. 1999; Yik, Russell and Barrett 1999). This is why many, including Watson and Tellegen themselves, refer to these two dimensions as positive and negative *activation* (see, Tellegen, Watson and Clark 1999). We draw on this perspective in our empirical application.

Figure 3: Two-Dimensional Perspectives of Emotion: Russell (Solid) versus Watson and Tellegen (Dashed)




*Note*: On the left we show Watson and Tellegen's model. On the right, we show Marcus, Neuman and MacKuen (2000)'s model. The bolded "v" is shows what section of Russell's and Watson and Tellegen's model refer to the anxiety and enthusiasm dimensions. Watson and Tellegen's work (see dashed lines) focuses on the diagonals of Russell's two-dimensional model (see vertical and horizontal solid lines).

## Theoretical Expectations

The theory of affective intelligence posits that when voters are anxious they search for information, but since this search is promoted by anxiety they tend to focus on negative information. This process has evolutionary value because it fosters active learning. From a political perspective, it leads to more thoughtful choices and ones that are less grounded in prior beliefs. This is particularly troublesome for incumbents, who would be the targets of retrospective appraisals. Under conditions of voter anxiety, incumbents are likely to be evaluated more harshly than challengers because incumbents are more often, and more plausibly, held responsible for the current state of affairs. Central to this interpretation is the assumption that the active learning that takes place under the Marcus, Neuman and MacKuen (2000) anxiety dimension most likely will be negatively valenced. Translating this expectation into Watson and Tellegen's model, *voters are less likely to vote for incumbents when they are in a high negative affective state*.

According to Marcus, Neuman and MacKuen (2000), "anxiety" is related to the surveillance system which assesses the current environment for novelty or potential threats. When activated, emotional responses help individuals efficiently respond to the perceived intrusion and return to equilibrium. The dispositional system monitors the habits or scripts which allow us to perform tasks without consciously considering them. When novelty or threats are not experienced, then individuals can safely rely on habitual responses to a variety of stimuli. This generates "enthusiasm," which reinforces voters' choices and lead them to rely more on heuristics like party identification. This bodes well for incumbents who are likely to be evaluated favorably when voters are enthusiastic, because this emotional condition should lead voters to accept the current state of affairs. Translating this expectation into Watson and Tellegen's model, *voters are more likely to vote for incumbents when they are in a high positive affective state*.

## Independent and Dependent Variables

Our dependent variable is a dummy variable capturing whether respondents voted for Barack Obama (1) or Mitt Romney (0) in the 2012 Presidential Election. Our main independent variable is the interaction between `Obama Valence` and `Obama Intensity`. The former was created by taking the number of things respondents liked about Barack Obama minus the number of things they disliked with positive values implying they had a more favorable opinion. Similar to the American National Election Studies (ANES), we only allowed respondents to list up to six likes and dislikes. `Obama Intensity` is derived from the measure of emotional intensity generated by the AESR system we developed for this study. Here, we took the average emotional intensity in audio files associated with things the respondent liked about Barack Obama minus the average emotional intensity in the audio files associated with things the respondent disliked. Positive values mean respondents spoke with more emotional intensity when describing what they liked about Barack Obama than when reporting what they disliked, and negative values mean the opposite. Looking to Figure 3, respondents have "High Negative Affect" towards Barack Obama when they list more dislikes than likes and score high on our audio-based measure of `Obama Intensity`. Respondents have "High Positive Affect" towards Barack Obama when they list more likes than dislikes and score high on our audio-based measure of `Obama Intensity`.

Because incumbents should be preferred when individuals are in a high positive affective state, the interaction between `Obama Valence` and `Obama Intensity` should be a positive and statistically significant predictor of whether respondents voted for Barack Obama. To help isolate this relationship, we also created a number of controls. `Democrat` is simply a dummy variable capturing whether the respondents identified themselves as being members of the Democratic Party. Dummy variables were also used for gender, race, and education. The `Female` variable records whether the respondents identified themselves as female. We also include a control for whether the respondent did (1) or did not (0) consider themselves Caucasian/White (`White`). A similar variable was created indicating whether the individual did (1) or did not (0) graduate college

(College Graduate). A seven-point Likert scale ranging from Very Strong Conservative (-3) to Very Strong Liberal (3) was used to capture respondents' ideology (Ideology). For Income, we created a similar scale which ranged from 0 to 4 with zero meaning respondents earned less than $25,000 and a four meaning they earned more than $100,000 a year. Finally, we include a continuous variable for respondents' age in years (Age).

## Results

We begin with data obtained from in-person interviews. These results are reported in Table 3. Here, we report simple logistic regressions predicting whether respondents voted for Barack Obama in the 2012 Presidential Election. Beginning with Model 1, which only includes our measure of emotional intensity generated from the audio files associated with descriptions of candidate likes/dislikes, we find that when individuals speak with more emotional intensity about things they like about Barack Obama as compared to dislike they are significantly more likely to vote for him ($p < 0.001$). This is consistent with Marcus, Neuman and MacKuen (2000)'s enthusiasm dimension, where high values should lead individuals to accept the status quo. A more direct test of our main hypothesis is found in Model 2. Here, we interact Obama Valence and Obama Intensity, an approach that more directly reflects both Marcus, Neuman and MacKuen (2000) and Watson and Tellegen's model. Ultimately, we find a similar result, but instead of emotional intensity having a direct effect on vote choice it now moderates the effect of valence.

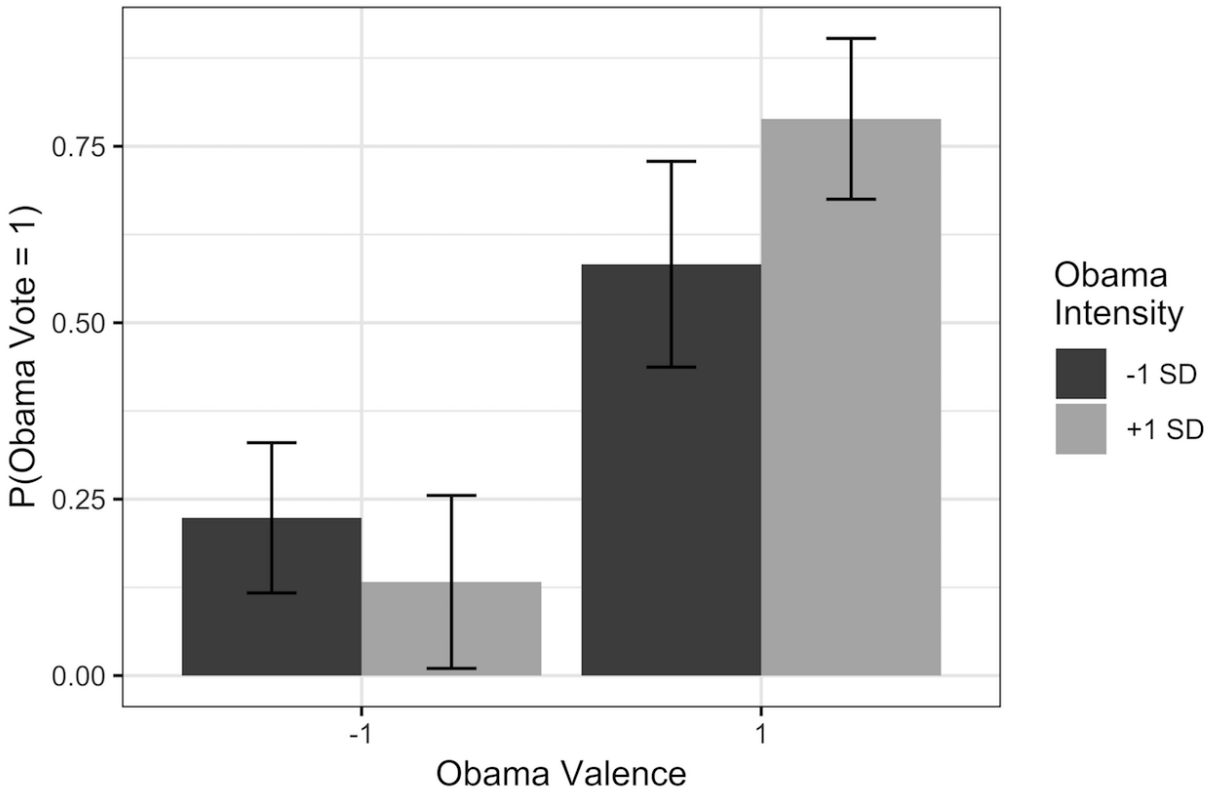
More specifically, Model 2 shows the interaction of Obama Valence and Obama Intensity is positive and statistically significant at the 0.001-level. The positive coefficient suggests as individuals speak with more emotional intensity about things they like about Barack Obama they are more likely to vote for him. This is again consistent with Marcus, Neuman and MacKuen (2000)'s enthusiasm dimension. Individuals who are generally pleased with the status quo are unlikely to actively think about incumbents which should them more likely to vote for them. This result holds even when controls are included for party identification, ideology, age, gender, race, education,

Table 3: Audio-Based Measures Obtained from In-Person Interview Responses Significantly Predict Vote Choice

| | *Dependent variable:* | | |
|---|---|---|---|
| | Voted for Obama | | |
| | (1) | (2) | (3) |
| Constant | $0.334^{**}$ | $-0.367^{*}$ | $-1.632$ |
| | (0.139) | (0.223) | (1.721) |
| Obama Intensity | $2.115^{***}$ | 0.226 | $-0.507$ |
| | (0.440) | (0.561) | (0.872) |
| Obama Valence | | $1.194^{***}$ | $0.832^{***}$ |
| | | (0.174) | (0.226) |
| Democrat | | | $2.668^{***}$ |
| | | | (0.691) |
| Ideology | | | $0.419^{**}$ |
| | | | (0.193) |
| Age | | | 0.036 |
| | | | (0.023) |
| Female | | | 0.031 |
| | | | (0.592) |
| White | | | $-1.347$ |
| | | | (1.207) |
| College Graduate | | | 0.512 |
| | | | (0.987) |
| Income | | | $-0.388^{*}$ |
| | | | (0.218) |
| Obama Intensity × Obama Valence | | $1.038^{***}$ | $1.040^{**}$ |
| | | (0.402) | (0.526) |
| Date Fixed Effects | | | ✓ |
| N | 241 | 241 | 241 |
| Log Lik | $-148.225$ | $-82.737$ | $-47.281$ |
| AIC | 300.451 | 173.474 | 124.561 |

*Note*: In all models, the dependent variable equals 1 when respondents voted for Barack Obama in the 2012 Presidential Election. These models report the results from simple logistic regressions. All variables are described on page 21. Checkmark (✓) indicates fixed effects. Levels of significance are reported as follows: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Standard errors are reported in parentheses.

Figure 4: Predicted Probabilities Showing How The Interaction Between Obama Valence and Intensity Influences Vote Choice (In-Person Interviews)




*Note*: This figure plots predicted probabilities using coefficients from Table 3, Model 2. On the $x$-axis, `Obama Valence` varies from -1 to 1 which suggests respondents have one more dislike and like, respectively. In the black and grey bars, we set `Obama Intensity` to $\pm 1$ standard deviation (0.39), respectively – meaning in the latter we assume the respondents speak with slightly more emotional intensity expressing their likes towards Barack Obama. Vertical lines represent 95-percent confidence intervals.

income and the when the interview took place.

Figure 4 plots probabilities using coefficients from Model 2. On the $x$-axis, we set `Obama Valence` to $-1$ and $1$ meaning on the far left we assume that respondents had one additional dislike of Barack Obama as compared to their likes. The reverse is true for the bars on the right, with the bars representing respondents who reported one more like than dislike. In the black and grey bars, we set `Obama Intensity` to $\pm 1$ standard deviation (0.39), respectively – meaning in the latter we assume the respondents speak with slightly more emotional intensity when expressing their likes towards Barack Obama. The vertical lines represent 95-percent confidence intervals.

Beginning with the bars at the far left, we find for respondents who list one more dislike about Barack Obama ($-1$) when they speak with slightly more emotional intensity ($+1$ SD), they are 9.09 percentage points *less* likely to vote for Barack Obama as compared to those who speak with slightly less emotional intensity ($-1$ SD). This suggests that emotional intensity – as measured only using the audio from responses – magnifies the relationship between candidate likes/dislikes and vote choice. We find a similar pattern in the bars on the far right. Here, we find for respondents who list one more like about Barack Obama ($+1$) when they speak with slightly more emotional intensity ($+1$ SD) they are 17.42 percentage points *more* likely to vote for Barack Obama as compared to those who speak with slightly less emotional intensity ($-1$ SD). This provides even more evidence that the way respondents speak when answering survey questions yields additional insights into their responses. The predictive power of expressed likes and dislikes varies as a systematic function of respondents' emotional states while voicing those likes and dislikes.

We now turn to the data we obtained from our telephone survey. These results are reported in Table 4. Similar to our previous results, we report simple logistic regressions in this table predicting whether respondents voted for Barack Obama in the 2012 Presidential Election. The only difference between Tables 3 and 4 is that the former includes daily fixed effects while the latter includes weekly fixed effects because the telephone survey was conducted over a three-week period.

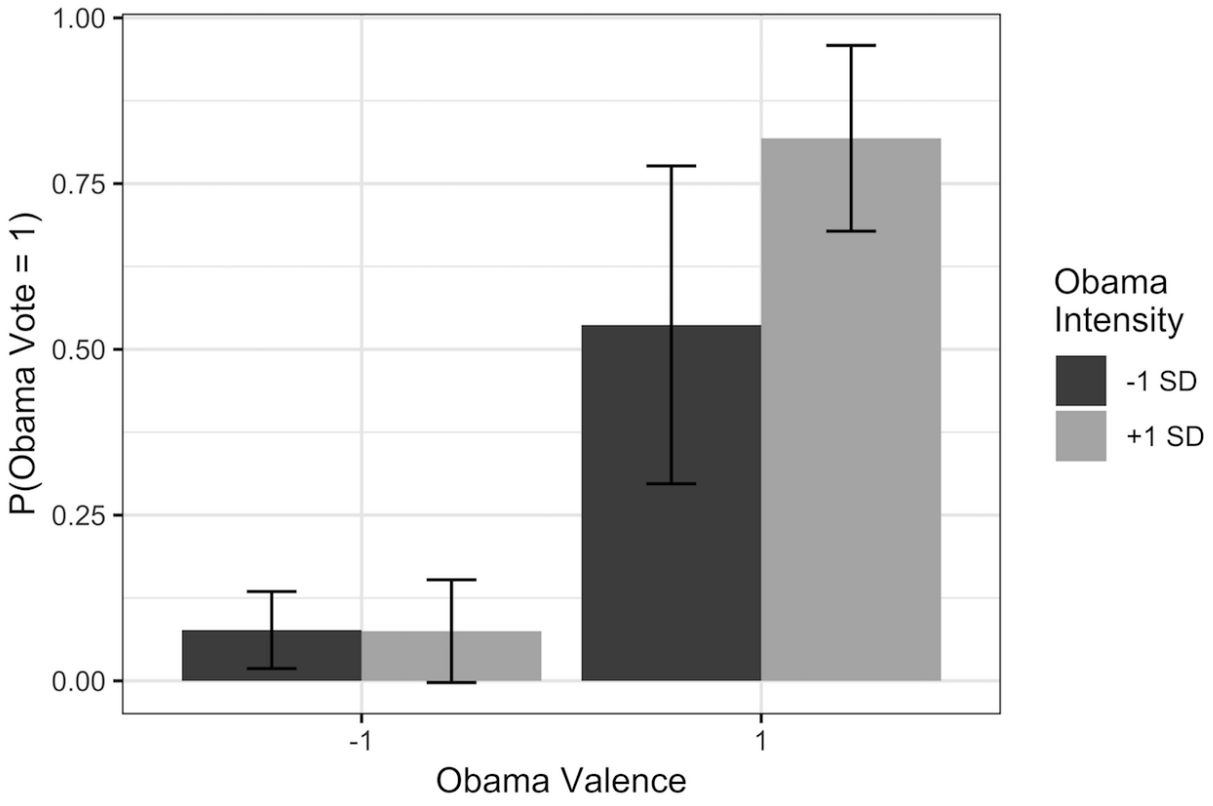
Beginning with Model 1, we again find that our measure of emotional intensity is a highly significant predictor of whether the respondent voted for Barack Obama ($p < 0.001$). As explained above, this is consistent with Marcus, Neuman and MacKuen (2000)'s enthusiasm dimension, but a more direct test is provided in Model 2. Although the coefficient associated with the interaction between `Obama Valence` and `Obama Intensity` is slightly smaller than what we found in Table 3, we still find that our audio-based measure significantly ($p < 0.05$) moderates the relationship between candidate likes/dislikes and respondent vote choice. Again, we find this result holds even when controls are included for party identification, ideology, age, gender, race, education, income and the when the interview took place. The coefficient associated with the interaction of

Table 4: Audio-Based Measures Obtained from Telephone Survey Responses Significantly Predict Vote Choice

| | *Dependent variable:* | | |
|---|---|---|---|
| | Voted for Obama | | |
| | (1) | (2) | (3) |
| Constant | $-0.666^{***}$ | $-0.837^{***}$ | $-1.709$ |
| | (0.157) | (0.237) | (1.942) |
| Obama Intensity | $2.599^{***}$ | 0.856 | $2.061^{*}$ |
| | (0.610) | (0.773) | (1.152) |
| Obama Valence | | $1.664^{***}$ | $1.135^{***}$ |
| | | (0.257) | (0.258) |
| Democrat | | | $2.013^{***}$ |
| | | | (0.606) |
| Ideology | | | $0.489^{***}$ |
| | | | (0.146) |
| Age | | | 0.011 |
| | | | (0.021) |
| Female | | | 0.220 |
| | | | (0.568) |
| White | | | $-0.173$ |
| | | | (1.468) |
| College Graduate | | | 1.026 |
| | | | (0.732) |
| Income | | | $-0.142$ |
| | | | (0.235) |
| Obama Intensity $\times$ | | $0.888^{**}$ | $1.053^{**}$ |
| Obama Valence | | (0.453) | (0.493) |
| Week Fixed Effects | | | $-0.946$ |
| | | | (0.734) |
| N | 206 | 206 | 206 |
| Log Lik | $-118.861$ | $-62.917$ | $-46.000$ |
| AIC | 241.722 | 133.834 | 118.000 |

*Note*: In all models, the dependent variable equals 1 when respondents voted for Barack Obama in the 2012 Presidential Election. These models report the results from simple logistic regressions. All variables are described on page 21. Checkmark (✓) indicates fixed effects. Levels of significance are reported as follows: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Standard errors are reported in parentheses.

Figure 5: Predicted Probabilities Showing How The Interaction Between Obama Valence and Intensity Influences Vote Choice (Telephone Survey)




*Note*: This figure plots predicted probabilities using coefficients from Table 4, Model 2. On the $x$-axis, `Obama Valence` varies from -1 to 1 which suggests respondents have one more dislike and like, respectively. In the black and grey bars we set `Obama Intensity` to $\pm 1$ standard deviation (0.39), respectively – meaning in the latter we assume the respondents speak with slightly more emotional intensity expressing their likes towards Barack Obama. Vertical lines represent 95-percent confidence intervals.

`Obama Valence` and `Obama Intensity` also becomes slightly larger when these controls are included while the main effect of `Obama Valence` becomes noticeably smaller.

Figure 5 plots probabilities using coefficients from Model 2 allowing `Obama Valence` and `Obama Intensity` to vary in the same way as Figure 4. The vertical lines represent 95-percent confidence intervals. Recall, in the predicted probabilities plot associated with our in-person interviews we found the moderating effect of `Obama Intensity` was higher when `Obama Valence` was set to $+1$ as compared $+1$. We again find this pattern in the telephone survey. Beginning with the bars at the far left of Figure 5, we find for respondents who list one more dislike

about Barack Obama ($-1$) when they speak with slightly more emotional intensity ($+1$ SD) they are 0.18 percentage points *less* likely to vote for Barack Obama as compared to those who speak with slightly less emotional intensity ($-1$ SD). In the bars on the far right of Figure 5, we find for respondents who list one more like about Barack Obama ($+1$) when they speak with slightly more emotional intensity ($+1$ SD) they are 18.19 percentage points *more* likely to vote for Barack Obama as compared to those who speak with slightly less emotional intensity ($-1$ SD). This provides some preliminary evidence our audio-based measure may be better able to capture Marcus, Neuman and MacKuen (2000)'s enthusiasm dimension as compared to their anxiety dimension.

Finally, in Table 5 we use data from both the in-person interviews and telephone surveys and find, not surprisingly, essentially the same results. Beginning with Model 1, we find that our measure of emotional intensity is a highly significant predictor of whether the respondent voted for Barack Obama ($p < 0.001$). In Model 2, the interaction between `Obama Valence` and `Obama Intensity` is statistically significant at the 0.001-level and this result holds when additional controls are included which is shown in Model 3. The coefficient associated with the interaction of `Obama Valence` and `Obama Intensity` again becomes slightly larger when these controls are included while the main effect of `Obama Valence` becomes noticeably smaller. These results provide consistent evidence that the audio associated with survey responses can yield additional insights into how respondents voted.

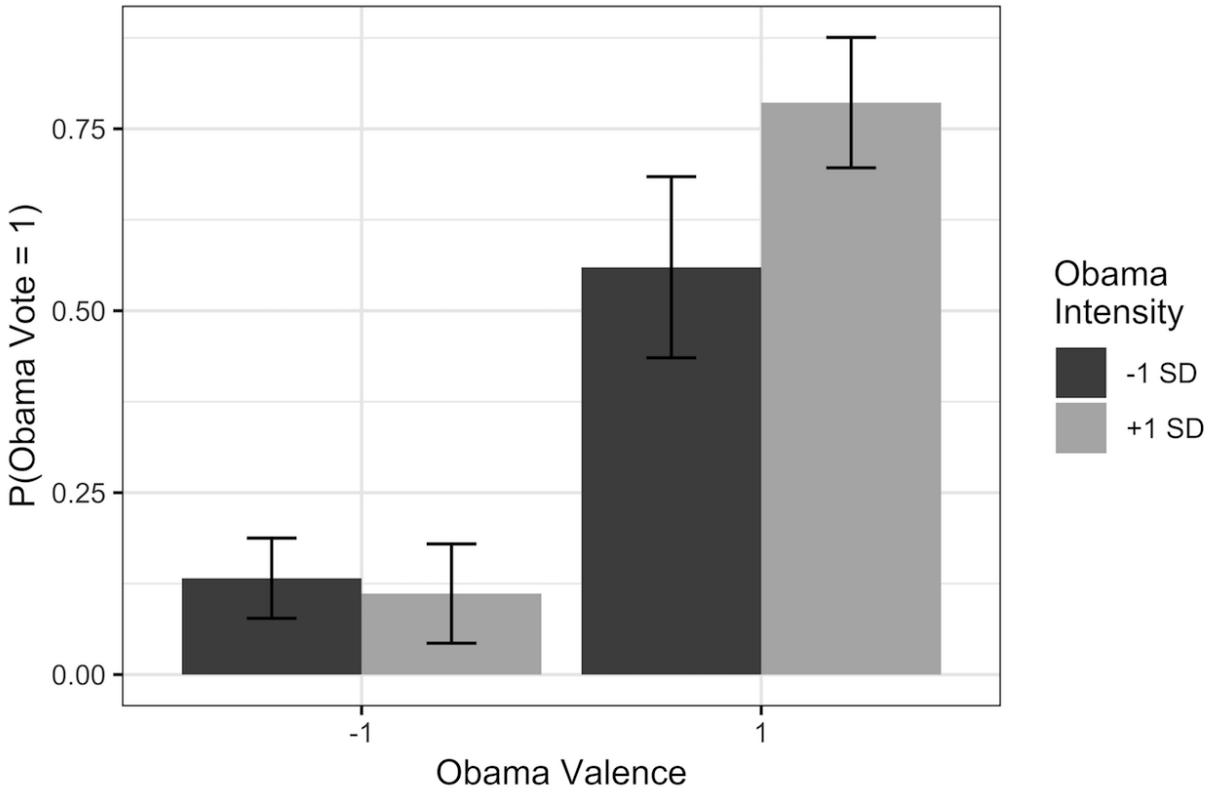
Figure 6 plots probabilities using coefficients from Model 2 allowing `Obama Valence` and `Obama Intensity` to vary in the same way as Figures 4 and 5. The vertical lines represent 95-percent confidence intervals. As before, we evidence our audio-based measure may be better able to capture Marcus, Neuman and MacKuen (2000)'s enthusiasm dimension as compared to their anxiety dimension. Beginning on the far left, we find for respondents who list one more dislike about Barack Obama ($-1$) when they speak with slightly more emotional intensity ($+1$ SD) they are 2.11 percentage points *less* likely to vote for Barack Obama as compared to those who speak with slightly less emotional intensity ($-1$ SD). On the far right, we find for respondents who list one more like about Barack Obama ($+1$) when they speak with slightly more emotional intensity

Table 5: Audio-Based Measures Obtained from Survey Responses Significantly Predict Vote Choice (Pooled Models)

| | *Dependent variable:* | | |
|---|---|---|---|
| | Voted for Obama | | |
| | (1) | (2) | (3) |
| Constant | −0.121 | −0.604*** | −1.389 |
| | (0.101) | (0.157) | (1.125) |
| Obama Intensity | 2.297*** | 0.555 | 0.610 |
| | (0.347) | (0.435) | (0.658) |
| Obama Valence | | 1.374*** | 0.963*** |
| | | (0.140) | (0.162) |
| Democrat | | | 2.177*** |
| | | | (0.413) |
| Ideology | | | 0.525*** |
| | | | (0.113) |
| Age | | | 0.019 |
| | | | (0.015) |
| Female | | | 0.114 |
| | | | (0.395) |
| White | | | −1.486* |
| | | | (0.812) |
| College Graduate | | | 0.988* |
| | | | (0.548) |
| Income | | | −0.267* |
| | | | (0.153) |
| Telephone Survey | | | −0.629 |
| | | | (0.648) |
| Obama Intensity × | | 0.809*** | 0.845*** |
| Obama Valence | | (0.231) | (0.277) |
| Date Fixed Effects | | | ✓ |
| N | 447 | 447 | 447 |
| Log Lik | −278.979 | −148.949 | −95.997 |
| AIC | 561.957 | 305.898 | 249.994 |

*Note*: All models report the results from simple logistic regressions. `Telephone Survey` indicates whether the respondent took the survey on the telephone. All other variables are described on page 21. Levels of significance are reported as follows: *p < 0.1; **p < 0.05; ***p < 0.01. Standard errors are reported in parentheses.

Figure 6: Predicted Probabilities Showing How The Interaction Between Obama Valence and Intensity Influences Vote Choice (Pooled Models)




*Note*: This figure plots predicted probabilities using coefficients from Table 5, Model 2. On the $x$-axis, `Obama Valence` varies from -1 to 1 which suggests respondents have one more dislike and like, respectively. In the black and grey bars we set `Obama Intensity` to $\pm 1$ standard deviation (0.39), respectively – meaning in the latter we assume the respondents speak with slightly more emotional intensity expressing their likes towards Barack Obama. Vertical lines represent 95-percent confidence intervals.

(+1 SD) they are 19.15 percentage points *more* likely to vote for Barack Obama as compared to those who speak with slightly less emotional intensity (−1 SD).

# Discussion and Conclusion

Despite the widespread use of telephone surveys for decades, the audio from these common data streams has received scant attention from political scientists. Yet, this data can yield important insights into the intensity with which respondents hold their opinions. The same can be said for in-person interviews which are frequently recorded.[7] However, to our knowledge no prior study has used the data from these recordings to gain insights regarding respondents' emotions. We develop the first Automatic Emotional Speech Recognition (AESR) system which can be used by future scholars to automatically measure the emotional intensity of respondents from both in-person interviews and telephone surveys. In doing so, we demonstrate the use of audio-as-data can extend beyond elite rhetoric and can actually be used to understand the mass public in real-time.

We argue emotions begin below conscious awareness, which is why scholars increasingly rely on physiological measures to measure them. However, physiological measures are very intrusive and cannot be used for large-N studies, especially those conducted over a telephone. With fMRI, respondents are encased in a confined horizontal space. When ERPs are used electrodes are placed at multiple locations on the respondent's scalp. Blink amplitude is measured with electrodes placed just below a person's eyes, and skin conductance is measured with sensors attached

[7]We recognize that, in bringing good news for what is perhaps a dying research mode, our findings with respect to the value of recording telephone surveys might be seen as too little, too late: very few telephone surveys in prior eras were recorded, and, relative to internet surveys, few telephone surveys may be conducted in the future. One intriguing possibility we would note is that future internet surveys could include at least some audio component. Especially for those survey organization such as YouGov that maintain their own respondent panels, it might be possible for respondents to be directed to turn on their computers' microphones and to record verbal answers to select items. The acoustic clarity of the resulting sound files would be a concern. For now, our point merely is that just because a survey is conducted via the internet does not mean it is impossible to obtain acoustic data.

to the respondent's fingers. Not only do these techniques require highly artificial settings, but all require expensive equipment and a laboratory to house it. Our AESR system can be used with any audio data obtained from in-person and telephone surveys which greatly reduces costs and make large scale data acquisition possible. Our central finding suggests that the emotional intensity of respondent answers – as derived from several audio measures – is a significant predictor of vote choice in the 2012 Presidential Election. We replicate this result for both in-person and telephone surveys. Taken together, these results in combination with the performance metrics associated with the machine learning models we estimated for this study serve as an important proof of concept. We show consistently that non-verbal content associated with survey responses provides information that cannot be captured by more overt measures traditionally used in political science. Our audio measure of emotional intensity, for example, cannot be explained by party identification or ideology. And, when additional controls are included in our models, the coefficients associated with our audio-based measure actually increase marginally which suggests the *way* respondents speak about their answers may yield important additional insights into their opinions.

By incorporating several audio variables into a common model, we also greatly expand the study of politically relevant audio beyond vocal pitch. Although previous scholars (e.g., Klofstad 2016; Dietrich, Enos and Sen 2018; Dietrich, Hayes and O'Brien 2018) have demonstrated the importance of this variable, there are countless other audio measures that could be applied to important political science questions. One of the main benefits of our AESR system is that it can be implemented by any scholar (eventually) using the accompanying source code. The (planned) plug-and-play system allows future scholars to simply upload their own audio files and receive emotional intensity scores with confidence ratings derived from the models we outline in this paper. Audio-as-data is a novel method, but we hope the (eventual) software we provide will help scholars learn more about their own audio files.[8]

Regardless of how future scholars use the tools created in this study, it is clear that respondents

[8]As this paragraph suggests, we are still in the process of putting the AESR system together, but we hope this gives you some insights into where we are heading with the project.

seem to deliver some response with more emotional intensity. Often, we think of survey responses as series of numbers, but current results show that the *way* those responses are delivered is also important. Voices carry meaning. The present study suggests that we can extract more politically relevant information from in-person interviews and telephone surveys which greatly expands the types of questions scholars can ask and add a new dimension to some of the variables that are traditionally used to study American political behavior.

# References


Amodio, David M, John T Jost, Sarah L Master and Cindy M Yee. 2007. "Neurocognitive correlates of liberalism and conservatism." *Nature neuroscience* 10(10):1246.

Ashworth, Scott. 2012. "Electoral accountability: recent theoretical and empirical work." *Annual Review of Political Science* 15:183–201.

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

Barrett, Lisa Feldman and James A. Russell. 1999. "The Structure of Current Affect: Controversies and Emerging Consensus." *Current Directions in Psychological Science* 8(1):10–14.

Bartels, Larry M. 2000. "Partisanship and voting behavior, 1952-1996." *American Journal of Political Science* pp. 35–50.

Bartels, Larry M. 2010. "The study of electoral behavior." *The Oxford handbook of American elections and political behavior* pp. 239–261.

Birnbaum, Dana W. and Bruce E. Chemelski. 1984. "Preschoolers' Inferences About Gender and Emotion: The Mediation of Emotionality Stereotypes." *Sex Roles* 10(7-8):505–511.

Birnbaum, Dana W., T. A. Nosanchuk and W. L. Croll. 1980. "Children's Stereotypes About Sex Differences in Emotionality." *Sex Roles* 6(3):435–443.

Bokhove, Christian and Christopher Downey. 2018. "Automated generation of ?good enough?transcripts as a first step to transcription of audio-recorded data." *Methodological Innovations* 11(2):2059799118790743.

Boudreau, Cheryl, Mathew D McCubbins and Seana Coulson. 2008. "Knowing when to trust others: An ERP study of decision making after receiving information from unknown people." *Social cognitive and affective neuroscience* 4(1):23–34.

Brody, Leslie R. 2006. "Gender Differences in Emotional Development: A Review of Theories and Research." *Journal of Personality* 53(2):102–149.

Campbell, Angus, Philip E Converse, Warren E Miller and Donald E Stokes. 1960. *The American voter*. University of Chicago Press.

Carmines, Edward G and James A Stimson. 1980. "The two faces of issue voting." *American Political Science Review* 74(1):78–91.

Cowie, Roddy, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz and John G Taylor. 2001. "Emotion recognition in human-computer interaction." *IEEE Signal processing magazine* 18(1):32–80.

Darwin, Charles. 1998. *The expression of the emotions in man and animals*. Oxford University Press, USA.

Dietrich, Bryce J. and Courtney L. Juelich. 2018. "When Presidential Candidates Voice Party Issues, Does Twitter Listen?" *Journal of Elections, Public Opinions, and Parties (Forthcoming)* pp. 1–41.

Dietrich, Bryce J, Matthew Hayes and Diana Z. O'Brien. 2018. "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech." Working Paper.

Dietrich, Bryce J., Ryan D. Enos and Maya Sen. 2018. "Emotional Arousal Predicts Voting on the U.S. Supreme Court." *Political Analysis (Forthcoming)* pp. 1–10.

Fairbanks, Grant and Wilbert Pronovost. 1939. "An experimental study of the pitch characteristics of the voice during the expression of emotion." *Communications Monographs* 6(1):87–104.

Ferejohn, John. 1986. "Incumbent performance and electoral control." *Public choice* 50(1):5–25.

Fiorina, Morris P. 1981. "Retrospective voting in American national elections.".

Funk, Carolyn L. 1999. "Bringing the candidate into models of candidate evaluation." *The Journal of Politics* 61(3):700–720.

Grabe, Maria Elizabeth and Erik Page Bucy. 2009. *Image bite politics: News and the visual framing of elections*. Oxford University Press.

Greene, Joshua D, R Brian Sommerville, Leigh E Nystrom, John M Darley and Jonathan D Cohen. 2001. "An fMRI investigation of emotional engagement in moral judgment." *Science* 293(5537):2105–2108.

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

Hayes, Danny. 2005. "Candidate qualities through a partisan lens: A theory of trait ownership." *American Journal of Political Science* 49(4):908–923.

Heilman, Kenneth M. 2000. Emotional Experience: A neurological Model. In *Cognitive Neuroscience of Emotion*, ed. Richard D. Lane and Lynn Nadel. New York, NY: Oxford University Press.

Heilman, Kenneth M., Robert T. Watson and Edward Valenstein. 2011. Neglect and Related Disorders. In *Clinical Neuropsychology*, ed. Kenneth M. Heilman and Edward Valenstein. New York, NY: Oxford University Press pp. 296–348.

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

Jones, Barbara E. 2003. "Arousal Systems." *Frontiers in Bioscience* 8:438–451.

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

Klofstad, Casey A. 2017. "Looks and Sounds Like a Winner: Perceptions of Competence in Candidates? Faces and Voices Influences Vote Choice." *Journal of Experimental Political Science* 4(3):229–240.

Klofstad, Casey A and Rindy C Anderson. 2018. "Voice pitch predicts electability, but does not signal leadership ability." *Evolution and human behavior* 39(3):349–354.

Klofstad, Casey A, Stephen Nowicki and Rindy C Anderson. 2016. "How voice pitch influences our choice of leaders." *American Scientist* 104(5):282.

Knoch, Daria, Alvaro Pascual-Leone, Kaspar Meyer, Valerie Treyer and Ernst Fehr. 2006. "Diminishing reciprocal fairness by disrupting the right prefrontal cortex." *science* 314(5800):829–832.

Knox, Dean and Christopher Lucas. 2017. "A General Approach to Classifying Mode of Speech: The Speaker-Affect Model for Audio Data." *Unpublished Manuscript* 23:31–33.

Koch, Michael and Ulrich Ebert. 1993. "Enhancement of the Acoustic Startle Response by Stimulation of an Excitatory Pathway from the Central Amygdala/Basal Nucleus of Meynert to the Pontine Reticular Formation." *Experimental Brain Research* 93(2):231–241.

Koenigs, Michael, Liane Young, Ralph Adolphs, Daniel Tranel, Fiery Cushman, Marc Hauser and Antonio Damasio. 2007. "Damage to the prefrontal cortex increases utilitarian moral judgements." *Nature* 446(7138):908.

Koo, Terry K and Mae Y Li. 2016. "A guideline of selecting and reporting intraclass correlation coefficients for reliability research." *Journal of chiropractic medicine* 15(2):155–163.

Lang, Peter J., Mark K. Greenwald, Margaret M. Bradley and Alfons O. Hamm. 1993. "Looking at Pictures: Affective, Facial, Visceral, and Behavioral Reactions." *Psychophysiology* 30(3):261–273.

Luo, Wei, Dinh Phung, Truyen Tran, Sunil Gupta, Santu Rana, Chandan Karmakar, Alistair Shilton, John Yearwood, Nevenka Dimitrova, Tu Bao Ho et al. 2016. "Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view." *Journal of medical Internet research* 18(12).

Marcus, George E. 2003. "The psychology of emotion and politics." *Oxford handbook of political psychology* pp. 182–221.

Marcus, George E. 2010. *Sentimental citizen: Emotion in democratic politics*. Penn State Press.

Marcus, George E, W Russell Neuman and Michael MacKuen. 2000. *Affective intelligence and political judgment*. University of Chicago Press.

Mauss, Iris B. and Michael D. Robinson. 2009. "Measures of Emotion: A Review." *Cognition and Emotion* 23(2):209–237.

Mutz, Diana C. 2007. Political psychology and choice. In *The Oxford Handbook of Political Science*.

Mutz, Diana C and Byron Reeves. 2005. "The new videomalaise: Effects of televised incivility on political trust." *American Political Science Review* 99(1):1–15.

Neuman, W. R., G. E. Marcus, A. N. Crigler and M. MacKuen. 2007. *The affect effect: Dynamics of emotion in political thinking and behavior*. University of Chicago Press.

Owren, Michael J and Jo-Anne Bachorowski. 2007. "Measuring emotion-related vocal acoustics." *Handbook of emotion elicitation and assessment* pp. 239–266.

Oxley, Douglas R, Kevin B Smith, John R Alford, Matthew V Hibbing, Jennifer L Miller, Mario Scalora, Peter K Hatemi and John R Hibbing. 2008. "Political attitudes vary with physiological traits." *science* 321(5896):1667–1670.

Paulhus, Delroy L. and Douglas B. Reid. 1991. "Enhancement and Denial in Social Desirability Responding." *Journal of Personality and Social Psychology* 60(2):307–317.

Rabinowitz, George and Stuart Elaine Macdonald. 1989. "A directional theory of issue voting." *American political science review* 83(1):93–121.

Robinson, Michael D. and Gerald L. Clore. 2002. "Episodic and Semantic Knowledge in Emotional Self-Report: Evidence for Two Judgement Processes." *Journal of Personality and Social Psychology* 83(1):198–215.

Rosen, Jeffery B., Janice M. Hitchcock, Catherine B. Sananes, Mindy J. D. Miserendino and Michael Davis. 1991. "A Direct Projection from the Central Nucleus of the Amygdala to the Acoustic Startle Pathway: Anterograde and Retrograde Tracing Studies." *Behavioral Neuroscience* 105(6):817–825.

Russell, James A. 1980. "A Circumplex Model of Affect." *Journal of Personality and Social Psychology* 39(6):1161.

Russell, James A. 2003. "Core Affect and the Psychological Construction of Emotion." *Psychological Review* 110:145–172.

Sanfey, Alan G, James K Rilling, Jessica A Aronson, Leigh E Nystrom and Jonathan D Cohen. 2003. "The neural basis of economic decision-making in the ultimatum game." *Science* 300(5626):1755–1758.

Scherer, Klaus R. 1986. "Vocal affect expression: A review and a model for future research." *Psychological bulletin* 99(2):143.

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

Simon-Thomas, Emiliana R, Dacher J Keltner, Disa Sauter, Lara Sinicropi-Yao and Anna Abramson. 2009. "The voice conveys specific emotions: evidence from vocal burst displays." *Emotion* 9(6):838.

Ślot, Krzysztof, Łukasz Bronakowski, Jaroslaw Cichosz and Hyongsuk Kim. 2009. "Application of Poincare-Mapping of Voiced-Speech Segments for Emotion Sensing." *Sensors* 9(12):9858–9872.

Spezio, Michael L and Ralph Adolphs. 2007. "Emotional processing and political judgment: Toward integrating political psychology and decision neuroscience." *The affect effect: Dynamics of emotion in political thinking and behavior* pp. 71–95.

Stewart, Patrick A, Bridget M Waller and James N Schubert. 2009. "Presidential speechmaking style: Emotional response to micro-expressions of facial affect." *Motivation and Emotion* 33(2):125.

Streeck, Jürgen. 2008. "Gesture in political communication: A case study of the democratic presidential candidates during the 2004 primary campaign." *Research on Language and Social Interaction* 41(2):154–186.

Sullivan, Denis G and Roger D Masters. 1988. ""Happy Warriors": Leaders' Facial Displays, Viewers' Emotions, and Political Support." *American Journal of Political Science* pp. 345–368.

Tellegen, Auke, David Watson and Lee Anna Clark. 1999. "On the Dimensional and Hierarchical Structure of Affect." *Psychological Science* 10(4):297–303.

Tigue, Cara C, Diana J Borak, Jillian JM O'Connor, Charles Schandl and David R Feinberg. 2012. "Voice pitch influences voting behavior." *Evolution and Human Behavior* 33(3):210–216.

Titze, Ingo R. 1989. "Physiologic and acoustic differences between male and female voices." *The Journal of the Acoustical Society of America* 85(4):1699–1707.

Watson, David and Auke Tellegen. 1985. "Toward a Consensual Structure of Mood." *Pscyhological Bulletin* 98(2):219–235.

Watson, David, David Wiese, Jatin Vaidya and Auke Tellegen. 1999. "The Two General Activation Systems of Affect: Structural Findings, Evolutionary Considerations, and Psychobiological Evidence." *Journal of Personality and Social Psychology* 76(5):820–838.

Welte, John W. and Marcia Russell. 1993. "Influence of Socially Desirable Responding in a Study of Stress and Substance Abuse." *Alcoholism: Clinical and Experimental Research* 17(4):758–761.

Westen, Drew, Pavel S Blagov, Keith Harenski, Clint Kilts and Stephan Hamann. 2006. "Neural bases of motivated reasoning: An fMRI study of emotional constraints on partisan political judgment in the 2004 US presidential election." *Journal of cognitive neuroscience* 18(11):1947–1958.

Xiao, Zhongzhe, Emmanuel Dellandrea, Weibei Dou and Liming Chen. 2010. "Multi-stage classification of emotional speech motivated by a dimensional emotion model." *Multimedia Tools and Applications* 46(1):119.

Yik, Michelle S. M., James A. Russell and Lisa Feldman Barrett. 1999. "Structure of Self-Reported Current Affect: Integration and Beyond." *Journal of Personality and Social Psychology* 77(3):600–619.

Zeng, Zhihong, Maja Pantic, Glenn I Roisman and Thomas S Huang. 2009. "A survey of affect recognition methods: Audio, visual, and spontaneous expressions." *IEEE transactions on pattern analysis and machine intelligence* 31(1):39–58.

Ziman, Kirsten, Andrew C Heusser, Paxton C Fitzpatrick, Campbell E Field and Jeremy R Manning. 2018. "Is automatic speech-to-text transcription ready for use in psychological experiments?" *Behavior research methods* pp. 1–9.