Multimodal Techniques for the Study of Affect in Political Videos

Ludovic Rheault[†] and Sophie Borwein[‡]

Abstract

An ever-increasing number of political events are being recorded and archived in video format, including real-time footage of parliamentary proceedings, candidate debates, media interviews and press conferences. While textual transcripts are often used in analysis, political scientists have recently considered methods for modeling audio and video signals directly. This paper looks at what can be achieved using machine learning models trained on each type of input—textual transcripts, audio, and video signals—for the automated recognition of emotion in political speeches. We draw on a newly-collected dataset containing 3,000 video recordings of sentence-long clips from speeches made by Canadian and American politicians, annotated for two primary dimensions of emotion and for anxiety, an emotion of substantive relevance to political science. We present three sets of findings. First, in line with previous work, we validate that methods using audio and visual signals improve upon the detection of emotion in political speeches when compared to methods that use only textual data. Second, we introduce a new approach to modeling audio signals that accounts for the unique characteristics of each politician's speaking style. Finally, we propose a simple procedure to preprocess video data and target the politicians of interest, using tools from the field of computer vision. Results from deep convolutional networks using images extracted with this procedure appear particularly promising for future research on emotion recognition.

[Preliminary version prepared for the 2019 PolMeth Conference, MIT, Cambridge, MA, July 18-20, 2019.]

[†]Assistant Professor, Department of Political Science and Munk School of Global Affairs and Public Policy, University of Toronto. Email: ludovic.rheault@utoronto.ca

[‡]PhD Candidate, Department of Political Science, University of Toronto. Email: sophie.borwein@mail.utoronto.ca

Introduction

The study of political speeches has too often been a missed opportunity. While political science experienced a recent surge in the development of research methods based on text documents, the textual record of political speeches represents only a tiny fraction of the data made available to the research community every day. To gain an appreciation of this information loss, consider President Obama's famous statement at the 2004 Democratic National Convention: "There's not a liberal America and a conservative America; there's the United States of America." Using a modern encoding standard, the transcript of this sentence represents about 96 bytes of data. The corresponding audio track makes 200 thousand bytes. Its video signal, without sound, around one million bytes.[1] The reach of textual analysis, roughly speaking, sums up to less than 0.01 percent of the total quantity of data encoded in a typical video. More importantly perhaps, in our example, the textual transcript would fail to translate many of the contextual clues—the tone, pace, and gestures—that made Obama's statement memorable.

In this paper, we examine the performance of three modes of inquiry—text-as-data, audio-as-data, and image-as-data—for the task of automated emotion recognition. For this purpose, we introduce a novel collection of political videos containing sentence-long utterances from more than 500 politicians in two different countries (Canada and the United States), recorded in various contexts, and labeled by human coders for their emotional content. Our coding scheme comprises two primary dimensions of affect, emotional arousal (or activation) and valence, as well as a specific emotion of theoretical importance for political research—anxiety. We rely on machine learning to compare the predictive power of each type of data input, and assess their potential for empirical research in the domain of politics.

Our study makes several contributions that may help inform future research in multimodal analysis. First, our results validate recent claims that audio and visual signals can enhance the automated analysis of political content, beyond what is feasible using textual transcripts (see e.g. Dietrich, Enos, and Sen 2019; Dietrich, Hayes, and O'Brien 2019; Knox and Lucas 2018; Hwang, Imai, and Tarr 2019; Torres 2018). We show that text-as-data is reliable for quantifying emotional valence (positive versus negative sentiment) in political speech, while audio signals are particularly useful for improving the levels of predictive accuracy in the modeling of activation and anxiety. Second, we introduce a model of speech emotion recognition that relies on path-breaking advances in voice synthesis to account for discrepancies in individual speaking styles. Our approach, which can be adapted to a variety of deep learning architectures, relies on abstract representations of a speaker's voice, which we refer to as *speaker embeddings*. These embeddings can be computed when making predictions for speakers not observed during the training stage, and offer improvements in accuracy. Third, we address some of the difficulties involved in the

[1] These calculations compare the text digitized in Unicode characters with the UFT-8 standard, an audio waveform in 16 bit encoding at a 16,000 sampling rate, and a video format with a 720 pixel height at 30 frames per second.

processing of visual data, an area of political methodology that is sure to grow in importance in the discipline. We build upon modern tools from the field of computer vision to extract face-centered images from our video collection, and show that deep convolutional networks based on these visual signals can match the predictive accuracy achieved with either text or audio data.

The following section begins with a survey of recent contributions from the field of political methodology, before moving to theoretical considerations in the study of emotion. We then introduce our corpus of videos, and present the modeling strategies that we considered for each of the three modalities under scrutiny. The penultimate section reports our empirical results, followed by a final discussion of the relative strengths of each modality.

The Automated Study of Emotion in Political Science

Our focus on emotion recognition in this paper is motivated by the current gulf between theoretical ambitions on the topic and the methods available to fulfill them. Studying the limits of rationality and the role of emotions in human behavior is one of political science's core research agendas, and there is now a vast literature linking emotion to disparate outcomes such as the formation of policy preferences, voting, election campaigning and messaging, responses to war and terrorism, and social movement organization (for a review, see Brader and Marcus 2013). Yet enthusiasm for this field of research is beset by the challenges involved in measuring emotions accurately. In spite of the spectacular advances in the field of artificial intelligence, there are still very few (if any) real-world applications able to detect and respond to our emotions (see Schuller 2018). The slow pace of progress is a testament to the difficulty of the task. As noted by Knox and Lucas (2018), the challenge is likely compounded when studying political elites, who often exercise considerable emotional control over their speech, as compared to everyday conversations.

The branch of methodological research that has witnessed the most steady progress in this area is textual analysis. Researchers have applied both dictionary-based and supervised learning methods for sentiment analysis to a wide array of text documents pertaining to politics, from politically-oriented social media tweets on the Twitter platform (Bollen, Mao, and Pepe 2011; Mohammad et al. 2015), to records of parliamentary debates dating back to the early 20th century (Rheault et al. 2016).[2] Most instances of political communication, however, are not generated in textual format; legislative proceedings, campaign debates, and other key speeches by political actors are generally first delivered in spoken form. Consequently, scholars applying sentiment analysis to political speeches are most often working with textual transcriptions of originally spoken word. As illustrated with the example utterance in our introduction, by studying only

[2]For more general overviews of the political literature in text-as-data, see Grimmer and Stewart (2013), Wilkerson and Casas (2017), and Benoit (2019).

the textual transcriptions of such events, a large proportion of the available data remains hidden from view.

Recognizing these limitations, a recent body of work has begun to explore how audio and/or visual data might contribute to the better modeling of affect in politics. Dietrich, Enos, and Sen (2019) and Knox and Lucas (2018) use novel techniques for audio data to study oral arguments before the U.S. Supreme Court, showing that acoustic signals provide information about Justices' emotional states and attitudes that is inaccessible using textual transcripts alone. Dietrich, Hayes, and O'Brien (2019) have recently extended the analysis of acoustic signals to U.S. Congressional floor debates, demonstrating how changes in vocal pitch relative to a speaker's base level can provide information about legislators' broader issue positions. But for a few studies, audio signal processing—now commonly used for emotion recognition in disciplines such as engineering and computer science (see El Ayadi, Kamel, and Karray 2011; Poria et al. 2018; Schuller 2018)—has been little integrated into the study of emotion in political science.

Political scientists have also been tentative in their adoption of automated methods using the visual signal to study affect. In this case, the sheer volume of data encoded in videos poses an additional burden in terms of computing time. Yet, as Torres (2018) points out, humans process information in the world first with their eyes, responding emotionally to visual stimuli before consciously processing what they see. Torres (2018) uses computer vision and image retrieval techniques drawn from computer science to show that conservative and liberal media outlets portray protest movements in visually disparate ways, with conservative outlets reporting protests as more dangerous (associated with darker and more nocturnal settings) than liberal outlets. Recent work by Hwang, Imai, and Tarr (2019) also applies computer vision techniques to political science, using these methods to automate the coding of politically-salient variables in campaign advertisement videos, an endeavor with promising implications for applied research.

The Representation of Emotions

This section provides a brief review of the two most widely adopted approaches for representing emotions in computational studies, namely Ekman's (1999) model of six basic emotions and the related facial action coding system (FACS) (Ekman and Friesen 1971; 1978), and Russell's (1980) circumplex model of affect.[3] We argue that the former suffers from shortcomings for applications in political research, even though the FACS itself provides a useful starting point for developing models based on visual data. Next, we identify the visual and acoustic traits that may help guide the process of feature engineering in automated speech recognition, and conclude with a discussion of theoretical studies on anxiety.

[3] An alternative choice is Plutchik's (1980) wheel of emotions, used for instance by Mohammad, Kiritchenko, and Zhu (2013), which we do not cover for simplicity.

Our natural starting point is the psychological literature on emotions. Psychologists have proposed a number of different approaches to classifying human emotion, with the most common being either categorical or dimensional (Cambria, Livingstone, and Hussain 2012). Categorical models of emotion are often underpinned by the idea that there are a series of "basic" or primary emotions that underpin human emotional life. Of these, Ekman's (1999) model of the six basic emotions—sadness, happiness, surprise, disgust, anger, and fear—is a common choice in many studies in speech emotion recognition. According to Ekman and Friesen (1971; 1978), these six emotions are universal to human beings, and can be found across cultures. More controversially, some scholars suggest that all other emotions emerge out of mixing these primary emotions (Cowie and Cornelius 2003). Ekman and contributors also pioneered the Facial Action Coding System (FACS), an exhaustive categorization of muscles movements occurring in various regions of the face, which can be used to map a wide range of reactions to stimuli extending beyond the six basic emotions (Ekman and Friesen 1978).

While the FACS appears indeed useful for guiding research based on visual data, the basic emotions categorization is problematic for political research applications. Simply put, we find that few of these six emotions are common in political videos. To demonstrate this point, approximately 1,400 video utterances from our datasets, which we discuss in greater detail in the next section, were annotated for basic emotions. The emotion most frequently observed by our human coders—anger—appeared in fewer than 14 per cent of these videos. Other basic emotions were virtually never observed (for example, surprise, sadness or disgust), such that we would require overly large collections of data before they could be considered in the development of machine learning applications. In contrast, more than one third of the speakers in our videos were coded as anxious, and even more were coded as emotionally aroused. Manifestations of emotions are indeed frequent in political speeches, but they do not seem to fit easily in the six basic categories. Our empirical analysis provides additional evidence to assess this specific point on measurement.

A second theoretical approach to emotion representation is dimensional, perhaps best exemplified by Russell's (1980) circumplex model. Dimensional approaches suggest that emotions can be mapped onto a finite number of dimensions. Although models with varying numbers of dimensions have been proposed, the two-dimensional representation—valence and activation—is often retained in computational studies (Fernandez 2004). Activation is a measure of emotional arousal, the amount of energy involved in expressing an emotion (El Ayadi, Kamel, and Karray 2011). Valence (often called sentiment) refers to the negative or positive orientation of the emotion being expressed. Discrete emotions can be placed along the continuum of these two dimensions. Figure A1 in this paper's Appendix reproduces Russell's two-dimensional model, and shows where some of the aforementioned basic emotions are located on each dimension. In this model, anxiety (the emotional response to stress) is located in the upper-left quadrant,

associated with emotional arousal and negative sentiment. This position is consistent with the empirical correlations observed in our dataset and a body of literature focusing on anxiety (see e.g. Gray and McNaughton 2000; Marcus, Neuman, and MacKuen 2000, Ch. 2).

Both the activation and valence dimensions of emotion have distinct manifestations in humans that allow us to form expectations for our three data modalities—textual, audio, and visual. Activation is often associated with measurable physiological changes that we expect audio recordings of speech to best capture. When humans are activated, the sympathetic nervous system becomes aroused. Heart rate and blood pressure increase, and respiration becomes deeper (Williams and Stevens 1972). Other physiological changes such as mouth dryness or muscle tremors may also be present. This triggers speech that is louder, faster, and "enunciated with strong high-frequency energy, a higher average pitch, and wider pitch range" (El Ayadi, Kamel, and Karray 2011, 573). Conversely, when de-activated, the parasympathetic nervous system dominates, which slows heart rate and blood pressure. The result is speech that is relented, lower in pitch, and with reduced high frequency energy (El Ayadi, Kamel, and Karray 2011).

While it may be well captured in acoustic signals, activation alone does not allow for differentiation among discrete emotions. For instance, happiness and anger are both associated with high arousal, and similar acoustic information (Sobin and Alpert 1999). The second "valence" dimension, which measures whether an emotion is positive or negative, is also important for classifying emotion. Unlike for activation, the positive/negative sentiment of a speech utterance is likely to be discernible from the linguistic message of the speaker. Consider, for example, that a textual transcript as short as President Obama's "yes we can" refrain provides the information needed to recognize that his message is positive. As discussed in previous sections, the ability of textual data to capture valence is demonstrated in the success that scholars have had in harnessing these methods to measure negative/positive sentiment in a variety of textual data relevant to politics.

Finally, there is a strong theoretical basis for using the visual signal contained in political videos. The human face is the primary visual cue in human interaction (Harrigan and O'Connell 1996), and there is now a well-established literature showing that human emotions can be identified through the muscular activity of the face. Brief changes in facial appearance, such as the raising of a brow, wrinkling of the nose, or puckering of the lips, combine to create the expression of different emotions (Ekman and Friesen 1978; Ekman and Rosenberg 2005). An emotion with a negative valence such as fear, for example, is associated with a raised and straightened eyebrow, open eyes, and dropped jaw (though the face may also be in a neutral position). For disgust, the upper lip will appear raised, changing the appearance of the nose; and the cheeks will be raised, causing lines below the eyes (Ekman and Friesen 1978). Even when choosing to rely on a dimensional representation of emotions, these visual features should prove particularly helpful to associate images with the target classes.

The Special Case of Anxiety

In addition to the two dimensions discussed in the previous section, our study also considers anxiety as a discrete emotion category. This emotion has arguably generated the most extensive literature in political science, due to its prevalence and implications for decision-making behavior. In particular, scholars have linked anxiety to a wide range of political phenomena, including terrorism and foreign policy decisions (Huddy, Feldman, and Weber 2007; Rheault 2016), negative attitudes toward immigration (Brader, Valentino, and Suhay 2008), partisanship defection in voting, and ethnic cleansing and genocide (Marcus, Neuman, and MacKuen 2000).

Anxiety is closely related to fear, yet the two expressions refer to different concepts. Freud (1920) was among the first to establish this distinction; whereas fear is a response to immediate and present danger, anxiety is a more generalized feeling of concern about "potential, signaled, or ambiguous threat" (Blanchard et al. 2008, 3). Therefore, in a literal sense, we should not expect politicians to experience fear, unless perhaps they are faced with a sudden and unexpected danger (for instance, if a natural disaster or an active gunman disrupts their speech). Although the terms are often conflated in common usage, more often than not, when we speak of fear in politics, we actually mean anxiety.

Anxiety in the human voice should be closely associated with emotional activation. As is true for emotional arousal, the sympathetic nervous system becomes activated when a person is anxious (Harrigan and O'Connell 1996). Studies examining the acoustic correlates of fear, anxiety's most closely related emotion, show that this emotion is characterized by an increased pitch, greater pitch variance, and a faster speech rate (Sobin and Alpert 1999; Banse and Scherer 1996). However, these acoustic characteristics are also found in other emotions related to activation. A study by Sobin and Alpert (1999), for example, shows that fear and anger share all but two acoustic features, which poses a challenge for multi-class problems in speech emotion recognition. The same authors show that, as compared to other basic emotions, acoustic indicators appear to be less important for human listeners in identifying fear in speech.

Facial traits should also provide useful reference points when modeling anxiety with visual signals. Harrigan and O'Connell (1996) examined the correlates between facial features (including the FACS discussed above) and human-coded annotations of anxiety in video recordings of experimental subjects. They report that facial expressions related to the partial (rather than full) expression of fear are visible on the faces of anxious speakers. These facial muscle movements include the horizontal pulling of the mouth, raising of the brows, and an increased rate of blinking. Meanwhile, facial expression associated with the more acute expressions of fear, such as wide eyes, and more pronounced brow raising, were not observable in the videotapes. These previous findings suggest that visual clues can be particularly helpful in differentiating anxiety, not just from other emotions, but from fear as well.

Video Collection

This study relies on three datasets of political videos annotated for their emotional content. The first dataset contains just under 1,000 utterances spoken by Members of Parliament during Question Period in the Canadian House of Commons (Cochrane et al. 2019). To build the corpus, researchers at the University of Toronto randomly selected ten time points from every third Question Period between January 2015 and December 2017. Video snippets were cut around each time point to include a full sentence, using punctuation marks from the official transcripts of the debates. Three independent coders watched and annotated the video clips for both valence and activation on 0-10 scales, with 10 being the most positive valence, and most activated.[4] For consistency with the rest of our video collection, we recoded valence and activation into binary 0-1 variables from the average coder score computed on the original 11-point scale. In this first dataset, approximately 33 percent of utterances were labeled as activated, while 67 percent were annotated as not-activated. For valence, 61 percent of videos were annotated as conveying negative sentiment, while 39 percent were coded as conveying positive sentiment.

The second dataset, collected in the spring of 2019, contains 1,000 video utterances selected at random from three sources: a collection of all Question Periods in the Canadian House of Commons between January 2015 and December 2017 (the original, full-length collection of videos described above); a collection of all floor proceedings from the United States House of Representatives between March 21st, 2017 and May 23rd; and, the 30 most watched hearings of the Senate Judiciary Committee posted by CSPAN between January 2015 and June 2019. We relied on a custom script to detect silences and randomly select segments between 5 and 20 seconds—the average duration of a sentence being 10 seconds. Videos were then individually inspected to discard segments with multiple speakers. Each video was labeled by three annotators for valence, activation, and anxiety on Amazon MTurk's crowdsourcing platform. To improve the quality of our annotations, we restricted the task to workers with a "Master's" badge, indicating a high rate of approval on the site. In this second dataset, about 35 percent of videos were labeled as activated, and 65 percent were coded as non-activated. 62.5 percent of videos were annotated as conveying a negative sentiment, while 37.5 percent were annotated as conveying a positive sentiment. Anxiety was present in 33 percent of videos.

We collected our third dataset in the summer of 2019. In contrast to the first two, the purpose of this dataset was to oversample speeches from specific politicians to further examine the consequences of speaker heterogeneity in speech emotion recognition tasks. As a result, we created this dataset by manually curating a list of recent speeches from ten high-profile political actors, five men and five women (see Table 3 for included speakers). This selection method is admittedly not as robust as random sampling (the availability of videos may be determined by media

[4]Three separate coders also annotated textual transcriptions of the data, but we do not draw on these annotations in this study. These videos were also labeled for the six basic emotions categories.

coverage choices), but it allowed us to accumulate extended examples from prominent speakers. Just over 1000 utterances—approximately 100 videos for each speaker—were compiled, selected from videos that were openly available in high resolution on the YouTube platform. Each video was once again annotated by three workers on the MTurk crowdsourcing platform. Consistent with the distributions observed in the previous two datasets, 36 percent of videos were labeled as activated. For valence, 60 percent of videos were coded as conveying negative sentiment, while 40 percent were coded as conveying positive sentiment. Annotators marked the speaker as anxious in 30 percent of the videos. In total, the three datasets comprise 2,982 video utterances with emotion annotations, and featuring 502 different political actors.

Methods

This section introduces the principal concepts and models used to process the three modalities at the core of this project: text-as-data, audio-as-data, and image-as-data. We use the term "modality" when referring to each communication channel, following the terminology used in other disciplines (see e.g. Baltrušaitis, Ahuja, and Morency 2018). We start by introducing the textual models that we use as a basis for empirical comparisons, although this discussion is deliberately brief given the breadth of literature already available on this modality. Next, we discuss the modeling of audio signals and describe the challenges involved with speech emotion recognition, in particular speaker heterogeneity. We then outline the methodology that we propose to address some of these difficulties. Finally, we repeat these steps for the visual modality.

Text Modality

Our empirical analysis starts by establishing a benchmark for the accuracy of emotion recognition models using textual input.[5] We fit machine learning classifiers on the text from each video utterance in our collection, using the three emotion labels described in the previous section. We rely on a methodology that had long represented the state of the art for text classification, namely long short-term memory (LSTM) recurrent neural networks fitted on sequences of tokens converted into numerical vectors using pre-trained word embeddings. The latest trend in the field is to rely on transfer learning using base models accounting for more complex linguistic structures (see, for example, Devlin et al. 2018; Howard and Ruder 2018). For simplicity, we report results only with the first approach.

Our classifiers rely on word embeddings computed using the GloVe algorithm (Pennington, Socher, and Manning 2014) to convert each word into numerical vectors of dimension $M =$

[5]We retrieved the textual content of videos from the last two of our datasets using the Google Cloud Speech-to-Text API. The model adapted for video recordings generates transcriptions of high fidelity, and its extensive dictionary helped to identify proper nouns, even challenging ones, such as the names of electoral districts and foreign politicians.

300. We use padding to normalize each utterance to T tokens. As a result, each input in the classifier is encoded as a $T \times M$ tensor. Given the short size of each utterance, we did not apply text preprocessing. Natural language being organized as sequences of words, recurrent neural networks (RNNs) represent a natural choice as they are capable of learning from current and previous elements in a sequence (Goodfellow, Bengio, and Courville 2016, Ch. 10). Specifically, we rely on LSTM networks, which include a "memory cell" that learns to either remember or forget prior contextual information associated with inputs as information moves through the network (see Greff et al. 2016). Our empirical section reports results from classifiers with an identical architecture for the three emotion labels, using two LSTM layers with 100 and 50 nodes followed by two dense layers.

Audio Modality

Digital audio signals are a data type that differs from the typical input formats used in social science research. As a result, we start with a brief introduction of three concepts: the speech signal, the waveform, and spectral analysis. Next, we address the question of speaker heterogeneity—the presence of repeated instances of the same speakers in a dataset—and propose a novel methodology to account for differences in the baseline voice patterns of each speaker. We conclude with a summary of the proposed model architecture for the audio modality.

A speech signal—like the sound of a voice in a natural environment—is produced through vibrations, or cyclical patterns occurring rapidly over time (see Rabiner and Schafer 2011, Ch. 1). Relevant for the following discussion is the fact that higher sounds are produced with faster vibrations, hence more cycles per seconds. The number of cycles per second is called the frequency and is measured in hertz (Hz). Thus, a politician with a higher pitch speaks at a higher base frequency. Louder sounds, on the other hand, produce larger waves for a given frequency. This is the amplitude, measured for instance in decibels. Because of the digital conversion of natural sound, the units of measurement for amplitude are not always preserved on a known scale of reference—volume can be adjusted, amplifying all waves, or mixed to amplify/attenuate sounds in specific ranges of frequencies (see Boersma 2014, 379). What matters for the purposes of signal processing is that higher amplitudes be measured with larger numbers, regardless of the scale.

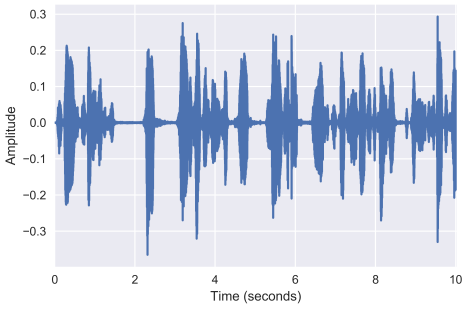
A central technology involved in sound recording converts the air pressure waves to an electric signal, called the waveform (Rabiner and Schafer 2011; Boersma 2014). The raw waveform is a unidimensional array of integers—the encoding in our set of videos uses a signed, 16-bit integers format—where each number represents the amplitude in discrete time. The sampling rate, also measured in hertz, is the frequency at which each integer is recorded.[6] With a sufficiently

[6]Modern videos typically have sampling rates of 44,100 or 48,000 Hz that can be recorded simultaneously on different channels (e.g. two in the case of stereo recordings). For our analyses, we convert the sample rate to 16,000 Hz and use a single channel.

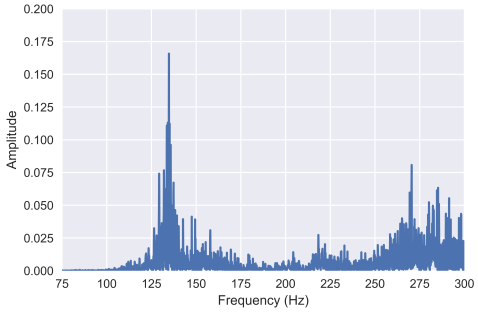
high sampling rate, the waveform can reproduce accurately the pulses of air pressure as they occur in the original speech signal. Figure 1a depicts the raw waveform of a speech from our dataset, a ten-second utterance from Barack Obama during the third 2012 presidential debate. Noticeable are the frequent pauses characteristic of Obama's speech pattern, during which the vibrations stop. Since the signal was converted to a 16KHz sampling rate, there are 16,000 data points every second; zooming in on parts of this figure would reveal elaborate patterns of pulses representing the various vocal sounds produced when uttering words.[7]

Figure 1: Digital Signal Processing of a Recorded Sentence from Barack Obama

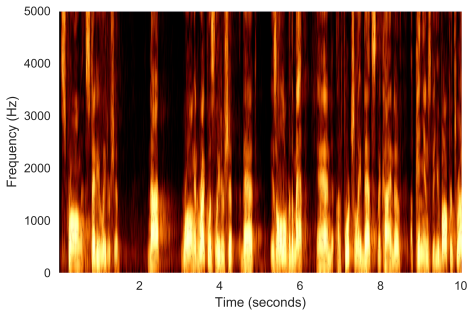
(a) Waveform



(b) Frequency Spectrum



(c) Spectrogram



The last building block in audio processing is spectral analysis, which includes various ways to express the waveform in the frequency domain. The aim is to represent sound amplitudes as a function of each frequency, usually with Fourier transforms (Boersma 2014). The resulting spectrum (or periodogram) helps us to understand which frequency ranges characterize a particular speech signal, for instance whether a speaker tends to produce sound at higher or lower frequencies. Figure 1b is the frequency spectrum of the aforementioned utterance from Barack Obama, computed with a Hamming window and displaying positive values in the 75 to 300Hz range. The first observable peak is the *fundamental frequency*, which corresponds to perceived pitch, and is followed by harmonics across the rest of the frequency range (Rabiner and Schafer 2011, Ch. 10). An additional transformation is the *spectrogram* (illustrated in Figure 1c for the

[7]The study of such patterns is the field of phonetics, a topic addressed in more depth elsewhere (see e.g. Laver 1994; Boersma 2014).

same example), which converts back the frequency spectrum in the time domain. The spectrogram depicts time on the x-axis, frequencies on the y-axis, and amplitudes using a color code. Our audio signal models rely on features derived from the waveform and from methods for spectral analysis. The most common audio feature for machine learning is the mel-frequency cepstral coefficients (MFCC), which are calculated by further transforming the audio spectrum (see, e.g. Zheng, Zhang, and Song 2001).

Dealing with Speaker Heterogeneity

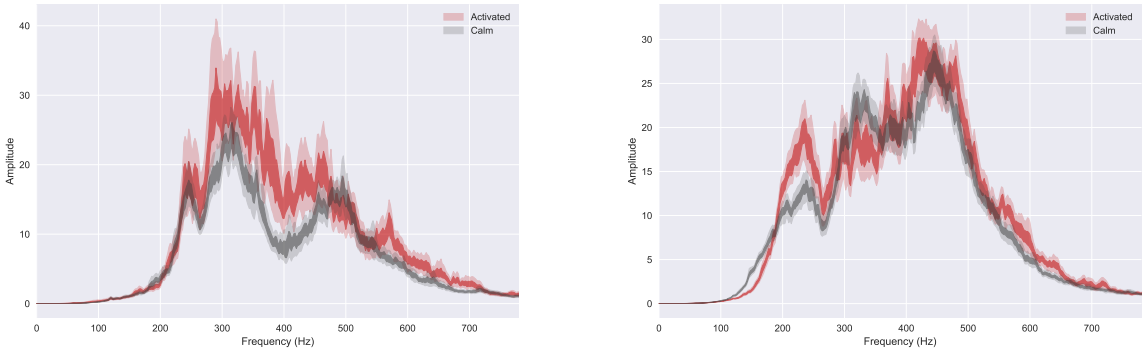
Acoustic signals from political speeches should contain many clues revealing emotional states, but their analysis causes complications due to the distinctiveness of each voice. This problem parallels the heterogeneity biases that may arise in inferential statistics when dealing with panel or hierarchical data. For example, without any correction, estimating changes in pitch as a measure of emotional activation makes sense for comparisons based on samples from the same speaker. However, every voice being unique, the pitch of an emotionally subdued person may fall just about anywhere in comparison to that of a second, emotionally aroused speaker. To address this problem, Dietrich, Hayes, and O'Brien (2019) relied on measurements of the baseline pitch level for each of the speakers in their dataset, before normalizing relative to the baseline. In effect, this achieves a similar goal as the within transformation in panel data analysis, where group-level means are subtracted from the variables. As far as we can tell, many applications of speech emotion recognition have simply ignored the presence of repeated examples from the same speakers in the training data.

We illustrate this problem by comparing the periodograms of female and male speakers in the first of our datasets. The plots in Figure 2 take advantage of the additive property of Fourier transforms and the randomized selection procedure to compute bootstrapped estimates of the average spectrum for subgroups of speakers. Such techniques are commonly used for spectral analysis in other disciplines (see, for example, Zoubir and Iskander 2004). The Figure illustrates average differences in frequencies between activated and calm speakers, broken down by gender. As can be observed, while the typical voice of politicians from either gender is transformed when emotionally activated, the voices of women and men also exhibit different patterns. For men, a range of low to mid-level frequencies have a higher amplitude when emotionally aroused, whereas the change tends to occur in higher frequencies for women. This suggests that emotions alter the timbre—or the range of harmonics—heterogeneously across groups of speakers. As a result, without knowing the gender of speakers a priori, predictive models are already at risk of confounding this attribute with the variation associated with emotions. Consider the extreme case of a sample in which speakers from a given gender are observed for only one emotional category; a classifier may learn to predict emotions based on the gendered attributes of a voice, and

predictions made on new data would reproduce the bias that was present in the training data.[8] Gender, however, is not the only source of ambiguity in audio data. Heterogeneity originates to a large extent at the level of individual speakers, and may be caused by other contextual factors at the time of recording.

Figure 2: Activated vs Subdued Speakers, by Gender (Frequency Domain)

(a) Women (b) Men



Mean frequency spectrum by groups of speakers for the audio samples in our first dataset from the Canadian House of Commons, computed using a bandpass filter to attenuate frequencies outside the 100-500 Hz range. The outermost error bands represent 95% bootstrap confidence intervals.

Our approach builds on methods developed in the field of automated speaker recognition (ASR), which is concerned with identifying the voice profiles of speakers. Early techniques in ASR include cepstral mean and variance normalization, which consists of centering the features extracted from each audio waveform (Viikki and Laurila 1998). Normalizing the acoustic features was found to improve predictive performance by eliminating noise associated with the recording conditions of the original speech signal. Applying a similar transformation at the speaker level before fitting a predictive model would be one possible way to deal with heterogeneity, and amounts to extending Dietrich et al.'s approach of normalizing the pitch of speakers to a broader range of mel-frequency cepstral coefficients. Modern approaches in ASR have led to alternatives such as i-vectors and x-vectors, embeddings representing the properties of a voice, which are commonly used in tasks such as speaker diarization (Snyder et al. 2018).

The method that we propose to deal with heterogeneity relies on the embeddings from a voice encoder, which we have found to generate surprisingly efficient representations of politicians' voices. These embeddings have been originally developed for speech synthesis, and correspond to the output layer of pre-trained models designed for speaker identification (the implementation we rely on is described in Wan et al. 2018; Jia et al. 2018).[9] This model has served as a

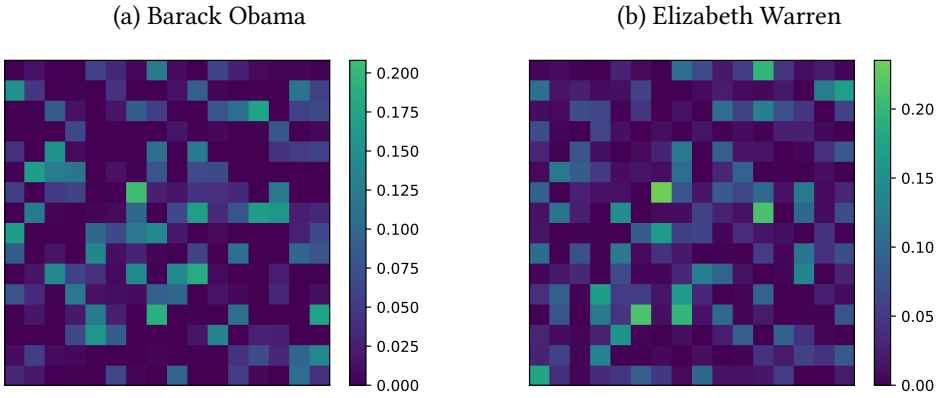
[8] A stream of research has emphasized the existence of similar biases caused by the choice of training samples (Caliskan, Bryson, and Narayanan 2017). In inferential statistics, we would also speak of an omitted variable bias in the previous example.

[9] The source code and pre-trained models we used are released in the following repository:

building block for text-to-speech applications such as Google's Tacotron (Jia et al. 2018). Our implementation relies on models trained using data from the LibriSpeech (Panayotov et al. 2015) and VoxCeleb (Chung, Nagrani, and Zisserman 2018) corpora, primarily designed for the development of speaker recognition machines.

Our approach consists of using the combined audio of all speeches available for each speaker in our dataset to compute what we call *speaker embeddings*, an abstract numerical representation of each speaker's voice.[10] We use the pre-trained weights of the voice encoder described above to generate embeddings of 256 numbers. Figure 3 plots these embeddings for two politicians using a two-dimensional heatmap to illustrate the methodology. The subtle differences in values reflect the idiosyncratic traits of each voice. Our goal is to augment predictive models of speech emotion recognition using these speaker embeddings, hence accounting for for the baseline voices of politicians. Again, a parallel can be made with models in inferential statistics that include unit-specific intercepts (e.g. fixed or random effects), with the difference that these speaker embeddings are continuous-valued and multidimensional.

Figure 3: Examples of Speaker Embeddings



Visualizations of two speaker embeddings using a heatmap. Each embedding is an abstract summary of the unique identifiers associated with a speaker's voice.

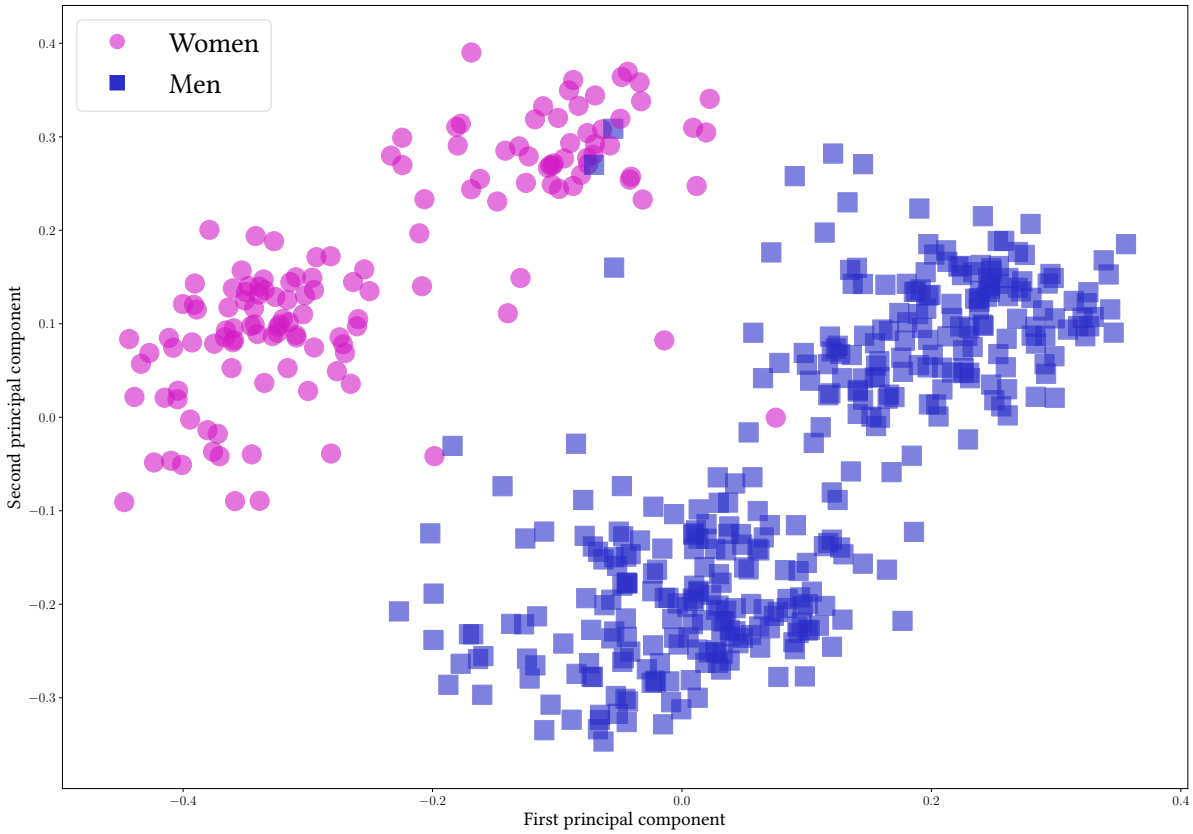
To assess the face validity of this approach, we plot the generated speaker embeddings on two dimensions using principal components analysis, and report the results in Figure 4. We use a color code to distinguish between male and female speakers. The projection reveals how these embeddings properly distinguish between female and male voices, each being clustered together in one region of the figure.[11] More generally, speakers with similar voices will be located closer together in the vector space.

https://github.com/CorentinJ/Real-Time-Voice-Cloning.

[10] We also verified that these speaker embeddings do not retain emotional characteristics per se, by computing them on a separate training set and making predictions on a holdout set. For simplicity, we report only of set of results in this study. In fact, the voice encoder requires only a relatively short sample of each speaker's voice to generate reliable embeddings.

[11] The female voices closer to the male cluster are deeper voices with more "masculine" traits, and vice-versa.

Figure 4: Projections of Speaker Embeddings in the Consolidated Dataset



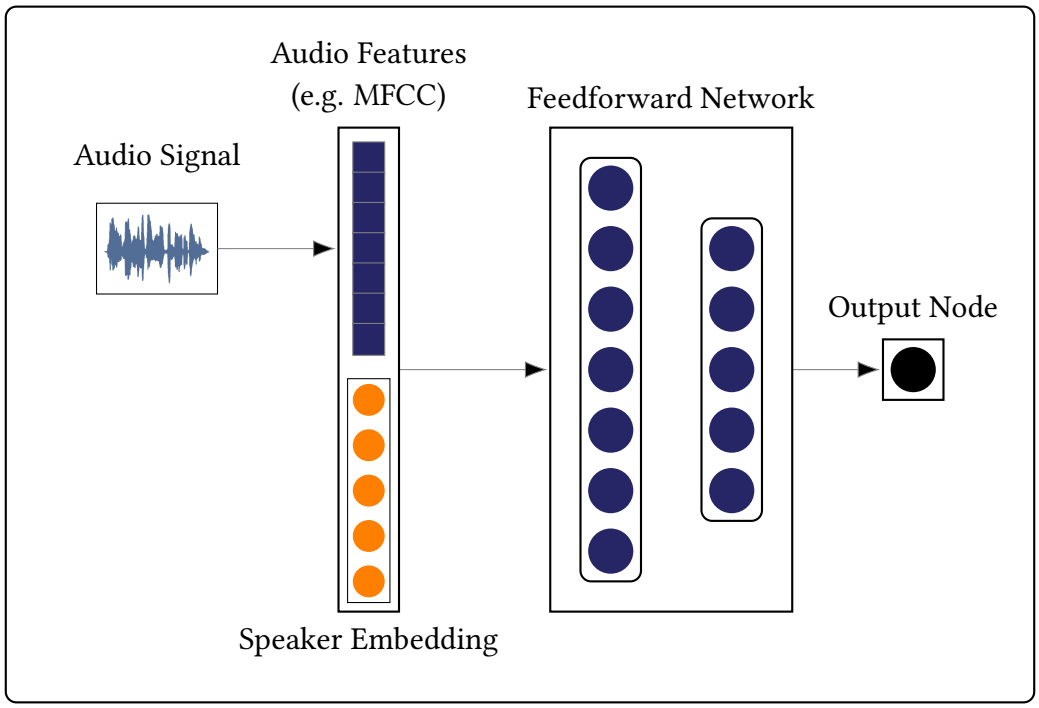
Distribution of all speaker embeddings in a two-dimensional space using principal components analysis. Voices from the same gender tend to resemble each other and are clustered together in the vector space.

The benefits of this approach are threefold. First, the embeddings can be generated for new examples, never seen before during the training stage of an emotion recognition application. Indeed, the voice encoder can be applied to any new speaker using a sample of their voice. An emotion recognition model trained with a sufficient number of examples can then rely on learned information about speakers with similar voices to predict the emotional state of new speakers more accurately. Second, the embeddings account not only for individual baselines, but also for other attributes that may affect predictive accuracy, for instance gender. Third, these speaker embeddings offer a lot of flexibility: they can perform the same purpose—speaker-level normalization—using other modalities since they are unique to each speaker, and be reshaped in different ways to match the input size of neural network layers.

Finally, we implement our methodology using a neural network architecture for speech emotion recognition in which acoustic features are augmented with these speaker embeddings. We depict this architecture in Figure 5. The first stage involves converting the waveform input of a speech into an array of features, which is then concatenated with the speaker embedding for the particular politician uttering that speech. Features are extracted from the waveform from each

audio file. For this paper, we rely on the full set of 68 mid-term audio features computed using the *pyAudioAnalysis* library (Giannakopoulos 2015) applied to non-overlapping windows of 0.25 seconds. The concatenated array of features constitutes the input of a multilayer perceptron, with the output layer being the emotional class that we aim to predict.

Figure 5: Emotion Recognition with Speaker Embedding



Schematic depiction of a deep neural network for speech emotion recognition, including speaker embeddings concatenated with the input features.

Visual Modality

This section introduces the visual models used to complete our analysis. We start with a discussion of computer vision toolkits available to preprocess images and locate the face of each politician in the video collection. We argue that this stage serves an important purpose by reducing extraneous elements from the frame. Next, we discuss the selection of a deep learning architecture appropriate for images, which have a different shape than the data types associated with the previous modalities.

Ignoring sound, a video is just a rapid sequence of images—or *frames*—chronologically appended together. Similar to digital waveforms, the visual signal is recorded at a variable sampling rate representing the number of frames captured per second. This rate is usually around 30 frames per second in modern standard formats found on the web. The dimensions of each frame, the resolution, may vary and affect the quality of the recording. Most of our videos, with the exception of the HouseLive collection, have a width of 1280 and a height of 720 pixels, a common

resolution for high definition videos. Thus, each frame in the video file is a multidimensional array (or tensor) of integer values with dimensions width $\times$ height $\times$ channels. The channels are the three basic colors, or put another way, each pixel is represented by a tuple of three integer values corresponding to red, green and blue.

Many open libraries from the field of computer vision facilitate the processing of video signals. Existing models have become spectacularly efficient for tasks such as face recognition and facial landmark detection, which offers a useful entry point for advancing research based on visual data in political science. In particular, we rely heavily on face recognition to process our videos. The external tools for this task depend on two core libraries in computer vision: OpenCV and *dlib*. The latter is a general-purpose library for deep learning, which contains the source code for our face recognition algorithm. Moreover, our empirical section briefly discusses facial landmark features extracted using Baltrušaitis et al.'s (2018) OpenFace C++ library, which is itself built on top of OpenCV and *dlib*.[12]

We choose to focus mainly on the face given the previously surveyed literature establishing a relationship between facial features and human emotions (Harrigan and O'Connell 1996; Ekman and Rosenberg 2005). Of course, there are several visual attributes of interest for speech emotion recognition, such as hand gestures and other, non-face related corporal motion. Based on observations made on our video collection, however, we argue that the face is a reliable starting point for modeling political videos. Depending on the context—close-up shots, speaker standing in front of a podium—the hands are often obstructed from view, whereas full body movements are also constrained for speeches made from a seated position, as is the case, for example, in Senate hearings. The face, on the other hand, is focused on systematically by camera workers.

Our main preprocessing step aims to identify and isolate the face of the speaker within the frames of each video, to reduce other sources of noise. We did this by first collecting external pictures of politicians in our dataset—in nearly all cases, using official portraits available on Canadian and US parliamentary websites. Each picture serves as the target for face recognition. We then processed each video using a face recognition algorithm from the *dlib* library,[13] to match faces present in each frame of our videos with the politician actually making the speech. The library works by measuring the similarity of the encoding for a face detected in the video with the encoding of the target image. This preprocessing step appears particularly important given that many of our videos include multiple faces per frame. Without efforts invested in this disambiguation, the accuracy of learning models can be affected by extraneous information present in the background of each video, which our coders were instructed not to consider. After a successful match with the target face, we crop the frame to include only the target face, and create a new

[12]The source code for all these models is released publicly, although OpenFace has restrictions for commercial usage.

[13]The algorithm was ported to Python and is available openly at: https://github.com/ageitgey/face_recognition. It achieves near perfect accuracy on benchmarks datasets for face recognition.

Figure 6: Preprocessing Pipeline and Model Architecture for Visual Modality

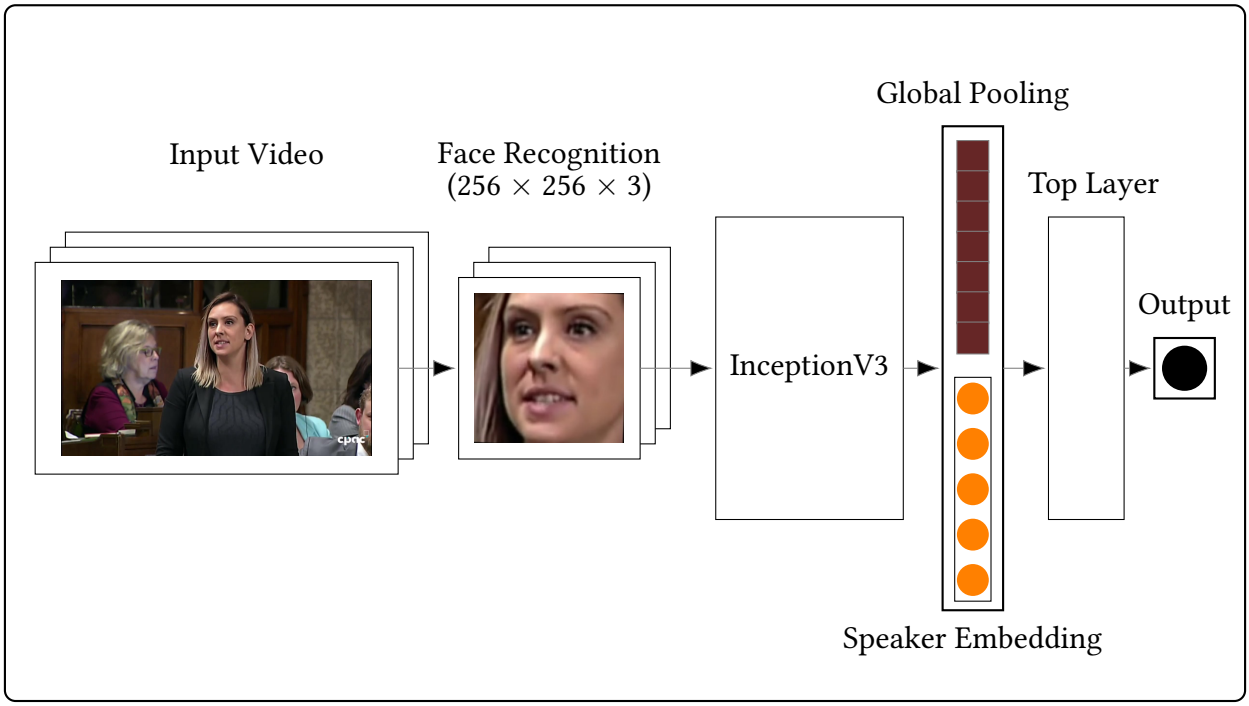


Illustration of the face recognition stage and model architecture, with an example video for our corpus featuring Canadian MP Ruth-Ellen Brosseau. Our main model uses the InceptionV3 ConvNet architecture (Szegedy et al. 2016), including a concatenation with speaker embeddings at the global pooling stage, and three additional dense layers.

video, now resized to a constant 256 x 256 x 3 format. We perform this operation every tenth frame, so that individual images can be used as input data while avoiding redundancies.[14] Figure 6 illustrates this preprocessing technique with an actual example from a video in our corpus.

Since images are three-dimensional, we require a model adapted to this data format. We rely on deep convolutional neural networks, specifically the modern Inception architecture proposed by Szegedy et al. (2016), widely used for image processing tasks. We also consider the methodology introduced in the previous subsection—correcting for heterogeneity in the corpus—in this case by concatenating the global pooling layer at the top of the Inception base module. We juxtapose three layers on the top of this architecture, plus an output layer using a sigmoid function to predict the binary emotion classes. As we did with the audio data, we will compare results achieved both with and without speaker embeddings. We use the embeddings discussed previously, based on voice recognition. Because they represent abstract numerical fingerprints, these embeddings can be used for a similar objective by integrating a baseline for each politician in the sample.[15]

[14] At a rate of 30 frames per second, any pair of consecutive frames tends to be very similar.

[15] We also computed visual speaker embeddings using the encodings from the above-mentioned face recognition algorithm. For simplicity, we discuss only one sets of results in the next section.

Finally, we rely on the the OpenFace toolkit mentioned earlier, as a validation probe to examine the association between facial features and emotions. The toolkit itself is based on deep learning models in computer vision, trained to recognize facial landmarks as well as action units from Ekman and Friesen's (1978) coding system. The features extracted comprise the pose, glaze direction, facial landmark location, and 18 action units from the FACS. Figure A2 in the appendix illustrates the output of this feature extraction algorithm applied to one of our videos. This software library appears particularly promising for applied research with image-as-data, although we should note that it requires proper calibration to deal with noisier videos in which multiple faces are present. Lower resolution videos also yielded problematic results. For this reason, our main models do not include the OpenFace features, and we use them as a complementary source of information to support our analysis in the next section.

Empirical Results

This section presents the key findings based on models fitted using each of the three data types, following the same order as in the previous section. Starting with the textual modality allows us to establish a frame of reference, although note that our main interest is in the improvements achievable with the other two modalities. For each type of input data, we report accuracy results for the main emotion classes: valence (or sentiment), activation, and the specific emotion of anxiety. Finally, we discuss the key findings and compare the potential of each modality for future research on political videos.

Text Models

The starting point for this paper was the observation that automated methods for emotion detection using textual data are effective at extracting negative/positive sentiment, but less capable of capturing other facets of emotion such as activation, and emotions related to activation. To highlight this point, we fit models predicting sentiment, activation, and anxiety using the textual transcriptions of the speech snippets in our datasets. The models reported here use word embeddings and recurrent neural networks, each with two long short-term memory layers (of length 100 and 50, respectively). Models are fitted with the Adam optimizer (with a learning rate of 0.001) over 15 epochs, such that the training corpus passes through the training examples 15 times. Regularization with dropout is used, with 1 in 5 neurons randomly removed from the network during training, to avoid overfitting the model.

Table 1 reports cross-validation metrics for textual models of sentiment, activation, and anxiety in our datasets. Since we are dealing with unbalanced classes, we report the percentage of observations in the modal category along with the accuracy score and the proportional reduction in error (PRE). Reported metrics are calculated on a holdout set representing 20 per cent

of the overall dataset. As anticipated, they show that textual models make reasonable predictions for sentiment—the models presented here accurately classified sentiment for 77 per cent of documents, reaching a reduction in error of approximately 39 percent. Textual models for activation and anxiety, meanwhile, perform poorly enough that guessing the modal category would actually yield a better accuracy rate. We tested a number of parameterizations, but the key substantive conclusion remains.

Table 1: Metrics for Textual Models

Emotion	Accuracy(%)	Modal Category(%)	PRE (%)
Valence	77.2	62.5	39.3
Activation	64.5	67.3	−8.7
Anxiety	66.1	68.4	−7.7

These results confirm a driving hypothesis of this paper—that textual data is useful for detecting negative/positive valence in text, but not the dimension of emotion associated with activation. This parallels recent findings in a study by Cochrane et al. (2019) that assessed the ability of human coders to detect emotional content—sentiment and activation—in the written versus video record of Canadian parliamentary debates. The authors show that human coders reliably agree on the sentiment of an utterance, whether they watched the video or read the textual record. In contrast, coders annotating activation from textual versus video records demonstrated little consistency with one another. As Cochrane et al. (2019, 2) conclude, "in short, the sentiment of the speech is in the transcript, but the emotional arousal is not."

Audio Models

We now turn to the audio signal to asses whether information gains are feasible relative to models based on text data. We begin by discussing associations between common audio signal features and our emotion categories. Next, we provide a detailed illustration of the problem of heterogeneity in speech recognition. We start by fitting predictive models using a single speaker at a time, and proceed by comparing these results with a pooled model, combining over 500 politicians. Finally, we report results from our main model enriched with speaker embeddings, and discuss the impact on predictive accuracy.

To begin, Table 2 reports average values of common acoustic features for each of the 2,982 speeches in our consolidated dataset, across the labels for activation and anxiety. We also report difference in means t-tests along with their p-values. The features include the energy (the normalized sum of squared values from the waveform); the pitch, or estimates of the fundamental frequency, using two different toolkits; the pitch standard deviation; and the speech rate, mea-

sured with the number of syllables over the speaking duration.[16] For both activation and anxiety, the differences are consistent with theoretical expectations. Activated and anxious politicians speak with a higher pitch, a larger pitch variance, and slightly faster than confident speakers (significantly so in the case of activation). Unsurprisingly, activated speakers produce more energy, meaning that they tend to generate a larger amplitude.

Table 2: Audio Features by Emotional Category

	Feature	Activated	Calm	t	p-value
Activation	Energy	0.020	0.018	2.918	0.004
	Pitch (Reaper)	189.527	154.971	21.038	<0.001
	Pitch Std. Dev. (Reaper)	54.229	41.889	15.461	<0.001
	Pitch (Praat)	202.117	165.16	19.277	<0.001
	Pitch Std. Dev. (Praat)	43.018	34.102	15.385	<0.001
	Speech Rate	3.935	3.833	3.587	<0.001

	Feature	Anxious	Non-Anxious	t	p-value
Anxiety	Energy	0.017	0.018	−2.182	0.029
	Pitch (Reaper)	174.705	154.854	9.290	<0.001
	Pitch Std. Dev. (Reaper)	44.492	35.099	11.986	<0.001
	Pitch (Praat)	186.088	160.628	11.823	<0.001
	Pitch Std. Dev. (Praat)	42.342	33.501	12.401	<0.001
	Speech Rate	3.784	3.750	1.070	0.285

Summary statistics and mean difference tests for a subset of audio features computed using the Parselmouth and pyAudioAnalysis libraries, as well as pitch estimates obtained with the Reaper algorithm. The dataset comprises 2,982 speeches annotated for activation, and 2,057 for anxiety.

Next, we show that deep neural networks (DNN) trained on individual speakers achieve a reasonable level of accuracy in the task of activation prediction (Table 3). Our models are relatively simple, and consist of four dense layers with ReLu activation functions, fitted with batches of 16 examples with an Adam optimizer to minimize the binary cross-entropy (with a learning rate of 0.001). We stopped the training after 50 epochs—that is, after 50 passes over the full set of training examples. Accuracy statistics are calculated based on a random holdout set representing 30% of the examples for each speaker. To fit the models, we rely on a data augmentation strategy that consists of using non-overlapping chunks (of 0.25 second each) from the original waveform

[16]For the voice pitch, we report estimates using both the default algorithm of the Praat software (Boersma 1993) and the Reaper model (https://github.com/google/REAPER). The rest of our analysis relies on two main libraries for digital signal processing in Python: Parselmouth (Jadoul, Thompson, and De Boer 2018), which is a Python wrapper for Praat, and pyAudioAnalysis (Giannakopoulos 2015).

of each input, each of which is associated with the original label observed at the utterance level. A similar approach has been used in learning tasks involving audio samples (see e.g. Hershey et al. 2017). Table 3 reports the percentage of observations in the modal category along with the accuracy score and the proportional reduction in error (PRE). Overall, when focusing on a single speaker at a time, our models reduce the error rate by about 30% compared to guessing the majority category.

Table 3: Accuracy Results for Speaker-Specific Models (Activation)

Politician	Accuracy (%)	Modal Category (%)	PRE (%)
Alexandria Ocasio Cortez	74.6	61.5	34.0
Barack Obama	72.9	61.2	30.3
Brett Kavanaugh	82.4	76.4	25.8
Christine Blasey Ford	76.8	65.2	33.2
Donald Trump	72.3	61.2	28.5
Elizabeth Warren	76.6	64.8	33.6
Hillary Clinton	81.8	77.6	18.6
Jody Wilson Raybould	79.3	68.8	33.6
Justin Trudeau	80.4	69.6	35.4
Mitt Romney	72.2	61.1	28.4

Pooling all speakers together in the same model architecture results in a large drop in predictive accuracy (see first rows of each panel of Table 4). We would normally expect the opposite, since the pooled model contains many more training examples, which should improve the performance of machine learning classifiers. Moreover, the model specification is identical to the one used for individual speakers. Yet, accuracy drops under 70% for activation, with a modest 15% reduction in error. This number is lower than the average of individual models reported in the previous table. This result helps to emphasize one of the contentions made earlier regarding the impact of speaker heterogeneity. Especially when models include multiple speakers, which will often be the case in research on political videos—in our case, we have 502 different voices, most of which appearing multiple times—accounting for heterogeneity appears desirable for training reliable classifiers.

Table 4 contrasts the previous result with models concatenating speaker embeddings, as described in the previous section. The rest of the model architecture is the same. For activation, the inclusion of a speaker baseline increases the accuracy above the level achieved with separate speakers. Model performance is also improved for anxiety, where the reduction in error now reaches approximately 47%. These results also suggest that, contrary to models based on textual data, audio signals perform better at detecting activated emotional states, as opposed to

valence. These two types of data inputs, therefore, appear to complement each other, and future applications in emotion recognition may stand to gain from the consideration of both modalities.

Table 4: Accuracy Results for Audio Models

Emotion	Model	Accuracy (%)	Modal Category (%)	PRE (%)
Valence	Pooled	61.5	60.1	3.6
	Speaker embeddings	73.4	60.1	34.3
Activation	Pooled	69.4	63.9	15.1
	Speaker embeddings	77.8	64.5	37.3
Anxiety	Pooled	71.1	67.6	10.7
	Speaker embeddings	80.3	62.7	47.2

The table reports accuracy statistics for models run with the three main binary classes measuring affect: sentiment (valence), activation, and anxiety. The models are deep, feedforward neural networks (multilayer perceptrons) computed with non-overlapping 0.25 second segments. The speaker embedding models use the exact same specification as the pooled ones, except for a concatenation with the pre-trained speaker embeddings at the input layer.

Visual Models

Finally, we explore the models based on visual data. For this analysis, we restrict our sample to the third dataset of videos described previously. This is done so that all videos analyzed are in high definition (1280 x 720 pixels or above), which improves the accuracy of the face recognition algorithm included in our processing pipeline. We observed that the algorithm's accuracy rapidly declines when using videos distributed with a lower resolution, for instance via the HouseLive website. Applying face recognition to these videos may be feasible, but would require a much more elaborate investigation. The third dataset still contains 1,057 videos, from which we extracted 32,566 frames with a positive match on the speaker, using a loop that cycled over every tenth frame (that is, limiting the number of images to three per second). These images constitute our data inputs in the analyses that follow.

We start by assessing the face validity of visual models by examining the prevalence of facial landmark features, broken down by the same three emotion categories used throughout this paper. Table 5 reports mean difference tests for features of theoretical relevance, based upon Ekman and Friesen's (1978) FACS. At least some of these statistical associations are consistent with expectations. For instance, the happiness facial action units (AUs) are more prevalent in the frames of videos coded as having a positive valence. Both the facial movements associated with anger and fear are related to activation, consistent with the dimensional representation of emotions. On the other hand, we find no significant relationship between fear action units and the emotion of anxiety. Like Harrigan and O'Connell (1996), however, we find that composites

of the action units associated with fear correlate with anxiety, when taken individually. Eye movements, in particular, appear promising for the detection of this emotion, as exemplified with a custom category combining eyebrow motions characterizing frowns.

Table 5: Facial Landmark Features by Emotion Category

	Feature	Positive	Negative	t	p-value
Valence	Frown (AU4 + AU9)	0.336	0.359	−7.438	<0.001
	Fear AUs	0.376	0.368	3.838	<0.001
	Anger AUs	0.394	0.394	0.133	0.894
	Happiness AUs	0.218	0.187	11.429	<0.001
	Feature	Activated	Calm	t	p-value
Activation	Frown (AU4 + AU9)	0.382	0.330	16.386	<0.001
	Fear AUs	0.396	0.356	17.432	<0.001
	Anger AUs	0.440	0.366	24.465	<0.001
	Happiness AUs	0.197	0.201	−1.211	0.226
	Feature	Anxious	Non-Anxious	t	p-value
Anxiety	Frown (AU4 + AU9)	0.366	0.341	7.590	<0.001
	Fear AUs	0.369	0.372	−1.223	0.221
	Anger AUs	0.405	0.388	5.242	<0.001
	Happiness AUs	0.194	0.202	−3.008	0.003

The table reports mean comparisons across the categories of the three emotion variables, using a total of 32,566 images from our third dataset, along with t scores for mean difference tests and their associated p-values. The features are extracted using the OpenFace toolkit based on cropped videos isolating the face of each speaker. Feature aggregations are computed by taking the average predicted score of facial units according to Ekman and Friesen's (1978) system, predicted by OpenFace on a five-point scale. The fear action units comprise AU1, AU2, AU4, AU5, AU7, AU20, and AU26; anger comprises AU4, AU5, AU7 and AU23; happiness comprises AU6 and AU12.

Next, we proceed by evaluating the performance of the deep convolutional architecture to predict emotion labels. As was done for the audio data, we compare pooled models that use only the images with models augmented with speaker embeddings. Table 6 summarizes our findings. Despite the limitations in sample size, our visual models outperform the results achieved using both the audio and the text approaches when valence is examined. The reduction in error is close to 50% after 25 epochs of training, using our model enriched with abstract representations of speaker—the speaker embeddings. In fact, the visual models achieve relatively high levels of accuracy for the three emotion categories under consideration. On the other hand, with the exception of valence, we do not discern sizable differences for models with and without the inclusion of speaker embeddings. This could be explained, however, by the subsample of data used in this section, which contains only ten speakers, hence comprises a lower source of hetero-

geneity.[17] Despite the limitations with the sample size, these results appear promising for future research involving visual data in political science.

Table 6: InceptionV3 CNN for Visual Models

Emotion	Model	Accuracy (%)	Modal Category (%)	PRE (%)
Valence	Pooled	71.6	60.7	27.7
	Speaker embeddings	80.1	60.7	49.4
Activation	Pooled	77.9	61.9	42.0
	Speaker embeddings	74.6	61.9	33.4
Anxiety	Pooled	81.4	64.2	48.0
	Speaker embeddings	78.1	64.2	38.8

Accuracy statistics for deep convolutional neural networks modeled on the InceptionV3 architecture, using a total of 32,566 individual frames from the videos, cropped to the face of each speaker (with dimensions 256 x 256 x 3). Each model is fit with 25 epochs with an RMSprop optimizer computing at a rate of 232 seconds per epoch on a GeForce RTX 2070 GPU. Accuracy statistics are calculated using a 30% validation split. The modal category is the percentage in the most frequent class for each emotion.

Conclusion

Detecting emotion in speech is fundamental to our understanding of political behavior, yet despite major advances in speech recognition and machine learning, we have arguably not made large advancements toward addressing this challenge. The root of this problem may be related to the difficulty of measuring emotions reliably, even by human judgement, a topic falling beyond the scope of this paper. The findings we report, however, provide clues as to what types of improvements are feasible. First, we demonstrate that voice signals show promise for the detection of activated emotional states, a finding that aligns with recent findings in the field (Knox and Lucas 2018; Dietrich, Enos, and Sen 2019; Dietrich, Hayes, and O'Brien 2019). This contrasts with the textual modality, which appears better tailored to valence recognition. Second, we approached the problem of speech emotion recognition by establishing a parallel with panel data analysis and hierarchical modeling. Our evidence supports the claim that predictive models are undermined by speaker heterogeneity. We proposed a model that expands on the idea of normalizing utterances at the speaker-level, by using speaker embeddings to augment traditional architectures in deep learning. Such a methodology leads to improvements in the modeling of audio signals, and can be applied to speakers not encountered at the training stage, making it suitable for applications across contexts. Finally, our preliminary findings suggest that visual

[17] The next version of this paper will rely on a complete sample, after ensuring that the videos can be preprocessed in a consistent manner regardless of variations in resolution.

signals may represent the most promising modality for emotion recognition. Deep neural networks performed well, relatively speaking, in predicting three different emotion labels. For this last modality, we proposed a preprocessing strategy that builds upon existing tools from the field of computer vision to detect the face of politicians in each video. This approach is consistent with theoretical considerations linking the face to emotions, and it has the benefit of removing the noisiness common to videos of political speeches.

References

Baltrušaitis, Tadas, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. "OpenFace 2.0: Facial Behavior Analysis Toolkit." In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66.

Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. "Multimodal Machine Learning: A Survey and Taxonomy." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2): 423–443.

Banse, Rainer, and Klaus R Scherer. 1996. "Acoustic Profiles in Vocal Emotion Expression." *Journal of Personality and Social Psychology* 70 (3): 614–636.

Benoit, Kenneth. 2019. "Text as Data: An Overview." In *The SAGE Handbook of Research Methods in Political Science and International Relations*, ed. Luigi Curini and Robert Franzese. London: SAGE Publishing.

Blanchard, Robert J., D. Caroline Blanchard, Guy Griebel, and David J. Nutt, eds. 2008. *Handbook of anxiety and fear*. Number v. 17 *in* "Handbook of behavioral neuroscience" Oxford: Academic Press.

Boersma, Paul. 1993. "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound." In *Proceedings of the Institute of Phonetic Sciences*. Vol. 17 pp. 97–110.

Boersma, Paul. 2014. "Acoustic Analysis." In *Research Methods in Linguistics*, ed. Robert J Podesva and Devyani Sharma. Cambridge: Cambridge University Press.

Bollen, Johan, Huina Mao, and Alberto Pepe. 2011. "Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena." In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. pp. 450–453.

Brader, Ted, and George E. Marcus. 2013. *Emotion and Political Psychology*. Oxford University Press.

Brader, Ted, Nicholas A. Valentino, and Elizabeth Suhay. 2008. "What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat." *American Journal of Political Science* 52 (4): 959–978.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science* 356 (6334): 183–186.

Cambria, Erik, Andrew Livingstone, and Amir Hussain. 2012. "The Hourglass of Emotions." In *Cognitive Behavioural Systems*, ed. Anna Esposito, Antonietta M. Esposito, Alessandro Vinciarelli, Rüdiger Hoffmann, and Vincent C. Müller. Lecture Notes in Computer Science Springer Berlin Heidelberg pp. 144–157.

Chung, Joon Son, Arsha Nagrani, and Andrew Zisserman. 2018. "VoxCeleb2: Deep Speaker Recognition." In *Interspeech 2018 Conference*.

Cochrane, Christopher, Ludovic Rheault, Tanya Whyte, Michael W.-C. Wong, and J.-F. Godbout. 2019. "Comparing human and machine learning classification of written and video records of parliamentary debates." 2019 Annual Conference of the Canadian Political Science Association. Vancouver, B.C.

Cowie, Roddy, and Randolph R. Cornelius. 2003. "Describing the emotional states that are expressed in speech." *Speech Communication* 40 (1-2): 5–32.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of NAACL-HLT 2019*.

Dietrich, Bryce J, Matthew Hayes, and Diana Z O'Brien. 2019. "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech." *American Political Science Review*. Forthcoming.

Dietrich, Bryce J., Ryan D. Enos, and Maya Sen. 2019. "Emotional Arousal Predicts Voting on the U.S. Supreme Court." *Political Analysis* 27 (2): 237–243.

Ekman, Paul. 1999. "Basic Emotions." In *Handbook of Cognition and Emotion*, ed. Tim Dalgleish and Mick Power. New York: John Wiley & Sons.

Ekman, Paul and Erika L. Rosenberg, eds. 2005. *What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. Series in affective science 2nd ed ed. Oxford: Oxford University Press.

Ekman, Paul, and Wallace Friesen. 1978. *Facial Action Coding System: The Manual*. Palo Alto, California: Consulting Psychologists Press.

Ekman, Paul, and Wallace V Friesen. 1971. "Constants across Cultures in the Face and Emotion." *Journal of Personality and Social Psychology* 17 (2): 124–129.

El Ayadi, Moataz, Mohamed S. Kamel, M.S., and Fakhri Karray. 2011b. "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases." *Pattern Recognition* 44 (3): 572–587.

Fernandez, Raul. 2004. "A Computational Model for the Automatic Recognition of Affect in Speech." PhD Dissertation, MIT Media Arts and Science, Feb. 2004.

Freud, Sigmund. 1920. *A General Introduction to Psychoanalysis*. New York: Boni and Liveright.

Giannakopoulos, Theodoros. 2015. "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis." *PLoS ONE* 10 (12): e0144610.

Gray, Jeffrey A., and Neil McNaughton. 2000. *The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System*. Oxford: Oxford University Press.

Greff, Klaus, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. "LSTM: A Search Space Odyssey." *IEEE Transactions on Neural Networks and Learning Systems* 28 (10): 2222–2232.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297.

Harrigan, Jinni A, and Dennis M O'Connell. 1996. "How Do You Look When Feeling Anxious? Facial Displays of Anxiety." *Personality and Individual Differences* 21 (2): 205–212.

Hershey, Shawn, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold et al. 2017. "CNN Architectures for Large-Scale Audio Classification." In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135.

Howard, Jeremy, and Sebastian Ruder. 2018. "Universal Language Model Fine-tuning for Text Classification." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Huddy, Leonie, Stanley Feldman, and Christopher Weber. 2007. "The Political Consequences of Perceived Threat and Felt Insecurity." *The Annals of the American Academy of Political and Social Science* 614 (1): 131–153.

Hwang, June, Kosuke Imai, and Alex Tarr. 2019. "Automated Coding of Political Campaign Advertisement Videos: An Empirical Validation Study". Working Paper. Department of Government, Harvard University.

Jadoul, Yannick, Bill Thompson, and Bart De Boer. 2018. "Introducing Parselmouth: A Python Interface to Praat." *Journal of Phonetics* 71: 1–15.

Jia, Ye, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. "Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis." In *Advances in Neural Information Processing Systems*. pp. 4480–4490.

Knox, Dean, and Christopher Lucas. 2018. "A Dynamic Model of Speech for the Social Sciences." *2018 Annual Conference of the Society for Political Methodology*, Provo, Utah.

Laver, John. 1994. *Principles of Phonetics*. Cambridge: Cambridge University Press.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville, *Deep Learning*. Cambridge: MIT Press.

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

Mohammad, Saif M., Svetlana Kiritchenko, and Xiaodan Zhu. 2013. "NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets." In *Proceedings of the Seventh International Workshop on Semantic Evaluation Exercises (SemEval-2013)*, . Atlanta, Georgia, USA.

Mohammad, Saif M., Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. "Sentiment, Emotion, Purpose, and Style in Electoral Tweets." *Inf. Process. Manage.* 51 (July): 480–499.

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books." In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE pp. 5206–5210.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "Glove: Global Vectors for Word Representation." In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Plutchik, Robert. 1980. "A General Psychoevolutionary Theory of Emotion." In *Emotion: Theory, Research and Experience. Vol. 1, Theories of Emotion*, ed. Robert Plutchik and Henry Kellerman. New York: Academic Press.

Poria, Soujanya, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain. 2018. "Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines." *IEEE Intelligent Systems* 33 (6): 17-25.

Rabiner, Lawrence R, and Ronald W Schafer. 2011. *Theory and Applications of Digital Speech Processing*. Vol. 64 Upper Saddle River NJ: Pearson.

Rheault, Ludovic. 2016. "Expressions of Anxiety in Political Texts." In *Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science*, Austin, Texas pp. 92–101.

Rheault, Ludovic, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis." *PLoS ONE* 11 (12).

Russell, James A. 1980. "A Circumplex Model of Affect." *Journal of Personality and Social Psychology* 39 (6): 1161–1178.

Schuller, Björn W. 2018b. "Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends." *Communications of the ACM* 61 (5): 90–99.

Snyder, David, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. "X-Vectors: Robust DNN Embeddings for Speaker Recognition." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE pp. 5329–5333.

Sobin, Christina, and Murray Alpert. 1999. "Emotion in Speech: The Acoustic Attributes of Fear, Anger, Sadness, and Joy." *Journal of Psycholinguistic Research* 28 (4): 347–365.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. "Rethinking the Inception Architecture for Computer Vision." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2818–2826.

Torres, Michelle. 2018. "Give Me the Full Picture: Using Computer Vision to Understand Visual Frames and Political Communication." Working Paper. Washington University in St. Louis.

Viikki, Olli, and Kari Laurila. 1998. "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition." *Speech Communication* 25 (1-3): 133–147.

Wan, Li, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. "Generalized End-to-End Loss for Speaker Verification." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE pp. 4879–4883.

Wilkerson, John, and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20: 529–544.

Williams, Carl E., and Kenneth N. Stevens. 1972. "Emotions and Speech: Some Acoustical Correlates." *The Journal of the Acoustical Society of America* 52 (October): 1238–1250.

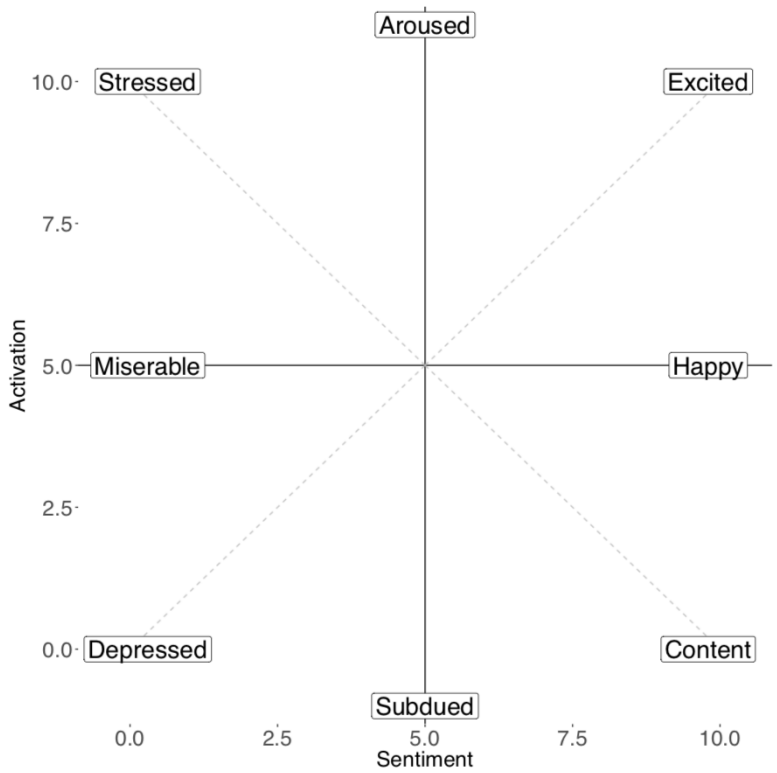
Zheng, Fang, Guoliang Zhang, and Zhanjiang Song. 2001. "Comparison of Different Implementations of MFCC." *Journal of Computer Science and Technology* 16 (6): 582–589.

Zoubir, Abdelhak M. and D. Robert Iskander. 2004. *Bootstrap Techniques for Signal Processing*. Cambridge: Cambridge University Press.

Appendix

Dimensional Model of Affect

Figure A1: Russell's Dimensional Model of Affect



James Russell. 1980. "A Circumplex Model of Affect." *Journal of Personality and Social Psychology* 39(6): 1161-1178.

Reproduction of Russell's original figure illustrating the dimensional model discussed in the theoretical section.

Figure A2: Facial Landmark Feature Extraction

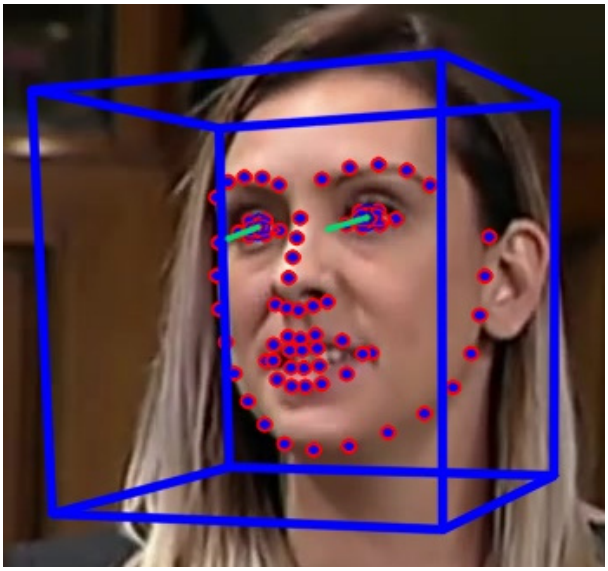


Illustration of facial landmark detection applied to a video from our dataset, using the OpenFace toolkit (Baltrušaitis, Ahuja, and Morency 2018).