Audio as Data

Ludovic Rheault
University of Toronto

Sophie Borwein
University of Toronto

Keywords: Audio as data, digital signal processing, hidden Markov models, pitch, speech analysis

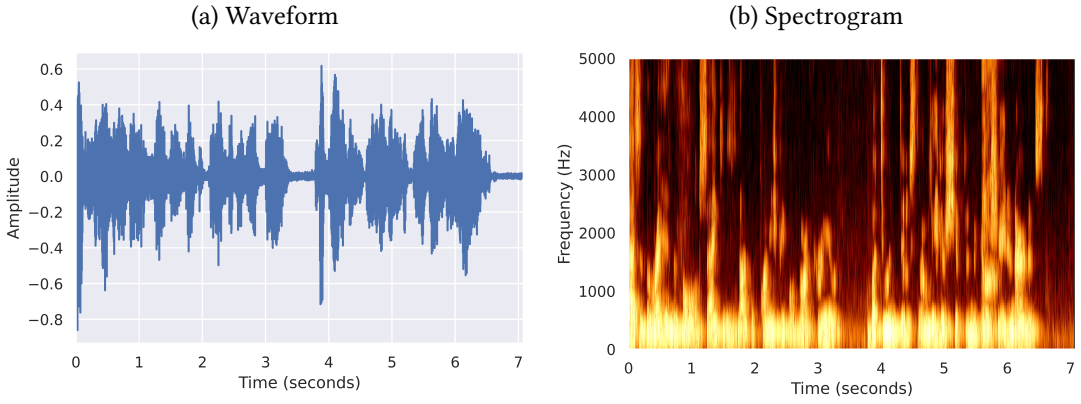
Introduction

Audio as data—the processing of raw audio recordings for empirical research—is an emergent field of political methodology that could spur significant advances in the study of human behavior. To put things in perspective, consider a thirty second speech made by a politician on the campaign trail. On average, the transcript of that speech will contain 75 words, amounting to 150 bytes of data. The audio recording of that same speech will contain approximately 1.3 million bytes of data, about 9,000 times more. The audio signal contains valuable clues about meaning, speaker attributes and context. Yet historically, and even today, researchers transcribing speeches or interviews into text incur a considerable loss of information about the phenomena they are studying. What if we could work with the full set of data contained in recorded speeches? Methods for audio as data aim to address precisely this problem.

Evidence suggests that the information lost by ignoring the audio is often relevant for political behavior. In a recent study, Cochrane et al. (2021) asked students to label political speeches. Some were given the transcript, while others watched the original video of the speeches. The authors find that transcripts are efficient at revealing one dimension of emotion, valence, but not other attributes such as the level of emotional arousal. This finding underscores the idea that human communication involves more than just the linguistic message. Audio features such as pitch, loudness, timing, voice quality and articulation all carry information about the underlying emotional state and disposition of a speaker (El Ayadi, Kamel, and Karray 2011). Sound recordings can also help to reveal subtle behaviors that would be difficult to detect without context, such as sarcasm, or the use of deception (Windsor et al. 2019). Depending on what theoretical constructs researchers are interested in, the choice of speech modality matters.

This entry provides a brief introduction to the field of audio as data. The next section emphasizes how researchers interested in studying audio recordings can build on established knowledge from the fields of digital signal processing and time series analysis. We then examine the studies that pioneered

Figure 1: Visual Representations of an Audio Recording of President Joe Biden



audio as data methods for political research in recent years, before outlining promising avenues for future research.

Modeling Sound as Data

For social scientists looking to work with audio recordings, a key challenge is gaining familiarity with a new data format. Simply put, a digitized audio file consists of a time series containing many numerical values per second. The number of data points recorded per second is called the frequency, measured in Hertz (Hz). Plotted as a time series, these data points constitute the waveform, a graph whose shape will be familiar to most (Figure 1a). Where audio as data differs from econometric time series is in the speed and consistency of oscillations, representing vibrations, or variations in air pressure created when producing sound (Knox and Lucas 2021). Faster (slower) oscillations produce a sound perceived as higher (lower) pitched. For any subset of the waveform, one can estimate the speed of these oscillations—also called frequency and also measured in Hz, which is a general metric for the number of cycles per second. The y-axis of the waveform, on the other hand, represents the amplitude: cycles reaching higher absolute values correspond to louder sounds, and vice-versa. The micro-patterns of sound oscillations are a formidable source of information about the human voice. Indeed, the data contained in the waveform are precise enough to characterize the timbre of a voice and detect the sounds composing specific words, in various languages. So much so that the existing state-of-the-art allows researchers to uniquely identify speakers from a short audio sample, detect a language, or automatically translate speech to text with high levels of accuracy (see Proksch, Wratil, and Wäckerle 2019).

The learning curve to start modeling audio data can be steep, and requires basic knowledge of a field called digital signal processing. Researchers would benefit from consulting full length introductory texts if they are to maximize the potential of their analysis (for instance, Rabiner and Schafer 2011),

and identify the most relevant quantities of interest. Pitch estimation—an attempt to measure the fundamental frequency of a voice—could be one of these quantities, as we review below. Another type of transformation is ubiquitous in audio analysis, and consists of representing the original waveform in the frequency domain (spectral density estimation), to better highlight the rates of oscillations. While the frequency domain is not as intuitive as a time series, a spectrogram can be used to visualize the frequencies observed in small subsections of the audio recording, called samples or frames, and to plot that information back in the original time domain. For instance, Figure 1b is the spectrogram for a recording of President Joe Biden. The lighter shades indicate that frequencies within the 0-500Hz range are prevalent when Biden speaks—the pitch of his voice would be the estimate of the dominant frequency within that range. Faster oscillations, at high frequency, are also visible. They represent the harmonics produced by the human voice. Yet another useful transformation consists of representing the data on the mel scale, which was designed to correspond more closely with changes in pitch as they are perceived by the human ear. Classically, mel frequency cepstral coefficients (MFCCs), derived from the mel scale, have been a go-to choice for input variables in audio modeling tasks (see Zheng, Zhang, and Song 2001).

We may classify modern modeling approaches into two principal families: hidden Markov models (HMM) and deep neural networks (see Hinton et al. 2012). HMMs are designed to model the data generation process of signals occurring over time, and have long represented a natural choice for tasks such as speech recognition. The second family of models, deep neural networks, has gained in popularity in all spheres of modern data analysis. Deep learning models can predict various attributes—the tone of a speech, the gender of a speaker, the word being uttered—by training a neural network where the researcher provides labeled measurements of the target attributes, and where the numerical features from the input audio files serve as predictive variables. A recent trend in deep learning is to avoid the transformations discussed above—such as MFCCs—and to use instead the raw numerical values from the original waveform as input data (Trigeorgis et al. 2016).

We should emphasize that progress in the field of audio processing has been spectacular in recent decades. As a result, political scientists can supplement their analysis of audio data using the output from pre-trained models, without having to reinvent the wheel every time. Automated speech-to-text transcription, language detection and speaker diarization (the identification of distinct speakers in a recording) are examples of tasks for which models are readily available. Major cloud services providers sell access to quality models at low cost, and researchers with programming skills can find open source equivalents from the academic community that replicate a similar level of performance.[1]

Applications to Political Research

A small but growing literature involves the modeling of audio signals to study political phenomena. Exploiting ready-made software for speech analysis, initial research in this area relied on audio data to understand how voter preferences are influenced by aspects of politicians' speech such as vocal pitch and pronunciation (Gregory and Gallagher 2002; Klofstad 2016; Klofstad and Anderson 2018; Klofstad, Anderson, and Peters 2012; Podesva et al. 2015; Tigue et al. 2012). A consistent finding in this stream of research, which relies extensively on experiments, is that voters prefer political candidates with lower-pitched voices, although voters do not find policy appeals from these candidates to be more persuasive (Klofstad and Anderson 2018).

More recently, political scientists have deployed audio as data methods in research studying the role of affect. These studies often focus on using vocal pitch to measure how emotional activation relates to political behavior, both among voters and political elites. Dietrich, Enos, and Sen (2019) use speech analysis software to measure Supreme Court Justices' vocal pitch in audio recordings of 3,000 hours of oral arguments before the Court, and show that judges' emotional arousal can predict case outcomes as well as models that use more traditional measures such as ideology and legal issue area. In the study of political elites, vocal pitch has further been shown to be a useful measure of politicians' issue commitments in Congressional floor speeches and presidential debates (Dietrich and Juelich 2018; Dietrich, Hayes, and O'Brien 2019). Among the broader public, Dietrich, Mondak, and Williams (2019) also show that features of vocal expression processed from the recordings of telephone survey responses, including but not limited to vocal pitch, can predict voting behaviour, even after controlling for standard predictors such as partisanship and demographics.

Audio as data methods have applications to various aspects of political communication, beyond affect. For instance, work by Neumann (2019) highlights the propensity of politicians to engage in phonetic "style-shifting" by adjusting their speech articulation—as measured using vowel space area—to different political settings. This approach is grounded in phonetics research; less distinct vowel articulation (a more compact vowel space) is associated with a lesser degree of articulation, and can be measured using formants, indicating which vocal organs are used in the generation of sounds. Using this approach, Neumann shows that politicians adjust their speaking style when addressing different audiences in settings such as Congress and on the campaign trail.

A separate set of studies has sought to build more general predictive models of the content—emotional or otherwise—of political speech. Knox and Lucas (2021) develop a semi-supervised HMM, which they refer to as the model of speech and audio structure (MASS), to infer tones of voice in political data, with particular attention to modeling the temporal dynamics of speech (i.e. inter-speaker interactions). Their goal is to provide researchers with a flexible model for predicting latent variables of interest—such as discrete emotions, but also other concepts such as "skepticism" or "decisiveness." Reflecting

this goal, their work differs from research using vocal pitch to measure emotional activation in the range of auditory features they use to model speech, including indicators based on the raw waveform and the frequency spectrum representation. Hwang, Imai, and Tarr (2019) also draw on audio features to predict the mood and negativity of televised political advertisements, in particular by making use of a classifier that identifies the mood of the music played during these ads. Their findings suggest that multimodal data (textual, audio, and visual) analysis tools could help to automate the process of coding variables in campaign advertisement databases.

Conclusion

The field of audio as data has important barriers to entry. Gaining familiarity with concepts in digital signal processing and learning to handle large data files efficiently requires an initial time investment. This cost should not be discounted. The payoff, however, may be well worth the investment. Audio as data opens the door to new opportunities for theory testing in political communication and comparative politics, ones that are currently beyond the grasp of textual analysis. Audio signals can reveal emotional states, contrasting speaking styles, unusual speech patterns, sarcasm, boredom, enthusiasm, and potentially other characteristics of language that are lost in translation whenever we reduce audio data to a transcript.

Two types of analysis involving audio as data seem especially promising for inferential statistics, the bread and butter of social sciences. First, the sequential arrangement of numerical data in digital recordings offers concrete benefits. For example, audio modeling could help tackle problems such as measuring emotional contagion, by tracking how the tone of one speaker affects the response of future speakers during a specific event. Similar methods could be deployed to study relationships of dominance during a dialogue, or how politicians strategically adapt their speaking style to their interlocutors. The type of model proposed by Knox and Lucas (2021) appears well suited for such endeavors. Second, audio recordings contain valuable information about each speaker. A recurring concern for causal inference is to buttress assumptions such as sequential ignorability; that is, we wish to rule out confounders when testing theories. An audio signal contains sufficient data to identify the unique characteristics of voices, thereby offering the potential to filter out unit-level attributes. Even when no external metadata are present, an audio recording allows researchers to detect whether a new speaker has taken the floor. It can reveal the likely gender, age, native language, among other useful control variables for empirical analysis. In short, leveraging the information contained in audio samples can help to bolster a fully fleshed-out empirical analysis of political language.

Notes

[1]See https://github.com/DolbyIO/awesome-audio to get started on available resources for audio as data. Examples of general purpose tools for audio analysis include the popular Praat software, the seewave library for R and the Parselmouth library for Python.

References

Cochrane, Christopher, Ludovic Rheault, Jean-François Godbout, Tanya Whyte, Michael W.-C. Wong, and Sophie Borwein. 2021. "The Automatic Analysis of Emotion in Political Speech Based on Transcripts." *Political Communication* (Forthcoming).

Dietrich, Bryce, Jeffery Mondak, and Tarah Williams. 2019. "Using the Audio from Telephone Surveys for Political Science Research." In *2019 Annual Meeting of the Society for Political Methodology.* Boston, MA.

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.

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* 113 (4): 941–962.

Dietrich, Bryce Jensen, and Courtney L. Juelich. 2018. "When Presidential Candidates Voice Party Issues, Does Twitter Listen?" *Journal of Elections, Public Opinion and Parties* 28 (2): 208–224.

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

Gregory, Stanford W., and Timothy J. Gallagher. 2002. "Spectral Analysis of Candidates' Nonverbal Vocal Communication: Predicting U.S. Presidential Election Outcomes." *Social Psychology Quarterly* 65 (3): 298–308.

Hinton, Geoffrey, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. 2012. "Deep Neural Networks for Acoustic Modeling in Speech Recognition." *IEEE Signal Processing Magazine* 29 (6): 82–97.

Hwang, June, Kosuke Imai, and Alex Tarr. 2019. "Automated Coding of Political Campaign Advertisement Videos: An Empirical Validation Study." In *2019 Annual Meeting of the Society for Political Methodology.* Boston, MA.

Klofstad, Casey A. 2016. "Candidate Voice Pitch Influences Election Outcomes." *Political Psychology* 37 (5): 725–738.

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., Rindy C. Anderson, and Susan Peters. 2012. "Sounds Like a Winner: Voice Pitch Influences Perception of Leadership Capacity in Both Men and Women." *Proceedings of the Royal Society B: Biological Sciences* 279 (1738): 2698–2704.

Knox, Dean, and Christopher Lucas. 2021. "A Dynamic Model of Speech for the Social Sciences." *American Political Science Review* (Forthcoming).

Neumann, Markus. 2019. "Hooked With Phonetics: The Strategic Use of Style-Shifting in Political Rhetoric." In *2019 Annual Meeting of the American Political Science Association.* Washington D.C.

Podesva, Robert J., Jermay Reynolds, Patrick Callier, and Jessica Baptiste. 2015. "Constraints on the Social Meaning of Released /t/: A Production and Perception Study of U.S. Politicians." *Language Variation and Change* 27 (1): 59–87.

Proksch, Sven-Oliver, Christopher Wratil, and Jens Wäckerle. 2019. "Testing the Validity of Automatic Speech Recognition for Political Text Analysis." *Political Analysis* 27 (3): 339–359.

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

Tigue, Cara C., Diana J. Borak, Jillian J. M. O'Connor, Charles Schandl, and David R. Feinberg. 2012. "Voice Pitch Influences Voting Behavior." *Evolution and Human Behavior* 33 (3): 210–216.

Trigeorgis, George, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. "Adieu Features? End-To-End Speech Emotion Recognition Using a Deep Convolutional Recurrent Network." In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 5200–5204.

Windsor, Leah, Alistair Windsor, Miriam van Mersbergen, George Deitz, Allison Sulkowski, and Lily Walljasper. 2019. "A Multimodal Approach to Analyzing Deception in Politics." In *International Studies Association 2019 Midwest Conference.* St. Louis, MI.

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