

a word error rate (WER)¹ of 0.05. As the authors caution, these validations are performed in laboratory conditions and do not account for levels of noise or crosstalk often present in political speech. Using the European Union State of the Union corpus for English, they find WER of 0.03 with the YouTube API and 0.21 with the Google API, with generally worse results for other languages, ranging from WER of 0.10 to 0.26 for French and German. They demonstrate that “automatic transcription tools have reached accuracy levels that make them useful for the study of parliamentary debates, campaign speeches, and intergovernmental deliberations” (Proksch, Wratil, and Wäckerle 2019, 357). However, they report that APIs performed differently depending on clip length and warn about legal issues related to data protection, as copies of the audio files are sent to YouTube or Google servers.² Recent sophisticated models, like OpenAI’s Whisper, have been analyzed for accent and speaker differences. Graham and Roll (2024) found Whisper performed best for U.S. and Canadian accents but poorly for Vietnamese and Thai accents in English, and for native but regional accents, such as the British Leeds accent, with a match error rate (MER)³ of almost 100%.

Our paper advances this conversation by addressing alignment, feature extraction, and classification of speech and emotion using modern audio embeddings. As a case study, we analyze televised U.S. presidential debates since first airing in 1960, until 2020, including vice-presidential debates. We obtained 47,150 sentences from the debate transcripts (38,649 from candidates and 8,501 from moderators, panelists, or audience members) initially unaligned to their respective videos. We obtained 243,023 s or 4,050 min or 67.5 h across 45 videos. In these videos, 34 candidates (16 democrats, 15 republicans, and 3 independent), 82 moderators and panelists, and 59 audience members appeared.

We begin by explaining core audio concepts relevant to political scientists (Section 3). We present four progressive applications to demonstrate ways of leveraging audio data for political science research, namely audio-text forced alignment (Section 4.1), analysis of individual speaking styles using low-level features (Section 4.2), construction of custom-made machine learning models for classification tasks (Section 4.3), and off-the-shelf machine learning models for emotion recognition (Section 4.4). We provide all replication codes in Python, transcripts, alignment data, and results in the Dataverse record Mestre and Ryan (2025) as well as our GitHub: <https://github.com/rafamestre/audio-as-data>.⁴

To fully understand the affordances of audio as data, it is necessary to appreciate the complex and multidimensional nature of audio signals. In speech processing and recognition, several features distinguish various aspects of audio signals. We explore two classical audio feature sets, low-level descriptors (LLDs) and Mel-frequency cepstral coefficients (MFCCs), as well as novel audio embeddings, which attempt to capture multidimensional representations of audio signals, discerning patterns, and relationships in a way akin to word embeddings. Table 1 introduces features’ key attributes, strengths, and limitations.

¹WER is a standard metric used in speech recognition and natural language processing that calculates the percentage of words in a transcribed text that were incorrectly recognized, compared with a reference transcription, taking into account substitutions, deletions, and insertions. Therefore, the smaller, the better.

²Other ethical concerns include transfer of copyright or violations of data protection laws if the data are held in servers from a different country. These issues can be especially relevant if the data to be aligned are not in the public domain but generated from fieldwork and contains sensitive information.

³MER is another common metric for the performance of ASR systems, which indicates the percentage of words that are incorrectly predicted. Therefore, a value closer to 0% is desirable.

⁴Due to copyright, we cannot make the videos of the debates freely available, although they are easily findable on YouTube. We emailed the broadcasters of the debates to ask for permission, but information was contradictory. The Commission for Presidential Debates was unsure who owned the rights of the videos and suggested contacting the networks, or if they were older the sponsors of the debates (such as the League of Women Voters). American Broadcasting Company (ABC) said they do not have the copyright for their broadcasted videos and suggested the Commission for Presidential Debates had it instead. Columbia Broadcasting System (CBS) said they would investigate but never got back to us, noting their licensing fee started at \$2,500. Public Broadcasting Service (PBS), who uploaded the videos to YouTube confirmed they do not have the rights beyond that and suggested contacting the Commission. National Broadcasting Company (NBC) outsourced licensing to Getty Images, who did not provide a useful reply. Fox News and Cable News Network (CNN) did not reply to queries.

Table 1. Summary of audio feature extraction techniques.

Features	Key attributes	Strengths	Limitations
LLDs	Basic acoustic features: pitch, energy, spectral contrast, zero-crossing rate	Simple, computationally inexpensive, interpretable	Capture only surface-level patterns; sensitive to noise
MFCCs	Derived from Mel-scaled spectrogram; models timbre and spectral envelope	Captures perceptual characteristics of human hearing	Fixed frame size; require preprocessing; less effective for context
Audio embeddings	Learned representations from models such as Wav2Vec, HuBERT, and Whisper	Captures complex, high-level attributes, adaptable to various tasks	Computationally expensive, requires large datasets

Abbreviations: LLD, low-level descriptor; MFCC, Mel-frequency cepstral coefficient.

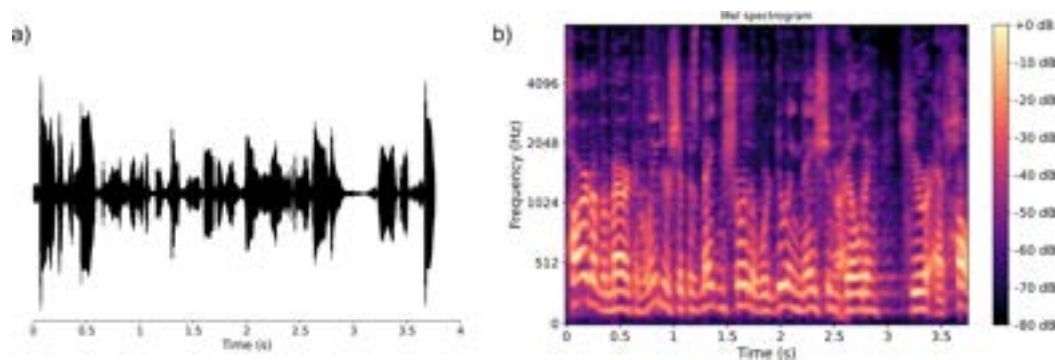


Figure 1. Representation of an utterance in different modalities. (a) As a discrete mathematical function with an amplitude changing over time. (b) As a spectrum of energy in decibels with respect to time.

3. Understanding Audio Features

An audio signal is a series of air vibrations measured over time. Its typical representation is a waveform like Figure 1a, which shows the amplitude (or strength) of vibrations at each point in time. This time-based view reveals little about the nature of the sound. Speech, music, or noise may look similar in waveform form but differ significantly in content. We often transform the signal into the frequency domain, producing a spectrogram (Figure 1b).

The y-axis represents the signal frequency and the z-axis (color bar) how much energy is present at each frequency over time. This spectral view decomposes the signal into a combination of simple waves at varying frequencies and amplitudes, making it more analytically useful.

3.1. Low-Level Descriptors

Derived from spectral representations, LLDs describe basic properties of an audio signal and serve as inputs for more complex feature extraction. Typical LLDs include root-mean-squared energy (also known as loudness or intensity), pitch, spectral contrast, spectral flux, zero-crossing rate, and various spectral shape descriptors, such as spectral centroid, spread, skewness, and kurtosis. We explore two of the most important LLDs for speech analysis: energy and pitch.

Energy refers to the loudness or intensity of an audio signal. Mathematically, for a discrete-time signal $x[n]$ of duration T , discretized into N points, and sampled at a rate/frequency $f_s = N/T$, the energy is calculated as $E = \sum_{n=0}^{N-1} x^2[n]$. This measure quantifies the “power” or “volume” of the signal,