Audio as Data

Benjamin W. Arold

December 20, 2024

Outline

Human Hearing is Interesting

- Hearing is a critical sense for communication and survival
- Like vision, it enables us to interact with the world without physical contact
- Various areas of the brain are associated with auditory processing
- Humans are **"analysts"** and **"generators"** of audio at the same time
- Human hearing has constraints:
  - Limited frequency range: 20 Hz to 20 kHz
  - Difficulty in locating low-frequency sounds
- Fascinating questions:
  - Fundamental: How does the auditory system decode sounds? (Beyond scope of this class)
  - Aggregate/social: What is the impact of sound on societal outcomes? (Central to our class)

Audio Analysis Definitions

- Many view it as a signal-processing challenge
- Julius O. Smith III: reconstructing auditory experiences
- This class explores techniques to analyze auditory data to study economics questions
- Focus on existing auditory data, not the mechanics of data capture

What is Sound?

- A sensory experience created by vibrations traveling through a medium (usually air)
- Analogy to ocean waves: sound waves propagate through air like ocean waves move through water
- As ocean waves interact with air, they create ripples in the atmosphere, analogous to **sound waves**

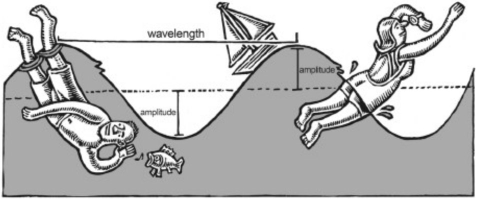

Figure: Ocean waves analogy to sound waves (Source: Sulzer, 2021)

Humans typically hear between 20 Hz and 20 kHz.

- We begin to hear low vibrations at around 20 Hz
- Teenagers can perceive frequencies up to about 20,000 Hz.
- Each octave doubles the frequency.
- Humans with good hearing can perceive a ten-octave range.

1. 20 Hz
2. 40 Hz (20 Hz $\times$ 2)
3. 80 Hz (40 Hz $\times$ 2)
4. 160 Hz (80 Hz $\times$ 2)
5. 320 Hz (160 Hz $\times$ 2)
6. 640 Hz (320 Hz $\times$ 2)
7. 1280 Hz (640 Hz $\times$ 2)
8. 2560 Hz (1280 Hz $\times$ 2)
9. 5120 Hz (2560 Hz $\times$ 2)
10. 10240 Hz (5120 Hz $\times$ 2)
11. 20480 Hz (10240 Hz $\times$ 2)

Figure: Human hearing range

Key Concepts: Hz, Octave, and Sampling Frequency

- **Hz (Hertz)**: Measures the number of cycles per second of a periodic waveform
  - Higher Hertz means a higher-pitched sound
  - If a tone has a frequency of 440 Hz, the sound wave completes 440 cycles every second (A4)
- **Octave**: The interval between one musical pitch and another with half or double its frequency
  - 440 Hz and 880 Hz
- **Sampling frequency**: The number of samples per second from a continuous signal to make a digital signal:
  - Standard CD sampling rate is 44.1 kHz...
  - ...meaning that the audio waveform is being sampled 44,100 times per second

The wavelength is an important sound wave characteristic

- What is a sound wave, to begin with?
  - A transfer of sound energy through a medium
  - It is transmitted via the vibration of particles within that medium
- Illustration for non-physicists:
  - When something vibrates (like a speaker's diaphragm), it pushes on the neighboring air particles, increasing the pressure in that region $\rightarrow$ this push is then transferred from particle to particle $\rightarrow$ when it reaches someone's ear, the ear interprets the varying pressures as sound
- The physical distance between identical points in consecutive cycles of a sound wave
  - Determines the pitch

The amplitude is another important characteristic

- Amplitude in a general wave context: the magnitude of change in the oscillating variable within the wave
  - $\rightarrow$ Larger amplitudes mean more energy and often translate to louder sounds when talking about sound waves
- Amplitude in digital audio signal processing : understood as instantaneous amplitudes that represent the "magnitude of change" in the pressure wave (the sound wave) from its equilibrium position at that particular moment
  - The specific value of the audio waveform at a given sample point

One-Tone Sound Wave

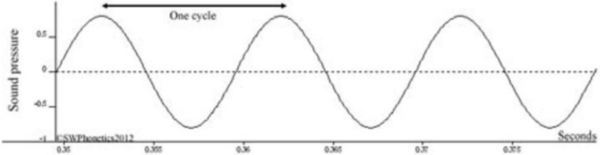

Figure: Sidney Wood and SWPhonetics, 1994-2024

A More Complex Sound Wave

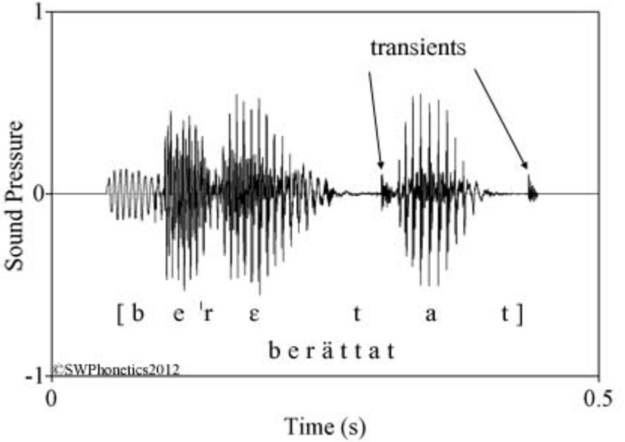

Figure: Sidney Wood and SWPhonetics, 1994-2024

Outline

Recap from computer vision: For classical ML, we (often) extract features explicitly.

- Recall the typical pipeline for classical machine learning:

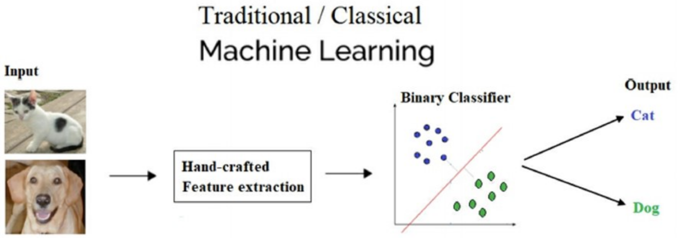

Figure: Dey (2018)[1]

[1] Dey, S. (2018). Hands-On Image Processing with Python: Expert Techniques for Advanced Image analysis and Effective Interpretation of Image Data. Packt Publishing Ltd.

We can use a similar workflow for audio data.

- Recall the typical pipeline for classical machine learning:

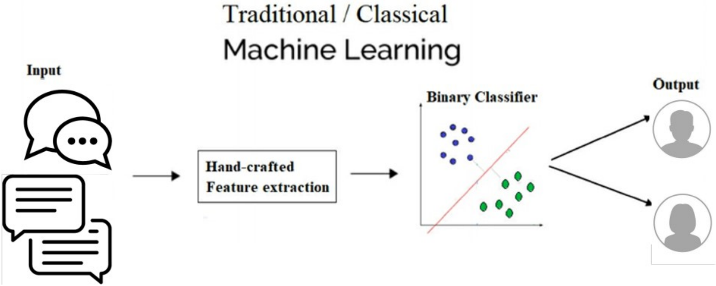

Figure: Adaptation by Widmer (2023) of Dey (2018)[2]

[2] Dey, S. (2018). Hands-On Image Processing with Python: Expert Techniques for Advanced Image analysis and Effective Interpretation of Image Data. Packt Publishing Ltd.

Can we use the raw audio input for classification tasks?

- Neural networks, such as CNN, can handle raw audio data $\rightarrow$ extract relevant features automatically (cf. images part)
- For very simple tasks, it may be possible with classical machine learning
- More generally, audio waveforms are high-dimensional data, especially for long clips
  - Computationally demanding
  - And often not necessary $\rightarrow$ suitable format depends on task
- For both classical and neural approaches, feature extraction is often useful

Spectral features capture important characteristics of a sound's frequency content

- Spectral features are widely used for audio tasks (e.g., music analysis, speech processing, audio classification)
- Typically, they encompass spectral centroid, spectral bandwidth, or spectral flatness
- The spectral centroid shows the frequency spectrum's "center of gravity"
  - Calculated as the weighted mean of the frequencies in the signal, with their amplitudes being the weights
  - Higher values reflect brighter sound
  - Technically, we cut the sound into frames and apply a Fourier Transform per frame
  - Hence, we transform the frame's time-domain signal into the frequency-domain

Spectral features beyond the spectral centroid

- The spectral bandwidth describes how wide the frequency band is
  - A wider bandwidth implies a broad range of frequencies
  - Conversely, a narrower bandwidth indicates a more tonally pure sound
  - It can help, for example, to identify the complexity of sound
- The spectral flatness indicates how "noise-like" a sound is instead of being tonal
  - Values close to 1 indicate a noise-like sound
  - Values near 0 indicate a more tonal sound
  - It can be helpful, for instance, to disentangle tones and environmental noises

Mel-Frequency Cepstral Coefficients (MFCCs) are often used for feature extraction

- Humans perceive sound frequencies non-linearly (rather, logarithmically)
- Human sensitivity is greater to changes in lower frequencies
- The Mel scale is a way to measure pitch that matches how we hear sounds
  - It is a perceptual scale of pitches judged to be equal in distance from one another
  - Originally derived from experiments with human listeners
- MFCCs represent the power spectrum of an audio signal more in line with human hearing
  - They use the Mel scale

MFCCs capture the short-term power spectrum of sound

- One typically begins by dividing the audio signal into short (overlapping) frames, for instance 20-40 ms
  - This allows assuming stationarity within each frame
  - Stationarity means that the statistical properties of the signal (like mean, variance) are constant over the frame's duration
- Several processing steps involved (including fast fournier transform)
- The bottom line is that we typically end up with 12-13 coefficients per frame
  - Empirically found to capture the most important features

Outline

Recall that feature extraction is used for many audio (classification) tasks

- Recall the typical pipeline for classical machine learning

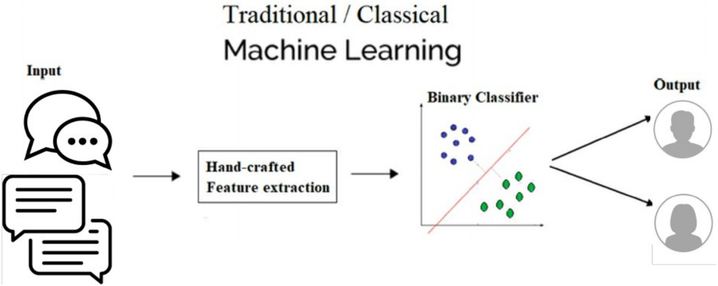

Figure: Adaptation by Widmer (2023) of Dey (2018)[3]

[3] Dey, S. (2018). Hands-On Image Processing with Python: Expert Techniques for Advanced Image analysis and Effective Interpretation of Image Data. Packt Publishing Ltd.

What models are used in audio analysis with deep learning?

- The audio analysis landscape is diverse
- CNN are used for classification with audio data
  - But less prominently than in CV
  - Audio data is often transformed into spectrograms (time-frequency representations) to use CNN
  - Examples: Emotions, speaker identity, accents
- Otherwise, varied, possibly hybrid approaches (RNN, transformers)
- In any case, audio data often requires pre-processing
- Other methods in audio analysis:
  - Classical ML techniques (e.g., MFCC)
  - Unsupervised and semi-supervised approaches: Useful with limited labelled data

Audio data is sometimes used with CNN or RNN

- Tailored for tasks with temporal dependencies
- For CNN, the conceptual foundation discussed in the computer vision part also applies here
- Recurrent Neural Networks (RNN) are “specialists” for sequential data
  - Sequential data examples are audio or text
  - RNN have a memory mechanism to remember previous inputs
  - They can, thus, “remember” context in sequences
  - For sequences, context is important to understand the overall pattern

Conceptually, how does an RNN proceed

- The data is typically pre-processed
  - Extract features: for example, MFCC by time t (often relatively short snippets)
  - Each feature at time t becomes input x_t to the RNN
- $\forall t \in \{0, 1, 2, \ldots, T\}$, x_t are processed by the RNN sequentially
- RNN computes new hidden state h_t for each time step t
- Typically: $h_t = \text{Activation}(W_{xh} \cdot x_t + W_{hh} \cdot h_{t-1} + b)$
- W_{xh}, W_{hh} are weight matrices; b is a bias vector
- RNN learns W_{xh}, W_{hh}, and b
- Hidden state h_t used for output (e.g., in classification)