AARHUS
UNIVERSITY

Class 9: Audio Basics

Theme: Audio

Computational Analysis of Text, Audio, and Images, Fall 2023

Aarhus University

Mathias Rask (mathiasrask@ps.au.dk)

Aarhus University

Today's Menu

Sound Theory

Digital Signal Representation

Audio Representations

Lab

Table of Contents

What's *Sound*?

Definition

Vibration of air molecules caused by air pressures
⇝ Sound waves

Vibrations can happen for many - sometimes cooccurring - events:

- animal calls
- traffic
- explosion

⇝ human speech

- . . .

Characterizing Sound Waves

A sound wave $s(t)$ is determined by two key characteristics: height and length

Height (amplitude)

- Maximum height at time t as a result of the wave
- Higher air pressure → higher vibration → higher height
- Perceived as loudness – a higher amplitude is perceived as a louder sound
- Usually measured in decibels (dB)

Length (frequency f)

- Length: Time of one cycle
- Frequency: Number of cycles per second (i.e. Hz)
- Longer air pressure → slower vibrations → lower frequency
- Perceived as pitch – a longer length is perceived as a lower pitch
- Inverse related to f: $1/f$ with f

Sine Waves

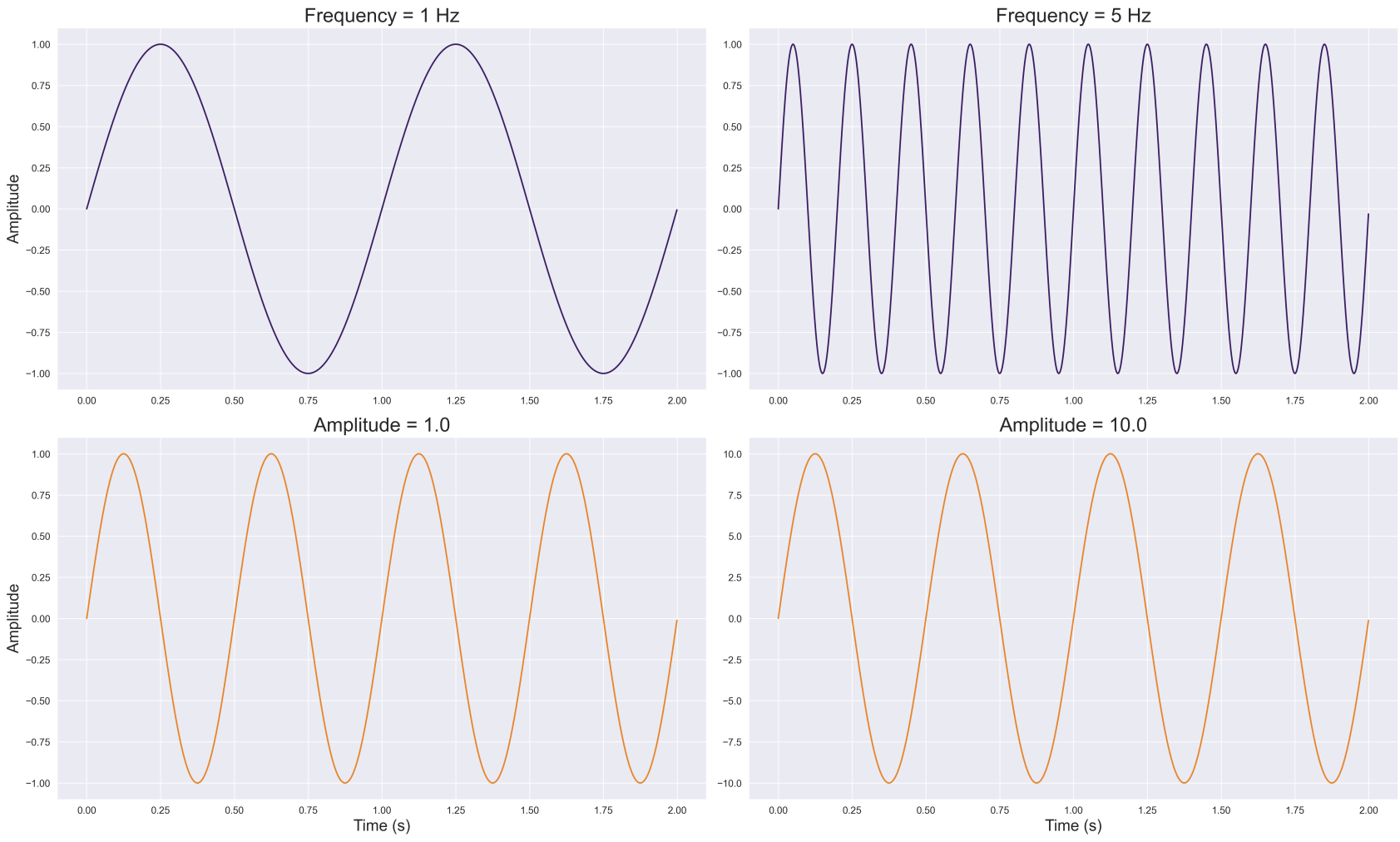
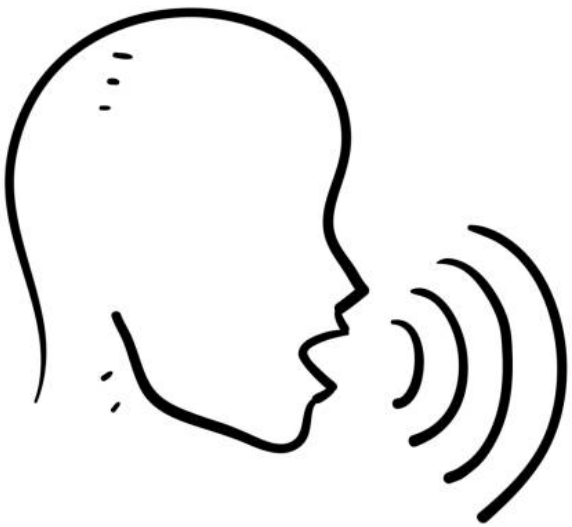

Frequency = 1 Hz
Frequency = 5 Hz
Amplitude = 1.0
Amplitude = 10.0
Amplitude
Time (s)

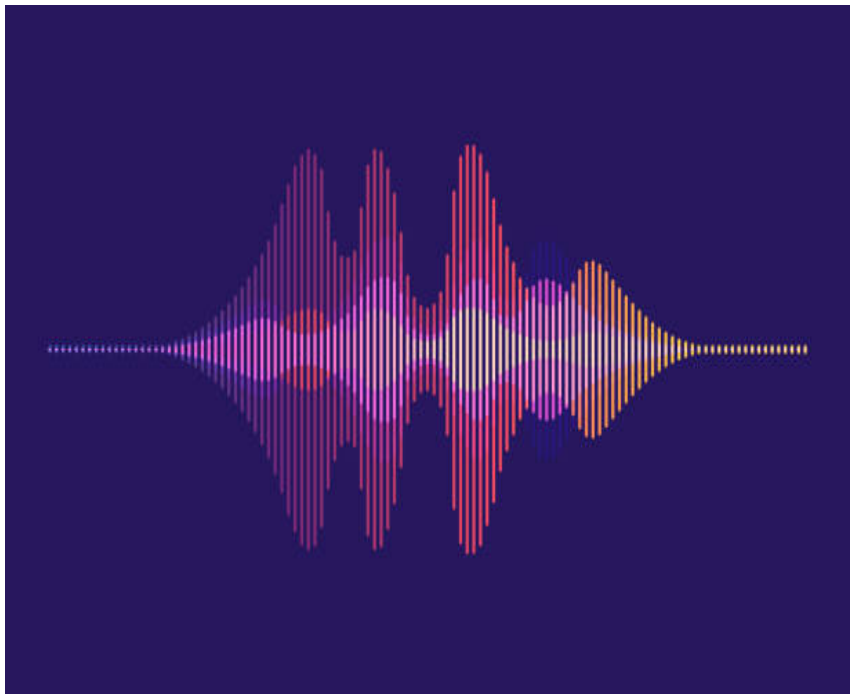
Table of Contents

From Analog to Digital Signals

Analog signal $x(t)$



Digital signal $y(t)$



Continuous Time-Signals

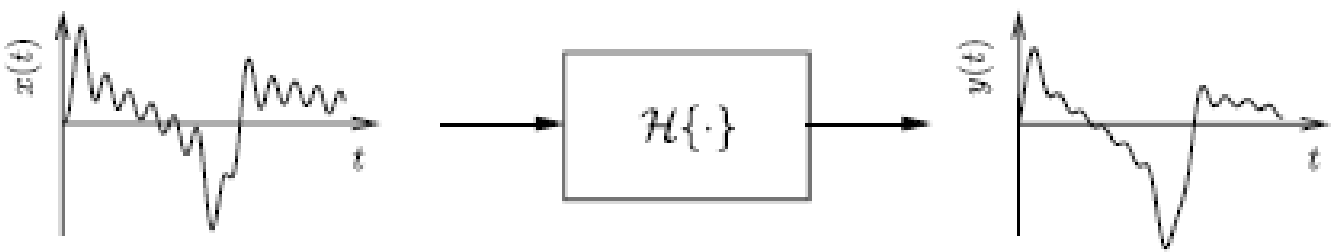
Most signals we are interested in are continuous in nature:

- Traffic
- Animals
- ⇝ Speech

A continuous time system maps an analog signal $x(t)$ to an output signal $y(t)$ with $t \in \mathbb{R}$ using a transformation $\mathcal{H}\{\cdot\}$:

$$y(t) = \mathcal{H}\{x(t)\}$$

Illustration:



Discrete Time-Signals

A discrete time signal is represented by a sequence of numbers $x(n)$ with $n \in \mathbb{Z}$:

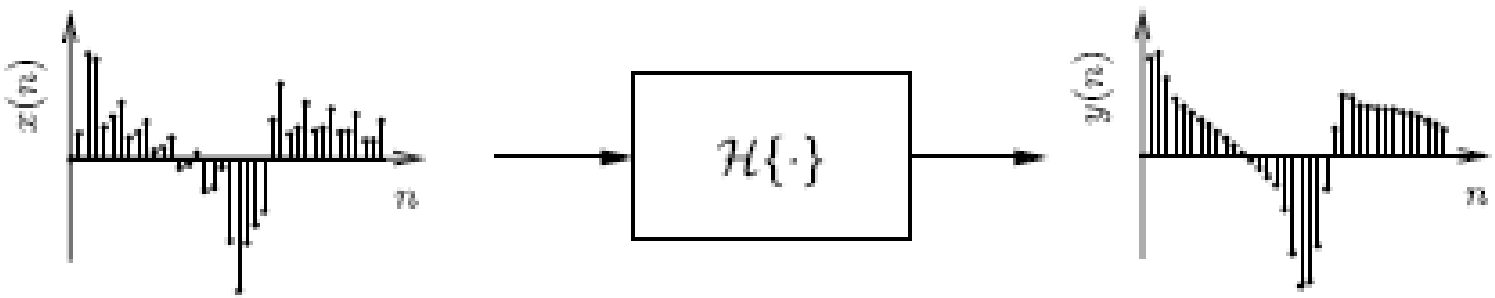
- Digital images
- Digital text

⇝ Digital audio

A discrete system maps an input sequence $x(n)$ to an output sequence $y(n)$ with $n \in \mathbb{Z}$:

$$y(n) = \mathcal{H}\{x(n)\}$$

Illustration:



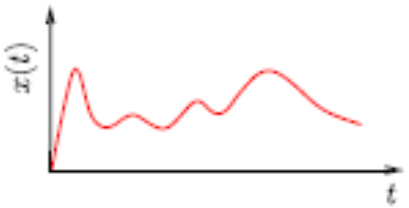
Sampling and Quantization

When the signal we analyze originates from a continuous time domain, it involves several building blocks before we can analyze the “data” computationally:

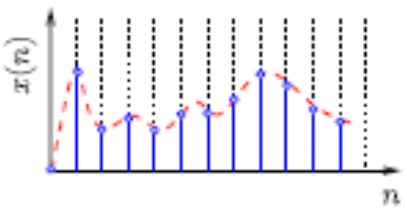
1. A/D:
    - Sampling
    - Quantization
2. Transformation
3. D/A
4. Low-pass filter

⇝ We care most about 1 and 2

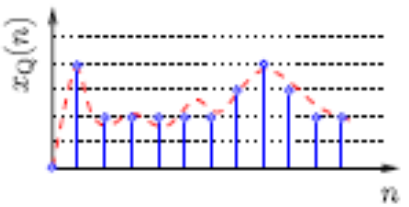
Digital Signal Generation (Figure 1.6 in Diniz, 2023)



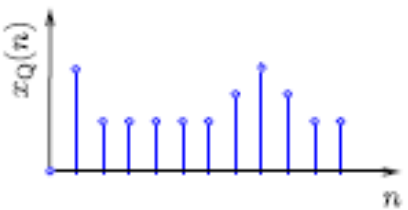
(a) Original continuous-time signal.



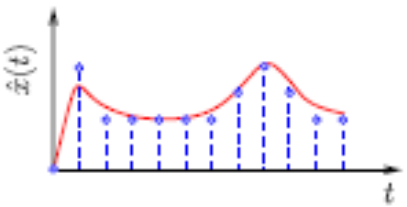
(b) A/D converter: sampling.



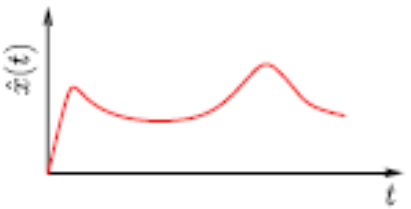
(c) A/D converter: quantization.



(d) Digital signal.



(e) D/A converter.



(f) Recovered continuous-time signal.

A/D Converter: Sampling

All digital audio recordings are sampled at equally spaced time intervals:

$$\text{sampling rate} = \frac{1}{f}$$

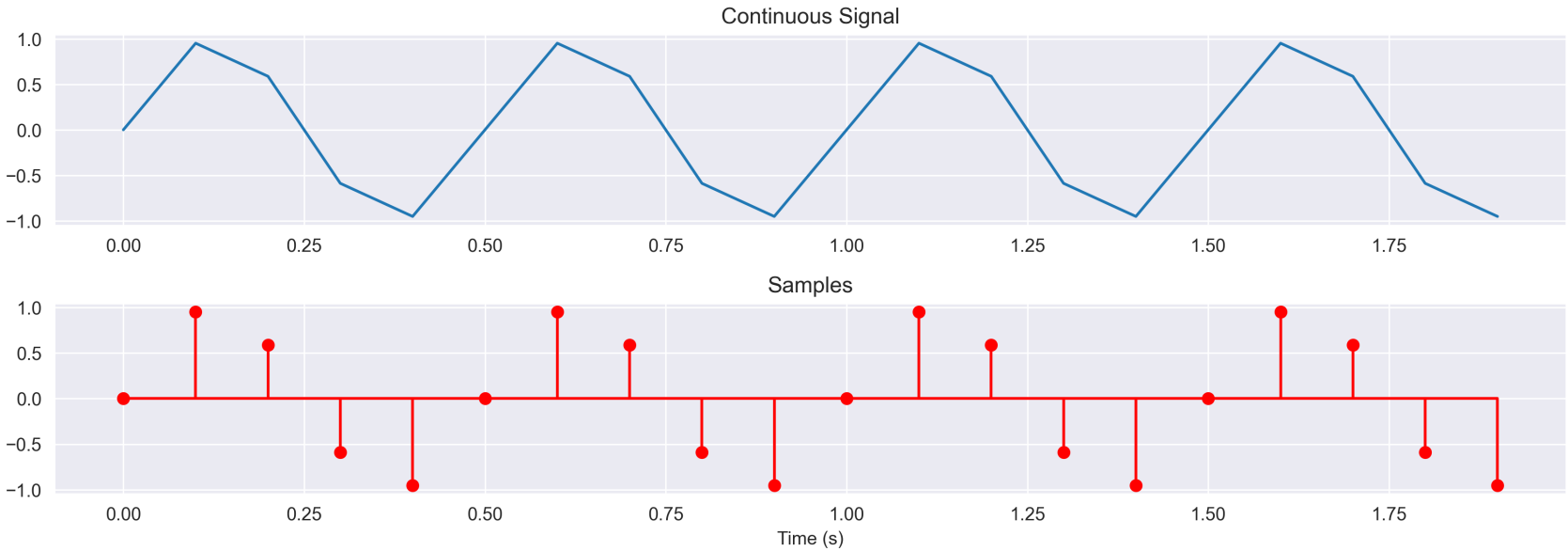
where f is the frequency (i.e. number of cycles per second). Common sampling rates:

- 8000
- $16,000$
- $22,050$
- $44,100$

Questions:

1. What's the unit of the sampling rate?
2. What's the time interval if using a sampling rate of $16,000$?

Sampling Illustration



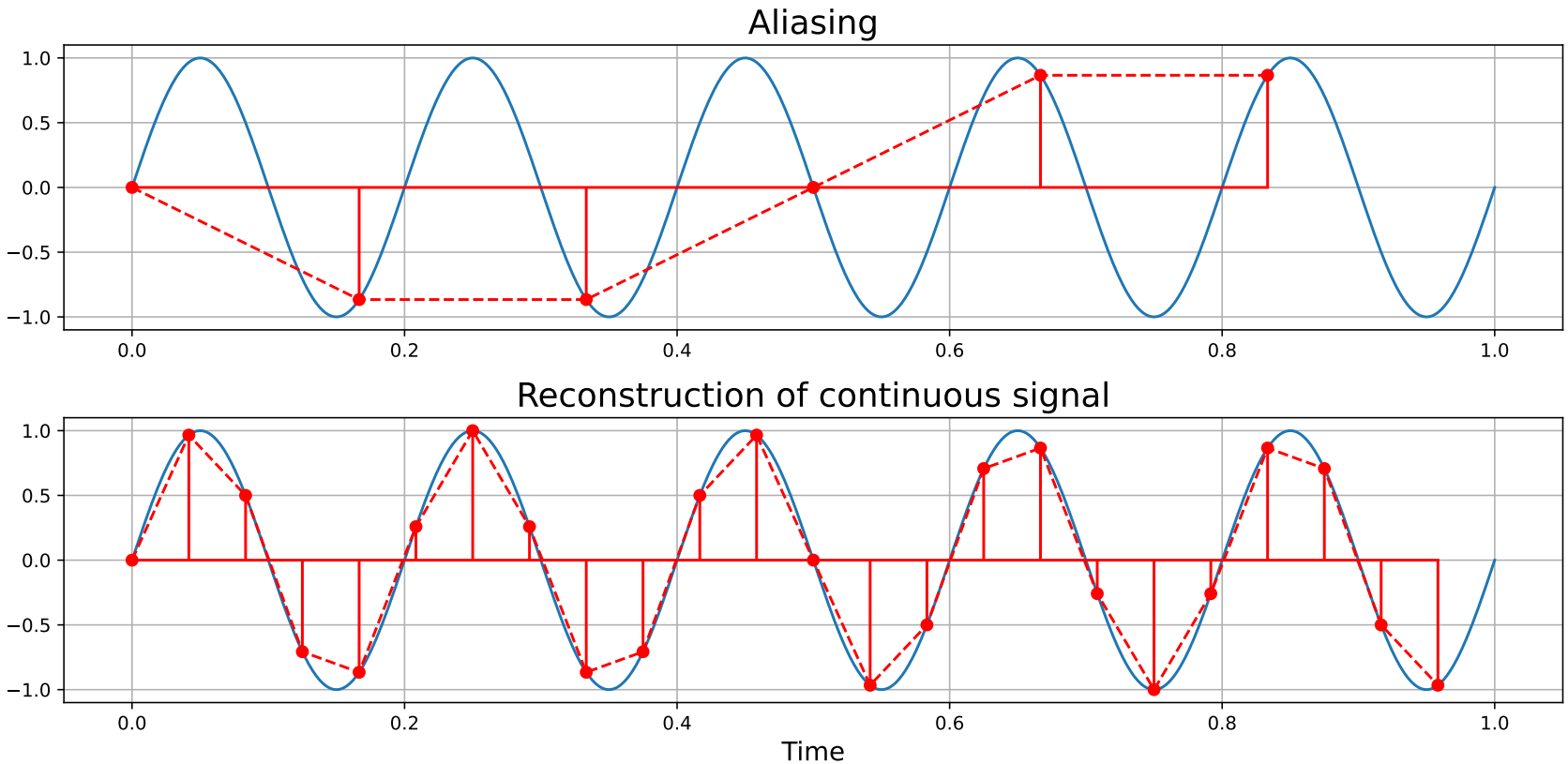
Questions:

1. What's the frequency in the top figure?
2. What's the sampling rate in the bottom figure?
3. What's the difference between frequency and sampling rate?

Sampling Rate vs. Frequency

The distinction between sampling rate and frequency is crucial for the validity of our later measurement. Why? *Aliasing*

⇝ The Nyquist Theorem: If a function $x(t)$ contains no frequencies higher than B Hz, then it can be completely determined from its ordinates at a sequence of points spaced less than $1/(2B)$ seconds apart



Quantization

Quantization amounts to representing each sample $x(n)$ in a binary form with a finite word length – we call this the *bit depth*

$\rightsquigarrow$ controls how fine-grained our samples are

How many possible values can a single sample $x(n)$ take?

$$\text{bit depth} = 2^b$$

$\rightsquigarrow$ increases exponentially Example: 3-bits:

$$= 2^3$$
$$= 8$$
$$= \{111, 110, 101, 100, 011, 010, 001, 000\}$$

(Very typically, we work with 2^{15} values in range $[-2^{15}, 2^{15} - 1)]$ = $[-32,768, 32,767]$

Quantization Illustration

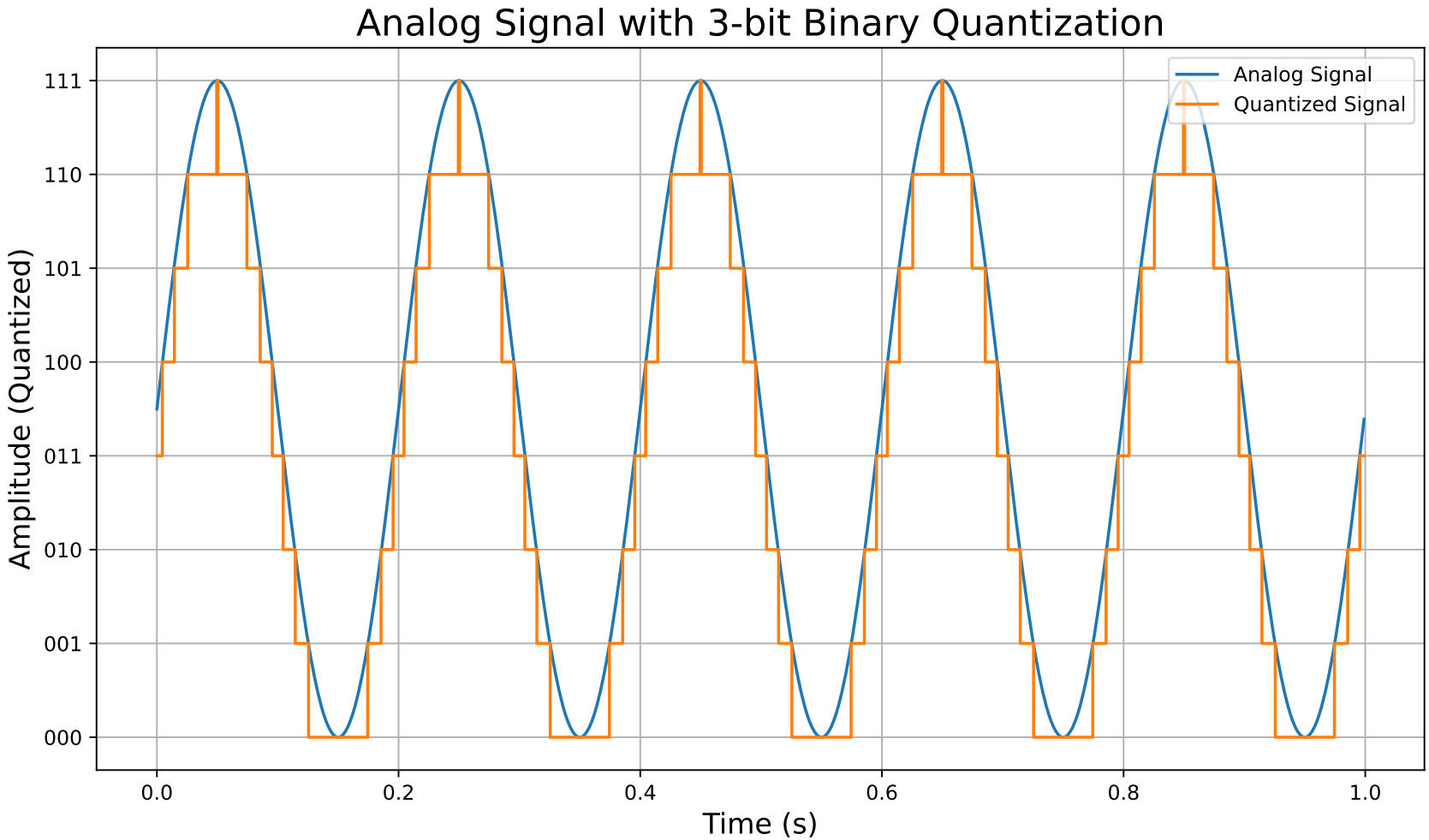
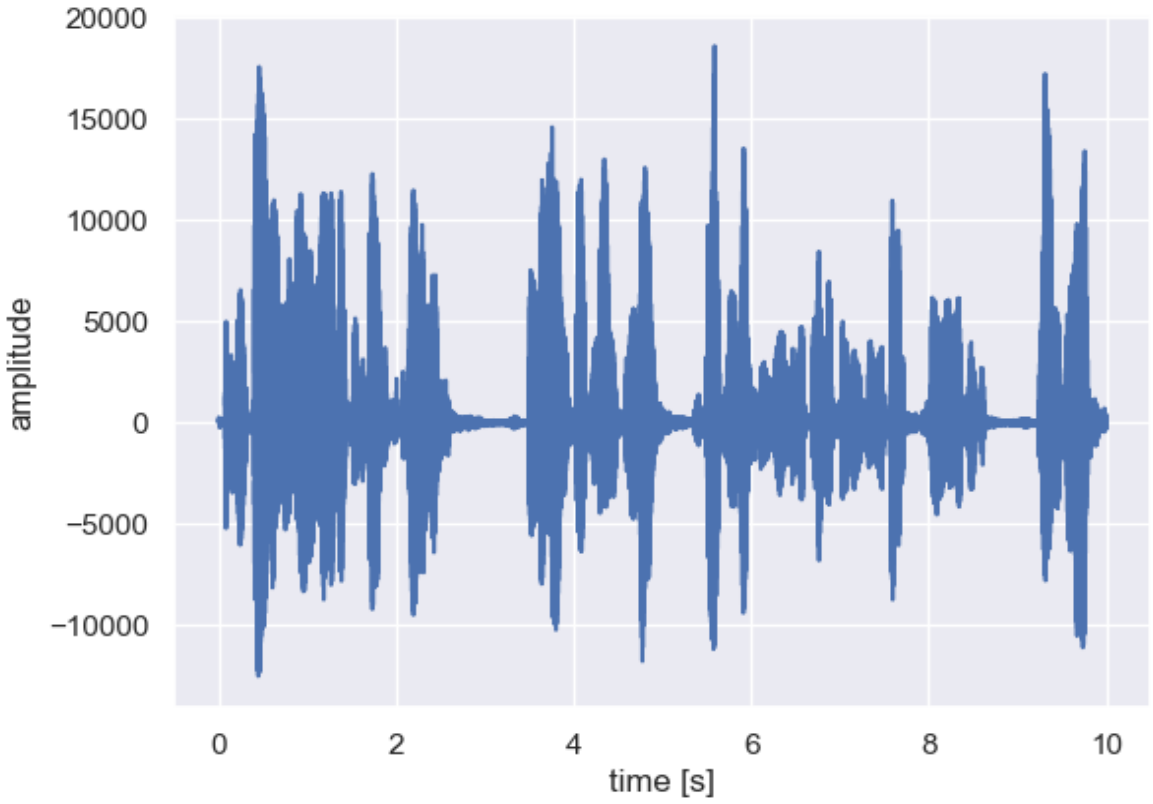


Table of Contents

Waveform of Human Speech

The most basic representation of digital audio is the *waveform*



Why does it look so messy compared to the illustrations we have seen earlier?

Decomposing a Waveform

’ A waveform is essentially an *univariate time series* $x(n) = x(nT)$ with samples every $t = nT$ and $x \in \mathbb{Z}^1$

- Example: $T = 0.0625$ ms with $f = 16,000$ Hz

Drawbacks:

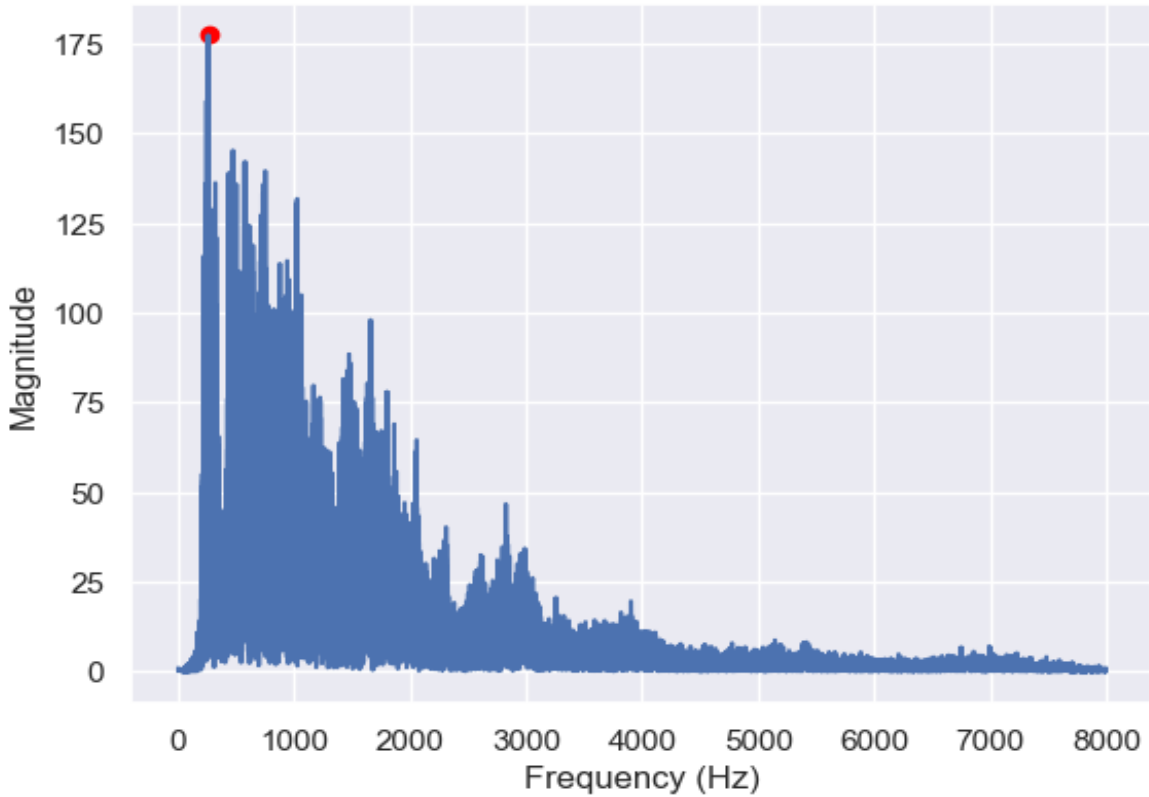
- Lack of frequency information: Signals where the variation happens at the spectral level such as speech can generally not be analyzed using time representations
- Separation: We can not tell sources apart

⇝ Frequency representation solves these issues

Frequencies in Human Speech

To cast a signal as a function of f we use a Fourier transform:

⇝ signal is divided into short windows - assume stationarity within each window

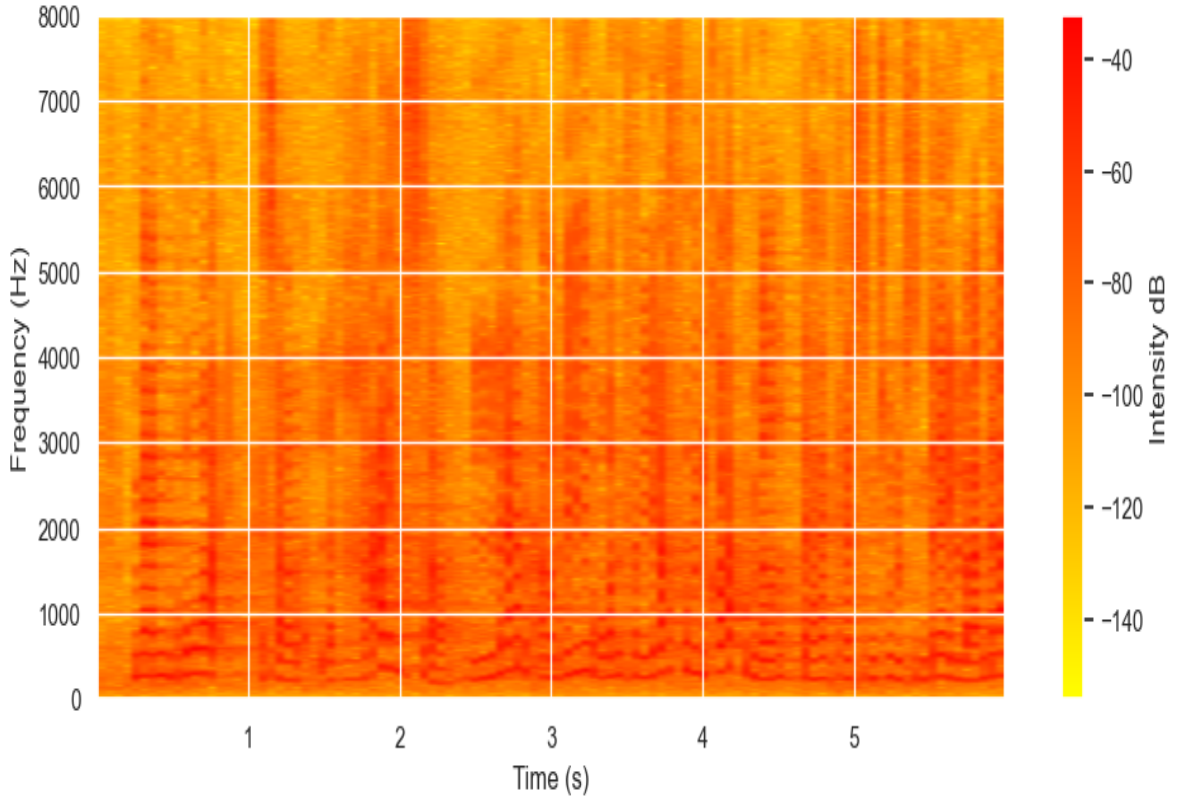


⇝ The first peak corresponds to the fundamental frequency $F0$ (in red)

⇝ Most frequencies are below $4,000$ Hz

Spectrograms

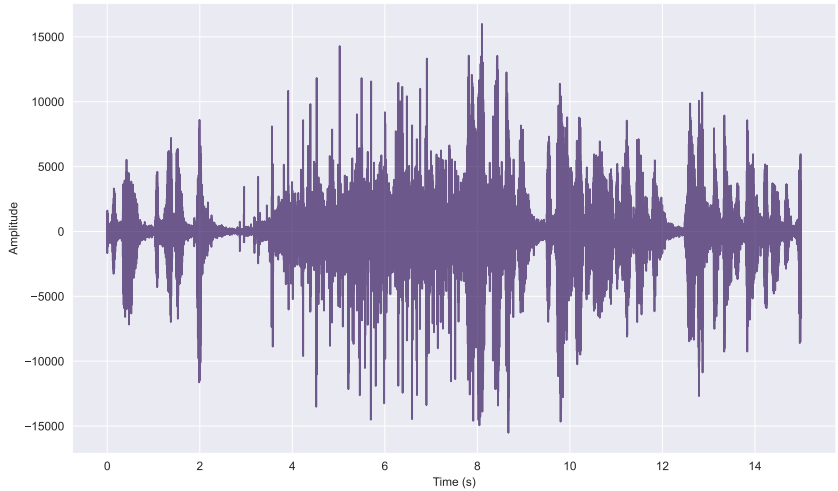
Spectrograms combine time and frequency information in a three-dimensional plot:



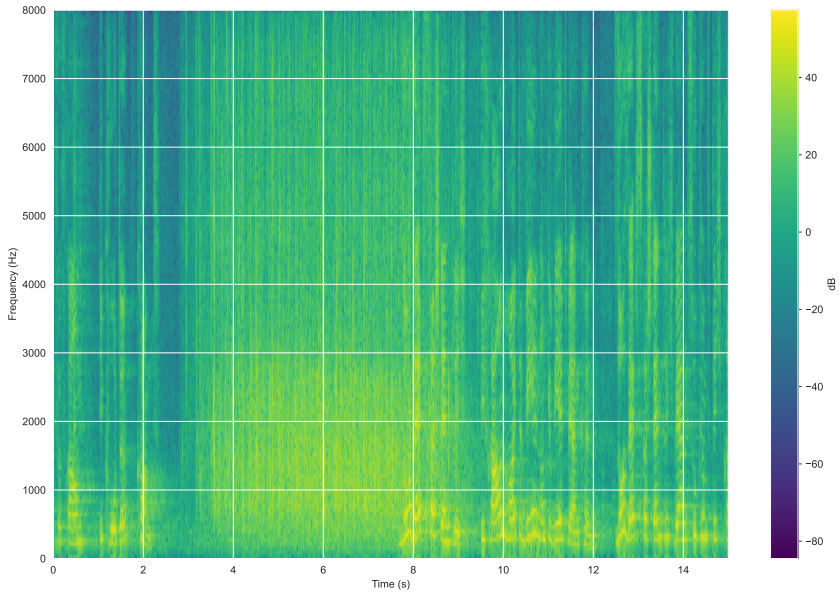
⇝ Still difficult to interpret!

Illustration: Speech vs. Applause

Waveform:



Spectrogram:



The Mel-Scale

- Raw spectrograms are not designed how human perceptions
- What we care about is concentrated in a narrow range of frequencies and amplitudes
- Humans perceive f as *pitch* - positive but nonlinear relationship
- We are more sensitive to differences in lower frequencies than higher:
  - 100Hz and 200Hz
  - $1,000$Hz and $1,100$Hz
  - $10,000$Hz and $10,100$Hz

  ⇝ Evenly distanced in Hz

  ⇝ Not evenly distanced in terms of percentage!
    - 100Hz and 200Hz: 100Hz increase and 100% increase
    - $10,000$Hz and $10,100$Hz: 100Hz increase and 1% increase

    ⇝ A log scale!
- Amplitude is perceived as loudness and also in a logarithmic manner - we account for this using the decibel (dB) scale

Table of Contents

See you next week!

Theme: Audio

Computational Analysis of Text, Audio, and Images, Fall 2023

Aarhus University

References i

[1] P. S. Diniz, *Signal Processing and Machine Learning Theory*. Elsevier Science & Technology, 2023.