Modern Audio Processing Techniques

From Fundamental to Advanced Models

Self-introduction

Dr. Zijiang **YANG**

Project Researcher @ **The University of Tokyo**

Co-founder, CTO @ **semo AI Co., Ltd.**

8 years of experiences in **Affective Computing**

with the **best lab** in Affective Computing with **Speech**

Emotional Speech Synthesis and **Emotional Voice Conversion**

Speech Emotion Recognition

Digital Health and Wellbeing

Deep learning and **Signal Processing**

Audio Processing is the *manipulation* and *analysis* of sound signals to *extract information*, *enhance quality*, or *transform characteristics*

From *voice assistants* to *music analysis*, audio processing is **everywhere** in our daily lives

Today...

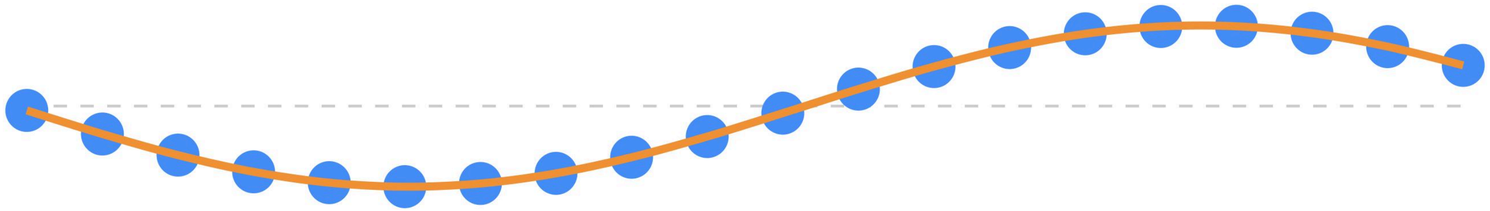
- from **Sound Physics** to **State-of-the-Art Models**
- **Examples**, **analogies**, and **real-world applications**

Content

Audio Basics

Sound is a **mechanical wave** that results from the back-and-forth **vibration** of **particles** in a **medium**

These vibrations create pressure variations that travel through *air*, *water*, or *solids*, but **not in vacuum**



1 Pitch

Human perception of **frequency**

Human hearing range: **20 - 20,000** *Hz*

We perceive pitch logarithmically, not linearly!

2 Loudness

Human perception of **amplitude**

Measured in decibels (*dB*)

A 10 *dB* increase is perceived as about twice as loud!

3 Timbre

The **quality** of “**colour**” of sound

Distinguish different **sound sources**

This is why a violin and a piano sound different!

4 Spatial Perception

Human perception of **location**

Interaural time and **level difference**

Our brain compares signals from both ears to locate!

Analog Signal

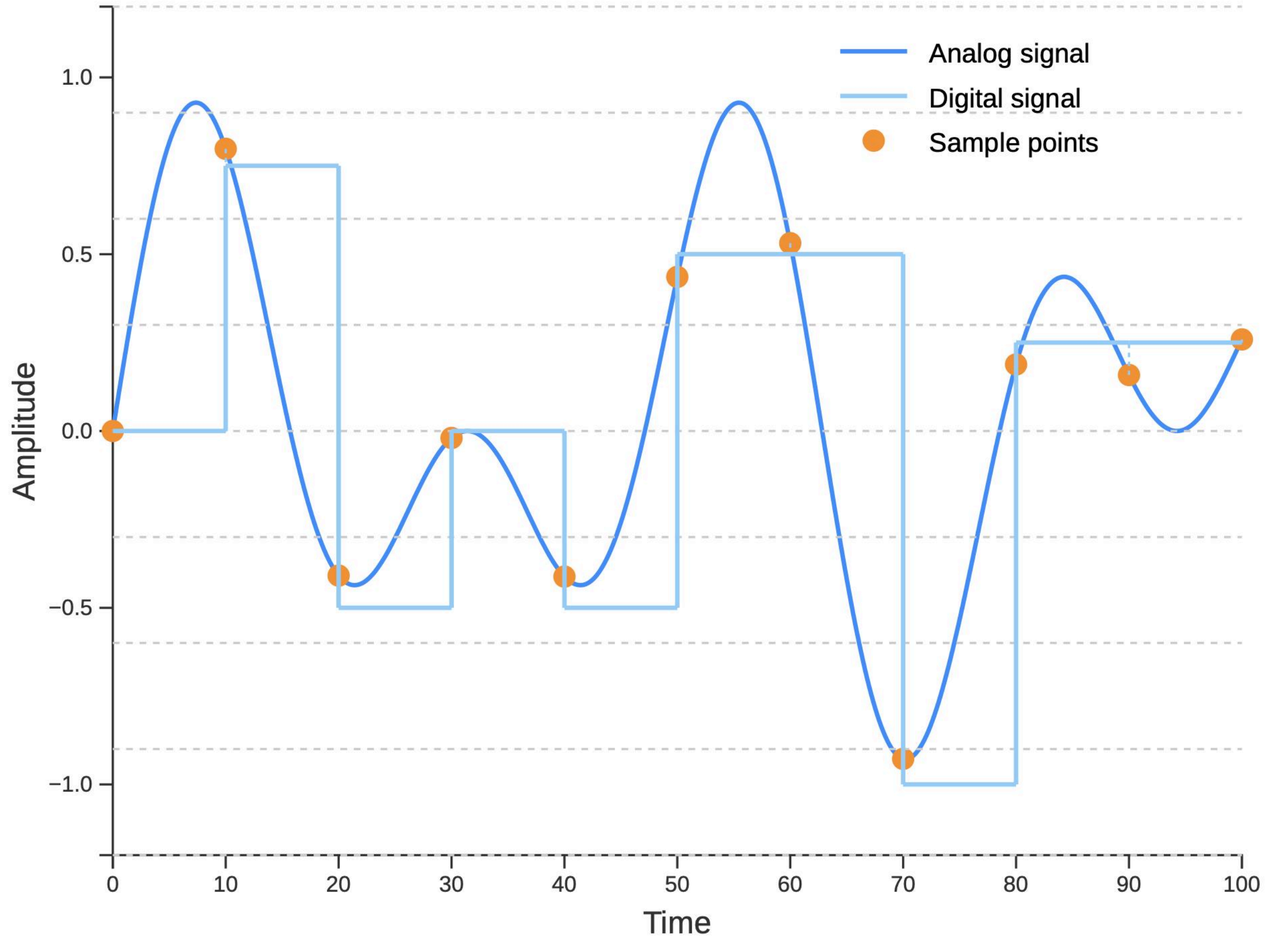
Continuous wave of varying physical measurements

Rich information, but **fragile**

Digital Signal

Discrete wave with a series of non-continuous values

Limited information, but **stable**



Converting analog sound to digital sound

1 Sampling

Measuring the amplitude of the sound wave at regular intervals

2 Quantisation

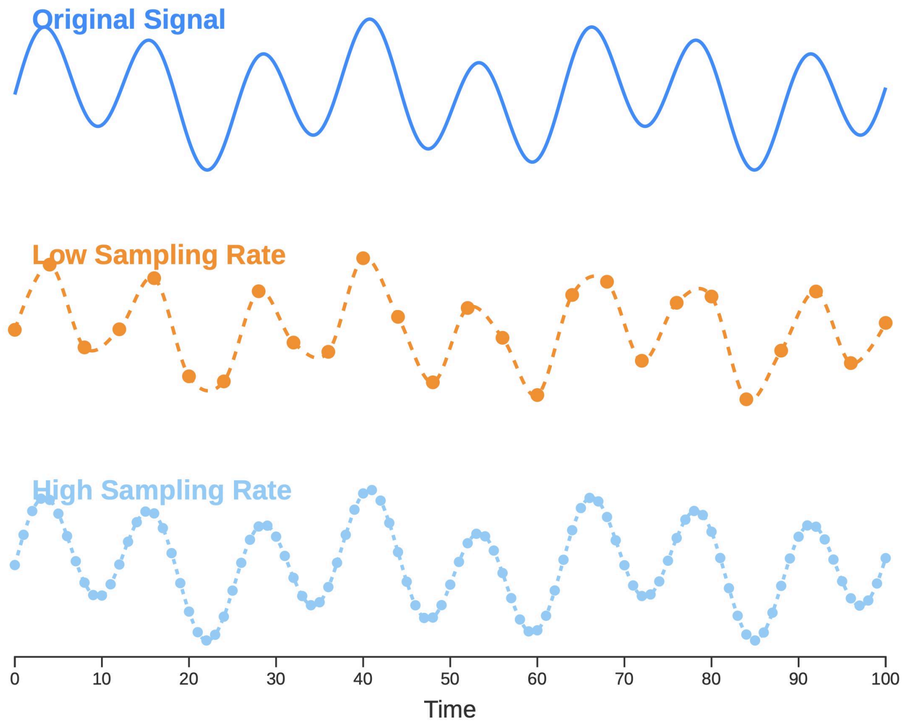
Assigning discrete numerical values to those measurements

3 Encoding

Storing these values in a specific format (e.g., PCM, MP3)

The process of measuring the amplitude of **a continuous signal** at **discrete time intervals**

Sampling rate is the number of **samples taken per second**, measured in **Hertz** (*Hz*)



8 *kHz* - Telephone speech

44.1 *kHz* - CD quality audio

96 *kHz* - High-resolution audio

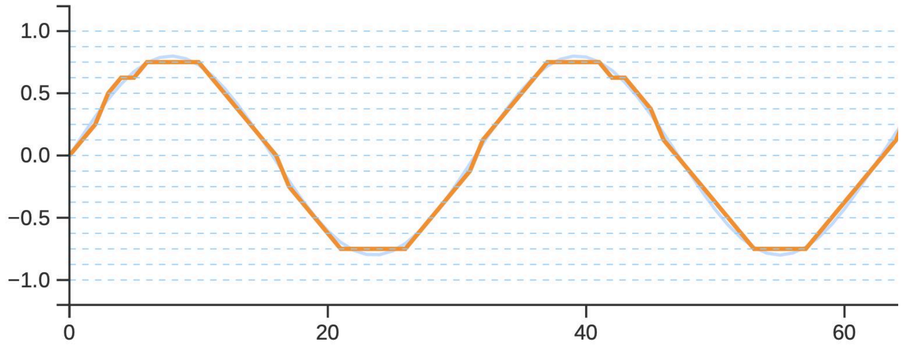
Nyquist-Shannon Sampling Theorem

To reconstruct a signal accurately, the **sampling rate** must be **at least twice** of the highest **frequency** in the audio

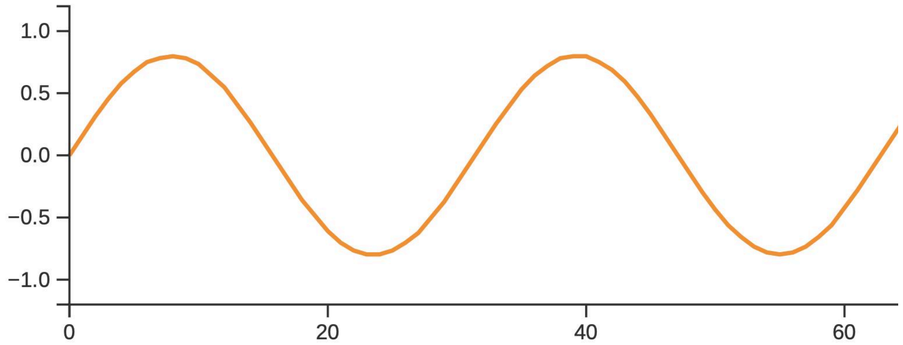
The process of mapping **continuous amplitude values** to a finite set of **discrete levels**

Bit depth is the number of **the discrete levels** (precision)

8-*bit* - 256 levels (old systems)
16-*bit* - 65,536 levels (CD quality)
24-*bit* - 16.7m levels (studio quality)



Low bit depth

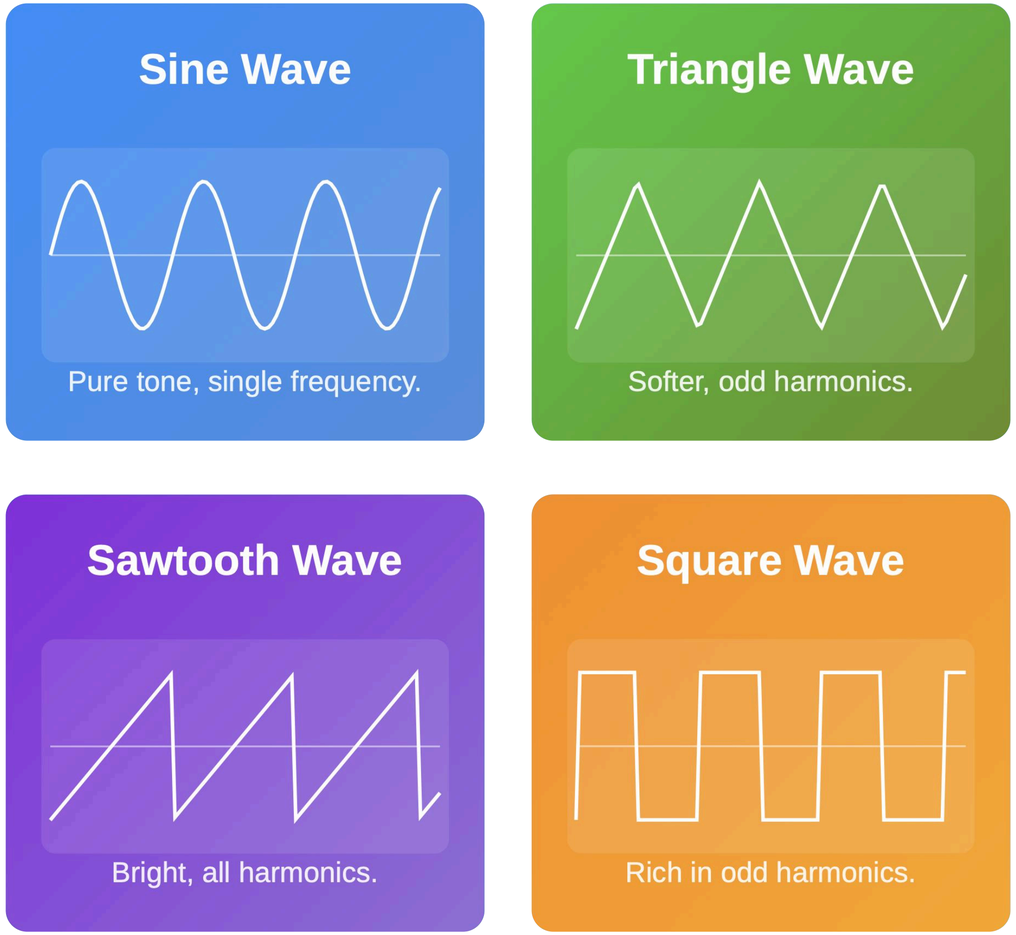


High bit depth

The time domain shows how the **amplitude** of a sound wave changes **over time**

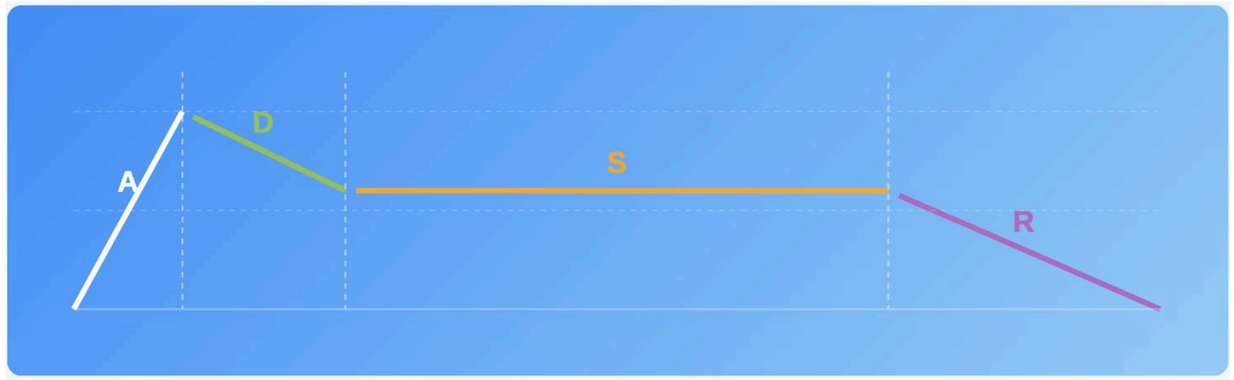
A **direct visualisation** of the sound signal as **human hears it**

Amplitude (loudness)
Duration (length)
Envelope (overall shape)



Describes how the **amplitude** of a sound wave changes **over time**

Envelope: Overall amplitude shape
Transients: Rapid amplitude changes
ADSR: Common envelope model



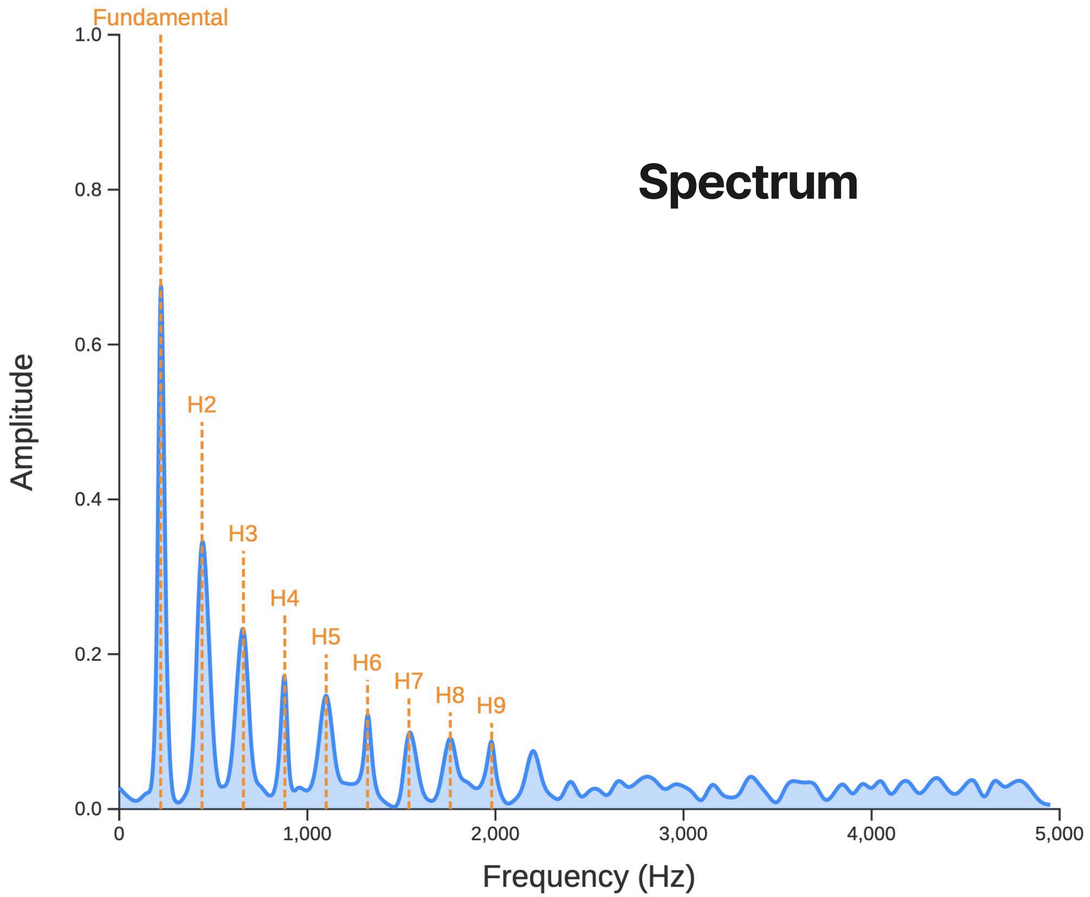
Attack: Initial rise from silence to peak
Decay: Fall from peak to sustain level
Sustain: Steady level but not silence
Release: Fade to silence at the end

The frequency domain shows the different **frequencies** present in a sound and their respective **amplitudes**

Fourier Transformation:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(t) \cdot exp(-2\pi i \cdot t\xi) dt$$

It helps us understand the **timbre** of a sound, revealing its **fundamental frequency** and **harmonics**

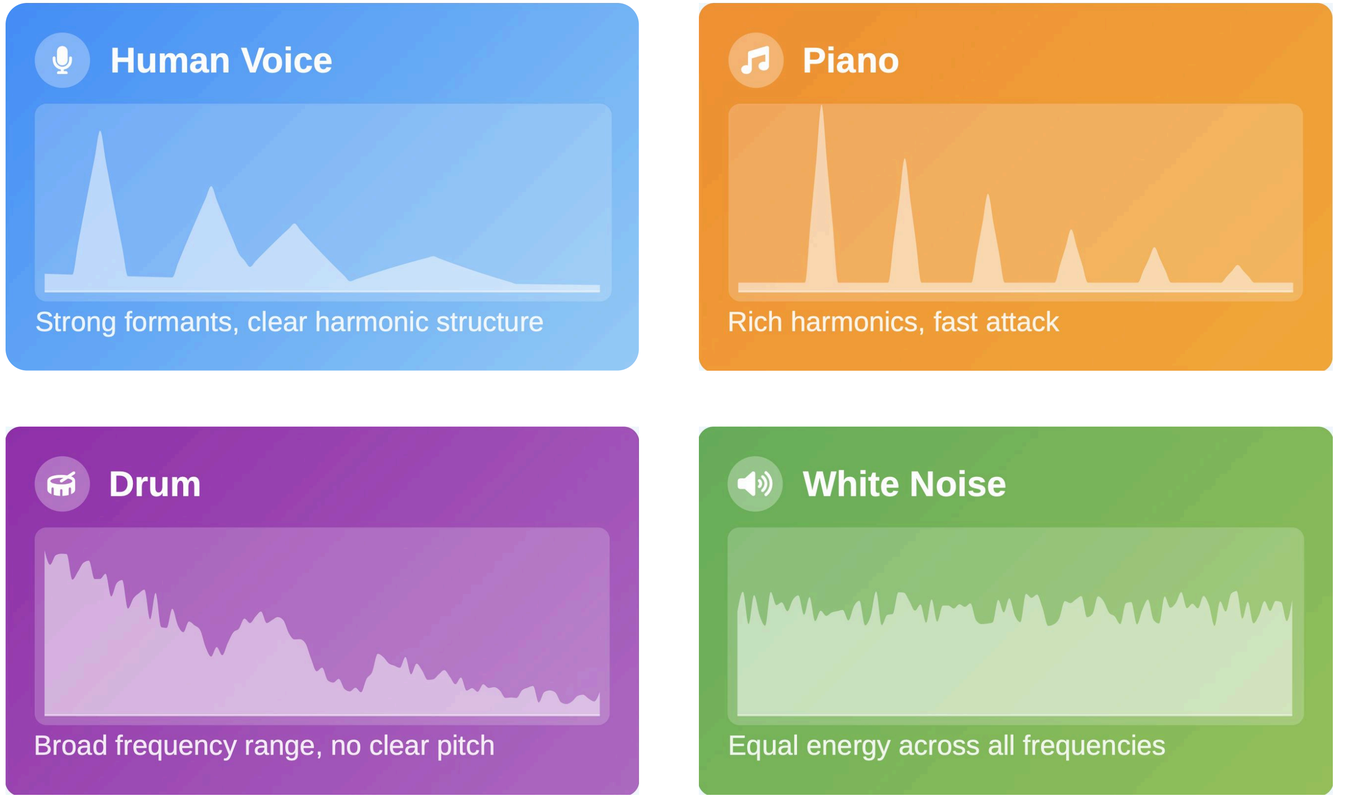


Different sounds have **unique spectral signatures** that help **identify** them

Instrument Recognition

Voice Activity Detection

Acoustic Scene Classification



Preprocessing

The process of **cleaning** and **transforming** *raw audio signals* before further analysis or processing

Why

Remove unwanted **noise** and **artifacts**

Normalise audio **features**

Extract **relevant parts** of the signal

Improve **model performance**

Reduce **computational** requirements

Challenges

Background **noise** and **interference**

Varying **recording conditions**

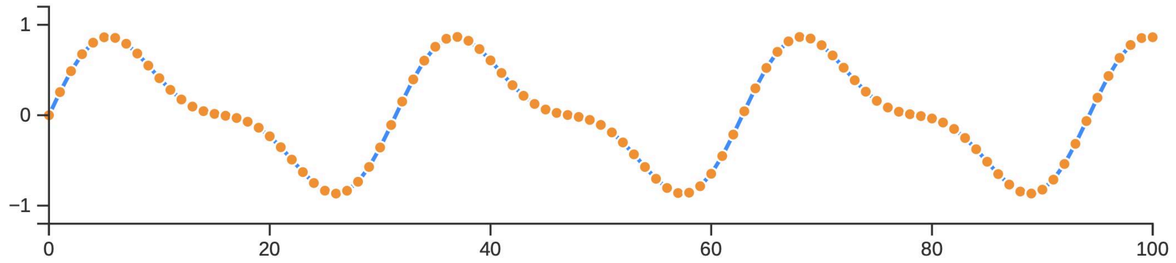
Different **sampling rates** and **formats**

Silence and **non-speech** segments

Reverberation and **echo**

Resampling / Noise Reduction / Voice Activity Detection / Normalisation

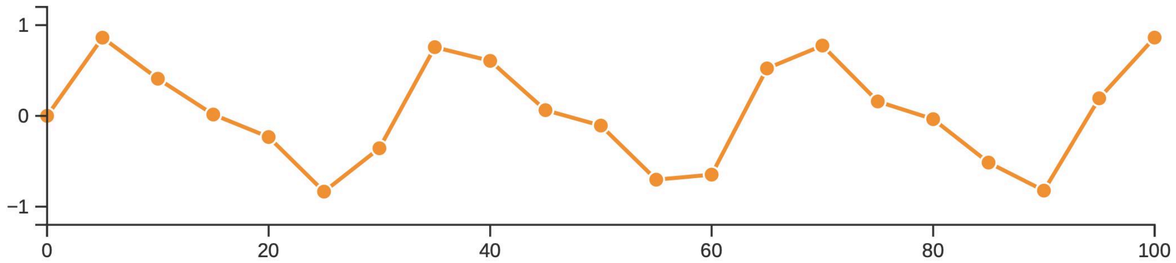
The process of changing the **sampling rate** of a digital audio signal



Downsampling: Decreasing sampling rate

Upsampling: Increasing sampling rate

Standardisation: Convert to a common rate



Don’t forget **Nyquist-Shannon Sampling Theorem**!

Noise Reduction

Remove unwanted sounds from an audio signal while **preserving the desired content**

Stationary: Fan noise, engine noise

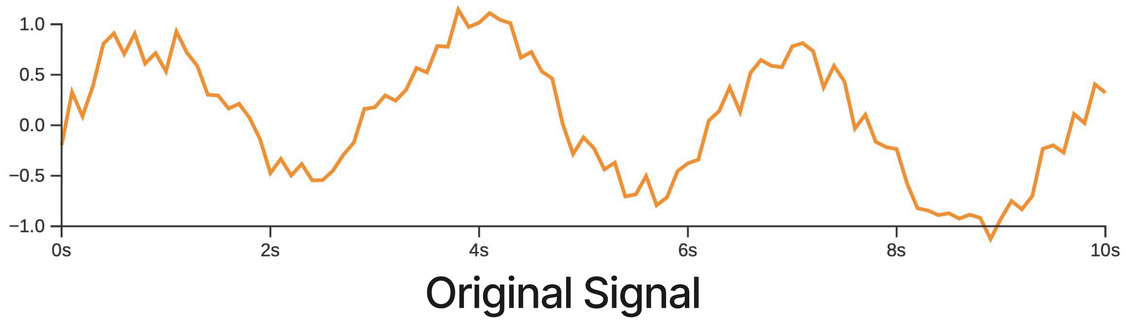
Non-stationary: Traffic sounds

Reverberation: Sound reflections in a space

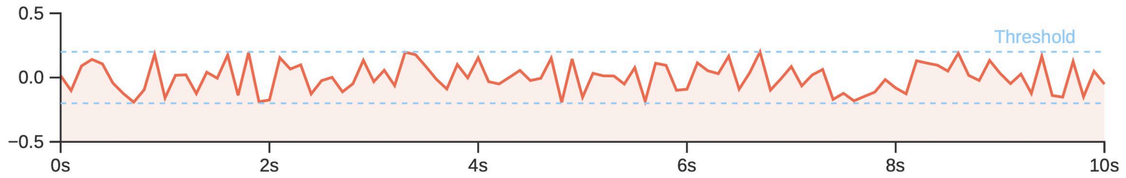
Spectral Subtraction

Filtering

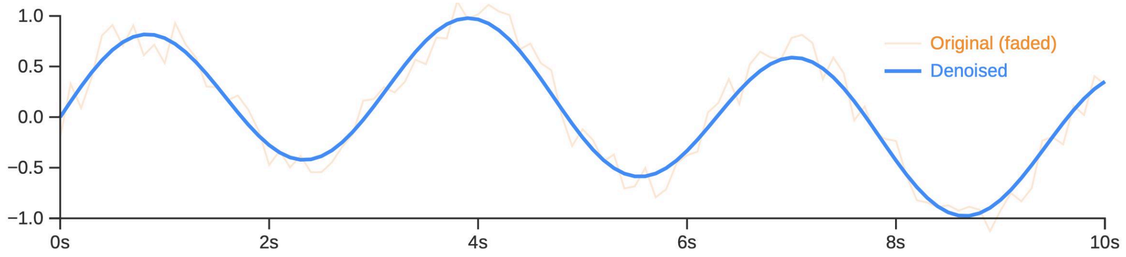
Deep-learning-based Methods



Original Signal

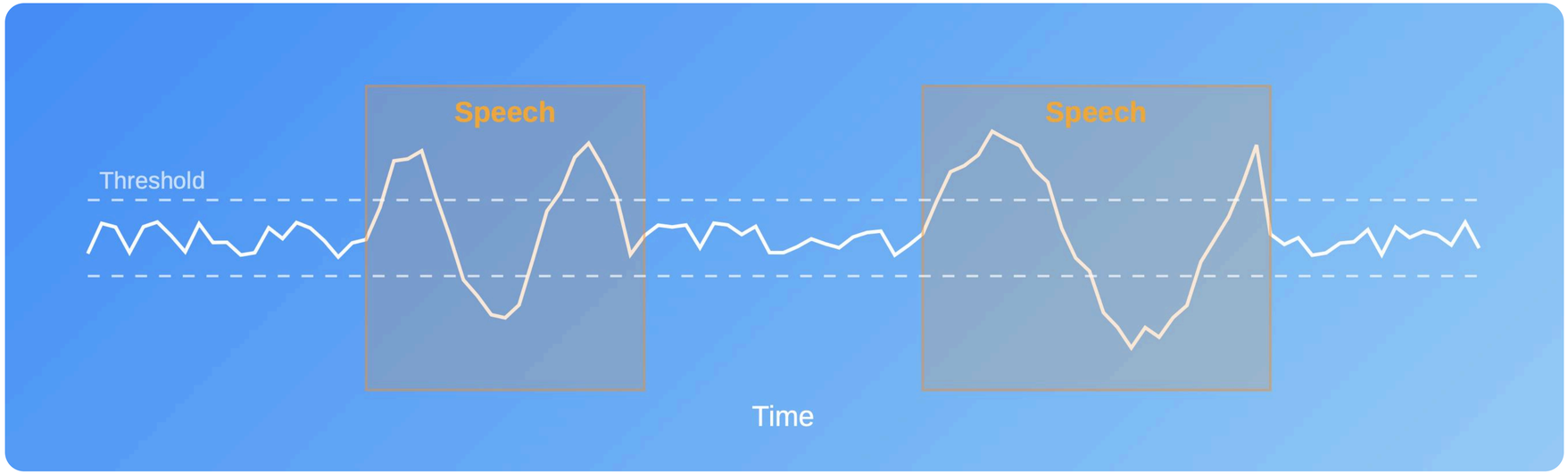


Noise Signal



Denoised Signal

Voice Activity Detection (VAD) identifies **speech segments** in audio signals



Reduces **processing time**

Improves **recognition accuracy**

Saves **storage** and **computational** complexity

Energy-based: Using amplitude to detect

Spectral-based: Using frequency to detect

Normalisation

The process of adjusting the **amplitude** of an audio signal to a **standard level**

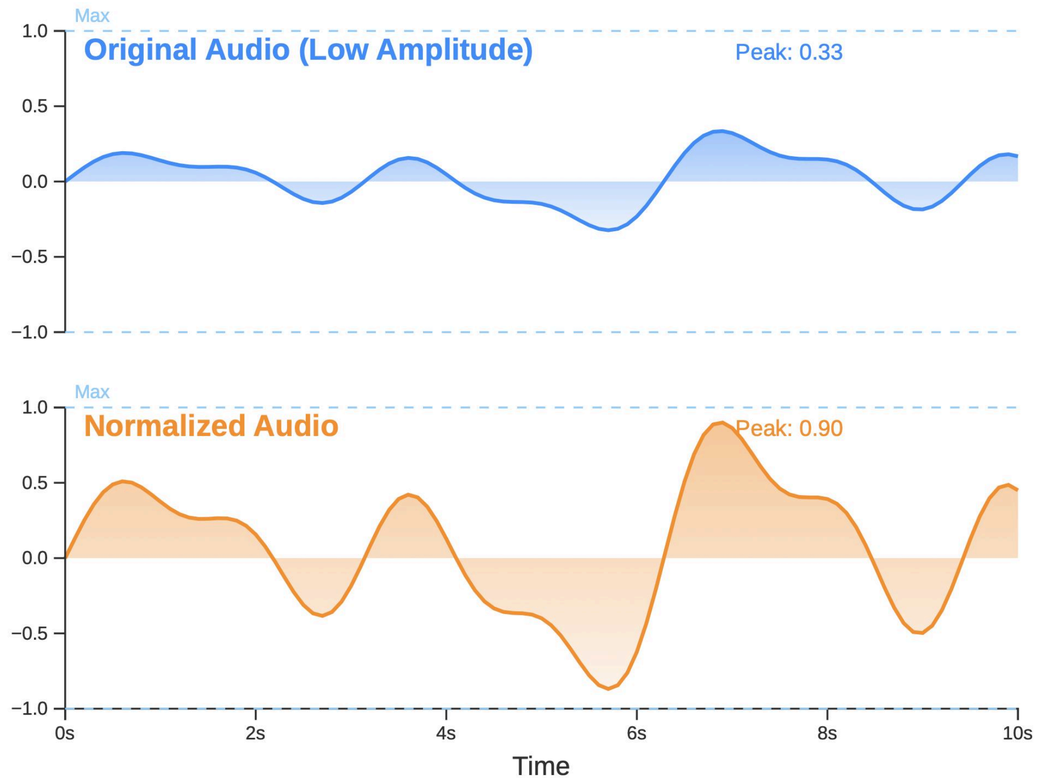
Consistent volume across recording

Improve **feature extraction** performance

Better and easier **modelling**

Min-max Normalisation

Mean-std Normalisation



Can be applied **in preprocessing** or **after feature extraction**

Feature Extraction

Audio features are **measurable characteristics** extracted from audio signals that **capture specific aspects** of the sound

Why

Reduce the **dimension** of raw audio

Capture perceptually **relevant** information

Enable **efficient** machine learning

Provide **interpretable** representations

Categories

Temporal features: Envelopes, energies

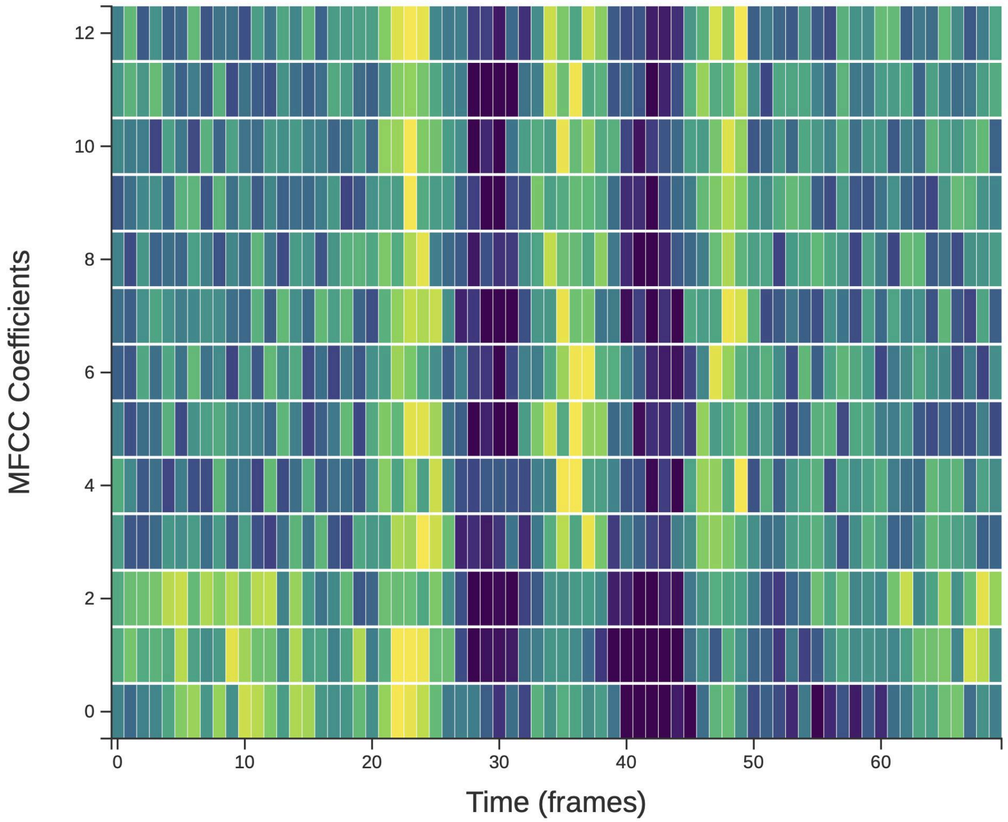
Statistical features: openSMILE

Spectral features: Spectrograms

Cepstral features: MFCCs

Mel-Frequency Cepstral Coefficients (MFCCs) are features that represent the **short-term power spectrum** of a sound

They are based on a **nonlinear Mel-scale** of frequency, which matches **human hearing perception**



A powerful and flexible **open-source toolkit** for extracting **acoustic features** from audio signals

Open-source **S**peech & **M**usic **I**nterpretation by **L**arge-space **E**xtraction

Comprehensive Feature Sets: From **low-level** descriptors to **high-level** statistical functionals

Standardised & Reproducible: **Pre-defined configurations** used in major academic challenges

High Performance: Written in C++ for fast, real-time processing suitable for live applications (also Python!)

Highly Configurable: **Custom** feature sets through simple **text-based** configuration files

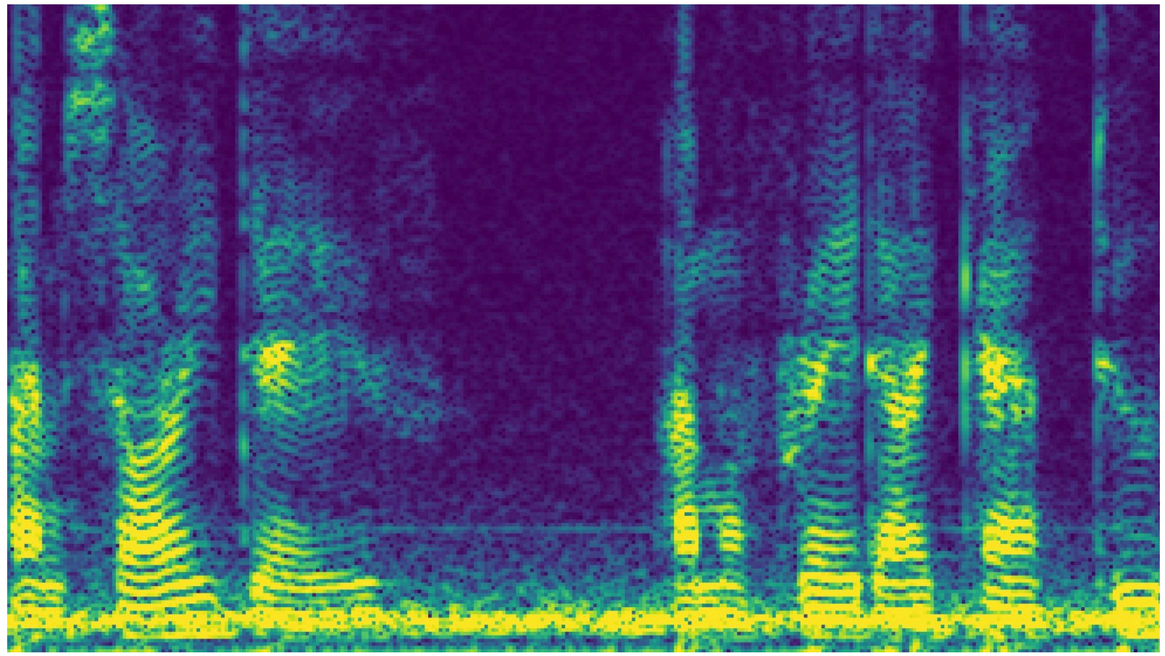
A feature to show **time**, **frequency** and **amplitude** at the same time?

Spectrogram

Less **information loss**

Commonly used in **deep learning**

Nearly lossless inverse process



Mel-spectrogram

Self-supervised learning frameworks for **speech representation** that learns from **unlabelled** speech

Pre-trained by large-size speech samples
Fine-tuned by **small amount** of labelled data
State-of-the-art on **ASR** tasks
Robust to different **accents** and **datasets**

Automatic Speech Recognition
Speech Translation
Speaker Identification
Speech Emotion Recognition

Commonly be applied as a **feature extractor** of speech samples

Applications

Audio Processing Applications

Speech & Languages

Automatic Speech Recognition

Speaker Diarisation

Language / Accent Recognition

Environment & Context

Acoustic Scene Classification

Sound Event Detection

Audio Surveillance

Music & Creativity

Music Information Retrieval

Music Source Separation

Automatic Music Transcription

Health & Biometrics

Voice Biometrics

Speech Emotion Recognition

Computational Paralinguistics

A journey from **audio physics** to **audio processing applications**

Preprocessing is the key: directly impacting the success of any analysis

Feature Extraction: help models can **understand**, **capturing** essential characteristics of sound

Self-Supervised Learning: learning **powerful** and **robust** representations

A World of Applications: a **wide range** of technologies we use in daily life

Questions?

semo

Empower every AI systems
to truly understand human emotions

Looking for More?

semo AI Co., Ltd.
info@semo.one