# MULTIMODAL AI FOR EDGE AI

## CVPR 2024 Tutorial

**The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024**

Seattle, WA, USA

# TUTORIAL AGENDA

1. Multimodal Perception
2. Gaze Correction
3. Hand Gestures Recognition
4. Sound Localization
5. Demos

GN

# GAZE CORRECTION

# INTRODUCTION AND CHALLENGES OF GAZE CORRECTION

# GAZE CORRECTION
## Overview

Gaze correction refers to using computer vision or artificial intelligence techniques to adjust the apparent direction of a person's gaze in digital video communication.

### Enhanced Engagement
It provides more natural and engaging interaction by simulating real eye contact

### Increased Attention
It can help maintain attention and focus during hybrid and virtual meetings

### Professional Appearance
It ensures that speakers appear to be addressing their audience directly in virtual presentations

### Improved Comprehension
Speakers may enhance the audience's ability to understand the discussed content

# VIRTUAL MEETING EXAMPLE

Original Video

# VIRTUAL MEETING EXAMPLE

## Gaze Correction


NVIDIA MAXINE


VIRTUAL MEETING EXAMPLE
Original Video vs. Gaze Correction
NVIDIA
MAXINE

# CHALLENGES IN EDGE AI

## Implementing Gaze Correction in Edge AI

Running complex AI models for gaze correction requires efficiently using the limited AI components resources without compromising performance

### Real-Time

It must process video in real-time to ensure the gaze is corrected without noticeable lag

### Model Optimization

It must be compressed and optimized without significant loss of accuracy

### User Diversity

The model must be trained on diverse datasets to work accurately across different ethnicities, genders and ages

### Integration

The gaze correction application must integrate seamlessly with existing video conferencing platforms and camera hardware

# CHALLENGE OF GAZE CORRECTION

Technical Challenges

# CHALLENGE OF GAZE CORRECTION

Technical Challenges

# HOW GAZE CORRECTION WORKS

## Deep Learning for Gaze Correction

## GAN

### Generative Adversarial Networks

GANs are used in gaze correction to generate realistic eye images that match the desired gaze direction.

## WARPING

### Warping Neural Networks

Warping techniques adjust the eye region in an eye region to redirect the gaze, creating the impression of eye contact.

# SYSTEMATIC REVIEW ON GAZE CORRECTION MODELS

## Recent Published Models

- *Isikdogan et al. (2020),* Eye Contact Correction using Deep Neural Networks.
- *Hsu et al. (2019),* Look at Me! Correcting Eye Gaze in Live Video Communication.
- *He et al. (2019),* Photo-Realistic Monocular Gaze Redirection using Generative Adversarial Networks.
- *Kononenko et al. (2018),* Photorealistic Monocular Gaze Redirection using Machine Learning.
- *Kaur and Manduchi (2021),* Subject Guided Eye Images Synthesis with Application to Gaze Redirection.

# JABRA GAZE CORRECTION MODEL FOR EDGE AI

# JABRA EYE CORRECTION

## Jabra PanaCast 20

Jabra Eye Correction model simulates direct eye contact between participants of online meetings, enhancing the sense of engagement and personal connection during remote interactions. The model can be deployed directly in our Jabra Business Collaboration products.

### Speed

The model runs at 30 frames per second with eye images of 64x48 resolution

### Eye Shape

The model generates good eye feature shape, especially iris contour and eyelid shape

### Color Composition

The model reconstructs the eye colors and skin color in the processed eye region

# HOW TO BUILD A GAZE CORRECTION APPLICATION?

## Edge AI Deployment

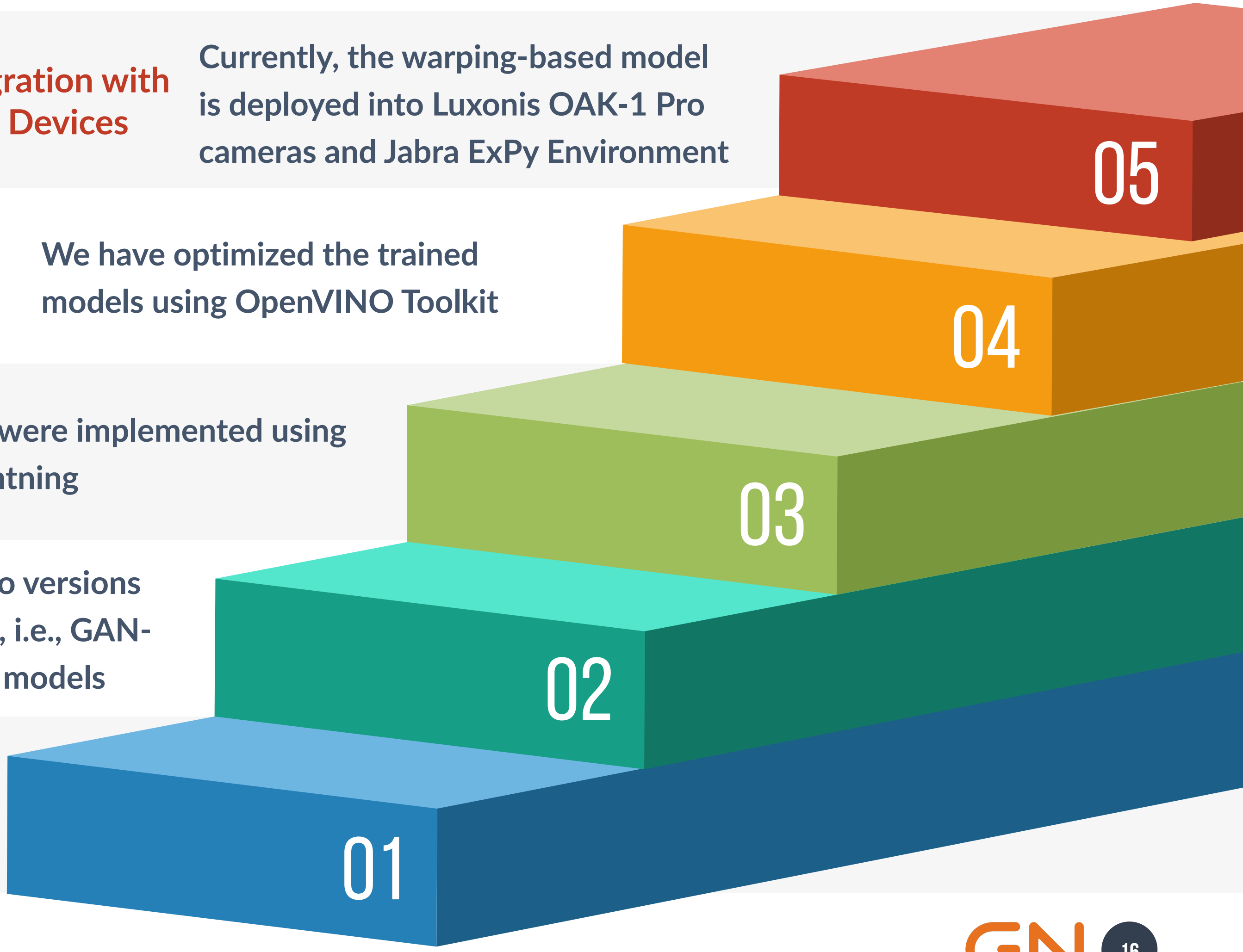
**Integration with Edge Devices** — Currently, the warping-based model is deployed into Luxonis OAK-1 Pro cameras and Jabra ExPy Environment

**Model Optimization** — We have optimized the trained models using OpenVINO Toolkit

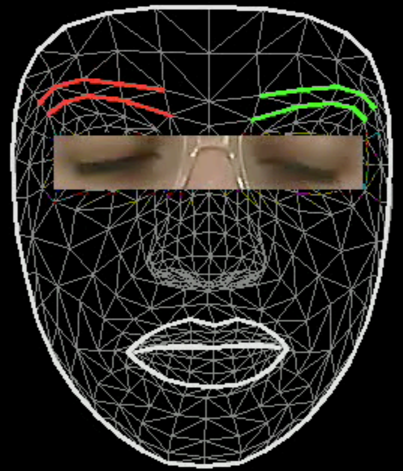
**Model Training** — The models were implemented using PyTorch Lightning

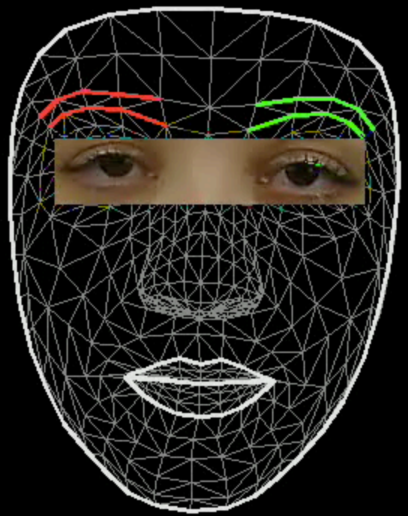
**Model Selection** — We have implemented two versions of gaze correction models, i.e., GAN-based and warping-based models

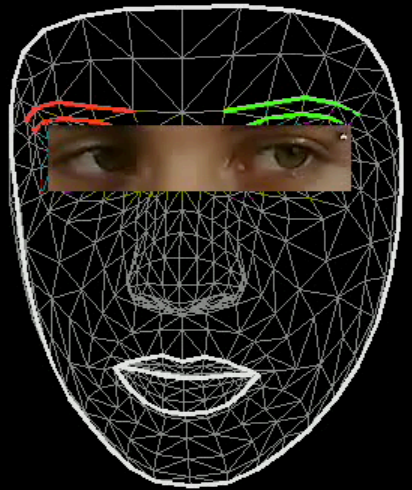
**Data Collection and Annotation** — A dataset from various angles, lighting conditions, and with different eye shapes and colors




UNITYEYES-JABRA DATASET
Eye Contact Dataset with Synthetic Data

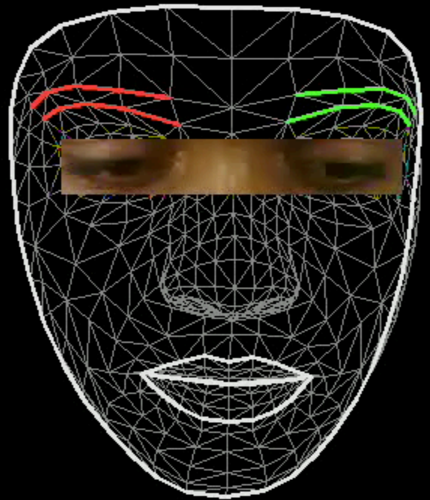


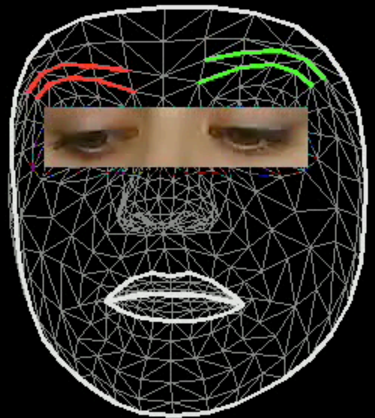


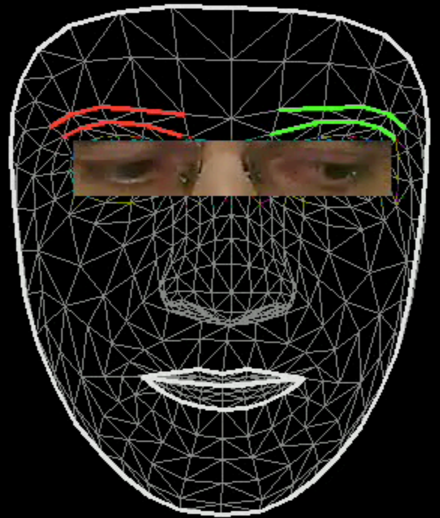


# IFSP-JABRA DATASET

## Eye Contact Dataset with Real Data

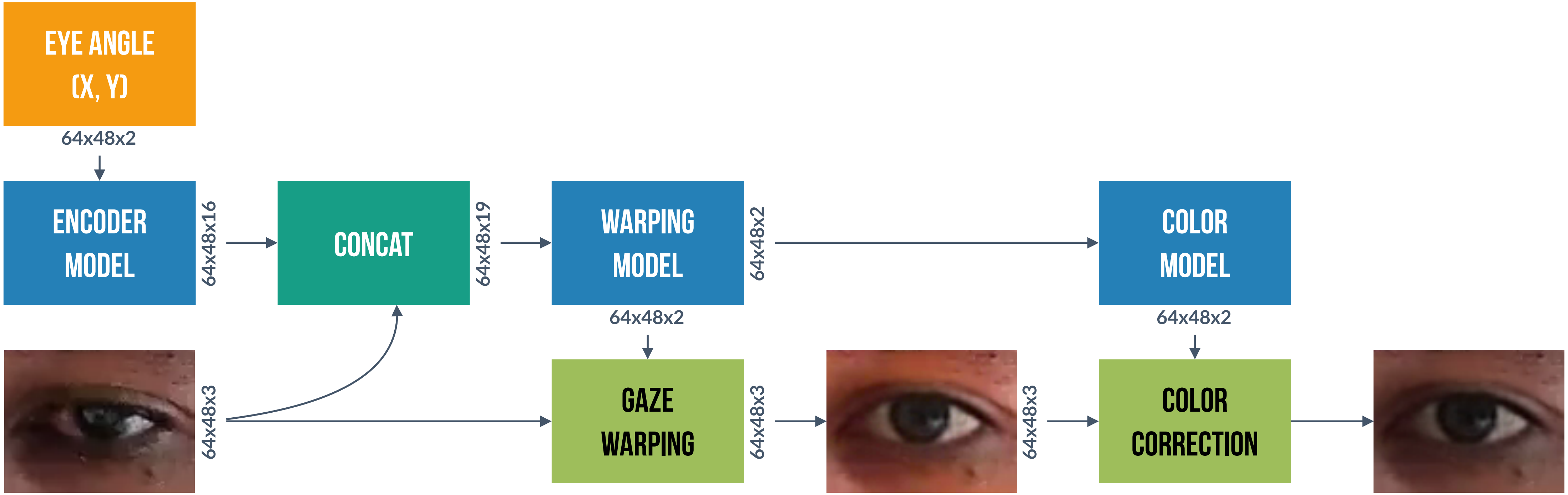






# JABRA GAZE CORRECTION MODEL ARCHITECTURE

JECModel




- ML Models
- PyTorch Methods
- CV Algorithm
- Input Data

# JABRA GAZE CORRECTION MODEL

Video Example (Gaze Correction)

# JABRA GAZE CORRECTION MODEL

## Video Example (Gaze Correction + Beautification)

# HAND GESTURES RECOGNITION

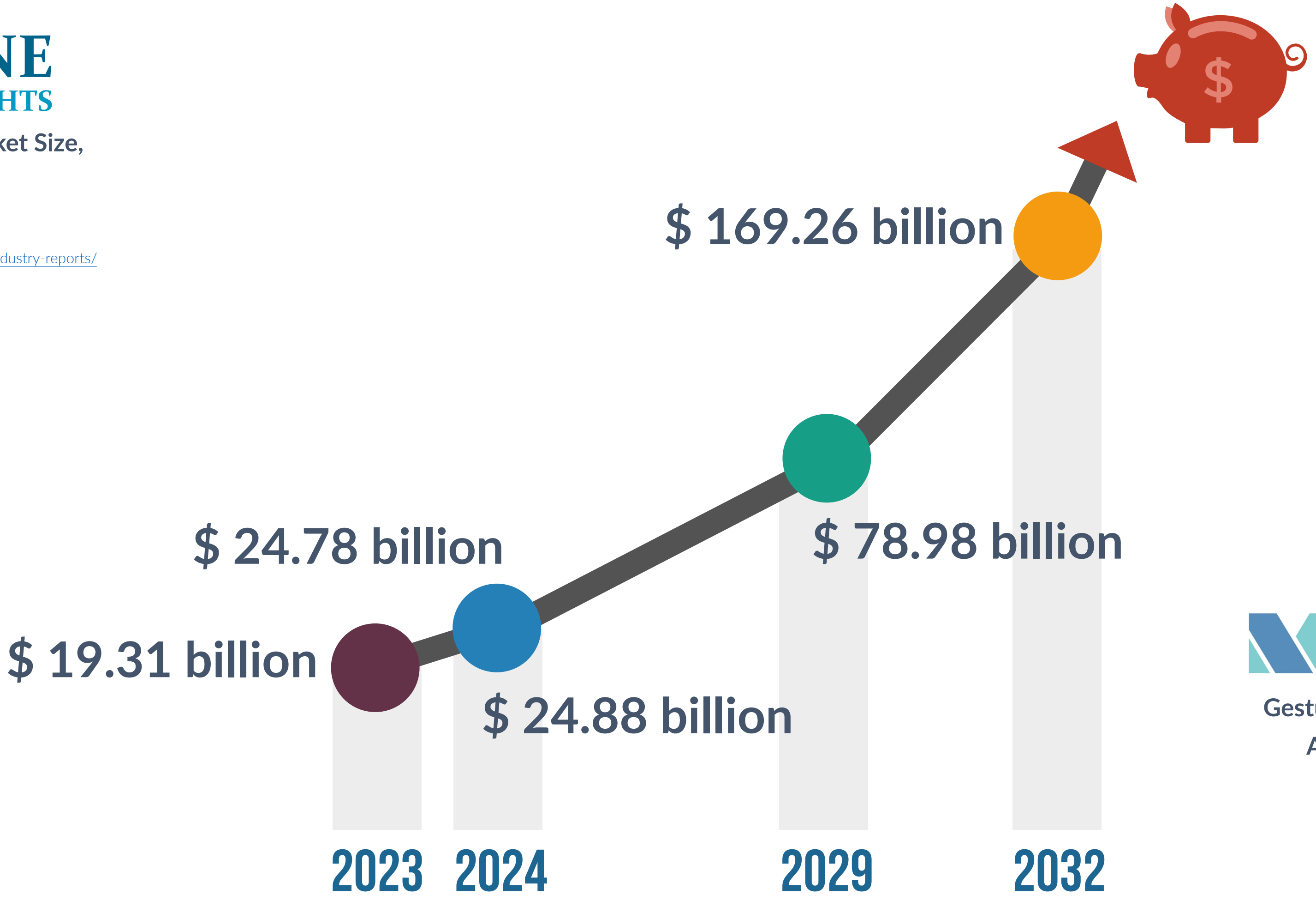
# WHY HAND GESTURES

## Is hand gesture recognition still a thing?

FORTUNE BUSINESS INSIGHTS

Gesture Recognition Market Size, Share & Industry Analysis (2024 – 2032)

https://www.fortunebusinessinsights.com/industry-reports/gesture-recognition-market-100235




https://www.mordorintelligence.com/industry-reports/gesture-recognition-market

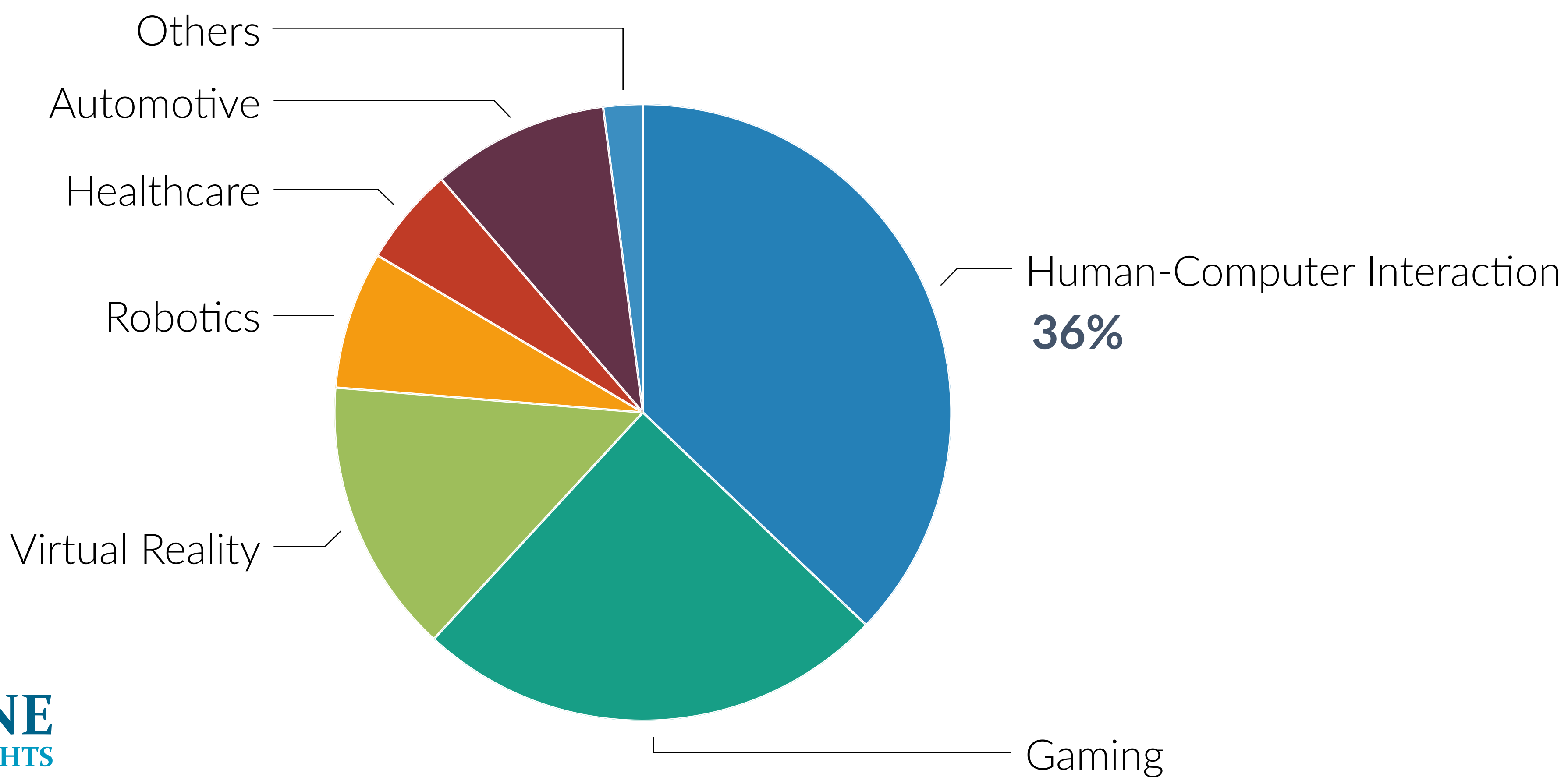
Mordor Intelligence

Gesture Recognition Market Size & Share Analysis - Growth Trends & Forecasts (2024 - 2029)

**Global gesture recognition market size estimation and projection**

# WHY HAND GESTURES

## Is hand gesture recognition still a thing?




FORTUNE BUSINESS INSIGHTS

**Global gesture recognition market share, by application segment, 2023**

# WHY HAND GESTURES

Hand gestures are everywhere

### AGRICULTURE AND INDUSTRY

Equipment Operation
On-Site Inspections

### HEALTHCARE

Surgical Assistance
Rehabilitation
Assistive Devices

### SMART HOME AND ENTERTAINMENT

Interactive Media Control
Home Automation and Control
Security System
Access Control

### ROBOTICS AND AUTOMATION

Service Robots
Collaborative Robots
Machine Control
Quality Inspection
Training Simulations

### HUMAN-COMPUTER INTERACTION

Assistive Technology
Interactive Learning
Virtual Classrooms
Special Education

### FITNESS AND SPORTS

Exercise Monitoring
Sports Training

### PUBLIC INTERACTION

Interactive Displays
Kiosks and Information Terminals
Exhibits and Museums
Transportations Hubs

### AUTOMOTIVE

In-Car Infotainment Systems
Driver Assistance

### VIRTUAL REALITY AND GAMING

Virtual Reality (VR)
Augmented Reality (AR)

### CONSUMER

Smartphones and Tablets
Virtual Assistants
Selfie cameras

Several Fields of Applications

# HAND GESTURE PRODUCTS

## Example in different industries

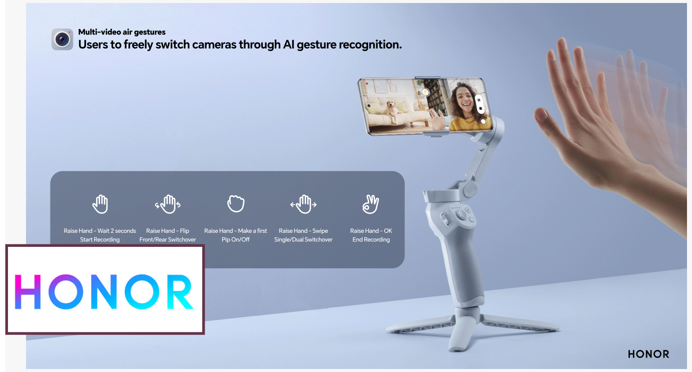

Leap Motion Controller 2


HoloLens 2


Echo Show


Gesture Control Armband


AIR Neo Selfie Pocket Drone




HONOR Cellphone Camera

# BENEFITS OF HAND GESTURES

## Overview

HG supports immersive experiences of entertainment and control by providing more natural and engaging ways to interact with digital environments, systems and devices.

### Enhances User Experience

Provides multimodal interaction methods, making systems more user-friendly and versatile.

### Promotes Accessibility

Offers alternative communication methods for individuals with disabilities, enhancing inclusivity and usability.

### Enables Touchless Control

Enables hygienic interaction by eliminating the need for physical contact, ideal for public and shared environments.

### Increases Efficiency

Allows for quick and efficient execution of commands through simple gestures, reducing reliance on traditional input devices.

# HAND-BASED TECHNOLOGY

## General view

Hand-based technology uses cameras or other sensors to capture the users’ hand gestures and movements.

Algorithms or Machine Learning models then analyze and interpret the hand poses or performances from the captured data.

HAND DETECTION → HAND TRACKING → GESTURE RECOGNITION

General View for Hand Gesture Technology

# HAND GESTURE RECOGNITION

## Looking into the pipeline process

HAND DETECTION — HAND TRACKING — GESTURE RECOGNITION

### DATA AQCUISITION

Capturing data using one or more sensors

### PREPROCESSING DATA

Processing the raw data to focus on the relevant part

### FEATURE EXTRACTION

Identifying and isolating key characteristics

### GESTURE CLASSISIFCATION

Classifying the extracted features into predefined gesture categories

### POST PROCESSING

Refining the classification results and mapping the gestures to actions

### INTERFACE FEEDBACK

Providing system feedback to the user

# HAND GESTURE RECOGNITION

General pipeline process

Cameras
Optical Sensors
Ultrasonic Sensors
Motion Sensors
Electromyography Sensors
Infrared Sensors

### TEMPORAL HAND GESTURE CLASSIFICATION (IMAGE BASED)

**STATIC**

Postures/Poses

Hand Shapes

**DYNAMIC**

Performed Gestures

Hand Movements

DATA AQCUISITION

Sensor types

PREPROCESSING DATA

FEATURE EXTRACTION

GESTURE CLASSISIFCATION

POST PROCESSING

INTERFACE FEEDBACK

# HAND GESTURE RECOGNITION

General pipeline process

Data Cleaning
Normalization
Segmentation
Data Augmentation
Color Space Conversion
Cropping and Resizing

## EXAMPLE PROCESS FOR PROCESSING DATA (IMAGE BASED)

### SUGGESTED STEPS

Noise Reduction
Background Subtraction
Thresholding
Contour Detection
ROI Extraction
Resizing
Grayscale Conversion

DATA AQCUISITION

**PREPROCESSING DATA**

Preparing data

FEATURE EXTRACTION

GESTURE CLASSISIFCATION

POST PROCESSING

INTERFACE FEEDBACK

# HAND GESTURE RECOGNITION

## General pipeline process

Shape-Based

Texture-Based

Keypoint Detection

Geometric

Motion-Based

Spatial and temporal

### COMMON LIBRARIES / FRAMEWORKS

#### FEATURE EXTRACTION

OpenCV

scikit-image

Keras

Tensor-Flow

MediaPipe

Kornia (for Pytorch)

dlib

MathLab (Image Processing Toolbox)

DATA AQCUISITION

PRE PROCESSING DATA

**FEATURE EXTRACTION**

Feature types/Techniques

GESTURE CLASSISIFCATION

POST PROCESSING

INTERFACE FEEDBACK

# HAND GESTURE RECOGNITION

## General pipeline process

- Deep Learning Ensembles
- Artificial Neural Networks (ANN)
- 3D Convolutional Neural Networks (3D-CNN)
- Support Vector Machines (SVM)
- Convolutional Neural Networks (CNN)
- Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)

### COMMON AVAILABLE MODELS

### FRAMEWORKS / LIBRARIES

Keras

Google Tensor-Flow

PyTorch

scikit-learn

MediaPipe

DATA AQCUISITION

PRE PROCESSING

FEATURE EXTRACTION

**GESTURE CLASSISIFCATION**

Inference models

POST PROCESSING

INTERFACE FEEDBACK

# HAND GESTURE RECOGNITION

## General pipeline process

### ENHANCE ACCURACY AND RELIABILITY OF THE RECOGNIZED GESTURES

**THECNIQUES / TOOLS**

Statistical Methods

Confidence Interval Calculation

Minimizing the mean of the squared error

Outlier Detection

Reinforcement Learning

- Filtering and Smoothing
- Error Correction
- Gesture Segmentation
- Gesture Mapping
- Event Detection Gesture
- Adaptation and Learning

DATA AQCUISITION

PRE PROCESSING

FEATURE EXTRACTION

GESTURE CLASSISIFCATION

**POST PROCESSING**

Recognized gestures

INTERFACE FEEDBACK

# HAND GESTURE RECOGNITION

## General pipeline process

### PERCEPTIVE RESPONSE TO THE USER

### MAPPED TO SPECIFIC ACTIONS

Interacting with the UI

Controlling hardware

Sending commands to other applications or devices

Confirm that the gesture has been recognized

Triggering Actions

Saving Data

Feedback Generation

DATA AQCUISITION

PRE PROCESSING

FEATURE EXTRACTION

GESTURE CLASSISIFCATION

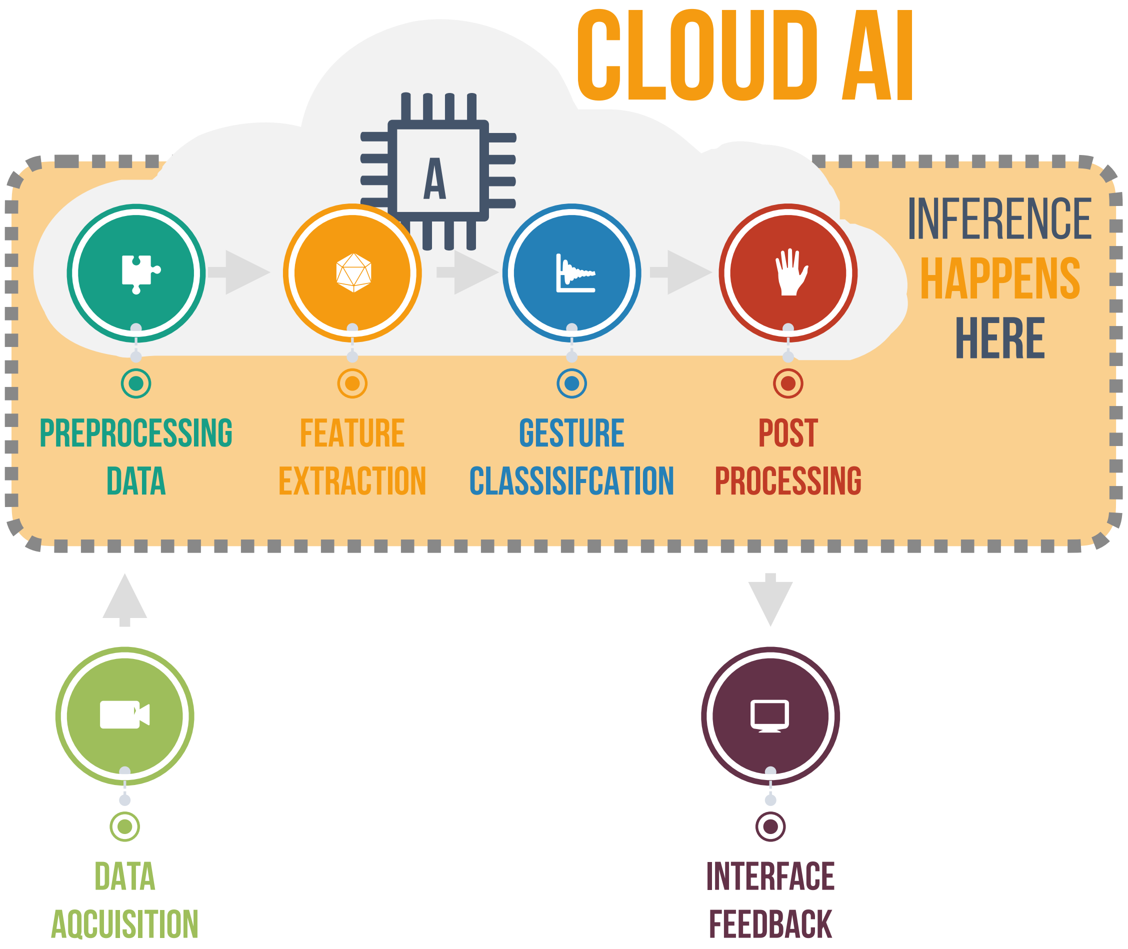
POST PROCESSING

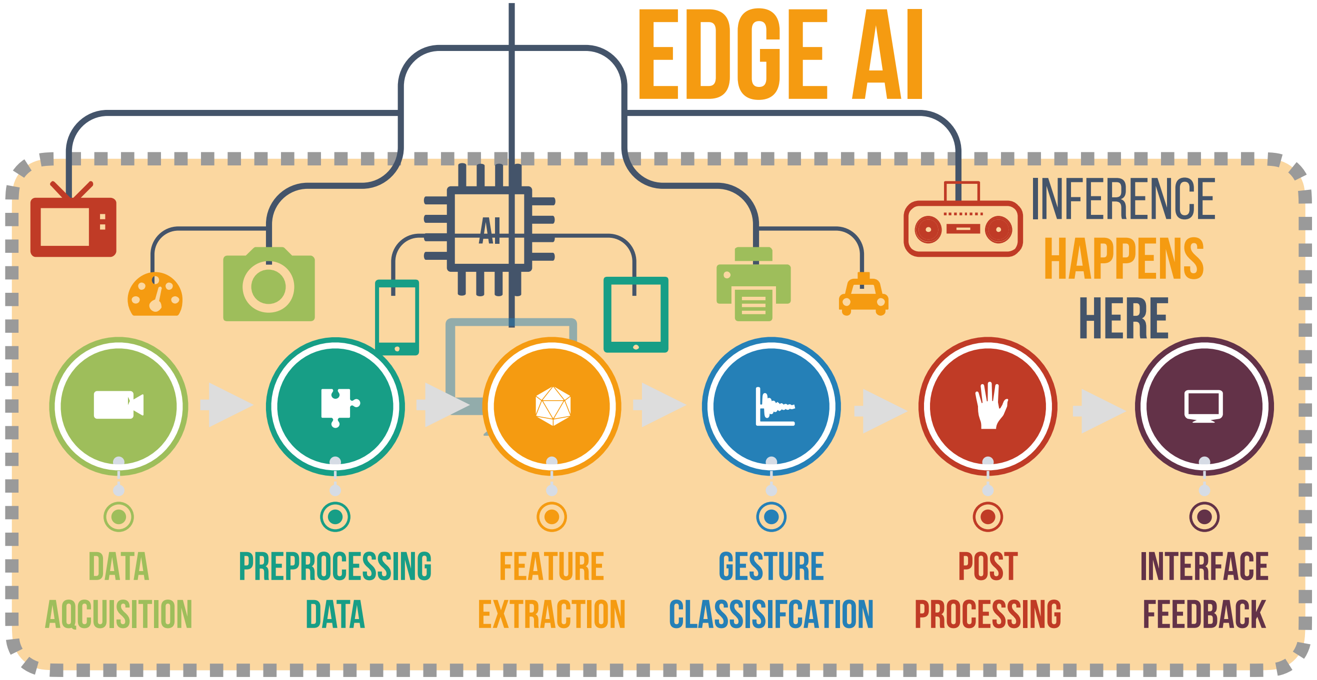
INTERFACE FEEDBACK

Interaction / Response

# HAND GESTURE RECOGNITION

Cloud versus Edge AI



# CHALLENGES IN HAND GESTURES

## Technical problems

Improving performance in these areas is essential for making hand gesture recognition systems more practical, reliable, and widely applicable in real-world scenarios.

### Datasets x Data Privacy

Ensuring datasets used for training gesture recognition models are diverse and representativity

### Model Size

It must be compressed and optimized without significant loss of accuracy

### Real-Time Processing

Low-latency processing to provide immediate feedback and smooth interaction in real-time applications

### Gesture Vocabulary

Common shared hand gestures vocabulary for contexts or systems actions

# CHALLENGES IN HAND GESTURES

## Cross-cutting problems

The most critical challenges in hand gesture recognition today include

### HG Education

Is it enough to rely on users' experience and intuitiveness?

### Fluidity

Depends on the perfect integration between the user and the system

### Cultural Prism

Hand gesture recognition must account for the cultural prism, as the meaning and interpretation of gestures can vary significantly across different cultures.

### Shared Vocabulary

A lack of shared vocabulary in hand gesture recognition can lead to inconsistencies and misunderstandings, as different systems and users may interpret gestures differently.

# HAND GESTURE EDGE AI DEMO

## Volume Control

**CAMERA**
Camera Luxonis OAK-1 MAX

**AI MODEL FORMAT**
MyriadX blob format

**PALM DETECTOR / HAND LANDMARK TRACKING**
Google MediaPipe (Blob)

# SOUND LOCALIZATION

# SOUND LOCALIZATION
## Motivation

Sound location models involve identifying the spatial position of sound-emitting objects within an image or a video to localize auditory cues.

### Current Experience
A sound location model that incorporates direction of arrival and head/body detection.

### ML Models
Developing a separate machine learning model tailored explicitly for audio and video.

### Multimodal Interaction
Find a multimodal solution or application with one neural network that inputs both audio and video components.

### Edge AI
Optimizing the sound location model to be specifically tailored for efficient and effective use on edge devices

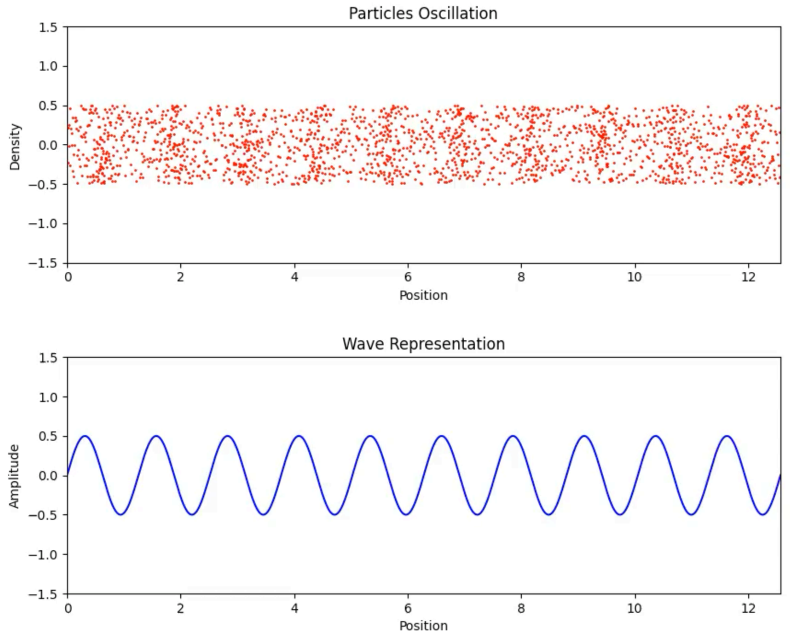
# AUDIO PROCESSING

## Modality

01 Sound is produced by vibration of an object.

02 This causes air molecules to oscillate.

03 Change in air pressure into the wave.

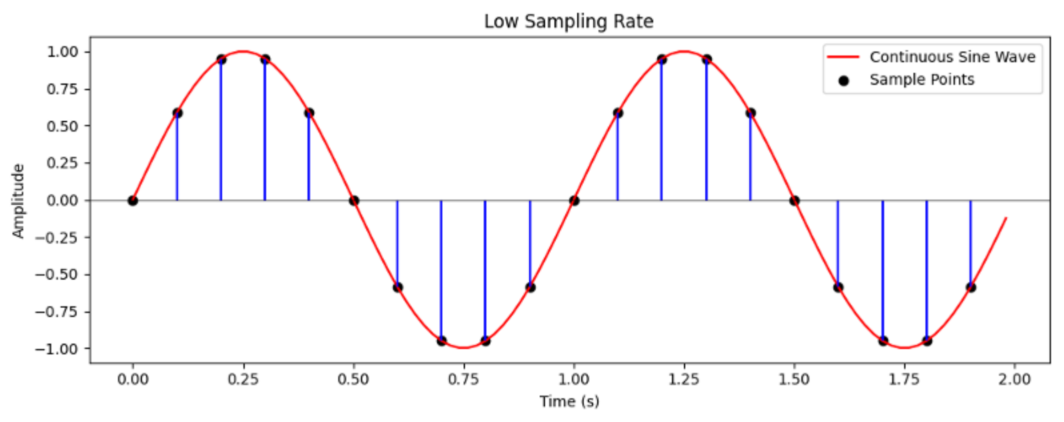


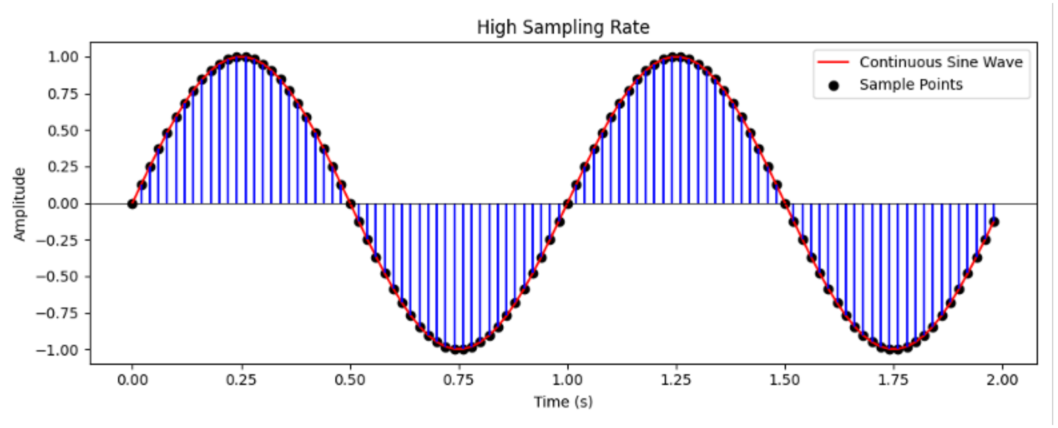

PARTICLES AND WAVE OSCILLATION

# AUDIO PROCESSING

## Sampling

"Sampling in audio processing involves capturing and converting continuous audio signals into discrete digital data points at regular intervals."

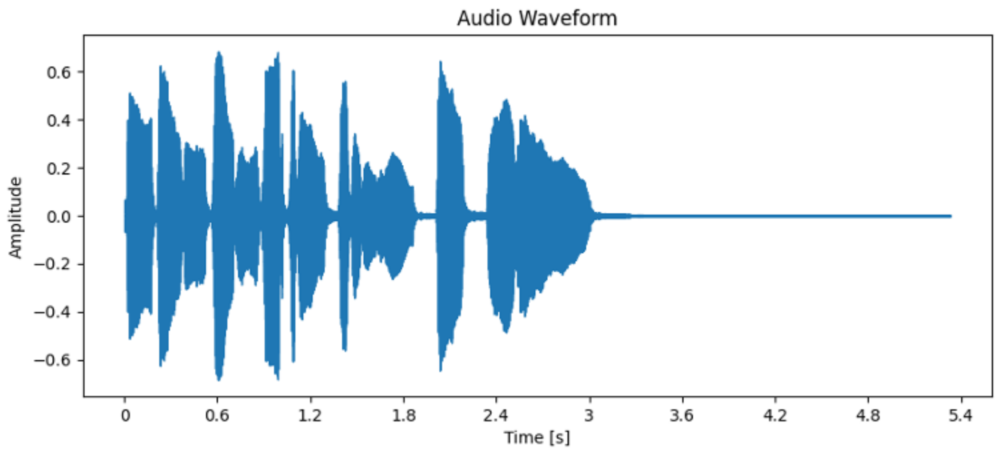


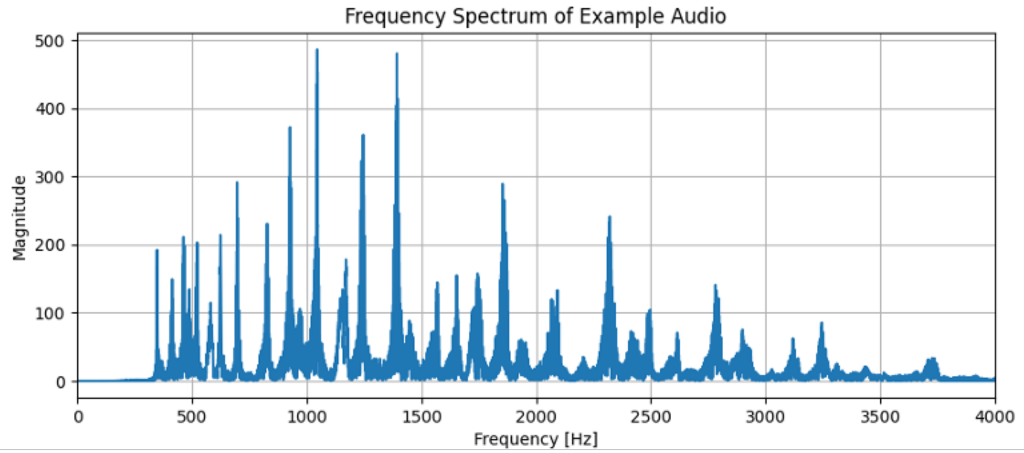




- Represented with sample points
- Better quality with higher sampling rate
- Similar to pixel in images
- Quantization

# AUDIO REPRESENTATION IN DIFFERENT DOMAINS

## Examples




### Waveform

Waveform is a graphical representation of the audio signal in the time domain.

### Frequency Spectrum

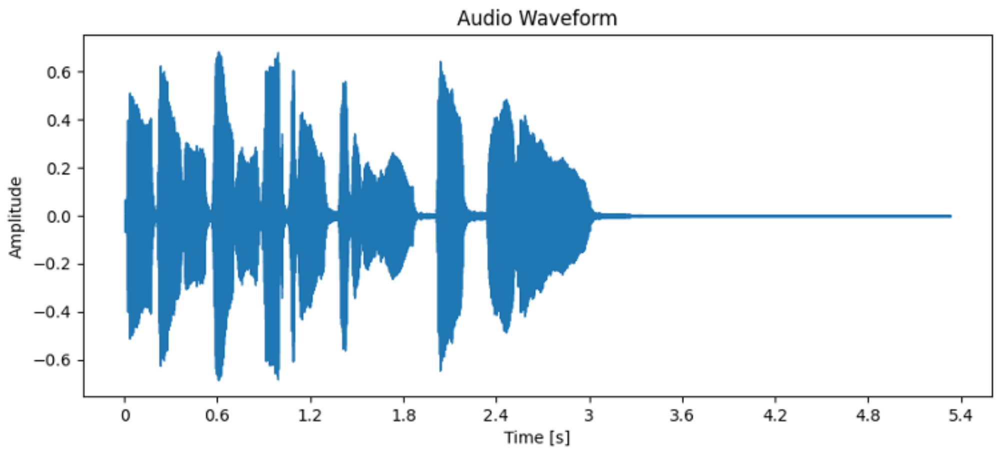
Frequency Spectrum is obtained using the Fast Fourier Transform.

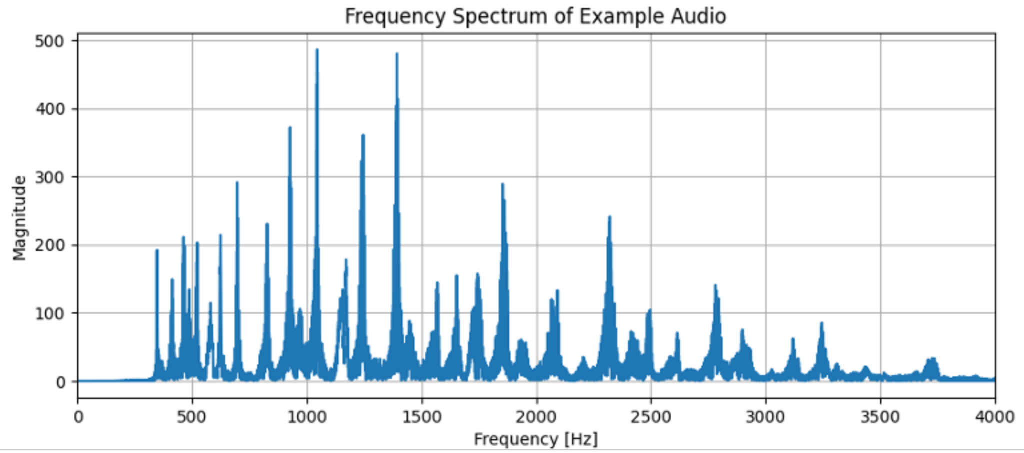
### FFT

Fast Fourier Transform is a representation in the Frequency domain.

# AUDIO REPRESENTATION IN DIFFERENT DOMAINS

## Examples




### Waveform

Waveform is a graphical representation of the audio signal in the time domain.

### Frequency Spectrum

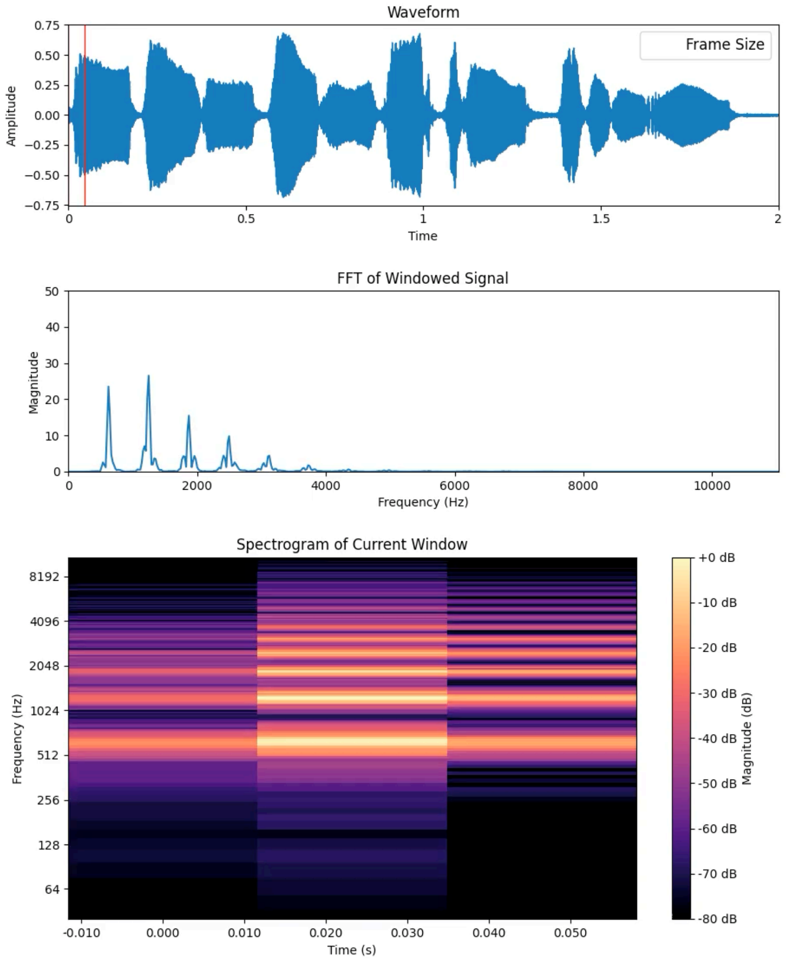
Frequency Spectrum is obtained using the Fast Fourier Transform.

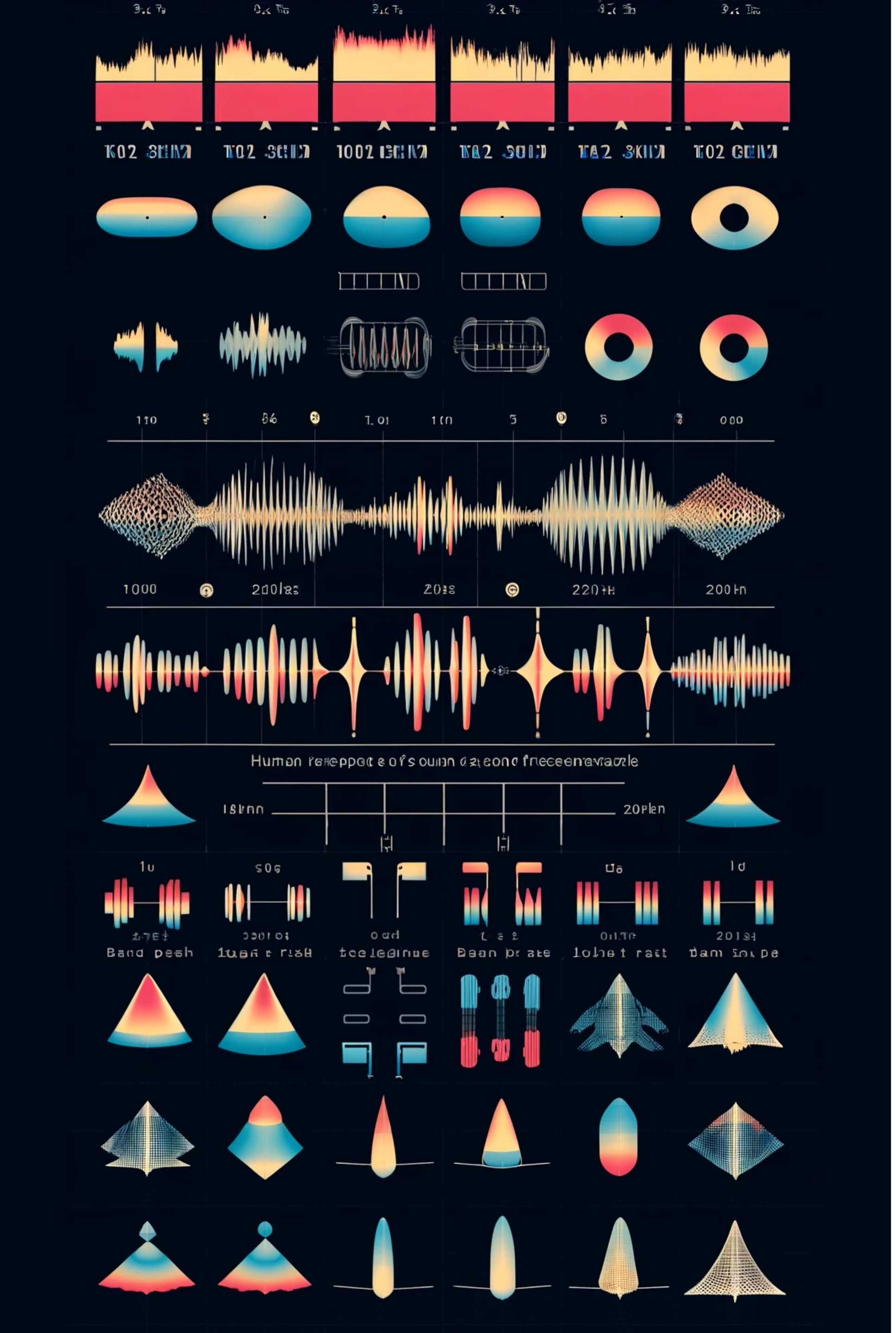
### FFT

Fast Fourier Transform is a representation in the Frequency domain.

# TIME-FREQUENCY DOMAIN

## Windowing to Spectrogram



# PERCEPTION OF SOUND

## Overview

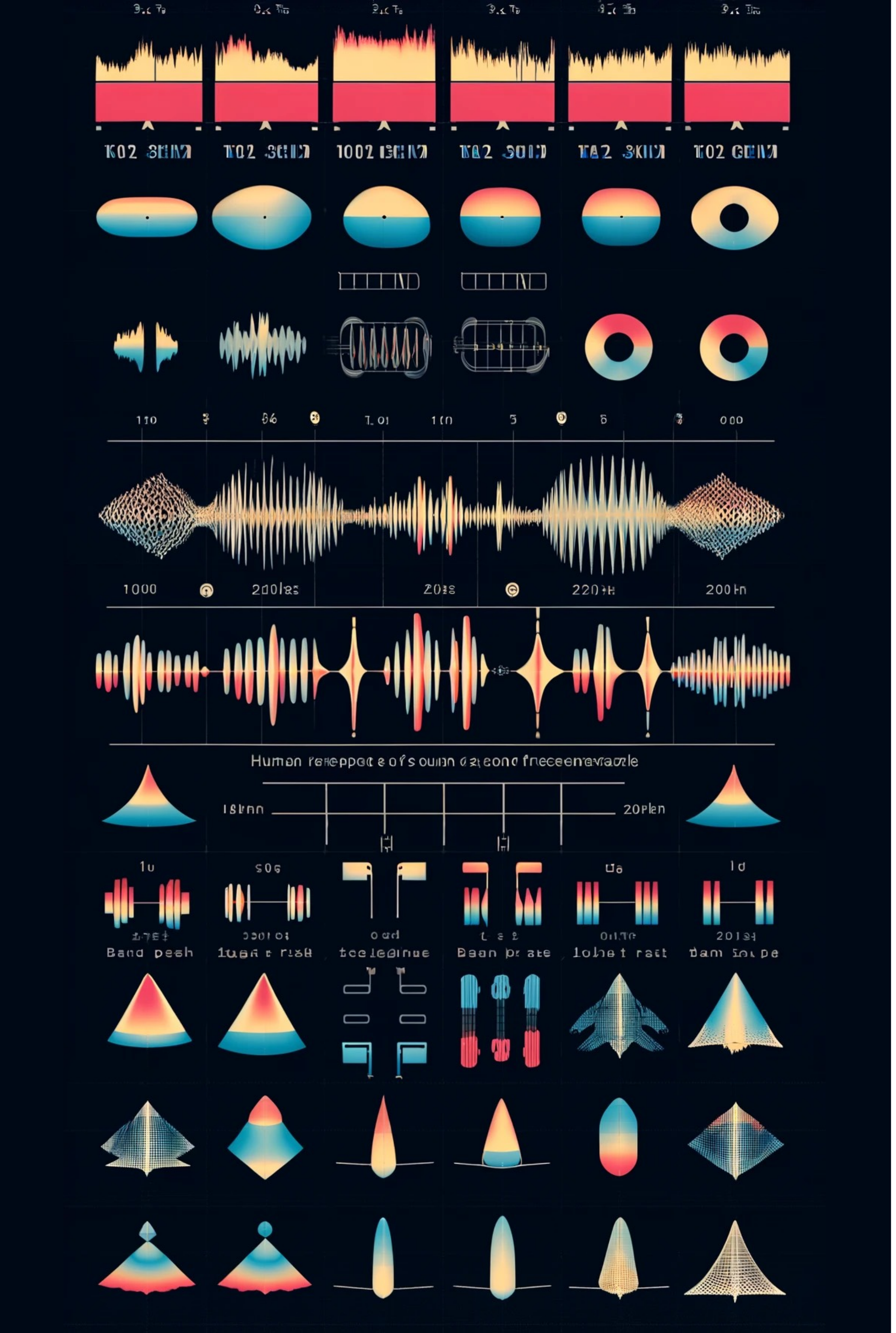
### THE NON-LINEAR SCALE OF THE HUMAN EAR

The human ear does not perceive frequencies on a linear scale. Instead, it perceives them on a logarithmic scale. This means that a change in frequency at lower frequencies is more noticeable than at higher frequencies.

---

### CAPTURING THE NON-LINEAR PERCEPTION OF FREQUENCY

The Mel scale is a perceptual scale of pitches listeners judge as equal in distance. It captures this non-linear perception of frequency.



# PERCEPTION OF SOUND

## Overview

### A COMPARISON OF AUDIO DISCRIMINATION

Humans can easily distinguish between 100Hz and 200Hz audio, but it will be tough to tell the difference between 2100Hz and 2000Hz audio.

### ACHIEVING SMOOTH MAGNITUDE SPECTRUM

The magnitude frequency response is multiplied by a set of triangular band-pass filters called Mel filter banks to attain a smooth magnitude spectrum.

# LOG-MEL SPECTROGRAM

## Spectrogram vs. Log-Mel Spectrogram

LOG-MEL IS DESIGNED TO MIMIC HUMAN PERCEPTION, WHICH IS MORE SENSITIVE TO DIFFERENCES IN LOWER FREQUENCIES THAN HIGHER ONES.

### Spectrogram

Spectrogram uses a linear/ logarithmic frequency scale.

### Linear

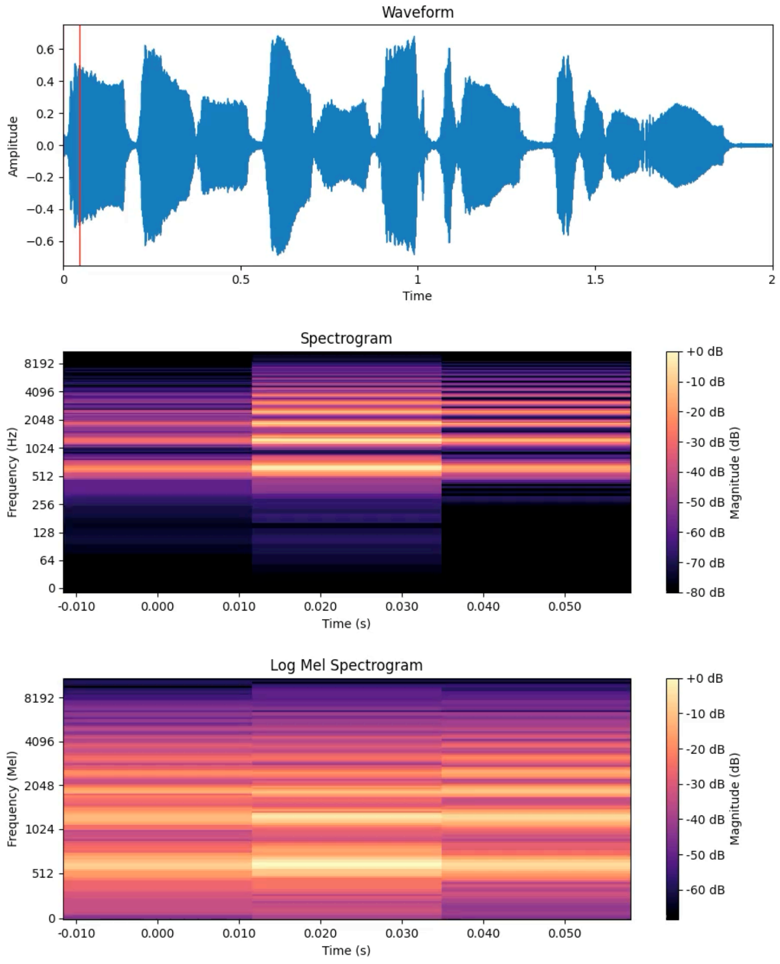
Direct frequency.

### Logarithmic

Lower Frequency better seen.

### Log-Mel Spectrogram

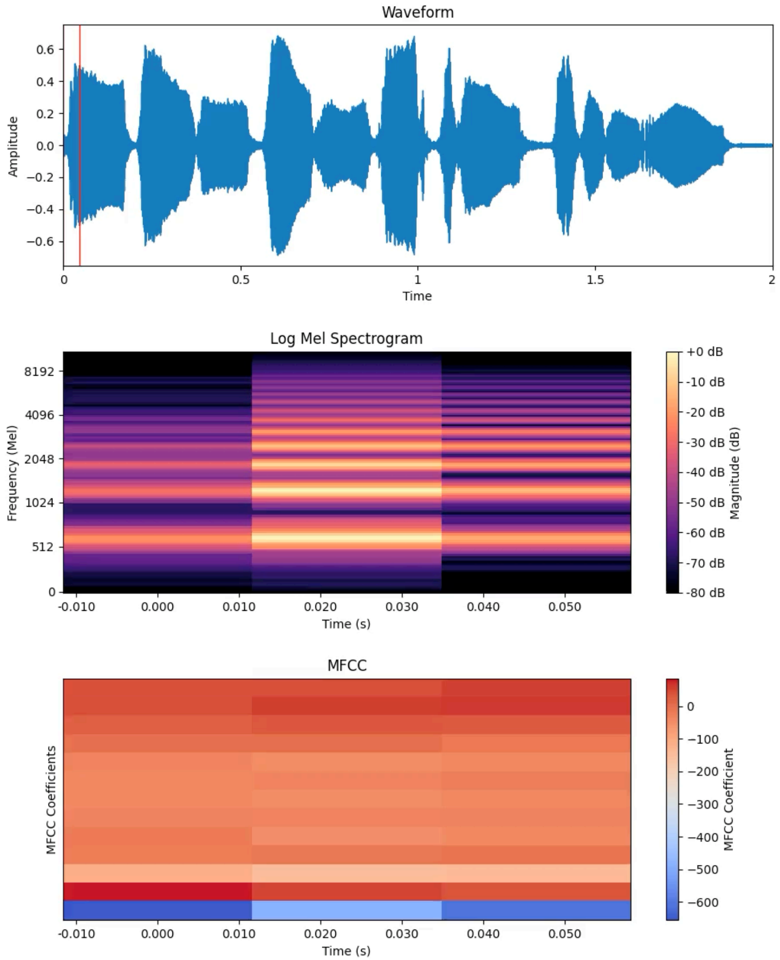
Uses the Mel scale for the frequency axis.



# MEL-FREQUENCY CEPSTRAL COEFFICIENT

## Log-Mel Spectrogram to MFCC

The speech signal's time power spectrum envelope represents the vocal tract, and MFCC (which is nothing but the coefficients that make up the *Mel-Frequency Cepstrum*) accurately represents this envelope.




### Usage

They are widely used in Speech Recognition.

### Transformation

Discrete Cosine Transform is done on Log-Mel to create Mel-Scale Coefficients.

### Influence

Sound generated by humans is determined by the shape of their vocal tract.

# LIBROSA PACKAGE

## A Python package for music and audio analysis

**01** All the math is simplified using Librosa

**02** Similar to OpenCV for Computer Vision

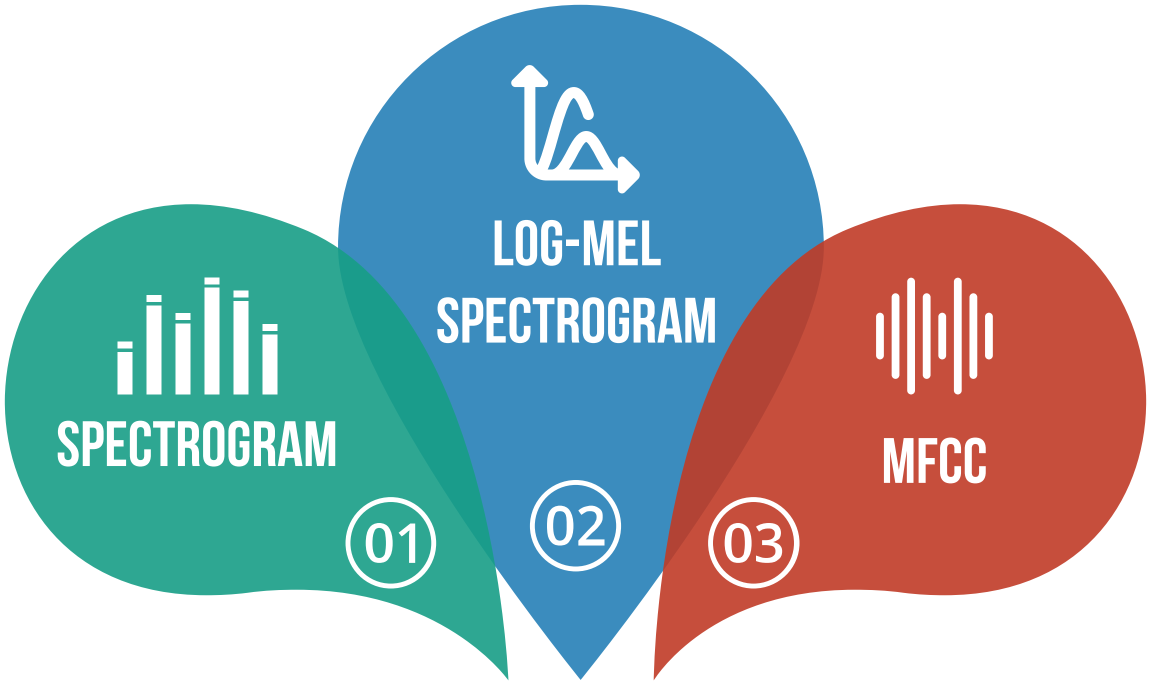
**03** The method librosa.load() loads audio like cv2.imread()

**04** To display waveforms, use *librosa.display.waveshow*

**05** All that math is a simple line of *librosa.feature.mfcc()*

**06** Generate spectrograms using *librosa.stft()*

# LIBROSA PACKAGE
## Key Differences




Option 01
**Spectrogram**
Represents the magnitude of frequencies over time.

Option 02
**Log-Mel Spectrogram**
Represents frequencies on the Mel scale, providing a more perceptually relevant frequency axis and using logarithmic magnitude scaling.

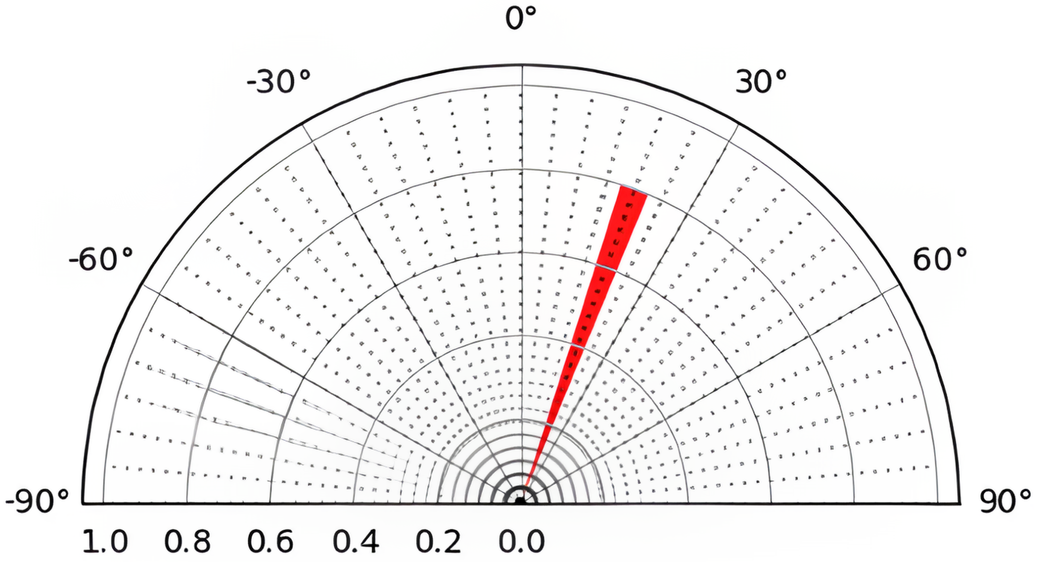
Option 03
**MFCC**
Represents the signal in a compact form, capturing the most important aspects of the power spectrum while reducing dimensionality.

# DOA ESTIMATION PLOT

Adavanne et al. (2021), Differentiable Tracking-Based Training of DL Sound Source Localizers

## Inputs

Real and imaginary components of spectrograms from 8 microphone channels.




## Outputs

Probabilities for 37 angles ranging from -90° to 90° in 5 degrees resolution.

# SOUND LOCALIZATION

## Owens et al. (2021), How to Listen: Rethinking Visualizing and Localizing Sound

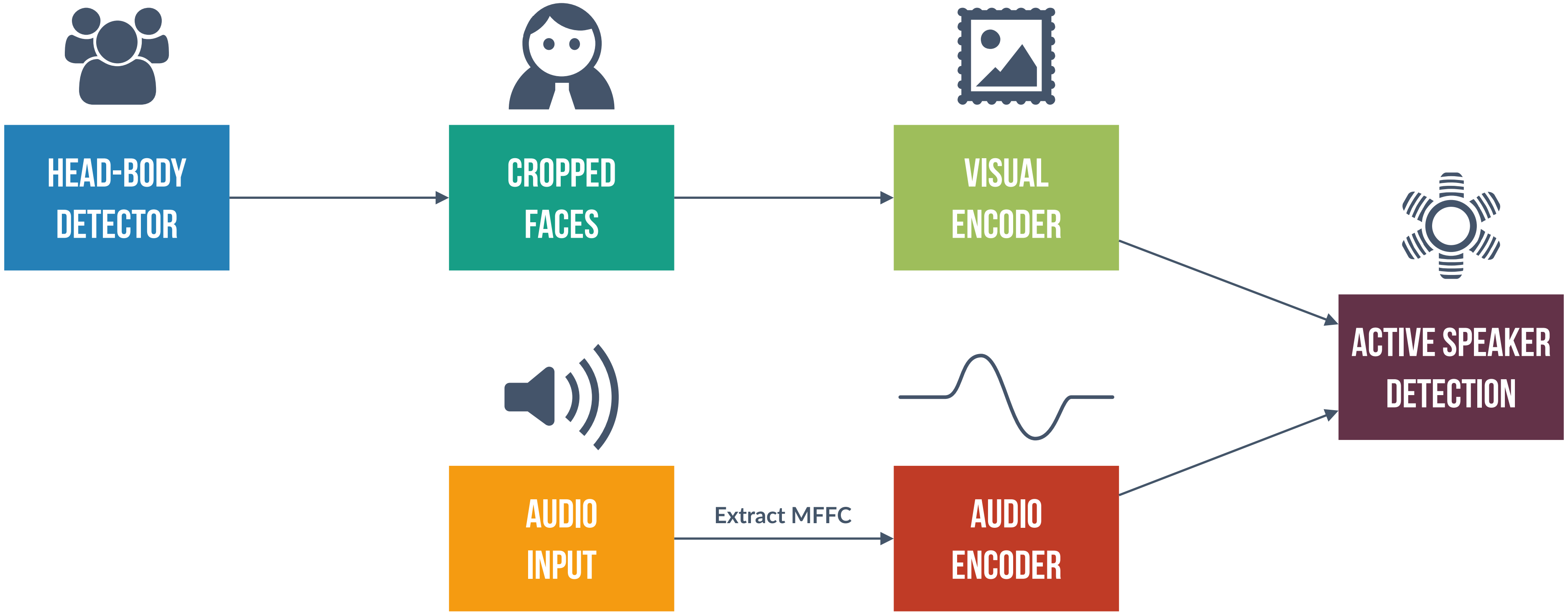
- Image Encoder CLIP
- Audio Encoder Wav2CLIP
- Contrastive Learning
- Uses Raw Audio Waveform

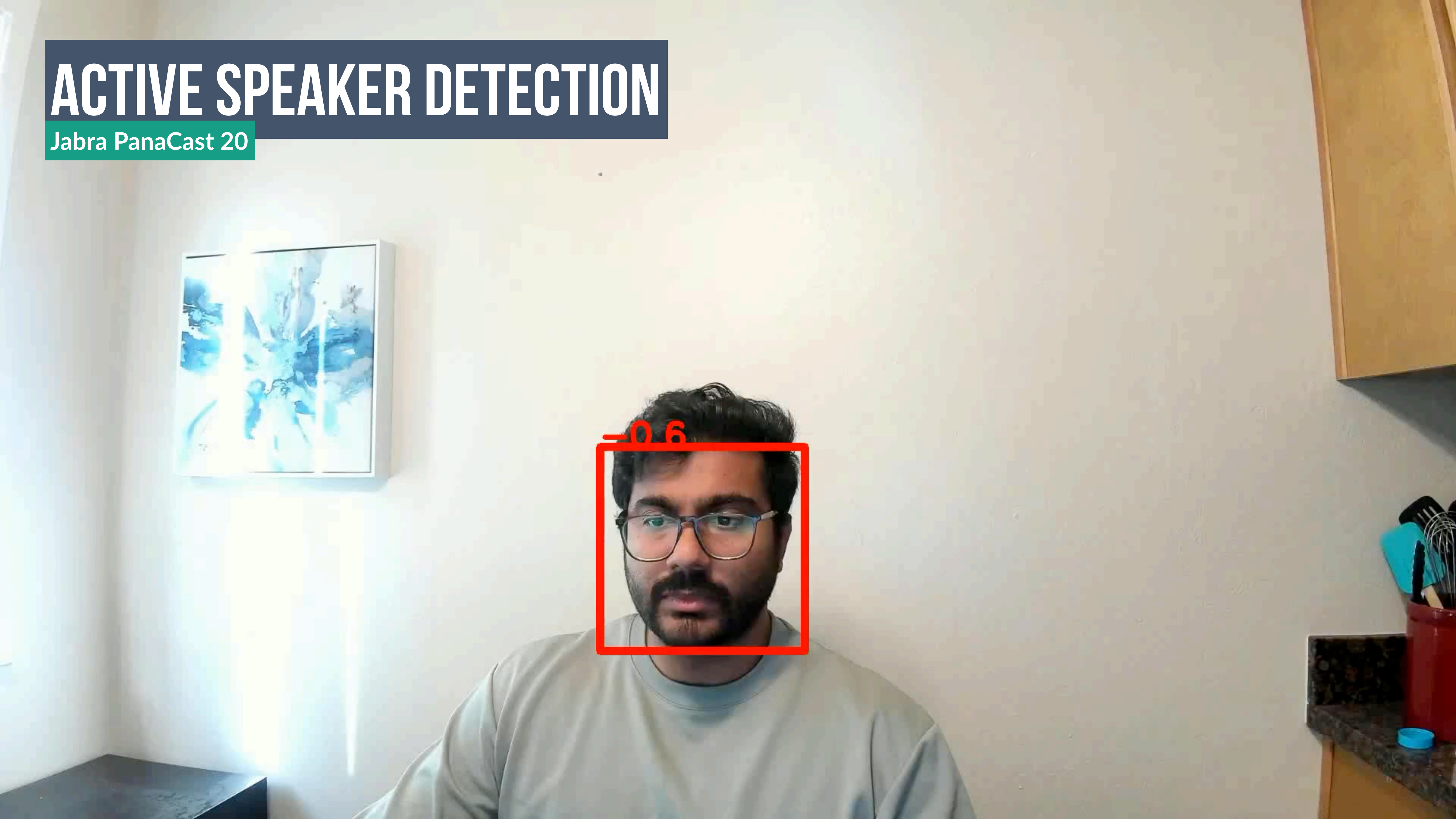
# SOUND LOCALIZATION

Owens et al. (2021). How to Listen: Rethinking Visualizing and Localizing Sound.

# ACTIVE SPEAKER DETECTION

Ruijie et al. (2021), Is Someone Speaking? Exploring Long-Term Temporal Features for Audio-Visual Active Speaker Detection




ACTIVE SPEAKER DETECTION
Jabra PanaCast 20
-0.6

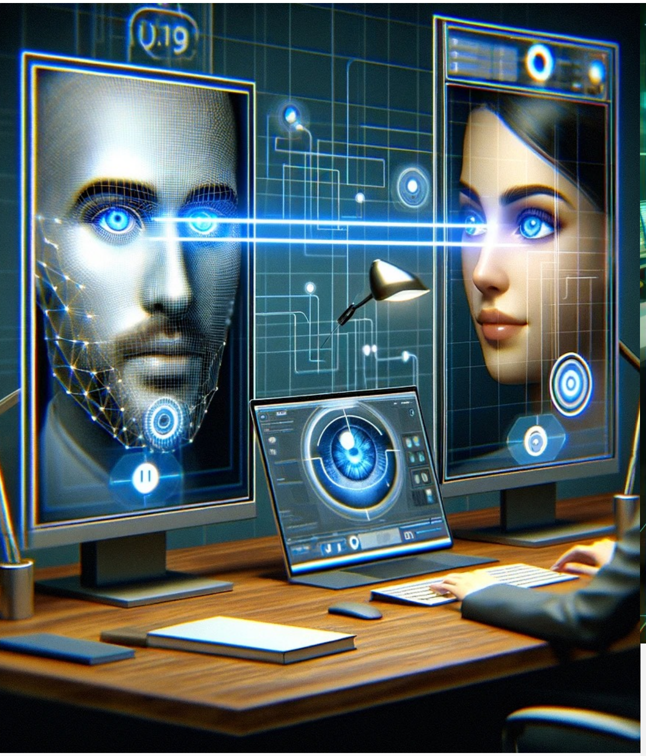
# ACTIVE SPEAKER DETECTION

Jabra PanaCast 50

# DEMOS

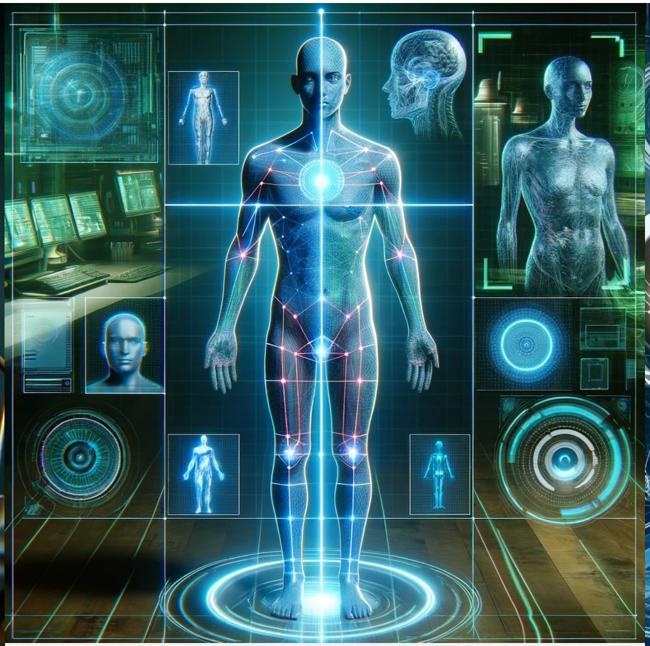
# JABRA COLLABORATION BUSINESS

## Try Our Multimodal Demos



**GAZE CORRECTION**

(JABRA EYE CONTACT)

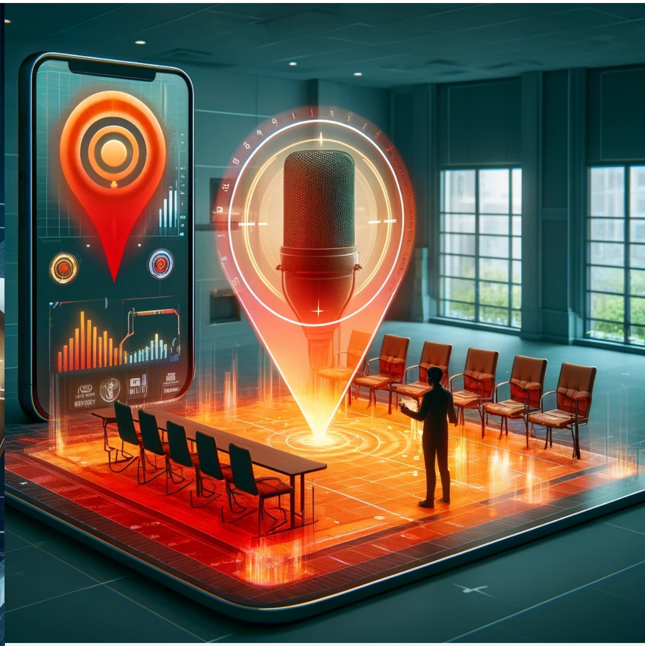


**BODY SEGMENTATION**

(JABRA PANACAST 20)

**HAND GESTURES RECOGNITION**

(EXPY)

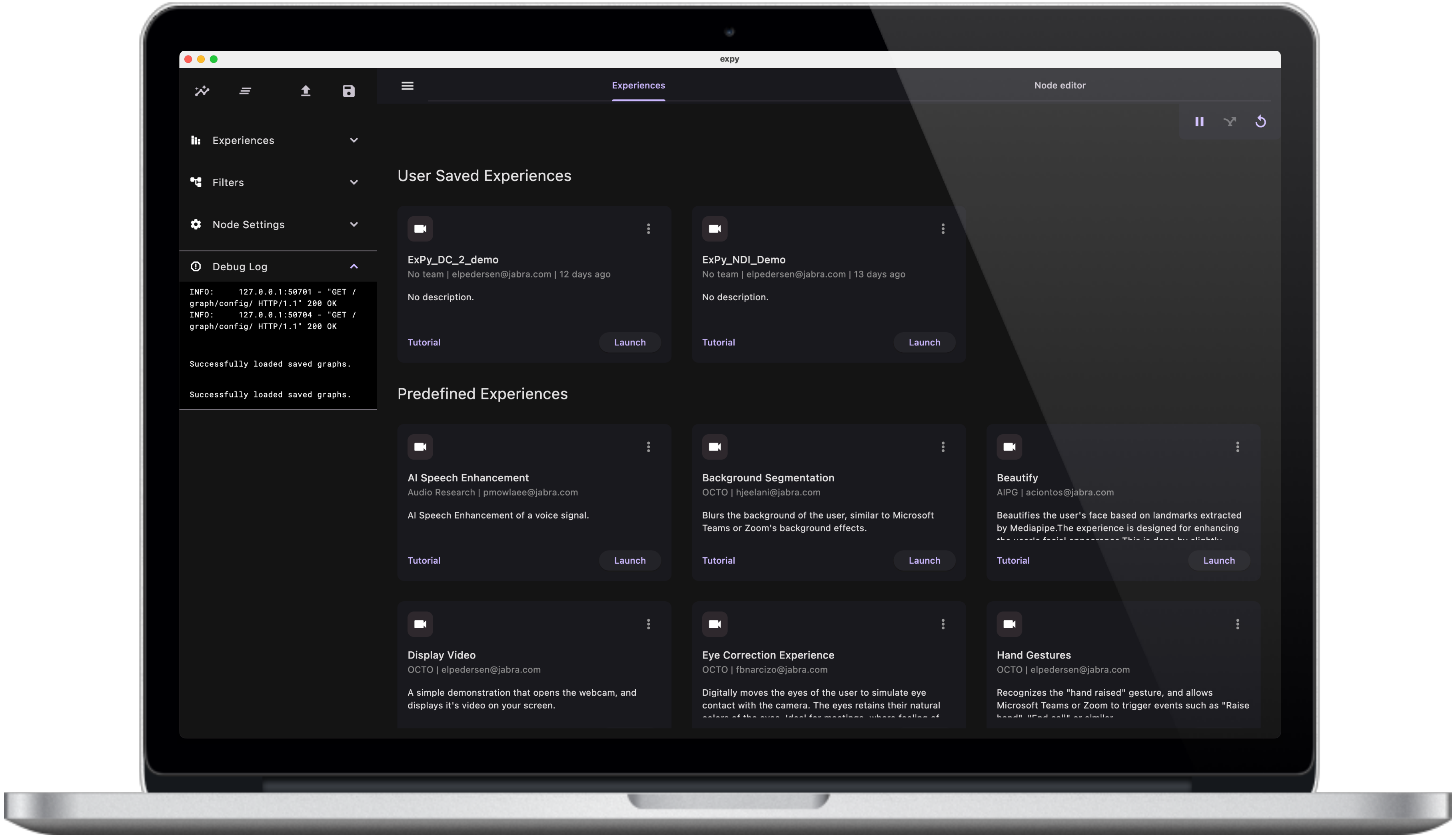


**SOUND LOCALIZATION**

(EXPY)

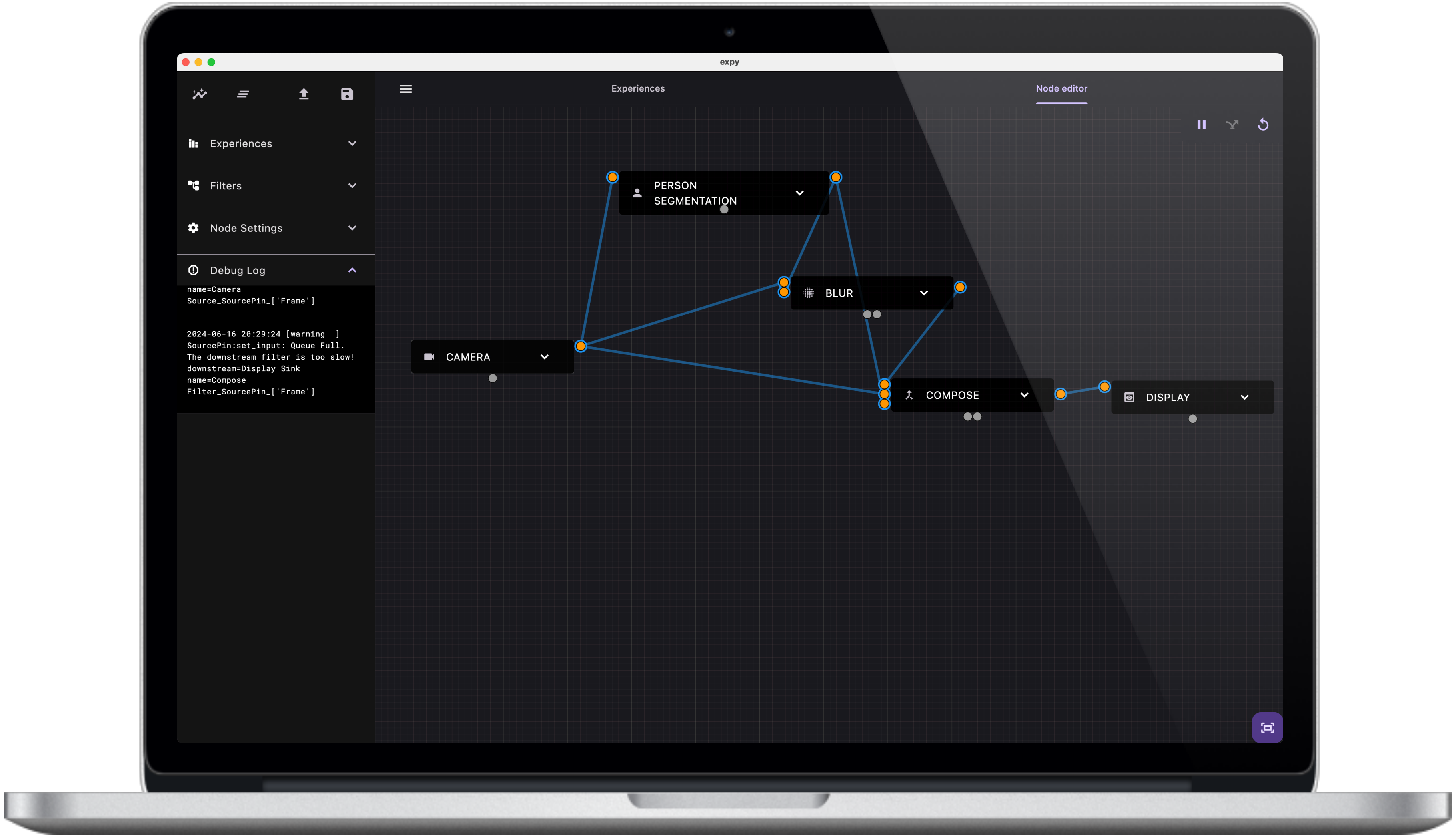
# JABRA COLLABORATION BUSINESS

## Expy Experience Platform



# JABRA COLLABORATION BUSINESS

## Expy Experience Platform



# QUESTIONS & ANSWERS

# THANK YOU!

# CLOSING REMARKS AND JOINT Q&A