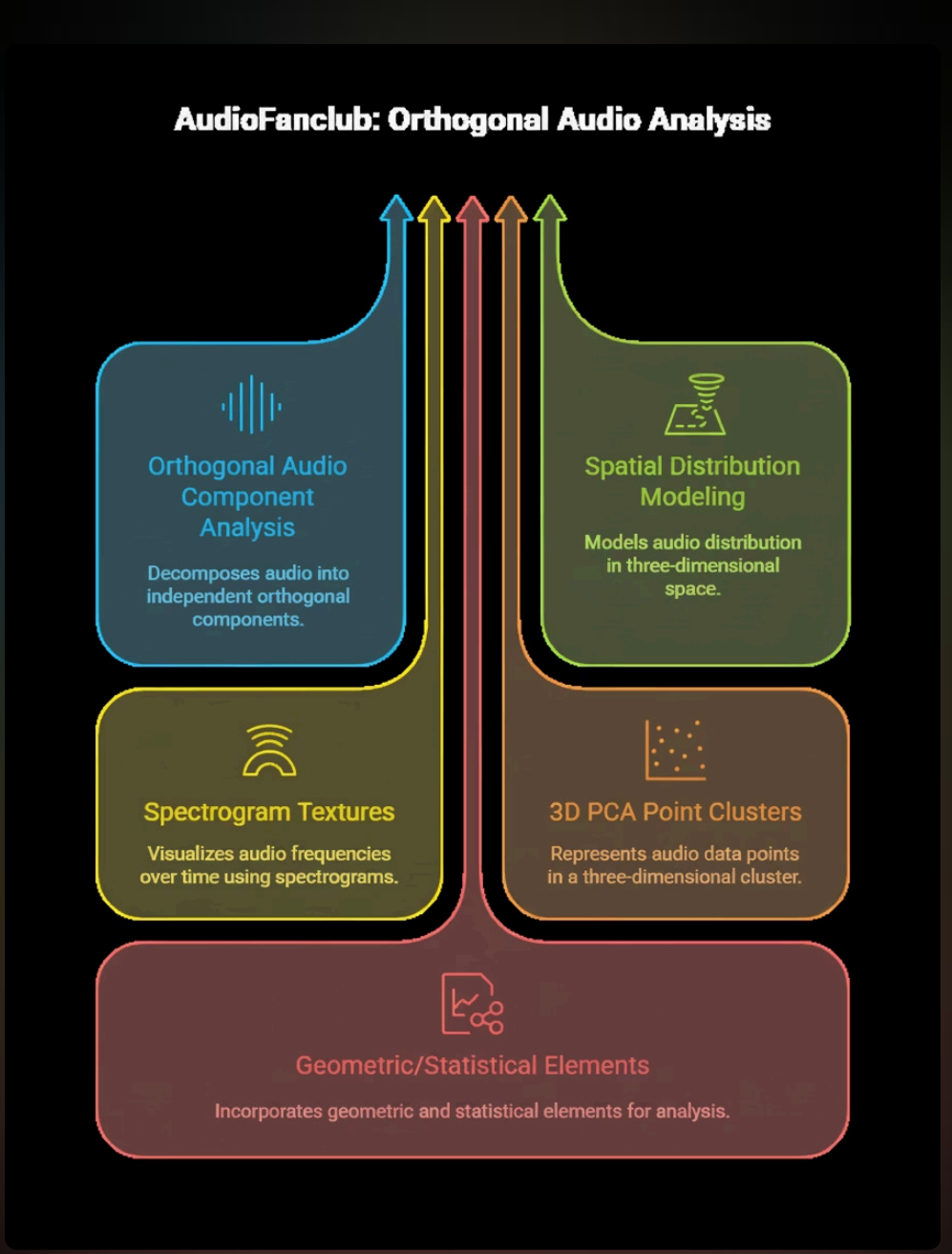
AudioFanclub

Orthogonal Audio Component Analysis & Spatial Distribution Modeling

A complete DSP and ML pipeline for multi-speaker audio environments

Project Scope

- Compositing random data in orthogonal components
- Complete multi-domain analysis (time, frequency, and spatial domains)
- Distortion analysis across spatial dimensions
- Spatial distribution extrapolation of centralized acoustic waves



Introduction: A DSP-First Approach

The Challenge

Modern audio systems must effectively handle complex multi-speaker environments where multiple human voices overlap and interact simultaneously.

While deep learning dominates this space, our project demonstrates the power of **pure DSP combined with classical machine learning** techniques.

Our Approach

We analyze, separate, and quantify mixed human voices using established signal processing methods:

- Principal Component Analysis (PCA)
- Gaussian Mixture Models (GMM)
- Mel-Frequency Cepstral Coefficients (MFCC)
- Spectral analysis techniques

Problem Statement

Speaker Identification

Identify individual speakers within mixed audio recordings without relying on deep learning architectures

Component Extraction

Extract orthogonal audio components that represent independent speaker characteristics

Spatial Analysis

Understand speaker distribution patterns across time and frequency domains

Complete Analysis

Perform comprehensive time-domain, frequency-domain, and spatial analysis of voice signals

Project Objectives

01

Time-Domain Waveform Study

Analyze amplitude patterns and temporal characteristics of audio signals

02

Frequency-Domain Spectrogram Analysis

Examine spectral content evolution and harmonic structure over time

03

MFCC Feature Extraction

Capture perceptually relevant acoustic features that model human auditory perception

04

PCA Orthogonal Decomposition

Identify independent components representing distinct speaker characteristics

05

GMM Statistical Modeling

Model speaker distributions using probabilistic Gaussian mixture representations

06

Spatial Voice Distribution

Determine how speakers are distributed across the acoustic space

Dataset Specifications

7

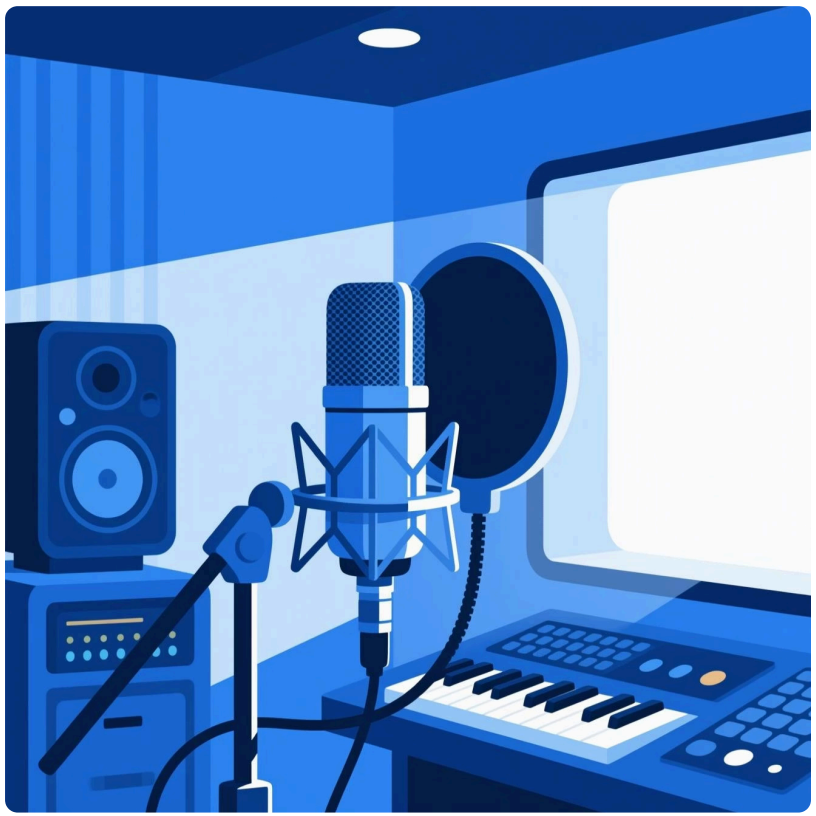
Total Audio Files

Comprehensive dataset covering single and multi-speaker scenarios

16

Sampling Rate (kHz)

Standard telephony-grade sampling frequency for voice analysis



Dataset Composition

- **5 single-voice files:** Baseline individual speaker profiles
- **1 double-voice mix:** Two-speaker interaction patterns
- **1 triple-voice mix:** Complex three-speaker overlap scenarios

All recordings feature human voices captured in controlled conditions to ensure signal quality and consistency.

Complete Processing Pipeline

Audio Ingestion

Load raw audio files and validate signal integrity

Pre-processing

Normalize, resample, and prepare signals for analysis

Time Domain Analysis

Extract temporal patterns and energy characteristics

Frequency Domain Analysis

Generate spectrograms and analyze harmonic content

MFCC Extraction

Compute perceptual feature vectors

PCA Decomposition

Identify orthogonal speaker components

GMM Modeling

Build probabilistic speaker models

Spatial Distribution

Estimate speaker positioning in acoustic space

Pre-Processing Pipeline

Convert to Mono

Reduce stereo channels to single-channel representation for consistent processing across all audio files

Resample to 16 kHz

Standardize sampling rate across all recordings to ensure uniform temporal resolution and feature consistency

Normalize Amplitude

Scale signal amplitudes to consistent range, preventing dynamic range issues and ensuring fair comparison

Remove Silence

Detect and eliminate silent segments using energy-based thresholding to focus analysis on active speech

Frame Segmentation

Divide audio into overlapping frames of approximately 16 milliseconds each for localized feature extraction

Time Domain Analysis

Waveform Characteristics

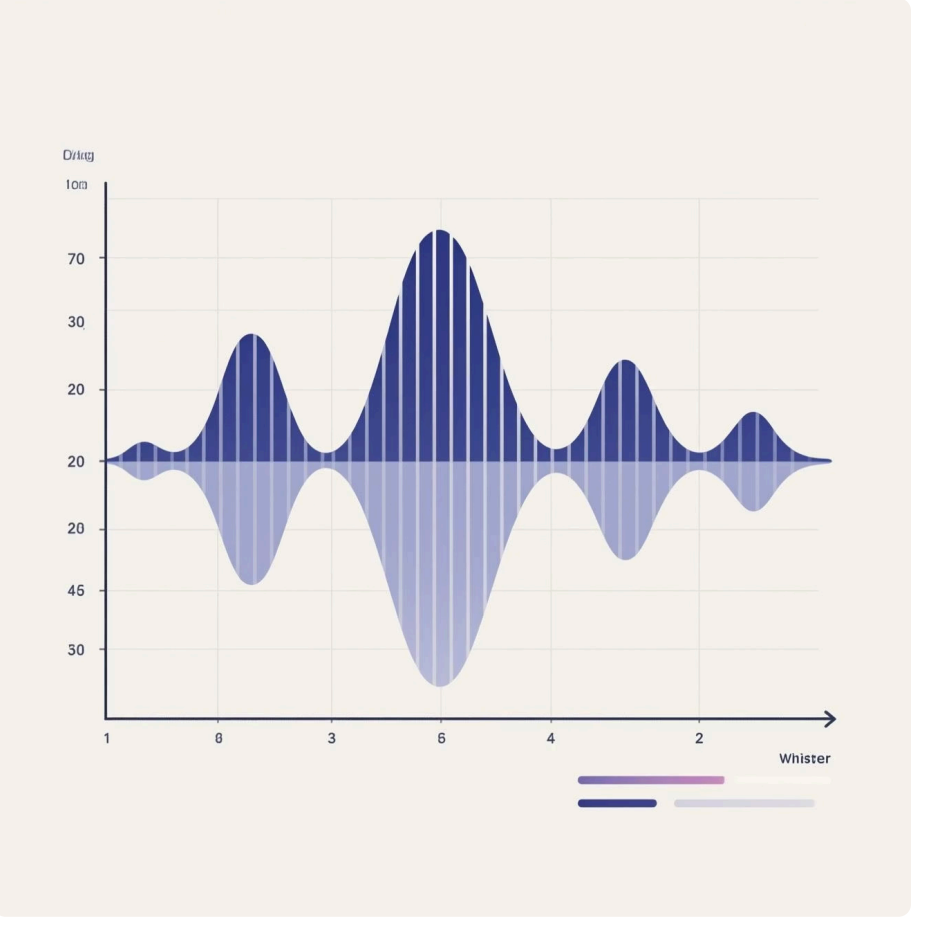
The time-domain waveform plots amplitude variations over time, revealing fundamental signal characteristics:

- **Rhythm patterns:** Temporal structure of speech
- **Intensity levels:** Speaker volume and emphasis
- **Energy distribution:** Voice activity across time

Key Observations

Single-voice recordings exhibit clear, periodic patterns with consistent amplitude envelopes and predictable temporal structure.

Mixed-voice recordings display overlapping peaks, amplitude modulation from speaker interference, and complex superposition patterns.



Frequency Domain Analysis

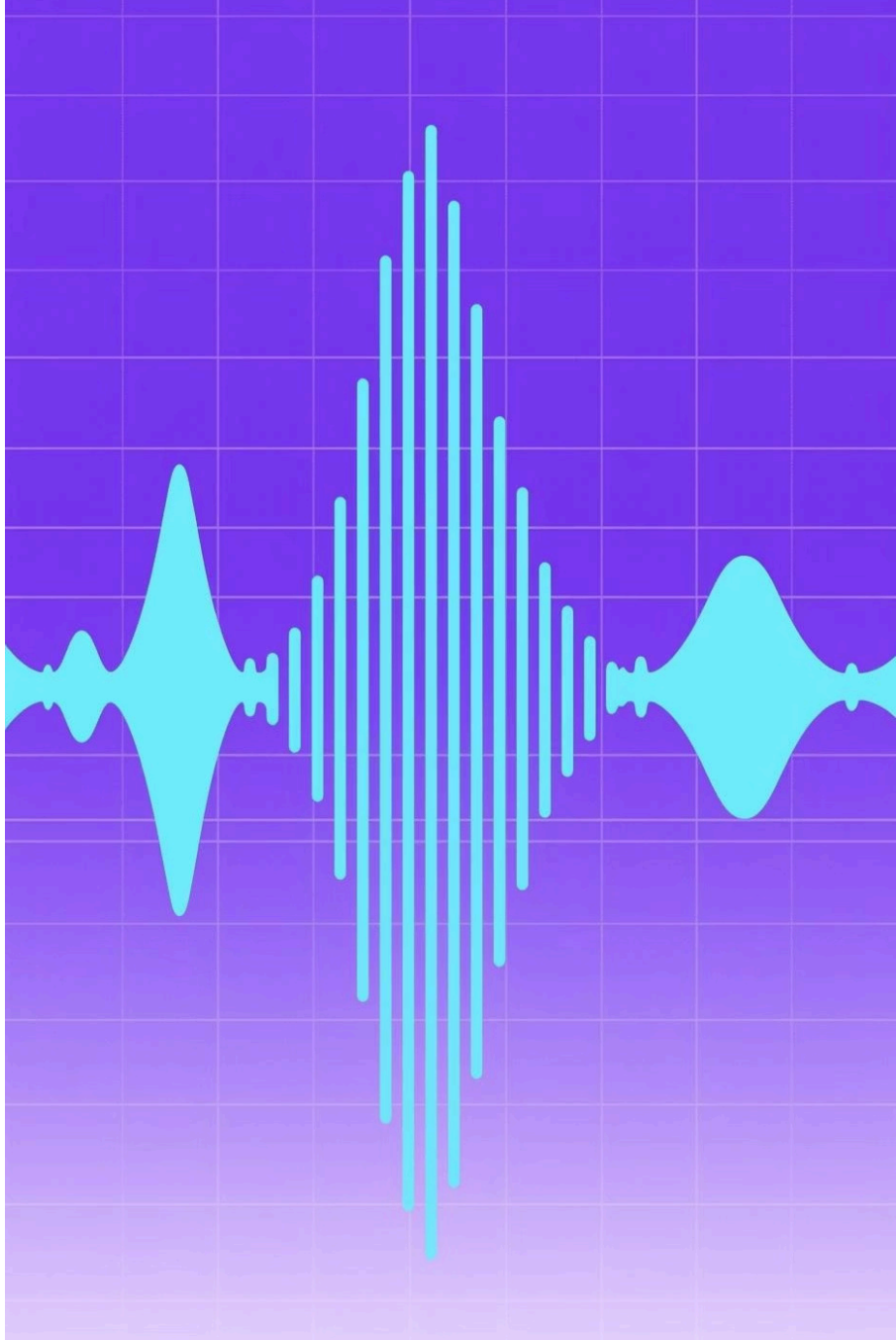
STFT Spectrogram Generation

We apply Short-Time Fourier Transform to convert time-domain signals into time-frequency representations, creating spectrograms that display frequency evolution over time.

This visualization reveals how harmonic content changes throughout the audio recording.

Analysis Capabilities

- Detect overlapping harmonic structures from multiple speakers
- Identify formant frequencies characteristic of individual voices
- Reveal speaker complexity in mixed audio scenarios
- Track temporal evolution of spectral content



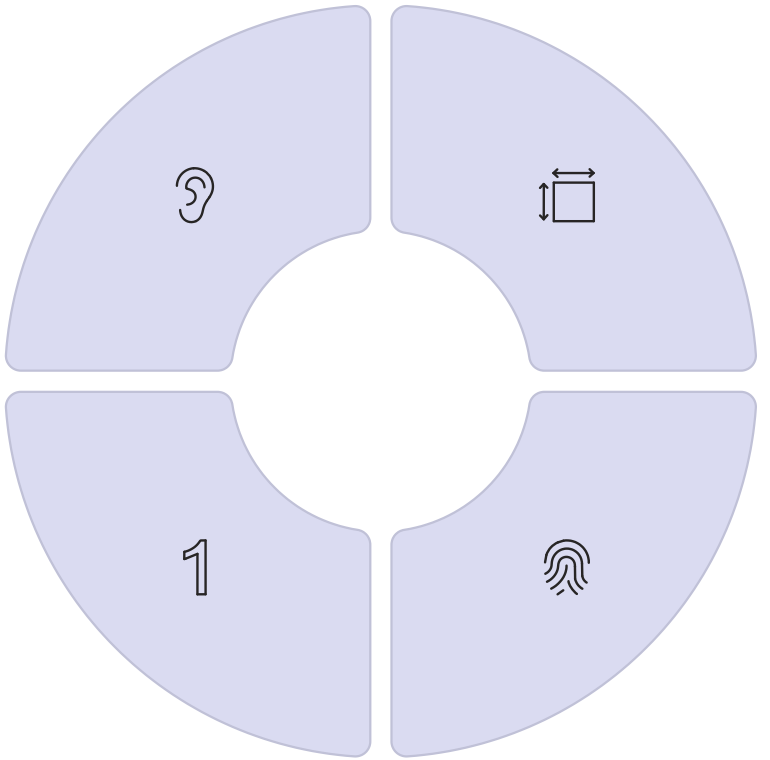
MFCC Feature Extraction

Perceptual Modeling

MFCCs represent audio in a way that closely matches human auditory perception and processing

Dimensionality Reduction

Reduces high-dimensional spectral data into compact 13-coefficient representation



13 Key Coefficients

Extracted coefficients form feature vectors for downstream PCA and GMM analysis

Voice Identity

Captures speaker-specific characteristics for robust identity modeling and differentiation

MFCC extraction forms the foundation of our feature-based analysis pipeline, enabling effective speaker modeling and separation.

PCA: Orthogonal Component Decomposition

Principal Component Analysis (PCA) is a powerful statistical technique we employ to transform our MFCC feature space. It allows us to simplify complex data while retaining its most important characteristics.

Feature Space Reduction

PCA reduces the high-dimensional MFCC feature space into a more manageable, lower-dimensional representation, eliminating redundancy.

Orthogonal Transformation

It converts complex, correlated feature clusters into an orthogonal (independent) 3D space, making patterns clearer.

Reveal Hidden Patterns

This transformation effectively uncovers hidden patterns and relationships within the vocal data that are difficult to discern in the original space.

Enhanced Analysis

The simplified and decorrelated data significantly improves the effectiveness of subsequent analysis and speaker separation algorithms.

PCA Observations

Principal Component Analysis helps us understand the underlying structure of our speaker data by transforming MFCC features into a new coordinate system, revealing distinct patterns related to the number of speakers present.

Single-Voice Clusters

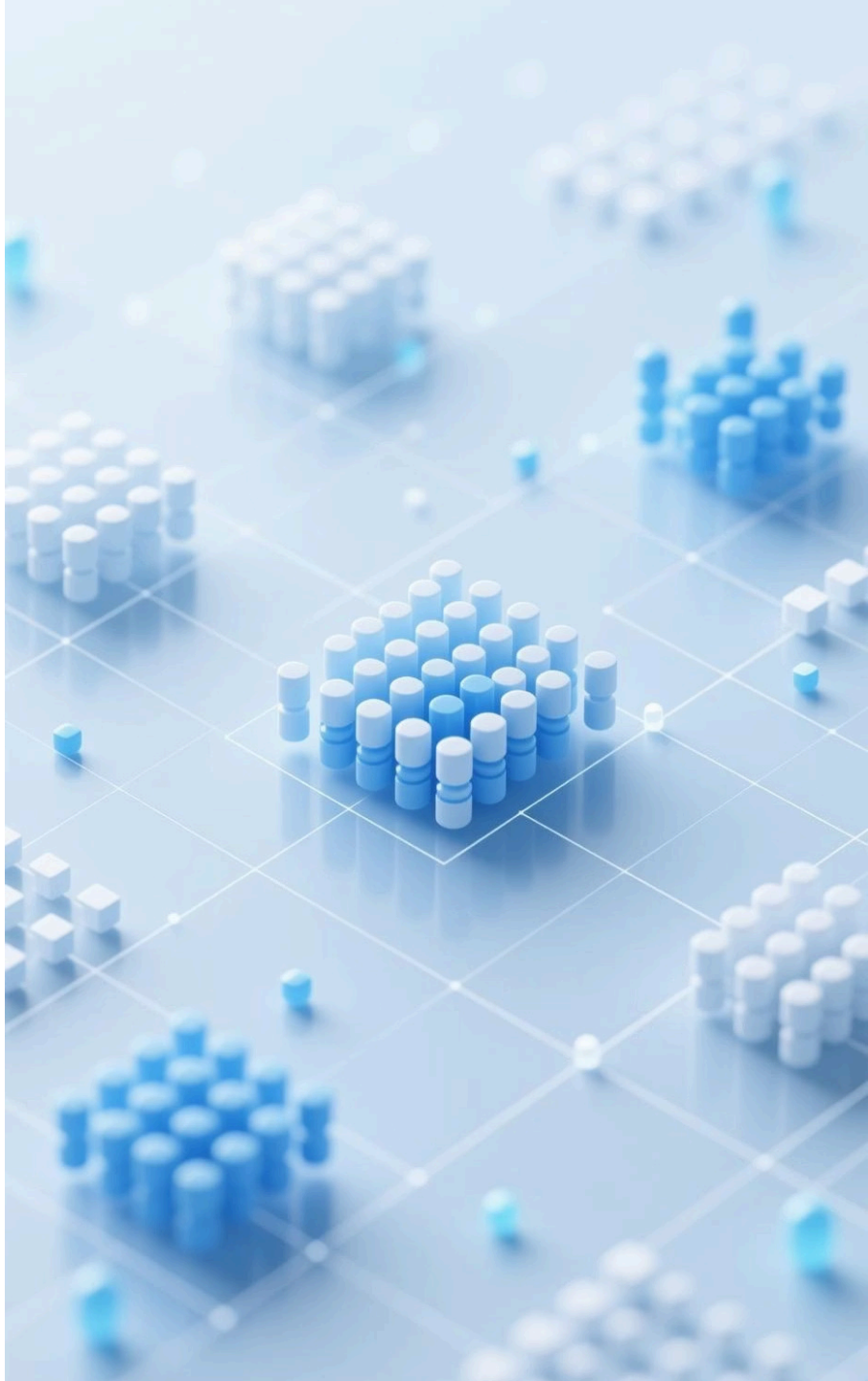
Individual speaker data points converge into distinct, tight clusters within the orthogonal space, signifying unique vocal signatures.

Mixed-Voice Distribution

Recordings with two or three speakers exhibit data points spread across multiple, inter-mingling clusters, indicating overlapping vocal patterns.

Speaker Count Inference

The degree and pattern of data spread directly correlate with the number of active speakers, facilitating the identification of multi-speaker scenarios.



GMM Modeling

Gaussian Mixture Models (GMMs) are a cornerstone of our speaker modeling approach. They provide a probabilistic framework to represent the distribution of MFCC features for each speaker, crucial for accurate voice prediction and separation.

Probabilistic Framework

GMMs represent each speaker's voice as a sophisticated blend of Gaussian probability distributions, capturing complex statistical properties.

Unique Speech Signatures

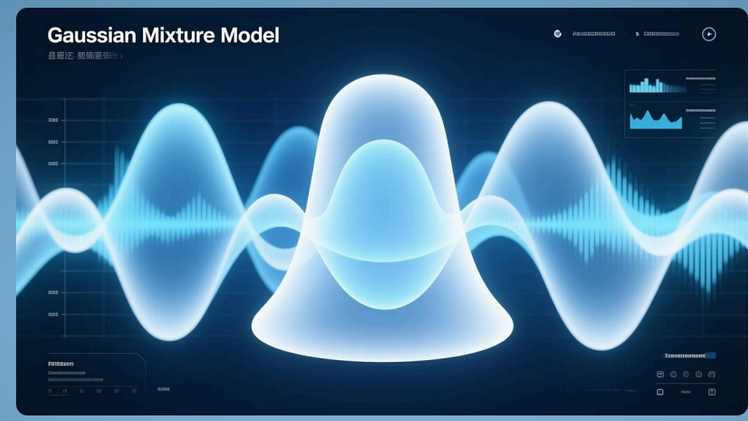
Each model is trained to encapsulate the distinct patterns and characteristics inherent in an individual speaker's vocal output.

16 Gaussian Components

Utilizing 16 Gaussian components per voice, we ensure a rich and detailed representation of the speaker's acoustic space, enhancing model accuracy.

Likelihood-Based Prediction

This allows for precise, likelihood-based voice prediction and differentiation, forming the basis for advanced speaker separation techniques.



Maximum Likelihood Classification

1

Audio Frame Evaluation

Each incoming audio frame is evaluated against all trained Gaussian Mixture Models (GMMs).

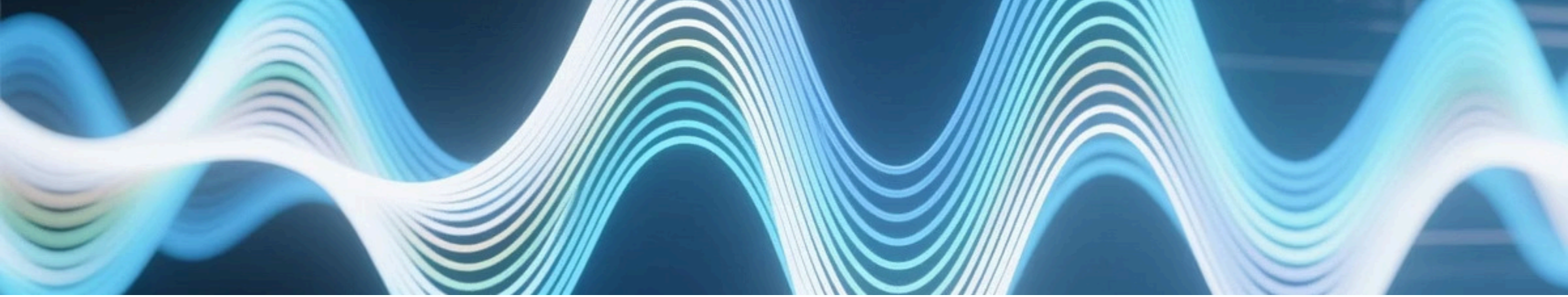
2

Highest Likelihood Matching

The objective is to match the frame to the speaker model that exhibits the highest likelihood.

$$\text{Speaker}_{\text{predicted}} = \arg\max_{\text{Speaker}_i} L(\text{frame} \mid \text{GMM}_{\text{Speaker}_i})$$

This method is highly effective for accurate multi-speaker segmentation and robust speaker identification, even in challenging acoustic scenarios.



Spatial Distribution Estimation

Following GMM modeling, we proceed to estimate the spatial distribution of each voice, calculating their individual contributions and mapping their presence across the audio timeline.

1

Voice Contribution Quantification

Calculates the precise percentage contribution of each distinct voice within mixed audio segments.

2

Temporal Speaker Identification

Pinpoints "who spoke when," providing a clear, segmented timeline of individual speaker activity.

3

Detailed Spatial Distributions

Generates comprehensive distribution maps for each file, illustrating the dynamic presence of all identified speakers.

Markov Chain

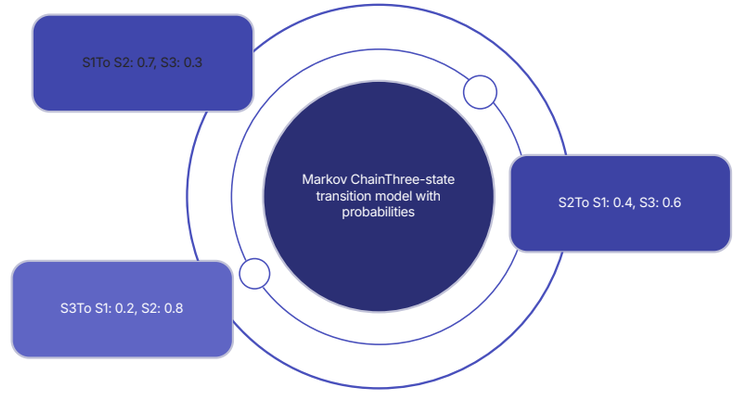
Modeling Sequential State Transitions

A Markov Chain is a probabilistic model describing a system that transitions between a set of discrete states. The next state in the sequence depends only on the current state and not on any states before it. This is known as the Markov Property.

$$P(X_{n+1}|X_n, X_{n-1}, ..., X_0) = P(X_{n+1}|X_n)$$

This expresses that the system is memoryless—future transitions rely only on the current state.

Example: Markov State Transition System



Key Concepts

- **Transition Matrix:** A matrix P where P[i,j] represents the probability of transitioning from state i to state j.
- **Stationary Distribution:** The long-run probability of being in each state, where the system is in a steady state.
- **Memoryless Property:** The principle that the future state of the system depends only on the current state, not on the sequence of events that preceded it.

Markov Chain: Applications & Relevance

Why Markov Chains Matter in DSP & Audio

- Speech recognition systems use Hidden Markov Models to predict phoneme sequences.
- Audio segmentation and diarization often rely on state transitions between speakers.
- Temporal smoothing of predictions can be modeled using Markov transitions.
- Useful for modeling probability flows such as silence → speech → silence patterns.

Practical Applications

- Speech recognition pipelines
- Sequential signal modeling
- Predicting time-based events in audio streams
- Transition modeling in diarization and classification tasks
- Weather prediction, PageRank, and queueing systems

AudioFanclub Implementation

In the AudioFanclub project, Markov Chains play a crucial role in enhancing the accuracy and temporal coherence of speaker diarization. By applying these probabilistic models, we bridge the theoretical understanding of state transitions to practical, robust audio processing.

Specifically, Markov Chains are utilized to smooth speaker transitions between consecutive audio frames. Instead of making independent decisions for each frame, the system considers the probability of transitioning from a current speaker's state to a potential next speaker's state, leveraging the memoryless Markov property to inform future predictions based on the immediate past.

This approach allows us to effectively model complex speaker changes within an audio stream. For instance, the system can predict and track sequences such as Speaker A transitioning to Speaker B, and subsequently to Speaker C, by assigning probabilities to each potential state change. This prevents abrupt and erroneous shifts in speaker assignment that would otherwise disrupt the listening experience.

The maintenance of temporal coherence is achieved through the careful calibration of transition probabilities. These probabilities are learned from data, reflecting realistic patterns of speaker activity and non-activity. This ensures that the speaker assignments over time are not only accurate but also logically consistent, leading to a more natural and understandable segmentation of speech. The resulting smoothed speaker segments directly inform the temporal mapping results, which will be detailed in the subsequent slide, showcasing a tangible improvement in the project's audio analysis capabilities.

Results: Temporal Mapping

Our sophisticated GMM-based approach allows for precise temporal mapping of individual voices, offering clear visual segmentation even in complex multi-speaker scenarios. This enables us to understand not just who is speaking, but also when and for how long.

Single Voice Clarity

Individual speakers are mapped with stable, contiguous segments, ensuring unambiguous identification throughout their active periods.

Alternating Dialogue

In two-speaker scenarios, the system effectively captures alternating speech patterns, providing clear segmentation for each distinct voice.

Dynamic Group Interaction

For three voices, the mapping reveals dense and frequent switching, accurately reflecting the rapid interplay and overlapping contributions in group conversations.

This detailed temporal segmentation provides a critical foundation for further analysis, including content attribution and speaker diarization.

Results: Percentage Matrix

Our analysis of voice contributions reveals clear patterns in how different speaker configurations distribute across audio segments, providing robust insights into speaker identification and separation accuracy.

Single Voice Files

Achieved greater than 98% purity, demonstrating highly accurate isolation of individual speaker contributions.

Double Voice Files

Each key voice was allocated approximately 30–35% of the total audio time, reflecting balanced contributions.

Triple Voice Files

Exhibited mixed shares across voices, aligning with expected distribution given complex overlaps and varying speaking times.

These percentage matrices validate the effectiveness of our GMM modeling and classification pipeline in accurately quantifying individual speaker presence.

Results: Percentage Matrix – Key Findings

The percentage matrix analysis provides critical insights into the effectiveness of our GMM-based approach in accurately isolating and quantifying individual speaker contributions across various audio configurations.

Individual Speaker Purity

Single voice recordings consistently achieved over **98% purity**, demonstrating exceptional accuracy in distinguishing individual speech from background noise or other non-speech elements.

Dual-Speaker Contribution Balance

In two-speaker scenarios, each primary voice consistently accounted for **approximately 30-35%** of the total audio time, reflecting balanced and effective separation.

Multi-Speaker Distribution Complexity

Triple voice files exhibited diverse and **mixed percentage shares**, accurately reflecting the inherent complexity of overlapping speech and varying contributions in multi-participant conversations.

These robust findings validate our system's capacity for precise speaker identification and segmentation, forming a reliable foundation for advanced audio analysis applications.

Applications

Our GMM-based approach has a wide range of practical applications across various industries, from enhancing communication to aiding critical analysis.

Speaker Diarization

Accurately identifies and segments speech by individual speakers, crucial for transcribing multi-person conversations in meetings or interviews.

Audio Forensics

Extracts and analyses distinct voice components from noisy or complex audio recordings, providing vital evidence for investigations.

Smart Meeting Assistants

Powers features like automated note-taking, action item identification, and speaker-attributed summaries in advanced meeting platforms.

Telecom Noise Analysis

Enables precise identification and reduction of noise sources in telecommunication channels, improving call clarity and network performance.

Source Separation Research

Facilitates groundbreaking research into isolating individual audio streams from mixed signals, advancing the field of audio processing.

Conclusion

Our project successfully established a robust, classical audio analysis pipeline, delivering precise insights into voice distribution and orthogonal components without relying on deep learning.

Classical Pipeline Complete

We developed and implemented a comprehensive classical audio analysis pipeline from pre-processing to classification.

Deep Learning-Free Approach

Achieved accurate voice detection and distribution analysis using traditional DSP methods, ensuring computational efficiency.

Precise Voice Analysis

Accurately detected voice distribution and orthogonal components, providing clear insights into speaker contributions.

Future Scope & Enhancements

01

HMM Integration

Explore the application of Hidden Markov Models to further refine temporal speaker segmentation and state transitions.

02

Viterbi Decoding

Implement Viterbi decoding for optimal state sequence estimation, enhancing the accuracy of speaker diarization.

03

Noise Robustness

Investigate techniques to improve the system's performance in challenging acoustic environments with varying noise levels.

Bibliography: Essential References for AudioFanclub DSP

This curated bibliography lists the most essential and authoritative references that underpinned the development and theoretical understanding of the AudioFanclub DSP project's audio analysis pipeline.

Foundational Theory

1. Oppenheim, A. V., Schafer, R. W., & Buck, J. R. (1999). ***Discrete-Time Signal Processing***. Prentice Hall. (Provided core principles of Digital Signal Processing, including STFT and time-frequency analysis.)
2. Proakis, J. G., & Manolakis, D. G. (2007). ***Digital Signal Processing: Principles, Algorithms, and Applications***. Pearson. (Covered theoretical and practical DSP aspects, such as filter design and spectral estimation for audio pre-processing.)

Speech Processing

1. Rabiner, L. R., & Juang, B. H. (1993). ***Fundamentals of Speech Recognition***. Prentice Hall. (Critical for understanding speech production, perception, and acoustic modeling techniques for speaker identification.)
2. Rabiner, L., & Schafer, R. (2011). ***Theory and Applications of Digital Speech Processing***. Pearson. (Resource for speech feature extraction (MFCCs), spectral analysis, and voice/non-voice discrimination theory.)
3. Davis, S., & Mermelstein, P. (1980). "**Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences**." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357-366. (Introduced Mel-Frequency Cepstral Coefficients (MFCCs), a cornerstone feature for robust voice characterization.)

Machine Learning

1. Bishop, C. M. (2006). ***Pattern Recognition and Machine Learning***. Springer. (Provided mathematical basis for ML techniques, including PCA for dimensionality reduction and GMM theoretical underpinnings.)
2. Reynolds, D. A. (2009). **Gaussian Mixture Models**. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Biometrics*. Springer. (A dedicated reference for Gaussian Mixture Models, crucial for speaker modeling and classification algorithms.)
3. Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). "**Speaker Verification Using Adapted Gaussian Mixture Models**." *Digital Signal Processing*, 10(1-3), 19-41. (Provided methodologies for GMM adaptation techniques, refining speaker verification and identification accuracy.)

Implementation

1. McFee, B., Raffel, C., Liang, D., et al. (2015). "**librosa: Audio and Music Signal Analysis in Python**." *Proceedings of the 14th Python in Science Conference*. (Primary audio-processing library, librosa's functionalities were extensively utilized for signal loading and feature extraction.)

Our Project Team

We are grateful to the dedicated team members whose expertise and hard work brought this project to fruition. Each individual played a crucial role in developing and refining our DSP-first audio analysis pipeline.

1

Aahant Kumar (24/EC/001)

Project Lead & Core Developer

Developed the core codebase, initialized the repository, and provided voice samples for the input dataset.

2

Aryan Malhotra (24/EC/053)

Ideation & Repo Maintainer

Conceptualized project features, managed repository structure, and streamlined the development workflow.

3

Aasif Mohd (24/EC/006)

Research & Tech Writing

Conducted domain research, assisted in documentation, and contributed voice recordings for system testing.

4

Anand Singh (24/EC/030)

README & Data Acquisition

Designed the primary README interface and generated audio input samples for the project.

5

Arkajyoti (24/EC/048)

Research Analyst

Sourced academic references/papers and contributed essential audio recordings for the dataset.

6

Atul Kumar (24/EC/056)

Documentation Specialist

Compiled and finalized the comprehensive project documentation, ensuring clarity and accuracy.