CentraleSupélec

université PARIS-SACLAY

Master Thesis Report

Dilbar ISAKOVA

dilbar.isakova@student-cs.fr

Intelligent Ambient Data Visualization on Non-Planar Displays: A Multi-Modal Machine Learning Approach for Spatial Audio-Visual Analytics

---

Erasmus Mundus Joint Master's Degree in Big Data Management and Analytics

Advisor: Anastasia BEZERIANOS, Full Professor, Université Paris-Saclay
anastasia.bezerianos@universite-paris-saclay.fr
Advisor: Petra ISENBERG, Research Director, Inria AVIZ
petra.isenberg@inria.fr
Advisor: Tobias ISENBERG, Senior Research Scientist, Inria Saclay
tobias.isenberg@inria.fr

August 31, 2025

Acknowledgments

I would like to express my sincere gratitude to my supervisors: Professor Anastasia Bezerianos from Université Paris-Saclay, Dr. Petra Isenberg from Inria AVIZ, and Dr. Tobias Isenberg from Inria Saclay. Their expertise, guidance, and constructive feedback were essential to the development and completion of this research.

I thank the tVISt project partners for their valuable input during project presentations, and my colleagues at INRIA who participated in surveys and data collection. Their contributions were crucial for the iterative design process.

Contents

Chapter 1

Introduction

This chapter presents a summary of the work done in this thesis. The following sections describe background knowledge and state of the art of the field, the problem to be solved, purpose of the study, methods, results, and their interpretation. The project is accessible on the GitHub repository[1], providing transparency and access to all implementation details described in this report.

1.1 Context and Motivation

The evolution of data visualization has shifted from specialized scientific tools to widely used technologies that influence both industry and society [50]. For decades, the primary medium for digital data visualizations has been the traditional flat, rectangular desktop display. While these displays have improved in size, resolution, and color representation since the 1990s, their geometric limitations have remained unchanged. This paradigm is now being challenged by the emergence of non-planar display technologies that offer new possibilities for how we experience and interact with information [44]. Collaborative environments such as meeting rooms bring specific challenges for systems that aim to support environmental awareness. Traditional visualization approaches, designed for individual desktop interactions, are not well suited to the social dynamics and awareness requirements of collaborative spaces [46]. Meeting participants must balance their primary focus on collaborative activities while maintaining peripheral awareness of environmental conditions that directly impact meeting effectiveness, such as acoustic dynamics, air quality, and occupancy levels. The concept of **ambient visualization** plays a key role in this setting. Unlike traditional visualization systems that demand explicit user attention and interaction, ambient visualization operates in the peripheral awareness space, providing environmental consciousness without disrupting primary activities [44]. This approach aligns with Weiser and Brown's principles of "calm technology," where information systems enhance human capabilities without overwhelming cognitive resources [49].

1.2 Statement of the problem

Contemporary meeting room environments present opportunities for ambient environmental visualization that existing technologies do not fully address. While there are individual technologies in related domains, their combination for meeting room applications represents limited investigation. Non-planar display research has focused primarily on interactive applications [44], while spatial audio processing focuses on computational analysis [42], but we explored combining these domains specifically for meeting room environmental

[1] https://github.com/isakovaad/sphereDisplay

awareness. Meeting analytics systems typically analyze data after meetings rather than providing real-time information [18], and visualization approaches designed for focused attention may not be optimal for the peripheral awareness during collaborative activities. While spatial audio processing research concentrates on speech recognition and source localization [42], we investigated how spatial audio characteristics could be meaningfully mapped to ambient visual displays on spherical surfaces. Situated visualization principles exist for individual displays and simple environmental parameters [11], but we explored extending these principles specifically for meeting room contexts using non-planar displays. These observations led us to investigate a specific research question: How can spatial audio data from dual-microphone systems be visualized on spherical displays to provide ambient meeting room environmental awareness? We address this question through prototype development and initial technical evaluation of spatial audio visualization for collaborative environments.

1.3 Objectives and Approach

We address these observations through prototype-driven exploration of ambient meeting room awareness systems using non-planar displays. Our research combines hardware sensing, audio signal processing, machine learning, and web-based visualization for meeting room environmental monitoring. In summary, this thesis aims to answer the following research questions:

1. Which display form factors best suit meeting room environments?
2. How can non-planar displays effectively support ambient environmental awareness in collaborative meeting environments?
3. How can machine learning classification enhance the meaningfulness of real-time meeting analytics for ambient display applications?

Research Objectives

We explore design approaches for ambient visualization on non-planar displays through iterative prototyping, building upon situated visualization theory while investigating applications in collaborative meeting environments. We develop a working prototype system integrating IoT sensor technologies, spatial audio processing, machine learning classification, and WebGL visualization techniques on spherical display. We conduct preliminary user research to understand basic requirements, preferences, and design considerations for ambient meeting room displays through survey and initial design feedback. Finally, we develop classification systems that transform raw environmental sensor data into meaningful visualization parameters through supervised learning approaches for meeting room contexts.

Methodological Approach

Our methodology combines user-centered design, prototyping, IoT development, and machine learning to address the problem. We begin with a user requirements analysis through

an online surveys targeting meeting room users. This is followed by iterative design space exploration using AI-assisted concept generation and form factor comparison. We then implement a 3D prototyping with Three.js and WebGL, allowing real-time exploration of non-planar display characteristics for ambient visualization. For hardware prototyping, we develop an IoT sensing infrastructure with ESP32 microcontrollers and dual-microphone spatial audio arrays, integrated via WebSocket protocols for real-time streaming. Finally, our machine learning pipeline supports audio classification through data collection, augmentation, feature engineering, and model training.

1.4 Thesis Structure

This thesis is organized as follows. Chapter 2 introduces the background knowledge and definitions across non-planar display technologies, situated and ambient visualization theory, spatial audio processing, and machine learning for audio classification. Chapter 3 reviews existing research in visualization beyond desktop displays, non-planar display technologies, situated visualization, and meeting analytics, identifying the context that motivates this work. Chapter 4 details the user research process including requirements analysis, design space exploration, and the development of 3D prototyping using Three.js and WebGL. Chapter 5 presents the complete technical implementation including hardware design, embedded software development, visualization system architecture, and real-time communication protocols. Furthermore, Chapter 6 describes the machine learning pipeline including data collection, augmentation strategies, feature engineering, and model training, with deployment integration representing future work. Finally, the thesis concludes with the discussion of results, limitations, and directions for future work.

Chapter 2

Background and Definitions

This chapter provides the technical background and vocabulary necessary to understand the interdisciplinary concepts used in this thesis. We define core terms and explain fundamental principles in non-planar display technologies, situated visualization theory, spatial audio processing, and machine learning for audio classification.

2.1 Non-Planar Display Technologies

This section defines non-planar displays and explains the mathematical coordinate systems used to describe their geometric properties.

2.1.1 Display Geometry Fundamentals

Displays with curved, spherical, cylindrical, and other three-dimensional surfaces break away from conventional flat screens, creating new possibilities for how we present information. To design effective visualizations for these non-flat displays requires to understand their geometric characteristics which has specific mathematical coordinate systems explained below.

Spherical displays utilize spherical coordinate systems where any point on the surface can be described using coordinates (r, θ, ϕ), where r represents the radius (constant for the display surface), θ is the azimuthal angle ($0 \leqslant \theta \leqslant 2\pi$), and ϕ is the polar angle ($0 \leqslant \phi \leqslant \pi$). The conversion between Cartesian and spherical coordinates is given by:

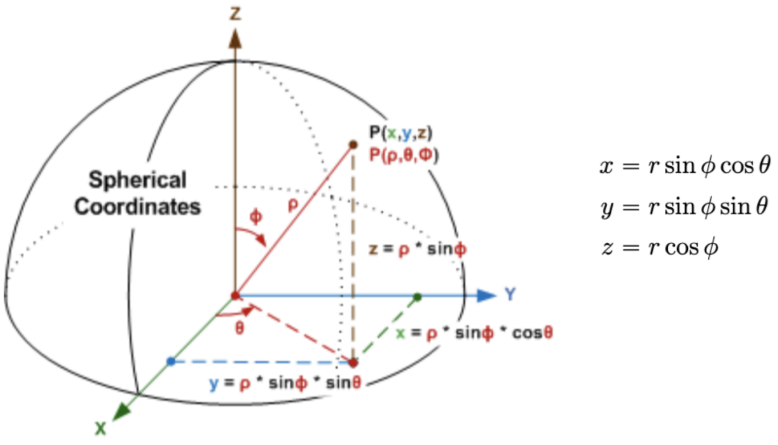
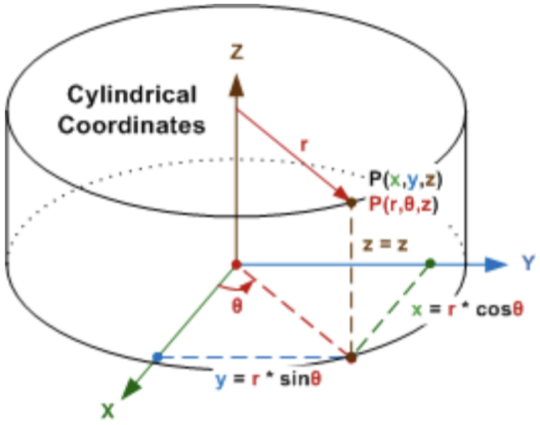


Figure 2.1: Cartesian to spherical conversion [37].

This coordinate system allows precise mapping of visual elements to specific locations

on the spherical surface, which is particularly important for environmental data visualization where spatial relationships must be preserved [20].

Cylindrical displays employ cylindrical coordinates (ρ, ϕ, z), where ρ represents the radial distance from the central axis, ϕ is the azimuthal angle around the axis, and z represents height along the cylinder. The coordinate transformation is:



$$x = \rho \cos \phi$$
$$y = \rho \sin \phi$$
$$z = z$$

Figure 2.2: Cartesian to cylindrical conversion [37].

2.1.2 Display Form Factor Categories

Form factors refer to the physical shape categories of non-planar displays:

Geometry	Description
Spherical	Complete 360° visibility in all dimensions
Cylindrical	Continuous horizontal presentation with vertical constraints
Conical	Hybrid geometry combining spherical and cylindrical characteristics
Flexible	Bendable surfaces with dynamic geometry changes

Table 2.1: Geometric display categories and their characteristics.

2.2 Situated and Ambient Visualization Concepts

This section defines situated visualization, ambient displays, and related theoretical concepts used throughout this thesis.

2.2.1 Core definitions

The key concept behind **situated visualization** is connecting data sources directly to their physical display locations, which results in more natural and contextually meaningful ways of presenting information. The theoretical foundation of situated visualization rests on the principle that physical context improves data interpretation capabilities. Therefore,

this concept is defined as visualization displayed in proximity to the phenomenon it represents, creating spatial coupling between data source and display location [11]. **Spatial coupling** describes the relationship strength between data source location and display location, characterized by physical distance, temporal alignment, and semantic relevance. The effectiveness of spatial coupling can be understood through proximity relationships. For environmental data visualization, the coupling strength S between a data source and display can be conceptualized as [17, 39]:

$$S = f(d_{\text{spatial}}, d_{\text{temporal}}, d_{\text{semantic}}) \tag{2.1}$$

where:

d_{spatial}: physical distance

d_{temporal}: temporal alignment

d_{semantic}: semantic relevance

Ambient displays represent a specialized category of situated visualization designed for peripheral awareness rather than focused attention. These systems provides environmental consciousness without disrupting primary activities. The theoretical foundation for ambient displays derives from the concept of **"calm technology,"** where information systems provide awareness without demanding explicit attention [49]. This principle is formalized through attention management models that distinguish between peripheral awareness (information perceived without focused attention), focal attention (information requiring deliberate cognitive processing), and transition mechanisms (smooth transitions between peripheral and focal modes) [44]. Human attention can be modeled as a finite resource distributed across multiple information channels. For ambient displays, the attention allocation A_{ambient} must satisfy:

$$A_{\text{ambient}} < A_{\text{threshold}} \tag{2.2}$$

where $A_{\text{threshold}}$ is the maximum attention that can be allocated to peripheral information without disrupting primary tasks. The effective ambient displays typically consume less than 10% of available cognitive attention [18].

Embedded Data Representation Framework

The embedded data representation framework describes how visualizations can be integrated into physical environments. The framework distinguishes between the logical world (where data processing occurs) and the physical world (where data is displayed and users interact). [50]. There are three key principles:

1. **Physical Integration**: Visualizations become part of the physical environment rather than separate information displays.

2. **Contextual Relevance**: Information directly relates to the immediate physical context where it is displayed.

3. **Natural Mapping**: Visual encodings leverage spatial and physical metaphors that align with human spatial cognition.

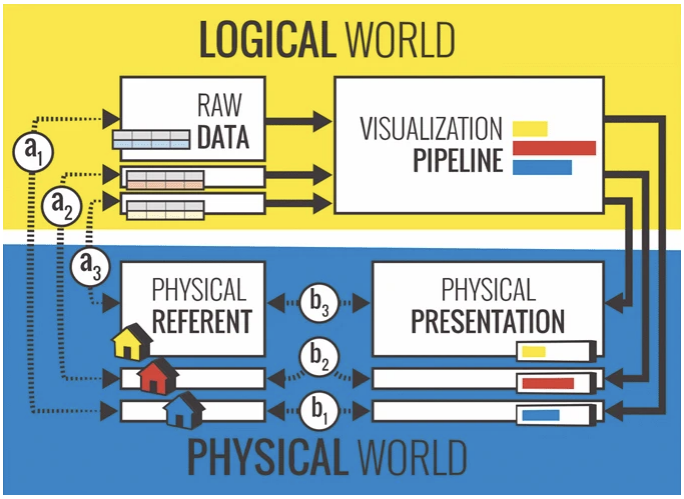


Figure 2.3: Embedded Data Representation Framework [50].

2.2.2 Environmental Data Visualization

Environmental data presents unique challenges for visualization due to its temporal variability, spatial distribution, and the need for real-time representation. Understanding these characteristics is essential for designing effective ambient environmental displays.

Environmental measurements have multiple temporal scales that must be accommodated in visualization design. These scales form a hierarchical structure of environmental awareness, each serving distinct informational purposes. **High-frequency variations** show as moment-to-moment fluctuations occurring over seconds to minutes, and representing immediate environmental changes that require responsive visualization updates to support real-time decision-making. The other scale, **medium-term trends** develop as gradual changes over hours or days, capturing developing environmental patterns that influence user comfort, workflow planning, and adaptive behaviors. Also, the **long-term patterns** emerge as seasonal or cyclical variations spanning weeks to months, providing essential context for understanding current conditions relative to historical norms and allowing predictive insights for future environmental management [40]. Environmental phenomena often shows spatial gradients and patterns that must be represented on display surfaces. For meeting room applications, spatial interpolation techniques allow mapping discrete sensor measurements to continuous surface representations. This creates smooth visualizations that accurately represent environmental conditions across the entire space rather than only at specific sensor locations. [13].

Multi-dimensional Data Integration

Environmental monitoring typically involves multiple simultaneous measurements including *temperature, humidity, noise levels,* and *air quality.* Ambient displays must integrate

these dimensions through layered encoding strategies that prioritize information based on relevance and urgency. Table 2.2 below explains the primary objectives of each encoding layer [5].

Channel	Function
Primary	Handles the most critical measurement for the context (e.g., noise level in meeting rooms where acoustic conditions affect communication).
Secondary	Encodes additional measurements using complementary visual attributes such as color variation, texture patterns, or subtle animation—informative but not overwhelming.
Tertiary	Provides access to detailed environmental data through attention transitions, preserving the ambient nature of the primary display.

Table 2.2: Encoding Layers of Ambient Information

2.3 Spatial Audio Processing Fundamentals

This section covers the signal processing concepts and mathematical foundations required for spatial audio analysis.

2.3.1 Digital Signal Processing for Audio Analysis

Digital audio processing forms the foundation for extracting meaningful information from meeting room acoustic environments. These fundamental principles are required for implementing effective spatial audio analysis systems.

Analog-to-Digital Conversion Principles

Audio signals must be converted from continuous analog waveforms to discrete digital representations for computational processing (Figure 2.4). The analog-to-digital conversion process involves two critical parameters: *sampling rate* and *quantization resolution* [31].

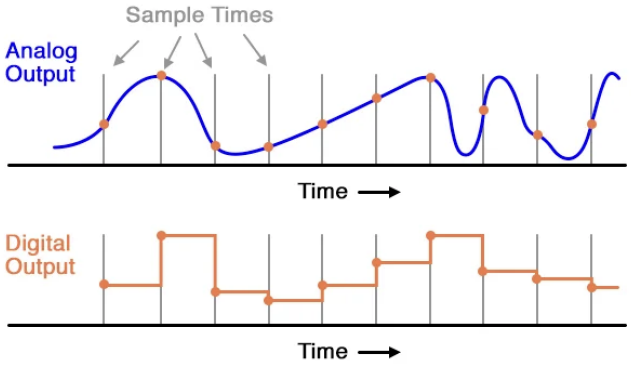


Figure 2.4: Analog-to-Digital conversion rate versus signal type [22].

Sampling Rate Requirements

The Nyquist-Shannon sampling theorem establishes the fundamental constraint for digital audio processing:

$$f\text{sampling} \geqslant 2 \times f\text{max} \tag{2.3}$$

Where $f$sampling represents the sampling frequency and $f$max represents the highest frequency component in the signal. For meeting room audio analysis, human speech typically contains frequencies up to 8 kHz, requiring minimum sampling rates of 16 kHz. However, practical implementations often use 44.1 kHz or 48 kHz to ensure adequate frequency response and avoid aliasing artifacts [31].

Quantization and Dynamic Range

Quantization resolution determines the amplitude precision of digital audio samples. The relationship between bit depth and dynamic range follows:

$$\text{Dynamic\_Range (dB)} = 6.02 \times \text{bit\_depth} + 1.76 \tag{2.4}$$

For meeting room applications requiring accurate volume level measurement, 16-bit quantization provides approximately 96 dB dynamic range, sufficient for most acoustic analysis tasks [12].

2.3.2 Fast Fourier Transform Theory and Applications

The Fast Fourier Transform (FFT) enables frequency-domain analysis of audio signals, converting time-domain waveforms into spectral representations that reveal frequency content and characteristics [34].

The Discrete Fourier Transform (DFT) for a sequence of N samples $x[n]$ is defined as:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j2\pi kn/N} \tag{2.5}$$

Where $X[k]$ represents the frequency-domain representation at frequency bin k. The FFT algorithm efficiently computes the DFT using a radix-2 decimation-in-time approach, as illustrated in the butterfly diagram, achieving computational complexity $O(N \log N)$ rather than $O(N^2)$ for direct computation. The diagram Figure 2.5 shows how the 8-point FFT decomposes the computation into smaller 4-point DFTs, with twiddle factors $W_N^k = e^{-j2\pi k/N}$ representing the complex exponential terms that combine partial results at each stage [34].

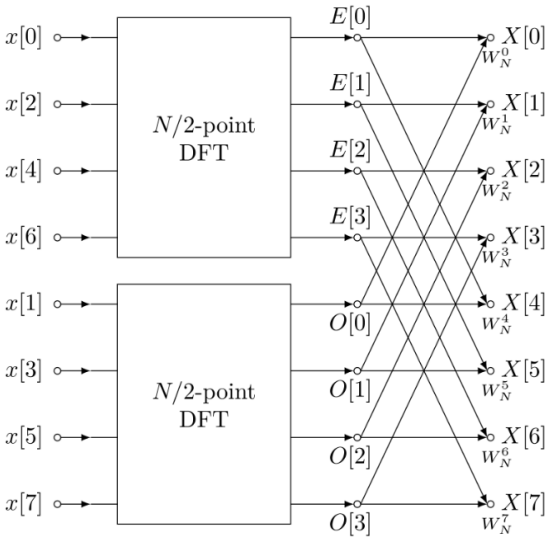


Figure 2.5: An example FFT algorithm structure, using a decomposition into half-size FFTs [29].

The real-world audio signals require windowing to minimize spectral leakage and improve frequency resolution [32], as illustrated in the frequency domain representation Figure 2.6. When signals contain sharp transitions or are truncated abruptly (high frequency characteristics shown on the left), significant spectral leakage occurs, spreading energy across multiple frequency bins. Windowing functions address this by smoothly tapering the signal edges, reducing the high-frequency artifacts and concentrating energy in the appropriate frequency bins (low frequency characteristics shown on the right).

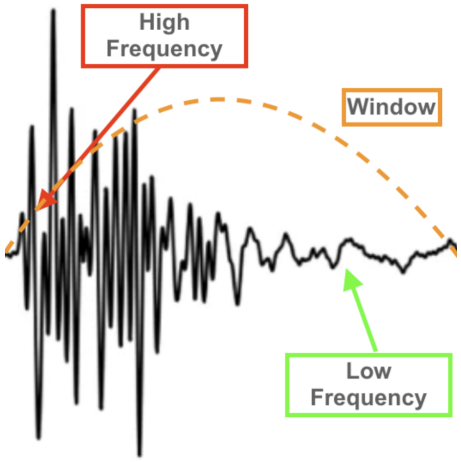


Figure 2.6: Understanding FFT and Windowing [28].

Common windowing functions include [34]:

- **Hamming Window:** $w[n] = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi n}{N-1}\right)$
- **Hann Window:** $w[n] = 0.5 \cdot \left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right)$

FFT analysis involves fundamental trade-offs between frequency resolution and temporal resolution. Frequency resolution Δf is determined by:

$$\Delta f = \frac{f_{\text{sampling}}}{N} \tag{2.6}$$

Where N represents the FFT size. The windowing process, while reducing spectral leakage, also affects this resolution trade-off by determining how much of the signal is analyzed in each time frame.

2.3.3 Spatial Audio Processing with Dual-Microphone Systems

Spatial audio processing enables systems to determine not only what sounds are present, but where they originate and how they move through space. Dual-microphone systems provide a practical approach to spatial audio analysis for meeting room applications.

Binaural Hearing Principles

Human spatial hearing relies on interaural differences between sounds reaching the left and right ears. These same principles apply to dual-microphone audio processing systems. Key spatial cues include [42]:

Interaural Time Difference (ITD)

$$\text{ITD} = \frac{d \cdot \sin\theta}{c} \tag{2.7}$$

Where d represents the distance between microphones, θ is the source angle, and c is the speed of sound (approximately 343 m/s at room temperature).

Interaural Level Difference (ILD)

$$\text{ILD} = 20 \cdot \log_{10}\left(\left|\frac{P_{\text{left}}}{P_{\text{right}}}\right|\right) \tag{2.8}$$

Where P_{left} and P_{right} represent the sound pressure levels at the left and right microphones respectively.

Cross-Correlation Analysis

Spatial localization using dual microphones often employs cross-correlation techniques to identify time delays between channels [45]:

$$R_{xy}(\tau) = \sum_{n} x[n] \cdot y[n+\tau] \tag{2.9}$$

Where $x[n]$ and $y[n]$ represent the left and right microphone signals, and τ represents the time delay.

Meeting room environments present characteristics that require specialized feature extraction approaches beyond traditional speech processing techniques [1]. The formulations below provide the foundations necessary for implementing effective spatial audio processing systems to extract meeting analytics in real-time from dual-microphone arrays.

Volume Level Analysis

Root Mean Square (RMS) analysis provides robust volume level estimation for meeting audio [14]:

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x[n]^2} \tag{2.10}$$

Converting to decibel scale:

$$\text{Level}_{dB} = 20 \cdot \log_{10}\left(\frac{\text{RMS}}{\text{Reference\_level}}\right) \tag{2.11}$$

Voice Activity Detection

Distinguishing speech from background noise requires energy-based and spectral-based approaches [30]:

Energy-based VAD:

$$\text{Energy}[n] = \sum_{k=\text{low}}^{\text{high}} |X[k]|^2 \tag{2.12}$$

Where the summation occurs over frequency bins corresponding to human speech (typically 300-3400 Hz).

Turn-Taking Pattern Analysis

Meeting dynamics can be characterized through temporal analysis of speaking patterns [24]:

$$\text{Speaking}_{\text{ratio}}[i] = \frac{\text{Total speaking time}[i]}{\text{Total meeting time}} \tag{2.13}$$

$$\text{Turn}_{\text{rate}} = \frac{\text{Number of speaker changes}}{\text{Meeting duration}} \tag{2.14}$$

$$\text{Overlap\_ratio} = \frac{\text{Overlapping speech time}}{\text{Total speech time}} \tag{2.15}$$

2.4 Machine Learning for Audio Classification

Machine learning for audio classification relies fundamentally on extracting meaningful numerical features from raw audio signals [33]. Feature engineering transforms high-dimensional time-series audio data into compact representations that capture relevant acoustic characteristics while remaining computationally tractable [23]. This section defines machine learning concepts and algorithms used for audio classification in this thesis.

2.4.1 Feature Engineering Fundamentals

Traditional audio feature extraction relies on descriptors. These descriptors are designed to represent acoustically and perceptually meaningful properties. Mel-Frequency Cepstral Coefficients (MFCCs) provide compact spectral representations based on human auditory perception model. The computation involves several steps [3]:

Mel-scale frequency warping:

$$\text{mel}(f) = 2595 \times \log_{10}(1 + f/700) \tag{2.16}$$

Cepstral analysis:

$$\text{MFCC}[n] = \sum_{k=1}^{K} \log(S[k]) \times \cos\left(\frac{\pi n(k - 0.5)}{K}\right) \tag{2.17}$$

Where $S[k]$ represents the mel-scale filterbank outputs and K is the number of filterbank channels.

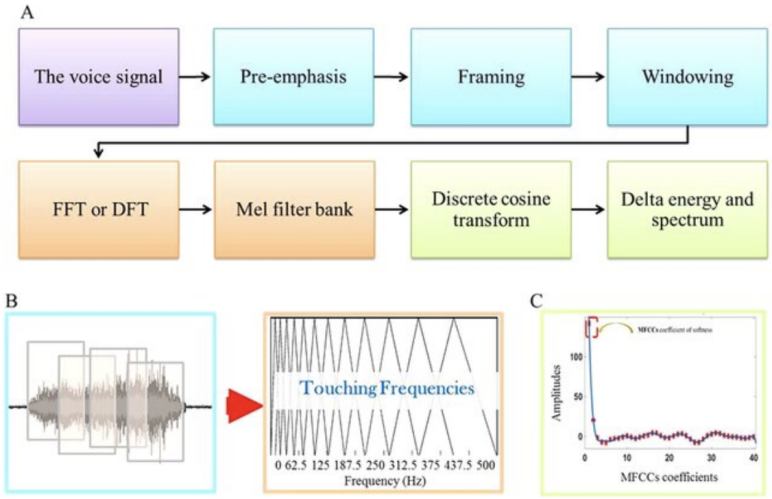


Figure 2.7: Mel Frequency Cepstral Coefficients algorithm [2].

In Figure 2.7, (A) Block diagram of the MFCC algorithm. Four phases can be identified. 1- Read the vibratory signal sample. 2- Split the audio signal into distinct "frames". 3-Compute the Mel-spaced filterbank; this is a triangular filter (illustration in B part) that were applied to the periodogram power spectral estimate from step 2. 4-Take the Discrete Cosine Transform (DCT) of the log filter bank energies to give cepstral coefficients (illustration in C part). (B) A Mel-filterbank adapted to correspond to the mechanoreceptors frequencies. (C) Cepstral coefficients of a vibratory signal. The highlighted coefficient is the sum of how much energy exists in the range of (0 to 500Hz) and is indicated in the softness coefficient.

2.4.2 Meeting-Specific Feature Engineering

Based on observations during the project development process, meeting room audio classification requires specialized features that capture collaborative dynamics and environ-

mental characteristics add to traditional speech processing applications. Therefore, the following classification features were studied.

Spectral Features

Spectral characteristics provide frequency-domain information about audio content [33]:

$$\text{Spectral Centroid} = \frac{\sum_k k \times |X[k]|^2}{\sum_k |X[k]|^2} \tag{2.18}$$

Spatial Audio Features

Dual-microphone systems enable extraction of spatial characteristics that indicate speaker positioning and acoustic scene structure [43]:

Stereo Difference Features:

$$\text{Level\_difference} = 20 \times \log_{10}(|L[k]|/|R[k]|) \tag{2.19}$$

$$\text{Phase\_difference} = \angle(L[k]) - \angle(R[k]) \tag{2.20}$$

Where $L[k]$ and $R[k]$ represent left and right channel frequency components.

Temporal Dynamic Features

Meeting dynamics exhibit temporal patterns that require specialized feature extraction [33]:

Volume Dynamics:

$$\text{Volume\_variance} = \text{var}(\text{RMS}[n]) \text{ over sliding window} \tag{2.21}$$

$$\text{Volume\_trend} = \text{slope}(\text{RMS}[n]) \text{ over analysis window} \tag{2.22}$$

$$\text{Speaker\_change\_rate} = \frac{\text{Number\_of\_changes}}{\text{Window\_duration}} \tag{2.23}$$

$$\tag{2.24}$$

2.4.3 Classification Algorithms

Ensemble learning methods address the multi-faceted nature of audio classification tasks, and they are effective in handling complex audio data which supports their adoption in meeting analytics contexts [41, 27]. Two primary algorithms form the foundation of our system: Random Forest for categorical classification and Gradient Boosting for continuous regression. The selection of these algorithms motivated us by their complementary strengths in handling heterogeneous feature sets, providing robust performance across diverse audio scenarios, and offering interpretable results through feature importance analysis [10].

Random Forest Classification

Random Forest serves as the primary classification algorithm for three of the four meeting analytics tasks. The algorithm addresses the inherent variability in meeting audio characteristics through ensemble diversity and bootstrap aggregation. Random Forest employs bootstrap aggregating (bagging) combined with feature randomization to create diverse decision trees that collectively provide robust predictions [19]:

Bootstrap Sampling: For each tree t in T total trees, a bootstrap sample is generated:

$$\text{Bootstrap\_sample}[t] = \text{random\_sample\_with\_replacement}(\mathcal{D}_{\text{train}}, n) \tag{2.25}$$

$$\text{Tree}[t] = \text{train\_decision\_tree}(\text{Bootstrap\_sample}[t]) \tag{2.26}$$

where $\mathcal{D}_{\text{train}}$ represents the training dataset and n is the sample size.

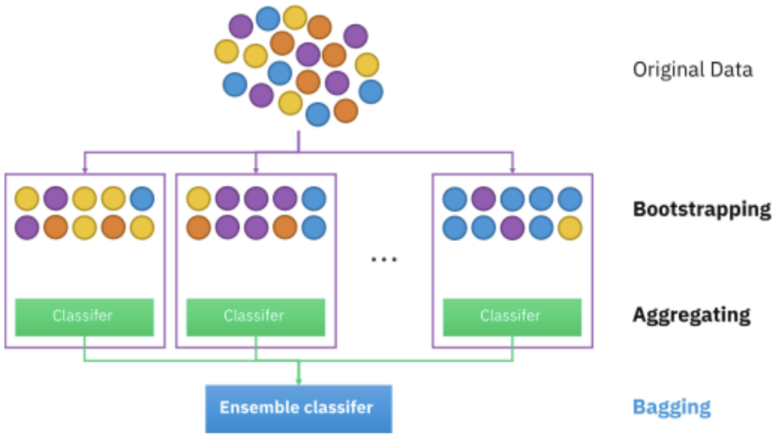


Figure 2.8: Bootstrap aggregating (bagging) process in Random Forest classification. The original training dataset is resampled with replacement to create multiple bootstrap samples, each used to train an individual decision tree classifier. The final ensemble classifier combines predictions from all individual trees through majority voting [21].

Feature Randomization: At each node split, only a subset of features is considered to ensure tree diversity:

$$\text{candidate\_features} = \text{random\_subset}(\mathcal{F}_{\text{all}}, \sqrt{|\mathcal{F}_{\text{all}}|}) \tag{2.27}$$

$$\text{best\_split} = \arg \min_{f \in \text{candidate\_features}} \text{Gini\_impurity}(f) \tag{2.28}$$

where $\mathcal{F}_{\text{all}}$ represents the complete feature set and the square root rule determines the number of candidate features per split.

Gini Impurity Calculation: The splitting criterion optimizes the Gini impurity measure:

$$\text{Gini}(S) = 1 - \sum_{i=1}^{C} p_i^2 \tag{2.29}$$

where S is the dataset subset, C is the number of classes, and p_i is the proportion of samples belonging to class i.

Prediction Aggregation

Final predictions combine outputs from all trees through majority voting:

$$\text{class\_prediction} = \text{mode}(\{\text{Tree\_prediction}[t]\}_{t=1}^{T}) \tag{2.30}$$

$$\text{confidence} = \frac{\max(\text{class\_votes})}{T} \tag{2.31}$$

$$\text{importance}[f] = \sum_{t=1}^{T} \sum_{n \in \text{nodes}[t]} \mathbb{I}(\text{feature}[n] = f) \cdot \Delta\text{impurity}[n] \cdot \frac{|\text{samples}[n]|}{|\mathcal{D}_{\text{train}}|} \tag{2.32}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\Delta\text{impurity}[n]$ represents the impurity decrease at node n, and the sample fraction weights the contribution of each split.

Hyperparameter Optimization

Grid search optimization ensures optimal model performance across the parameter space:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \Theta} \text{CV\_accuracy}(\text{RandomForest}(\boldsymbol{\theta})) \tag{2.33}$$

$$\Theta = \{n_{\text{estimators}}, \text{max\_depth}, \text{min\_samples\_split}\} \tag{2.34}$$

Parameter Space:

$$n_{\text{estimators}} \in \{50, 100, 200\} \tag{2.35}$$

$$\text{max\_depth} \in \{5, 10, \text{None}\} \tag{2.36}$$

$$\text{min\_samples\_split} \in \{2, 5, 10\} \tag{2.37}$$

The cross-validation accuracy serves as the optimization objective:

$$\text{CV\_accuracy} = \frac{1}{K} \sum_{k=1}^{K} \frac{\text{correct\_predictions}_k}{\text{total\_predictions}_k} \tag{2.38}$$

where K represents the number of cross-validation folds.

2.4.4 Gradient Boosting for Engagement Regression

Gradient Boosting Regression addresses the continuous nature of engagement scoring (fourth meeting analytics tasks), building strong predictors through iterative weak learner combination. This approach is well suited for the complex, non-linear relationships [16] between audio features and engagement levels.

Gradient Boosting

The algorithm iteratively improves predictions by fitting new models to residual errors [25]:

Initialization:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \tag{2.39}$$

where L is the loss function (squared error for regression) and γ is the initial constant prediction.

Iterative Improvement: For iterations $m = 1$ to M:

Step 1 - Compute Pseudo-residuals:

$$r_{i,m} = -\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \tag{2.40}$$

For squared loss: $r_{i,m} = y_i - F_{m-1}(x_i)$

Step 2 - Fit Weak Learner:

$$h_m = \arg\min_h \sum_{i=1}^{n} (r_{i,m} - h(x_i))^2 \tag{2.41}$$

Step 3 - Optimize Step Size:

$$\gamma_m = \arg\min_\gamma \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{2.42}$$

Step 4 - Update Model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{2.43}$$

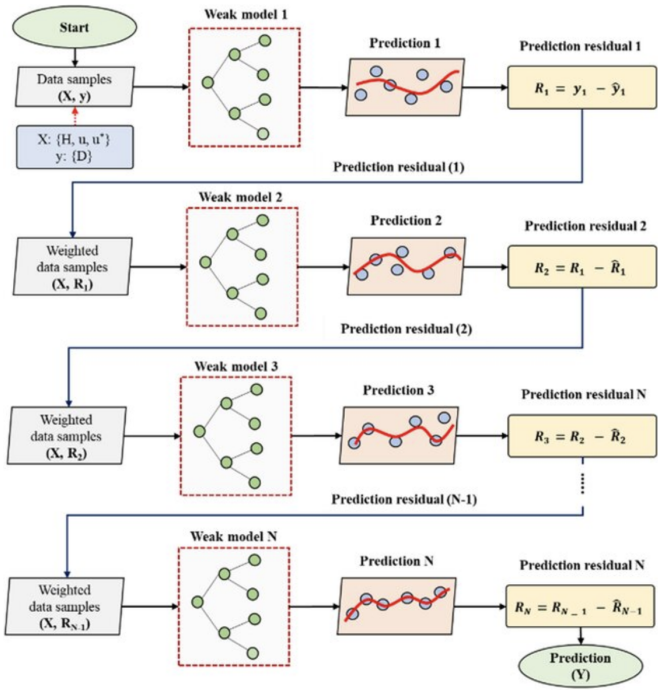


Figure 2.9: Gradient boosting algorithm for regression. Each weak model (decision tree) is sequentially trained on the residuals from the previous model's predictions, iteratively reducing prediction errors. The process continues until convergence, with each model contributing to the final ensemble prediction [4].

Chapter 3

Related Works

This chapter examines existing research across domains relevant to ambient meeting room awareness systems using non-planar displays. We organize the review into four main areas: visualization beyond desktop displays, non-planar display technologies, situated and ambient visualization approaches, and meeting analytics with spatial audio processing. For each area, we identify how prior work informed the design decisions and technical approaches used in this thesis.

3.1 Visualization Beyond Desktop Displays

This section reviews research on visualization technologies beyond traditional desktop displays, examining how prior work motivated our exploration of meeting room applications for non-planar displays.

The movement toward visualization beyond traditional desktop displays emerged from recognition that diverse display technologies require fundamental rethinking of visualization design principles. This shift recognizes that different contexts, form factors, and user needs demand specialized visualization approaches rather than simple adaptations of desktop metaphors [38]. The concept has evolved into "ubiquitous analytics", which integrates data visualization into everyday environments and activities. This approach emphasizes the idea of "interacting with big data anywhere, anytime" and represents a shift beyond the traditional constraints of workstation-based analysis [15]. Vermeulen et al.(2021) provide comprehensive analysis of the ubiquitous visualization landscape, identifying key challenges including adaptation to diverse form factors, context-aware information presentation, and the need for ambient versus focused attention modes [47]. These theoretical foundations served as a staring point to understand that non-planar displays require specialized visualization techniques rather than adaptations of desktop solutions. It also emphasized the importance of designing for peripheral awareness, where participants need to balance their focus on collaboration with maintaining environmental consciousness.

3.2 Non-Planar Display Technologies and Form Factor Research

This section reviews research on spherical and cylindrical displays, focusing on studies that informed our display form factor selection and visualization design.

Spherical Displays

Spherical displays have been explored primarily for interactive applications and public installations, with studies demonstrating fundamental interaction techniques for curved surfaces [6]. Comprehensive real-world evaluations through large-scale deployments have further assessed user interaction with spherical displays in public environments, providing insights on engagement patterns, optimal content types, and user behavior [51]. This finding supports our choice of spherical displays for meeting rooms with circular seating arrangements, where equal viewing access is essential. However, prior work has largely focused on interactive public installations rather than ambient information display applications, highlighting a research gap that our work addresses.

Cylindrical Displays

Cylindrical displays have received extensive research attention, particularly for public display applications. The studies on user behavior around cylindrical displays revealed how framing influences user positioning patterns, showing that physical display configuration affects how people interact with curved surfaces [8]. Longitudinal evaluations of audience behavior around large interactive cylindrical screens demonstrated that cylindrical geometry creates distinct engagement patterns compared to flat displays [7]. We consider these findings when determining optimal suspension height and positioning for spherical displays relative to seating arrangements. Their behavioral observations also suggest that curved displays can attract and sustain attention differently than planar displays, supporting the potential effectiveness of spherical displays for ambient awareness applications.

Form Factor Comparison and Selection Criteria

Vega et al. specifically addressed "visualization on spherical displays," identifying unique challenges including perspective distortion, optimal viewing distances, and information layout strategies [46]. Their work provided practical guidance for adapting visualization techniques to spherical geometries, directly informing our WebGL shader programming approach for handling curved surface rendering and our decision to use directional color mapping rather than traditional visualizations.

Spherical Display Characteristics	**Cylindrical Display Characteristics**
360° visibility suitable for circular meeting arrangements	Continuous horizontal presentation for linear configurations
Natural mapping for environmental phenomena	Efficient vertical space utilization
Uniform accessibility from all meeting positions	Simpler content adaptation
Complex projection requirements	Limited vertical viewing optimization

Table 3.1: Form factor comparison informing our display selection [8, 46].

This comparison, combined with our meeting room's circular architecture, led to our selection of spherical displays for the prototype implementation.

3.3 Situated and Ambient Visualization

This section reviews research on situated visualization and ambient displays,presenting how prior work informed our environmental data visualization approach.

Theoretical Foundations of Situated Visualization

Situated visualization represents a shift from traditional desktop-based data analysis toward context-aware information display, providing the theoretical foundation for our meeting room application. A comprehensive survey defines situated visualization as "visualization that is displayed in proximity to the phenomenon it represents," emphasizing the principle of spatial coupling [11], which applies to our work where environmental sensors - our dual MAX4466 microphone amplifiers are located in the same meeting room as the spherical display. This creates immediate spatial relevance between data source and visualization location. This concept is further formalized through the notion of "embedded data representations" (see Chapter 2), which demonstrate how visualizations can be integrated into physical environments to support situated understanding [50]. Following this framework, we integrate the spherical display into the meeting room's physical architecture rather than treating it as a separate information appliance.

Ambient Information Display Systems

Research on energy-harvesting situated displays has demonstrated how ambient information systems can be deployed sustainably in diverse environmental contexts, highlighting technical considerations such as power consumption, update frequencies, and visual attention management [18]. We draw upon these insights when designing our system architecture and visualization update strategies. The principle of ambient information display aligns with the concept of calm technology, where systems provide awareness without demanding explicit attention, making this approach well-suited for meeting environments where participants must focus on collaboration while remaining peripherally aware of environmental conditions that may influence meeting quality.

3.4 Meeting Analytics and Spatial Audio Processing

This section examines research on meeting room sensing and spatial audio processing, identifying how prior work informed our technical implementation decisions.

Meeting Room Sensing and Analytics

The integration of sensor technologies into meeting environments has enabled new approaches to understanding and enhancing collaborative work. Meeting analytics systems aim to capture, process, and interpret environmental and behavioral data to support meeting effectiveness and participant well-being. Foundational work in the recognition and

understanding of meetings demonstrated how multiple sensor modalities can be combined to automatically analyze meeting dynamics, showing that audio analysis provides particularly rich information about meeting structure, participant engagement, and collaborative patterns [36]. These findings directly support our focus on spatial audio processing as the primary sensing modality for meeting awareness. Subsequent research extended this direction through the automatic analysis of multimodal group actions in meetings, developing techniques for real-time understanding using audio, video, and environmental sensors [26] (see Figure 3.1). This multimodal approach highlights the value of integrating diverse sensing modalities to create comprehensive meeting analytics systems, informing our integration of spatial audio data with environmental sensing for ambient display applications.

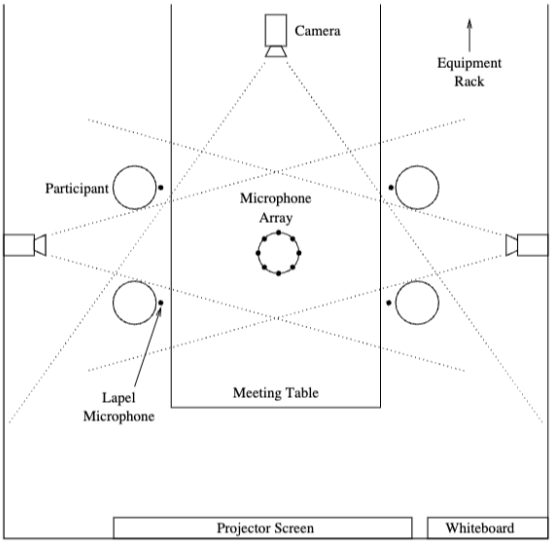


Figure 3.1: Meeting Room Configuration: the room has been equipped with fully synchronised multi-channel audio and video recording facilities [26]

Spatial Audio Processing Fundamentals

Spatial audio processing enables systems to understand not just what sounds are present, but where they originate and how they move through space. This capability is crucial for meeting analytics, as speaker location, turn-taking patterns, and acoustic dynamics provide rich information about meeting structure and participant engagement. Brandstein and Ward [9] establish the technical foundations for microphone array processing, demonstrating how multiple microphone inputs can be processed to extract spatial audio information. Their work on beamforming, source localization, and acoustic scene analysis provides the signal processing foundation for our dual-microphone approaches. Wölfel and McDonough [52] address processing challenges in meeting environments through their treatment of distant speech recognition. Their work guided our preprocessing approach, leading us to implement voice frequency filtering (300-3400 Hz) to focus on human communication and use Hamming windowing to reduce spectral leakage in our FFT analysis.eting audio processing.

3.5 Machine Learning for Meeting Audio Classification

This section reviews machine learning approaches for automatically analyzing meeting audio to extract behavioral insights such as speaker count, meeting type, and engagement levels. It establishes the academic foundation for using ensemble learning methods and synthetic label generation to transform raw audio features into meaningful meeting analytics for ambient visualization systems.

Ensemble Methods in Audio Classification

Traditional approaches to audio classification have relied primarily on single-model architectures, but Zhang et al. (2013) [53] establish that ensemble methods including Random Forest, Boosting, and Multiple Classifier Systems consistently outperform single classifiers in environmental audio classification. Their work demonstrates that ensemble approaches maintain robust performance even with limited training data, a critical consideration for meeting analytics where labeled behavioral data is expensive to obtain. Given the time constraints during the data collection we were able to obtain limited data too. Therefore, this finding motivated our decision to employ ensemble methods for meeting audio classification, combining Random Forest for speaker count detection, Gradient Boosting for engagement prediction.

Feature Engineering for Audio Behavioral Analysis

The extraction of meaningful features from audio signals forms the foundation of automated meeting analysis. Ren et al. (2025) [35] demonstrate advanced multi-dimensional feature extraction techniques, showing how temporal, frequency, global, and local audio characteristics can be weighted and combined to improve sound event detection by up to 9.94%. Their Group Feature Calibration approach validates the principle of weighted feature combination that underlies our engagement scoring methodology. We adopt their multi-dimensional framework, extracting volume variation patterns for temporal engagement indicators, stereo dynamics for spatial interaction analysis, and peak density measures for activity level detection.

Engagement Score Synthesis Through Weighted Feature Combination

Given the absence of engagement ground truth labels in our dataset, we developed a theoretically-grounded engagement scoring formula based on established principles of multi-dimensional audio feature analysis [35] and behavioral signal processing [48]. Following the multi-dimensional feature weighting approach demonstrated by Ren et al. 2025, where different audio characteristics are weighted according to their contribution to behavioral indicators, our engagement score synthesis combines five complementary audio features. The formula weights volume variation patterns to capture engagement indicators such as peak density measures for speaker activity levels, stereo dynamics for multi-participant interaction patterns, speaker balance metrics, and energy change patterns for conversation dynamics. This approach addresses the methodological challenge identified by Zhang et al. (2013) [53] regarding limited training data availability, providing an

alternative to manual annotation while maintaining consistency with established audio-to-behavior modeling frameworks. The weighted combination approach enables our system to generate continuous engagement metrics.

Engagement Score Formula:

$$E = (V_{std} \times 0.20) + (P_{density} \times 0.15) + (S_{variation} \times 0.25) + (B_{balance} \times 0.10) + (C_{changes} \times 0.20) \tag{3.1}$$

Where:

- V_{std}: Volume variation (standard deviation of audio levels)
- $P_{density}$: Peak count normalized by session length
- $S_{variation}$: Stereo channel variation patterns
- $B_{balance}$: Speaker balance between left/right channels
- $C_{changes}$: Rate of energy level transitions

This formula provides engagement scores normalized to a 0-100 scale, enabling consistent behavioral assessment across different meeting contexts and durations.

3.6 Research Opportunities and Our Approach

Based on this literature review, we identified specific opportunities that motivated our thesis work. While non-planar display research focuses on interactive applications and spatial audio research emphasizes computational analysis, we explored combining these domains specifically for ambient meeting room awareness. Existing meeting analytics present post-meeting analysis, while ambient display research typically uses simple environmental parameters, so we investigated real-time meeting audio visualization for ambient awareness during collaborative activities. Spatial audio research concentrates on speech recognition and source localization, while spherical display research focuses on interactive content, leading us to explore how dual-microphone spatial audio data could be meaningfully mapped to ambient spherical visualizations. Situated visualization principles exist for general environmental data, but we investigated these specifically for collaborative meeting environments with their unique social dynamics and attention requirements.

Our contribution lies in the technical integration and initial exploration of these research directions through working prototype development. We developed a complete system pipeline from ESP32 dual-microphone sensing through WebSocket communication to Three.js spherical visualization, combined with machine learning classification achieving 73-97% accuracy across meeting analytics tasks. This work represents initial exploration demonstrating technical feasibility and providing a foundation for future research in ambient meeting room awareness using non-planar displays.

Chapter 4

User-Centered Design Methodology

This chapter describes the user-centered design approach we employed to develop ambient meeting room awareness systems using non-planar displays. The methodology included user requirements analysis through surveys, iterative design space exploration, and form factor comparison through virtual prototyping. Our design process integrated human-centered design principles while adapting to the non-planar display applications.

4.1 User Requirements Analysis

4.1.1 Survey Design and Methodology

Our initial research focused on understanding user needs and preferences for public display systems within office laboratory environments. The target context involved public displays deployed in shared spaces of a research laboratory, where employees, PhD students, and researchers regularly collaborate and spend significant portions of their workday. These spaces include common areas like entrances, coffee lounges, break rooms, and meeting spaces where environmental information could enhance awareness and decision-making. The survey questions addressed key design decisions including spatial preferences, content priorities, context mapping, content gaps, utility assessment, engagement potential, and implementation concerns. The complete survey results with detailed respondent analytics are available through this Google Forms analytics page.

Question Category	**Survey Question**
Spatial preferences	Which of these areas of our building would a public display make sense for you?
Content priorities	What kind of information would you like to see displayed publicly in the building?
Context mapping	Where would you like to see each type of information displayed?
Content gaps	What kind of information would you like to see displayed that isn't listed above?
Utility assessment	Would you find such public display systems useful?
Engagement potential	Would you feel more engaged in your building if relevant data was shared through public displays?
Implementation concerns	Do you have any feedback, ideas, or concerns about installing public displays in shared building areas?

Table 4.1: Survey Questions

4.1.2 Participant Recruitment and Data Collection

The survey was distributed online using Google Forms through the INRIA laboratory team communication channels, targeting interns, PhD students, and researchers working within the laboratory environment. This recruitment approach ensured participants had direct experience with the physical spaces under consideration and understanding of the collaborative work activities that would be supported by ambient displays. Data collection occurred over a one-week period, yielding responses from 15 laboratory members.

4.1.3 Results Analysis and Key Findings

The survey results, Figure 4.1 revealed clear spatial preferences and content priorities that informed subsequent design decisions. Building entrance locations received the highest support (80% of respondents), while coffee lounge areas (66.7%) and bridges/cross-buildings/transition areas along with break rooms (both 53.3%) also showed strong preference. Meeting rooms received more limited support (26.7%), suggesting moderate interest in meeting-focused displays.

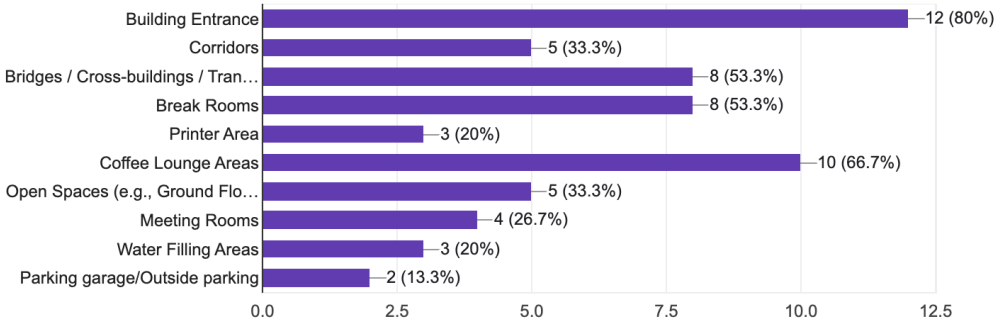


Figure 4.1: Results for the survey question: Which of these areas of our building would a public display make sense for you?

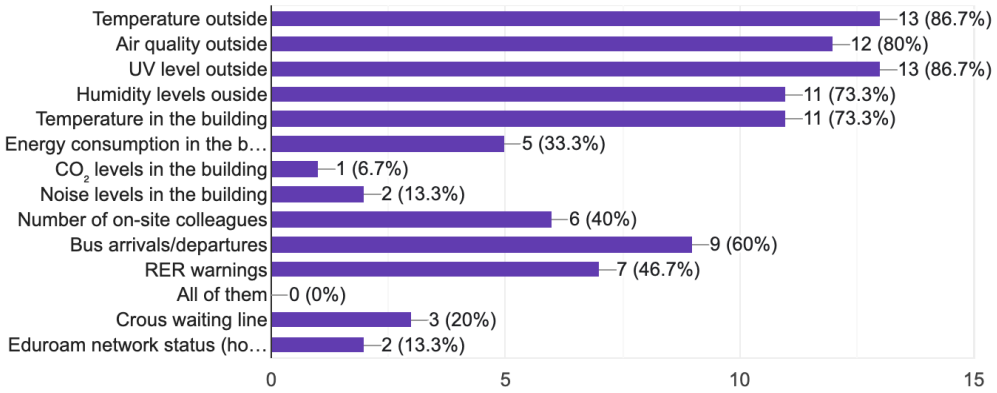


Figure 4.2: Results for the survey question: What kind of information would you like to see displayed publicly in the building?

Content analysis, Figure 4.2 revealed strong preference for environmental and contextual information. Outdoor environmental factors dominated preferences, with temperature outside, UV levels outside (both 86.7%), and air quality outside (80%) receiving highest priority. Transportation information such as bus arrival/departures (60%) and indoor temperature (73.3%) also showed significant interest. Critical environmental factors including CO_2 levels (6.7%) and noise levels (13.3%) received surprisingly low interest, despite their direct impact on indoor environmental quality and collaborative effectiveness. User engagement potential showed, Figure 4.3 positive reception, with 73.3% finding display systems useful and 93.3% (combining "yes" and "maybe" responses) indicating potential for increased building engagement through relevant data sharing. Only 6.7% of respondents expressed no interest in such systems, demonstrating strong overall support for the concept.

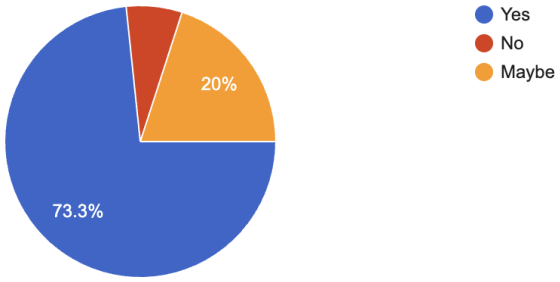


Figure 4.3: Results for the survey question: Would you feel more engaged in your building if relevant data was shared through public displays?

4.1.4 Survey Insights and Design Implications

The survey results initially suggested focusing on building entrance displays showing outdoor environmental data. However, after analysis and discussions with supervisors led us to make a strategic pivot toward meeting room applications for several key reasons.

Firstly, we chose to avoid displaying information that people already have access to through online sources or personal devices. Outdoor temperature, UV levels, and transportation data are readily available on smartphones and weather applications. Instead, we focused on environmental data that is typically "invisible" to users - CO2 levels and acoustic conditions that are not available through personal devices but directly impact meeting effectiveness.

While building entrances received high preference scores, people spend significantly more time in meeting rooms than in transition areas. This extended exposure provides better opportunities for studying the impact of ambient displays on user behavior and environmental awareness. Meeting rooms also offer controlled environments for evaluation and testing.

The circular architecture of our target meeting room presented a perfect fit for spherical displays, allowing us to take full advantage of the non-planar form factor. In such spaces, spherical displays can be positioned to be visible from all seating positions, maximizing

the benefits of 360-degree visibility.

The low interest in CO2 and noise levels in the survey was interpreted not as lack of importance, but as lack of awareness regarding these environmental factors. This presented an opportunity to make invisible environmental data visible and relevant, addressing a genuine need that users might not recognize they have.

4.2 Design Space Exploration

4.2.1 Initial Design Concepts and Form Factor Exploration

Building on survey insights and theoretical foundations from situated visualization research, our initial design exploration examined multiple form factors and application contexts. The design concept exploration encompassed cylindrical, rectangular, spherical, and cubic display geometries, each evaluated for their potential in different spatial contexts. This work grounded our understanding of how different geometries might support environmental data visualization. Two key transitions shaped our design direction during this phase. First, we shifted from dashboard-style explicit visualization to ambient display approaches, recognizing that meeting environments require displays that provide awareness without attracting focused attention away from the primary collaborative task. Second, we focused specifically on audio sensing for practical reasons - audio processing provides rich information about meeting dynamics while requiring simpler hardware setup compared to video-based approaches that raise privacy concerns. Figure 4.4 illustrates our initial design concepts, showing multiple form factors displaying air quality information with situated integration (positioned in relevant physical contexts) and color coding for environmental conditions. These concepts established three key design principles: multiple form factor options to match different architectural contexts, situated integration to provide spatial relevance for environmental data, and color coding strategies to create intuitive mappings between environmental conditions and visual representations.

Figure 4.4: Initial design concepts of multiple form factor options

4.2.2 AI-Assisted Design Generation

To explore the design space more comprehensively, we initially created AI-generated design concepts to visualize ambient environmental displays showing CO2 levels in realistic meeting contexts. Figure 4.5 shows these AI-generated concepts integrated into the actual

INRIA meeting room, demonstrating how spherical displays could provide environmental awareness through color changes responding to CO2 concentration levels. However, our design direction evolved significantly following a presentation to tVISt project partners (visualization researchers). During this collaborative session, we received valuable input that led us to transition from CO2 monitoring to spatial audio processing. The feedback highlighted that audio-based environmental sensing offered several advantages: richer real-time information about meeting dynamics, more immediate visual feedback opportunities, and fewer hardware deployment constraints compared to CO2 sensors which require careful calibration and positioning. This strategic shift from CO2 to audio sensing represented a pivotal moment in our design process, transforming our ambient displays from static environmental monitoring to dynamic meeting awareness systems that could respond to conversation patterns, speaker positioning, and acoustic energy levels in real-time.

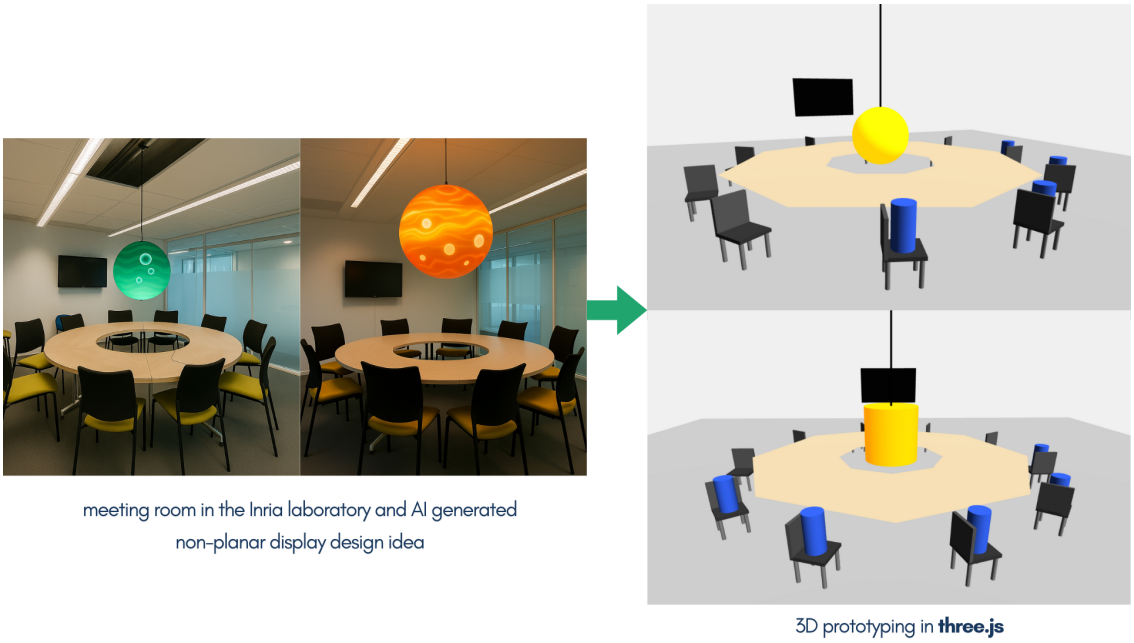


Figure 4.5: AI generated meeting room displays and initial prototyping framework developed using Three.js with WebGL

4.3 Three.js and WebGL-Based Prototyping Framework

Given the prohibitive cost of physical spherical displays, we developed a virtual prototyping framework using Three.js with WebGL rendering capabilities which is connected to a real data-stream capturing audio data.

4.3.1 Advanced Virtual Prototyping Architecture

The Three.js and WebGL implementation (see Figure 4.5) provided several critical advantages over traditional 2D prototyping approaches: hardware-accelerated 3D rendering enabling complex geometric transformations and lighting calculations, real-time shader

programming for sophisticated visual effects and environmental data encoding, accurate perspective rendering simulating actual viewing conditions around curved displays, and seamless integration with WebSocket protocols for live sensor data streaming.

4.3.2 Multi-Form Factor Implementation and Comparison

The Three.js framework allowed us to compare different non-planar geometries (Figure 4.6) through parametric modeling and real-time manipulation. We implemented three distinct display forms to explore their suitability for meeting room applications: spherical displays providing 360-degree visibility optimal for circular meeting arrangements, cylindrical displays offering continuous horizontal information presentation suitable for linear configurations, and conical displays combining characteristics of both geometries.

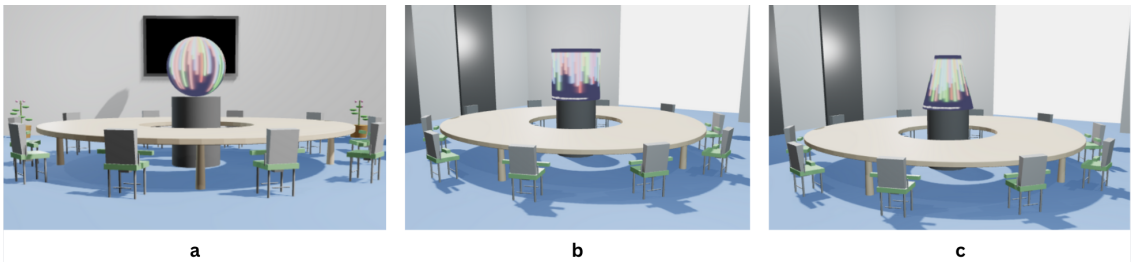


Figure 4.6: Different non-planar geometries: (a)spherical, (b)cylindrical, (c)conical

Spherical Display Implementation

The spherical display utilized high-resolution geometry (128×128 subdivisions) to ensure smooth surface rendering and accurate texture mapping. The implementation incorporated multiple visualization layers: a base spherical mesh with custom shader materials for environmental data representation, a wireframe overlay providing subtle geometric context without overwhelming the ambient visualization, internal particle systems and audio visualization lines contained within the sphere volume, and dynamic lighting systems responding to environmental conditions.

Cylindrical and Conical Display Implementation

The cylindrical alternative used optimized geometry suitable for horizontal information presentation, featuring parametric generation allowing for dynamic height and radius adjustment. Add to this, a seamless texture wrapping for continuous horizontal data display, and specialized lighting models optimized for curved vertical surfaces. While conical form factor provided a hybrid approach, combining the 360-degree accessibility of spherical displays with the directional characteristics of cylindrical forms.

4.3.3 Parametric Design Optimization and Height Configuration

Recognizing that display positioning significantly impacts both visibility and social dynamics in collaborative environments, the system incorporated comprehensive parameter testing capabilities, see Figure 4.7. The implementation included five distinct stand height

configurations, enabling systematic assessment of optimal positioning for meeting room

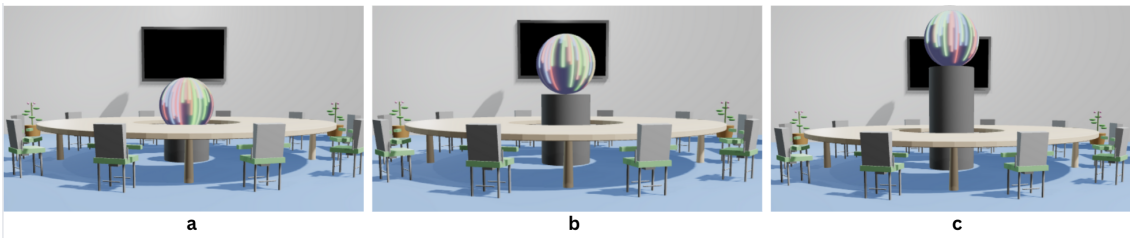


Figure 4.7: Distinct stand height configurations: (a)lowest, (b)medium, (c)highest

applications while considering both visual accessibility and interpersonal sight lines between participants. The height adjustment feature proved particularly valuable for understanding trade-offs between display visibility and social interaction preservation. Lower positions improved display visibility but potentially interfered with cross-table eye contact, while higher positions maintained interpersonal sight lines but reduced ambient display accessibility. This parametric approach provided empirical data for design optimization decisions.

Meeting Dynamics Simulation

The system included sophisticated meeting simulation capabilities enabling exploration of how environmental conditions change with varying participant numbers and activities. The “Add Person” functionality demonstrated real-time visualization changes as meeting occupancy increased, providing insights into how CO_2 levels, noise dynamics, and spatial audio characteristics evolve during meeting progression Figure 4.8. When the user clicks on the "Add Person" button the person (blue figure) appears on the chair and the display color changes gradually with each adding person. This simulation capability proved invaluable for understanding the relationship between meeting context and environmental visualization requirements. Stakeholders could observe how display behavior adapted to different meeting scenarios, from small discussions to larger collaborative sessions, providing concrete evidence of system responsiveness.

4.3.4 WebSocket Integration and Real-time Data Processing

The prototyping framework incorporated advanced real-time data communication enabling authentic ambient display behavior. WebSocket integration provided bidirectional communication between the ESP32 sensor system and the Three.js visualization environment, supporting multiple concurrent data streams. Details of WebSocket integration and hardware setup architecture are provided in Chapter 5.

Data Communication Architecture

The WebSocket implementation handled multiple data streams simultaneously: dual-microphone audio levels with spatial positioning information, environmental sensor readings including temperature and humidity data, machine learning classification results for

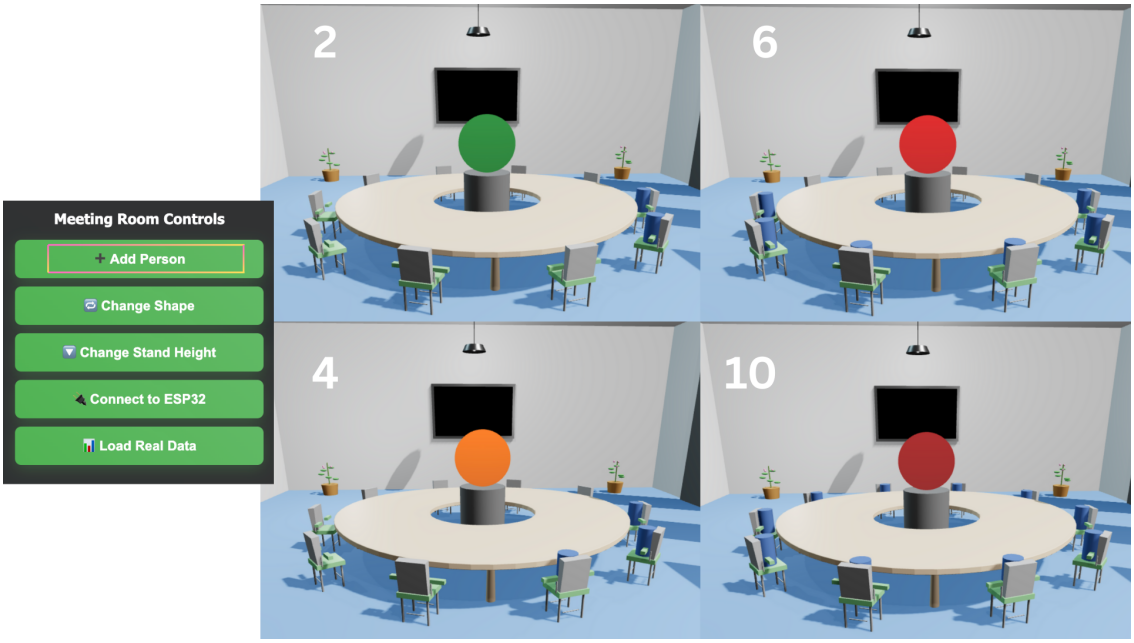


Figure 4.8: "Add Person" functionality to observe CO2 level changes.

meeting type and engagement analysis, and system status information ensuring reliable operation.

Listing 4.1: WebSocket Data Handling Implementation

```
// WebSocket data handling
webSocket.onmessage = function(event) {
    const audioData = JSON.parse(event.data);

    // Update shader uniforms with real-time data
    sphereMaterial.uniforms.leftAudio.value =
        normalizeAudioLevel(audioData.leftMic);
    sphereMaterial.uniforms.rightAudio.value =
        normalizeAudioLevel(audioData.rightMic);
    sphereMaterial.uniforms.time.value = Date.now() * 0.001;

    // Update visualization elements
    updateAudioVisualization(audioData);
};
```

Dual-Mode Data Integration

The system supported two distinct data input modes to accommodate different evaluation and demonstration requirements. Real-time ESP32 integration provided live environmental monitoring with immediate visualization updates based on actual meeting room conditions. The "Load Real Data" functionality (see Figure 4.9) enabled playback of pre-collected environmental datasets, providing consistent demonstration experiences and enabling detailed analysis of visualization behavior under controlled conditions.

4.3.5 Visualization Modes and Interactive Interface

The completed system featured four visualizations available via a unified selection interface, allowing users to investigate various methods for presenting environmental data on non-planar surfaces. The multi-modal environmental visualization mode selection bar provided access to specialized display modes tailored for different environmental monitoring applications. Table 4.2 below details these modes and their respective descriptions.

Visualization Mode	**Description**
Audio 3D	Three-dimensional spatial audio visualization showing sound source positioning and intensity patterns across the meeting space
Waves	Fluid wave animations reflecting acoustic energy patterns and sound propagation characteristics
Stereo Chart	Analytical visualization displaying left-right audio channel differences and spatial audio characteristics
Activity	Meeting engagement visualization showing participant activity levels and turn-taking patterns

Table 4.2: Visualization Modes for Environmental Data Representation

Interactive Control System

The comprehensive control interface enabled real-time manipulation of key design parameters during evaluation sessions. The Meeting Room Controls panel (see Figure 4.9) provided access to participant simulation ("Add Person"), form factor switching ("Change Shape"), height optimization ("Change Stand Height"), and connectivity management ("Connect to ESP32", "Load Real Data"). The Audio Data panel displayed real-time environmental monitoring information including left and right microphone levels, stereo difference calculations, connection status, and data source indicators. This dual-panel approach separated design exploration controls from environmental monitoring feedback, supporting both evaluation activities and demonstration requirements.

4.4 Evaluation Approach and Future Directions

Informal Feedback and Design Validation

We conducted informal feedback sessions with laboratory colleagues to assess design effectiveness and gather suggestions for improvement. The feedback focused on visualization clarity, informational utility, and ambient display characteristics. Stakeholders emphasized the need for more informative and useful visualizations that clearly demonstrated the relationship between participant behavior and environmental changes. They particularly appreciated the integration of machine learning models and data augmentation techniques, recognizing these as technical contributions that enhanced system intelligence. The overall design concept received positive reception, with particular support for appli-

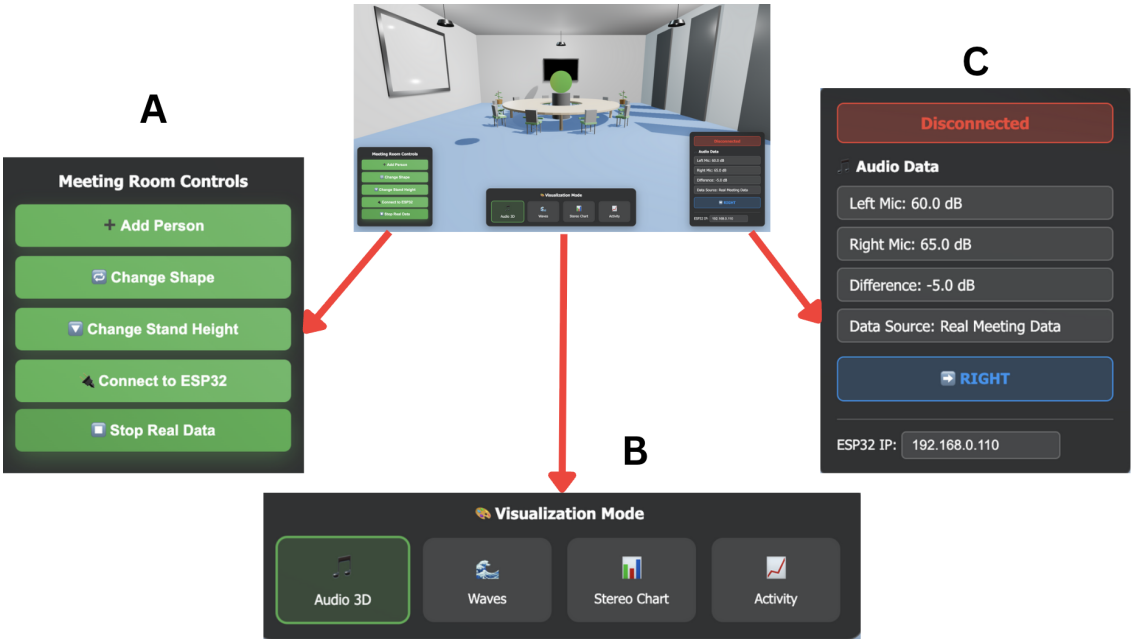


Figure 4.9: (A) - the Meeting Room Controls panel, (B) - Visualization Modes, (C) - the Audio Data panel

cation of spherical displays in meeting room contexts and the focus on making invisible environmental data visible and relevant.

Limitations and Future Evaluation Plans

Our current evaluation approach has limitations that we acknowledge. The informal feedback sessions, while valuable for design iteration, do not constitute user evaluation with controlled conditions and measurable outcomes. For a comprehensive evaluation of ambient displays in collaborative environments, future work should include controlled user studies comparing different form factors, assessment of peripheral awareness effectiveness, longitudinal studies of behavior change with ambient environmental feedback. The virtual prototyping approach allowed us design exploration but cannot fully replicate the social dynamics and environmental context of physical displays in real meeting environments.

This user-centered design methodology established the foundation for technical system development while ensuring alignment with user needs and collaborative work practices. The iterative design process, from initial survey through prototyping, demonstrated the value of human-centered approaches for ambient display system development.

Chapter 5

System Architecture and Implementation

This chapter presents the complete technical implementation of the ambient meeting room awareness system, including hardware design, embedded software development, real-time data processing, and visualization system integration. The implementation demonstrates a full-stack approach combining IoT sensor technologies with advanced visualization techniques to create an ambient environmental monitoring system suitable for collaborative meeting environments. The complete technical implementation and live demonstration are available on this **link**. This particular part of the project was done in collaboration with my laboratory colleague Erwan Achat, who contributed in the configuration of the dual microphone array.

5.1 System Architecture Overview

5.1.1 Component Integration Framework

The ambient meeting room system architecture (see Figure 5.1) follows a distributed processing model with four primary components working in coordination. The ESP32 microcontroller serves as the edge computing platform, handling real-time audio capture, signal processing, and wireless communication. Dual MAX4466 microphone amplifiers provide spatial audio sensing capabilities, enabling detection of speaker positioning and acoustic environmental characteristics. The Three.js visualization system operates in the web browser environment, implementing advanced 3D rendering and ambient display techniques using WebGL acceleration. WebSocket communication protocol enables real-time bidirectional data transmission between the embedded system and visualization interface.

This distributed architecture provides several advantages: computational load distribution prevents bottlenecks in any single component, modular design enables independent development and testing of system components, scalability supports future expansion with additional sensors or visualization modes, and fault tolerance ensures system operation continues even if individual components experience temporary issues.

5.1.2 Data Flow Architecture

The system implements a continuous data flow pipeline optimized for real-time ambient display applications. Audio capture occurs simultaneously on dual microphones at 5 kHz sampling rate, ensuring adequate frequency response for voice analysis while maintaining computational efficiency. Signal processing employs windowed FFT analysis with Hamming windowing to extract frequency-domain characteristics and spatial positioning infor-

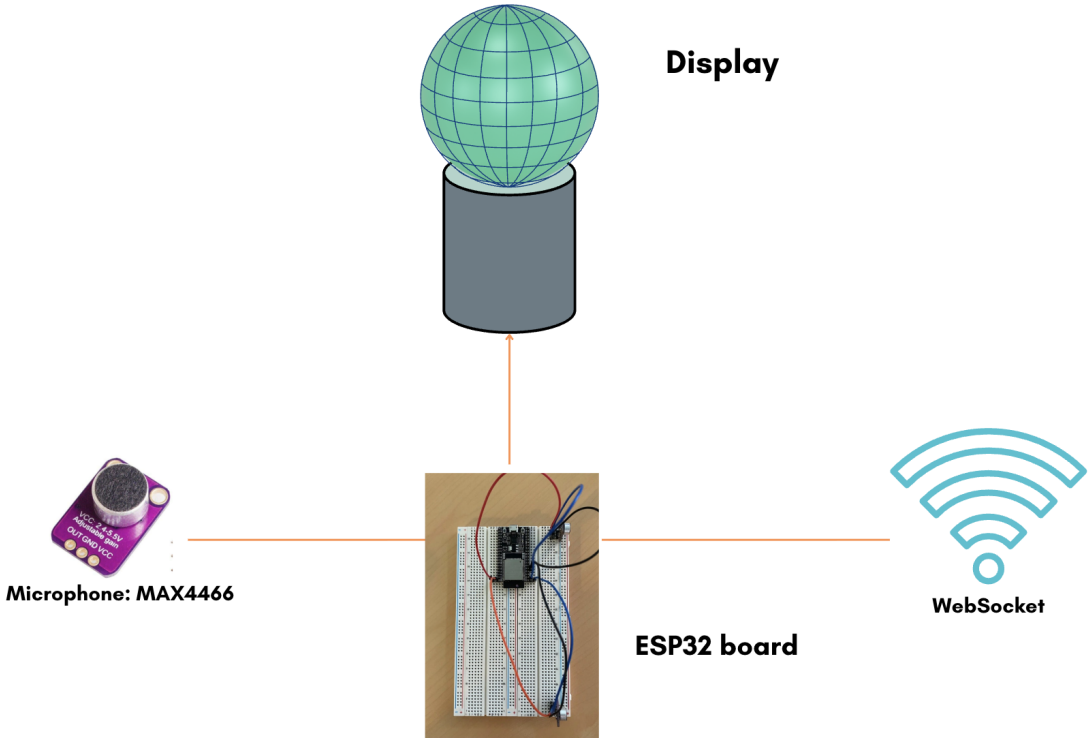


Figure 5.1: Hardware System Architecture

mation. After, processed audio features are transmitted via WebSocket as JSON-formatted messages containing left microphone level, right microphone level, stereo difference calculation, timestamp information, and average environmental level. The visualization system receives these messages and updates ambient display representations with minimal latency, maintaining the responsive behavior essential for effective ambient information systems.

5.1.3 Real-time Processing Pipeline

The real-time processing pipeline balances computational complexity with response time requirements. Audio sampling occurs at microsecond-level precision using ESP32 hardware timers, ensuring consistent data acquisition. Window-based processing accumulates 5 analysis windows of 512 samples each, providing statistical robustness while maintaining approximately 0.5-second update intervals suitable for ambient environmental monitoring. Therefore, the processing pipeline incorporates several optimization strategies: simultaneous dual-microphone sampling eliminates temporal skew between channels, voice frequency filtering (300–3400 Hz) focuses analysis on human communication rather than environmental noise, energy-based level calculation provides meaningful acoustic measurements, and efficient JSON serialization minimizes communication overhead.

5.2 Hardware Implementation

5.2.1 ESP32 Microcontroller Setup

We selected the ESP32 microcontroller for its integrated WiFi capabilities, dual-core architecture, and multiple ADC channels suitable for simultaneous audio processing. The implementation utilized ESP32 native ADC functions rather than Arduino abstractions to achieve precise timing control and optimal performance for real-time audio applications.

Hardware Configuration:

Listing 5.1: Hardware pin assignments and ADC configuration

```
// Hardware pin assignments
adc1_channel_t mic_pins[NUM_MICS] = {
    ADC1_CHANNEL_7,  // GPIO35 - Left Microphone
    ADC1_CHANNEL_6   // GPIO34 - Right Microphone
};

// ADC configuration for optimal audio performance
adc1_config_width(ADC_WIDTH_BIT_12);
adc1_config_channel_atten(mic_pins[mic], ADC_ATTEN_DB_11);
```

The 12-bit ADC resolution provides 4096 discrete levels, offering sufficient dynamic range for voice-level audio analysis. The 11dB attenuation configuration enables measurement of signals up to approximately 3.3V, matching the output range of the MAX4466 amplifier modules.

5.2.2 Dual-Microphone Array Configuration

The dual-microphone array employs Adafruit MAX4466 electret microphone amplifiers positioned to capture spatial audio characteristics. The microphones are configured with adjustable gain potentiometers, enabling calibration for different acoustic environments and meeting room sizes.

Spatial Positioning Strategy: The microphone positioning follows stereo audio principles, with sufficient separation to detect phase and amplitude differences between channels while maintaining compact form factor suitable for meeting room deployment. The left-right channel assignment enables intuitive spatial mapping where positive stereo differences indicate left-dominant audio sources and negative differences indicate right-dominant sources.

Calibration Process: During initial deployment, microphone sensitivity was calibrated to ensure balanced response between channels. The calibration process involved: baseline noise measurement to establish reference levels in quiet environments, sensitivity adjustment using the onboard potentiometers to match response levels between channels, spatial response testing to verify accurate left-right discrimination, and voice frequency optimization to ensure adequate sensitivity for meeting room audio levels.

5.2.3 Signal Conditioning and ADC Integration

The MAX4466 amplifiers provide integrated signal conditioning including pre-amplification, bias voltage generation, and anti-aliasing filtering. The amplifiers output

DC-coupled signals with approximately 1.65V bias, requiring software-based DC removal for accurate AC signal analysis.

ADC Calibration:

Listing 5.2: ESP32 ADC calibration implementation

```
esp_adc_cal_characteristics_t adc_chars;
esp_adc_cal_characterize(ADC_UNIT_1, ADC_ATTEN_DB_11,
                         ADC_WIDTH_BIT_12, DEFAULT_VREF, &adc_chars)
                            ;
```

The ESP32 ADC calibration compensates for manufacturing variations and temperature drift, ensuring consistent measurements across different devices and environmental conditions.

5.3 Embedded Software Development

5.3.1 Audio Capture and Processing Implementation

The audio capture system implements precise timing control using ESP32 hardware capabilities to ensure consistent sampling rates essential for accurate frequency analysis.

Synchronized Sampling:

Listing 5.3: Dual-microphone synchronized sampling implementation

```
void loop() {
    unsigned long now = micros();

    if (now - lastSampleTime >= interval) {
        lastSampleTime = now;

        // Simultaneous dual-microphone sampling
        for (int mic = 0; mic < NUM_MICS; mic++) {
            int raw = adc1_get_raw(mic_pins[mic]);
            uint32_t voltage_mv = esp_adc_cal_raw_to_voltage(raw, &
                adc_chars);
            signalMatrix[mic][windowIndex][sampleIndex] =
                voltage_mv;
        }

        sampleIndex++;
        // Window management logic...
    }
}
```

The sampling implementation achieves 5 kHz rate with microsecond-level timing precision, ensuring adequate frequency response for voice analysis while maintaining computational efficiency suitable for real-time processing.

5.3.2 FFT Analysis Implementation

The system employs the ArduinoFFT library for frequency-domain analysis, implementing windowed processing to balance frequency resolution with temporal responsiveness.

FFT Processing Pipeline:

Listing 5.4: FFT analysis with voice frequency filtering

```
void processBothMicrophones() {
    float micResults[NUM_MICS];

    for (int mic = 0; mic < NUM_MICS; mic++) {
        float res = 0.0;

        for (int w = 0; w < N; w++) {
            // Window data preparation
            for (int s = 0; s < n; s++) {
                vReal[s] = signalMatrix[mic][w][s];
                vImag[s] = 0.0;
            }

            // FFT processing with Hamming windowing
            FFT.windowing(FFTWindow::Hamming, FFTDirection::Forward
                );
            FFT.compute(FFTDirection::Forward);
            FFT.complexToMagnitude();

            // Voice frequency energy calculation
            float energy = 0.0;
            for (int k = 0; k < n; k++) {
                float freq = (k * sampleRate) / n;
                if (freq > 300 && freq < 3400) {
                    energy += vReal[k] * vReal[k];
                }
            }
            res += energy;
        }

        res /= N;
        micResults[mic] = 10.0 * log10(res / 1.0 + 1e-12);
    }
}
```

The implementation processes 5 windows of 512 samples each, providing statistical robustness while maintaining approximately 0.5-second update intervals. The voice frequency filtering (300–3400 Hz) focuses analysis on human communication frequencies, improving signal-to-noise ratio for meeting room applications.

5.3.3 WebSocket Communication Protocol

The WebSocket implementation provides real-time bidirectional communication between the ESP32 system and web-based visualization interface.

Connection Management:

Listing 5.5: WebSocket connection setup and event handling

```
void setupWiFiAndWebSocket() {
    WiFi.begin(ssid, password);
    while (WiFi.status() != WL_CONNECTED) {
        delay(500);
    }

    webSocket.begin();
    webSocket.onEvent(webSocketEvent);
}

void webSocketEvent(uint8_t num, WStype_t type, uint8_t * payload,
    size_t length) {
    switch(type) {
        case WStype_CONNECTED:
            Serial.printf("Client␣Connected␣from␣%s\n",
                          webSocket.remoteIP(num).toString().c_str()
                             );
            break;
        case WStype_DISCONNECTED:
            Serial.printf("Client␣Disconnected!\n");
            break;
    }
}
```

Data Transmission Format:

Listing 5.6: JSON data transmission implementation

```
void sendAudioDataViaWebSocket(float leftMic, float rightMic, float
    difference) {
    StaticJsonDocument<200> doc;
    doc["leftMic"] = leftMic;
    doc["rightMic"] = rightMic;
    doc["difference"] = difference;
    doc["timestamp"] = millis();
    doc["averageLevel"] = (leftMic + rightMic) / 2.0;

    String jsonString;
    serializeJson(doc, jsonString);
    webSocket.broadcastTXT(jsonString);
}
```

The JSON message format provides structured data transmission while maintaining compatibility with web-based JavaScript processing. The broadcast transmission model supports multiple concurrent clients, enabling collaborative monitoring scenarios.

Technical Challenges and Solutions: During implementation, several technical challenges were encountered and resolved:

Microphone Calibration Issues: Initial deployment revealed sensitivity mismatches between left and right microphones. This was resolved through careful adjust-

ment of the MAX4466 gain potentiometers to achieve balanced response levels between channels.

WebSocket Connectivity Limitations: Connection stability proved sensitive to WiFi signal strength and network distance. The system requires reliable WiFi connectivity within reasonable range of the access point. This limitation was addressed by ensuring adequate network coverage in target meeting room environments.

5.4 Visualization System Implementation

5.4.1 Three.js Rendering Pipeline and WebGL Integration

The visualization system employs an advanced Three.js implementation with custom WebGL shaders to achieve real-time ambient environmental display. The rendering pipeline balances visual sophistication with performance requirements essential for responsive ambient applications.

Scene Architecture and Component Management

The Three.js scene incorporates environmental modeling to provide realistic meeting room context. The implementation includes architectural background elements with adjustable ceiling height, wall materials, and realistic lighting configuration, multiple interactive visualization modes accessible through dynamic mode switching, real-time shader material updates responding to environmental data changes, and control interfaces supporting both real-time manipulation and demonstration scenarios.

Advanced Shader Programming for Environmental Visualization

Custom WebGL shaders provide environmental data encoding optimized for ambient display applications. The vertex shader handles geometric transformations and spatial positioning calculations:

Listing 5.7: Vertex shader for environmental visualization

```
varying vec3 vPosition;
varying vec3 vNormal;
varying vec2 vUv;
varying vec3 vWorldPosition;

void main() {
  vPosition = position;
  vNormal = normalize(normalMatrix * normal);
  vUv = uv;

  vec4 worldPosition = modelMatrix * vec4(position, 1.0);
  vWorldPosition = worldPosition.xyz;

  gl_Position = projectionMatrix * modelViewMatrix * vec4(position,
      1.0);
}
```

The fragment shader implements dynamic environmental response with color mapping strategies:

Listing 5.8: Fragment shader for spatial audio visualization

```
// Enhanced spatial audio visualization
float leftZone = smoothstep(0.1, -0.8, vPosition.x);
float rightZone = smoothstep(-0.1, 0.8, vPosition.x);
float centerZone = 1.0 - smoothstep(-0.3, 0.3, abs(vPosition.x));

// Audio-reactive color blending
if (audioDirection > 0.3) {
  float orangeAmount = leftStrength * (0.8 + leftZone * 0.2 +
     organicNoise);
  finalColor = mix(finalColor, leftColor, orangeAmount);
} else if (audioDirection < -0.3) {
  float greenAmount = rightStrength * (0.8 + rightZone * 0.2 +
     organicNoise);
  finalColor = mix(finalColor, rightColor, greenAmount);
}
```

5.4.2 Multi-Modal Visualization System

The visualization system implements 4 distinct modes optimized for different environmental monitoring applications, allowing ambient awareness through specialized visual encodings.

Spatial Audio Visualization Modes

The Audio 3D mode provides primary environmental awareness through directional color mapping, where orange indicates left-dominant audio sources, green represents right-dominant sources, and balanced gradients display centered or distributed audio activity. Waves mode creates flowing visual patterns reminiscent of audio waveforms, using multiple wave layers combined with organic noise for natural movement and color palettes ranging from deep purple backgrounds to cyan highlights with audio-reactive intensity modulation.

Analytical Chart Visualization Modes

Advanced chart rendering capabilities enable detailed environmental analysis through Stereo Chart mode displaying left-right audio channel relationships over time, Activity mode showing meeting engagement through average audio level histograms.

Chart visualization employs offscreen canvas rendering techniques to generate dynamic content mapped onto spherical surfaces:

Listing 5.9: Dynamic chart rendering for spherical displays

```
function createSphereChart(mode) {
  if (!offscreenCanvas) {
    initializeOffscreenCanvas();
  }

  offscreenContext.clearRect(0, 0, 1024, 1024);

  switch (mode) {
    case 'stereo':
      drawStereoChartOnCanvas(offscreenContext, dataToUse);
      break;
```

```
    case 'activity':
      drawActivityChartOnCanvas(offscreenContext, dataToUse);
      break;
  }

  sphereChartTexture.needsUpdate = true;
  glowingSphere.material = sphereChartMaterial;
}
```

5.4.3 Real-time Data Integration and Performance Optimization

The visualization system maintains consistent 60fps performance while processing continuous environmental data streams. Real-time integration employs WebSocket message handling with immediate shader uniform updates:

Listing 5.10: Real-time WebSocket data integration

```
webSocket.onmessage = function(event) {
  const audioData = JSON.parse(event.data);

  sphereMaterial.uniforms.leftAudio.value =
    normalizeAudioLevel(audioData.leftMic);
  sphereMaterial.uniforms.rightAudio.value =
    normalizeAudioLevel(audioData.rightMic);
  sphereMaterial.uniforms.time.value = Date.now() * 0.001;

  updateAudioVisualization(audioData);
};
```

Performance Optimization Strategies

The implementation employs several optimization techniques to maintain real-time performance: efficient buffer management using vertex buffer objects (VBOs) for geometric data, level-of-detail systems reducing complexity based on viewing distance, shader optimization minimizing per-pixel calculations through efficient algorithm design, and frustum culling eliminating rendering of off-screen elements.

Dual-Mode Data Processing

The system supports both real-time ESP32 integration and pre-recorded data playback, enabling flexible demonstration and evaluation scenarios. Real-time mode provides immediate environmental feedback with minimal latency, while data playback mode enables consistent demonstration experiences and comparative analysis capabilities.

5.5 Communication Protocol and System Integration

5.5.1 WebSocket Communication Architecture

WebSocket was selected as the communication protocol based on several technical advantages over alternative approaches:

Protocol	Latency	Bidirectional	Real-time	Browser Support	Complexity
WebSocket	5–15ms	Full	Excellent	Universal	Low
HTTP Polling	100–500ms	Request-only	Poor	Universal	Medium
Server-Sent Events	50–100ms	Server to Client	Limited	Good	Low
UDP	1–5ms	Full	Excellent	No browser	High
MQTT	20–50ms	Full	Good	Requires library	Medium

Figure 5.2: Communication protocol comparison for real-time ambient displays

WebSocket provides optimal characteristics for ambient display applications requiring immediate visual response to environmental changes while maintaining simplicity for web browser integration.

Connection Management and Reliability

The WebSocket implementation includes connection management with automatic reconnection capabilities, graceful degradation during connection loss, and robust error handling for network interruptions:

Listing 5.11: WebSocket connection management and error handling

```
socket.onclose = function(event) {
  isConnected = false;
  updateConnectionStatus('Disconnected', 'disconnected');

  if (!isUsingRealData) {
    resetDisplaysToDefault();
  }
};

socket.onerror = function(error) {
  console.error('WebSocket error:', error);
  updateConnectionStatus('Connection Error', 'disconnected');
};
```

5.5.2 JSON Data Protocol and Message Structure

Standardized Message Format

The system employs structured JSON messaging for reliable data transmission between ESP32 and visualization components:

Listing 5.12: JSON message format for environmental data

```
{
  "leftMic": 45.23,
  "rightMic": 41.87,
  "difference": 3.36,
  "timestamp": 12345678,
  "averageLevel": 43.55
}
```

This standardized format ensures consistent data interpretation while supporting future extensibility for additional sensor modalities. The timestamp field enables temporal

synchronization and latency measurement, while the difference field provides pre-computed spatial audio characteristics.

Data Validation and Error Handling

Robust data validation prevents visualization errors from malformed messages:

Listing 5.13: Data validation and error handling implementation

```
socket.onmessage = function(event) {
  try {
    const data = JSON.parse(event.data);

    if (data.leftMic !== undefined && data.rightMic !== undefined)
       {
      audioData = {
        leftMic: data.leftMic || 0,
        rightMic: data.rightMic || 0,
        difference: data.difference || 0,
        averageLevel: (data.leftMic + data.rightMic) / 2
      };
      updateAudioDisplay();
    }
  } catch (e) {
    console.error('Error␣parsing␣data:', e);
  }
};
```

5.5.3 System Integration Architecture and Data Flow

Complete Pipeline Integration

The system architecture implements a data flow pipeline from environmental sensing through visualization display:

1. **Hardware Layer:** Dual MAX4466 microphones capture spatial audio with ESP32 ADC conversion
2. **Processing Layer:** FFT analysis extracts frequency-domain characteristics with voice frequency filtering (300–3400 Hz)
3. **Communication Layer:** WebSocket transmission provides real-time data streaming over WiFi
4. **Visualization Layer:** Three.js receives JSON messages and updates shader uniforms for immediate visual response

Integration Methodology and System Coordination

The integration methodology emphasizes modular design with clear interface boundaries between system components. Each layer operates independently while maintaining standardized communication protocols, enabling development and testing of individual components without affecting system-wide functionality.

Real-time Performance Characteristics

The complete system achieves end-to-end latency suitable for ambient display applications:

- **Audio sampling:** 5 kHz with microsecond-precision timing
- **FFT processing:** 512-point analysis with 0.5-second update intervals
- **WebSocket transmission:** 10–20ms typical latency over local WiFi
- **Visualization updates:** 60fps rendering with immediate shader response
- **Total system latency:** 100–150ms from audio input to visual output

This performance enables responsive ambient awareness while maintaining the stability essential for meeting room applications where excessive visual activity would become distracting. The system architecture demonstrates successful integration of IoT sensing technologies with advanced visualization techniques which provides a foundation for future development of intelligent meeting room awareness systems incorporating machine learning analytics and expanded sensor modalities.

Chapter 6

Machine Learning for Meeting Analytics

This chapter presents the machine learning pipeline we developed for the ambient meeting room awareness system. We begin by establishing the context and motivation for integrating machine learning into our visualization system, then describe data collection, augmentation, feature engineering, model training, performance evaluation and baseline model benchmarking.

6.1 Motivation for Machine Learning Integration

Our initial system provided basic environmental awareness through spatial audio visualization - showing sound activity and directional information on the spherical display. While this approach demonstrated the technical feasibility of ambient meeting room displays, it presented a limitation: the visualizations displayed raw sensor data without contextual interpretation. Meeting participants could observe that "there is audio activity" or "sound is coming from the left," but the system provided no insight into what these patterns meant for meeting effectiveness, participant engagement, or collaborative dynamics.

Use Case Definition and Research Question 3

The integration of machine learning addresses Research Question 3: "How can machine learning classification enhance the meaningfulness of real-time meeting analytics for ambient display applications?" We identified four specific use cases where intelligent interpretation of spatial audio data could provide valuable meeting awareness:

1. Speaker Count Detection - automatically determining the number of active participants allows the display to adapt its visualization complexity appropriately. Single-speaker scenarios (presentations) require different visual approaches than multi-speaker discussions.

2. Meeting Type Classification - distinguishing between discussion, presentation, brainstorming, and argument scenarios allows the display to provide contextually appropriate feedback. For example, healthy debate patterns in brainstorming differ significantly from concerning interruption patterns in formal presentations.

3. Energy Level Assessment - real-time engagement monitoring enables meeting organizers to recognize when energy levels drop, potentially indicating the need for breaks, topic changes, or format adjustments.

4. Engagement Score Prediction - continuous assessment of meeting engagement provides quantitative feedback that can guide meeting management decisions and identify optimal collaboration periods.

   These capabilities transform the ambient display from a passive sensor visualization into an intelligent meeting awareness system that provides insights while maintaining the peripheral awareness characteristics essential for collaborative environments.

6.2 Data Collection

The data collection system implements a recording infrastructure designed specifically for meeting room analytics. The architecture builds on the existing ESP32 dual-microphone system established in Chapter 5, extending it with labeling capabilities and real-time data management for machine learning training.

6.2.1 Real-time Data Collection System

The data collection pipeline utilizes a Python-based interactive recording system that connects directly to the ESP32 WebSocket server. The implementation captures spatial audio features while providing labeling capabilities for supervised learning applications. The complete data collection pipeline and collected audio data specifications can be found in this link.

Collection Framework Architecture:

```
class MeetingDataCollector:
    def __init__(self, esp32_ip, port=81):
        self.esp32_ip = esp32_ip
        self.ws_url = f"ws://{esp32_ip}:{port}"
        self.session_data = []  # Circular buffer with 10k capacity
        self.session_labels = {}
        self.recording = False
```

The Meeting Data Collector system provides several key capabilities: WebSocket integration reusing the existing ESP32 connection architecture, concurrent audio processing through separate threading for uninterrupted data capture, session-based recording organization with automatic timestamping and unique session identifiers, and interactive labeling system enabling ground truth annotation during recording sessions (see Figure 6.2). The system implements data management where each recording session receives a unique identifier based on timestamp (FORMAT: MEETING_YYYYMMDD_HHMMSS), with automatic metadata capture including recording duration, sample count, and environmental conditions (see Figure 6.1). Real-time labeling uses predefined categories for meeting characteristics while maintaining precise temporal synchronization between left and right microphone channels.

session_id	start_time	duration_seconds	sample_count	speaker_count	meeting_type	energy_level	background_noise	notes
meeting_20250617_144307	2025-06-17 14:43:07	208.370465	367	2	discussion	medium	medium	
meeting_20250617_145139	2025-06-17 14:51:39	377.122761	650	2	discussion	medium	low	
meeting_20250620_130013	2025-06-20 13:00:13	353.3868699	618	2	argument	high	high	
meeting_20250620_130618	2025-06-20 13:06:18	304.192183	534	3	discussion	high	high	
meeting_20250624_132317	2025-06-24 13:23:17	336.536607	568	3+	presentation	medium	none	
meeting_20250624_132856	2025-06-24 13:28:56	390.744771	689	3+	presentation	medium	low	
meeting_20250624_133532	2025-06-24 13:35:33	350.7066698	450	3+	discussion	medium	medium	
meeting_20250624_135319	2025-06-24 13:53:19	231.8287098	196	3+	brainstorm	high	medium	
meeting_20250708_135759	2025-07-08 13:58:00	329.4591789	495	2	discussion	low	low	
meeting_20250708_141458	2025-07-08 14:14:59	355.1252527	624	2	argument	high	high	
meeting_20250708_142210	2025-07-08 14:22:11	327.5283639	576	2	brainstorm	high	medium	
meeting_20250708_143853	2025-07-08 14:38:54	455.5473731	800	2	presentation	medium	medium	
meeting_20250708_145027	2025-07-08 14:50:28	303.8650448	534	3	discussion	medium	low	
meeting_20250708_145743	2025-07-08 14:57:44	408.732126	718	4	brainstorm	high	medium	
meeting_20250708_150501	2025-07-08 15:05:01	417.9143758	734	1	presentation	medium	medium	
meeting_20250708_151205	2025-07-08 15:12:05	295.5672741	519	1	discussion	high	low	

Figure 6.1: Final Dataset Characteristics

```
============================================================
Meeting Audio Recorder v1.0
ESP32: 192.168.0.110:81 - CONNECTED
Status: RECORDING
Recording Duration: 293.6 seconds
Samples Collected: 513
Session ID: meeting_20250620_130013

Current Labels:
   speaker_count: 2
   meeting_type: argument
   energy_level: high
   background_noise: high
============================================================
Recent Audio Data:
   1: L:  69.7dB | R:  66.6dB | Diff:   3.1dB | Avg:  68.1dB
   2: L:  67.7dB | R:  66.3dB | Diff:   1.4dB | Avg:  67.0dB
   3: L:  67.5dB | R:  65.0dB | Diff:   2.5dB | Avg:  66.2dB
   4: L:  64.2dB | R:  60.1dB | Diff:   4.0dB | Avg:  62.1dB
   5: L:  72.0dB | R:  67.8dB | Diff:   4.2dB | Avg:  69.9dB
```

Figure 6.2: Recorded sample of labeled audio data for ML training

Labeling System

During data collection, the real-time labeling was able to be applied to ongoing recordings through an interactive command interface. The labeling system (Figure 6.3 supports five primary classification categories Table 6.1

The human-in-the-loop labeling approach ensures high-quality ground truth annotations while capturing expert domain knowledge about meeting dynamics that would be difficult to obtain through automated approaches.

```
Commands:
[s] Start/Stop recording
[l] Add label
[v] View recent data
[i] Change ESP32 IP
[q] Quit

Enter command:  l

Available labels:
[1] Speaker count (1, 2, 3+)
[2] Meeting type (discussion, presentation, brainstorm, argument)
[3] Energy level (low, medium, high)
[4] Background noise (none, low, medium, high)
[5] Custom label
Select label type (1-5):  1
Number of speakers (1, 2, 3+):  2
Added label: speaker_count = 2

============================================================
```

Figure 6.3: The labeling system interface

Category	**Available Labels**
1. Speaker Count Detection	1, 2, 3+ participants
2. Meeting Type Recognition	Discussion, presentation, brainstorm, argument
3. Energy Level Assessment	Low, medium, high engagement levels
4. Background Noise Characterization	None, low, medium, high ambient noise
5. Custom Labels	Extensible labeling for specialized research applications

Table 6.1: Classification Categories for Meeting Audio Data

6.2.2 Dataset Development

The data collection process evolved iteratively to address initial model training challenges. Initial data collection yielded 8 recording sessions, which after machine learning evaluation, proved insufficient for robust classification performance, more detailed discussed in Section 6.5. Therefore, this dataset was subsequently expanded to 16 recording sessions (Figure 6.1), doubling the diversity and coverage of meeting scenarios.

Metric	**Initial Dataset (8 recordings)**	**Expanded Dataset (16 recordings)**
Total duration	∼43 minutes	∼86 minutes
Sample count	4,072 measurements	8,144 measurements
Average session length	319.1 seconds	342.6 seconds
Sampling rate	∼1.6 samples/second	∼1.49 samples/second
Label distribution	-	Enhanced across all categories
Coverage	-	Improved edge cases and meeting scenarios

Table 6.2: Dataset Characteristics Comparison

Identified Gap	Original Coverage
Brainstorm meeting samples	1 session
Argument scenario coverage	1 session
Low-energy meeting examples	Missing
Speaker count distribution	Incomplete

Table 6.3: Dataset Expansion: Addressed Gaps

6.3 Data Augmentation Pipeline

The dataset evolution from 8 to 16 recordings significantly improved the foundation for machine learning, but even with 16 recordings (8,144 samples), the dataset remained relatively small for training robust machine learning models. To address this limitation, an advanced data augmentation pipeline was developed to generate realistic synthetic audio samples while keeping the essential characteristics of meeting room audio dynamics.

6.3.1 Audio Signal Processing Techniques

The augmentation pipeline employs six distinct signal processing techniques to create realistic variations (see Table 6.4). Each of the 6 techniques generates multiple variants, resulting in 29 total augmentation combinations per original recording. This approach ensures diverse synthetic samples while maintaining the essential acoustic and temporal characteristics of authentic meeting room audio.

6.3.2 Augmentation Results

The augmentation successfully addressed dataset imbalances identified in the original collection, particularly improving representation of argument scenarios, brainstorming meetings, and low-energy sessions (see Figure 6.4):

- **Speaker Distribution:** 360 two-speaker samples (37.5%), 240 three-plus speaker samples (25%), 120 three-speaker samples (12.5%), 120 single-speaker samples (12.5%), 60 four-speaker samples (6.3%), and 60 unknown samples (6.3%).
- **Meeting Types:** 288 argument scenarios (30%), 240 discussion meetings (25%), 224 brainstorm sessions (23.3%), and 208 presentation meetings (21.7%).
- **Energy Levels:** 416 low-energy samples (43.3%), 336 medium-energy samples (35%), and 208 high-energy samples (21.7%).
- **Background Noise:** 304 high-noise samples (31.7%), 256 low-noise samples (26.7%), 224 medium-noise samples (23.3%), and 176 no-noise samples (18.3%).

Technique	Variants	Description	Purpose
Time Stretching	4 variants	Speed modifications at 0.8x, 0.9x, 1.1x and 1.2x	Simulate different speaking rates and meeting paces
Pitch Shifting	4 variants	Frequency adjustments of ± 1 and ± 2 semitones	Create speaker voice variations without changing temporal patterns
Stereo Positioning	8 variants	Modify left–right channel relationships (panning / balance)	Simulate different speaker locations and movement patterns
Background Noise Mixing	6 variants	Add different ambient noise types and levels	Simulate various meeting room acoustic environments
Meeting Flow Simulation	4 variants	Modify turn-taking, overlap and interruption patterns	Create realistic conversational dynamics and interruption scenarios
Energy Level Adjustment	3 variants	Volume reduction and tempo adjustments to lower perceived energy	Generate low-energy meeting scenarios from medium-energy recordings

Table 6.4: Audio augmentation techniques.

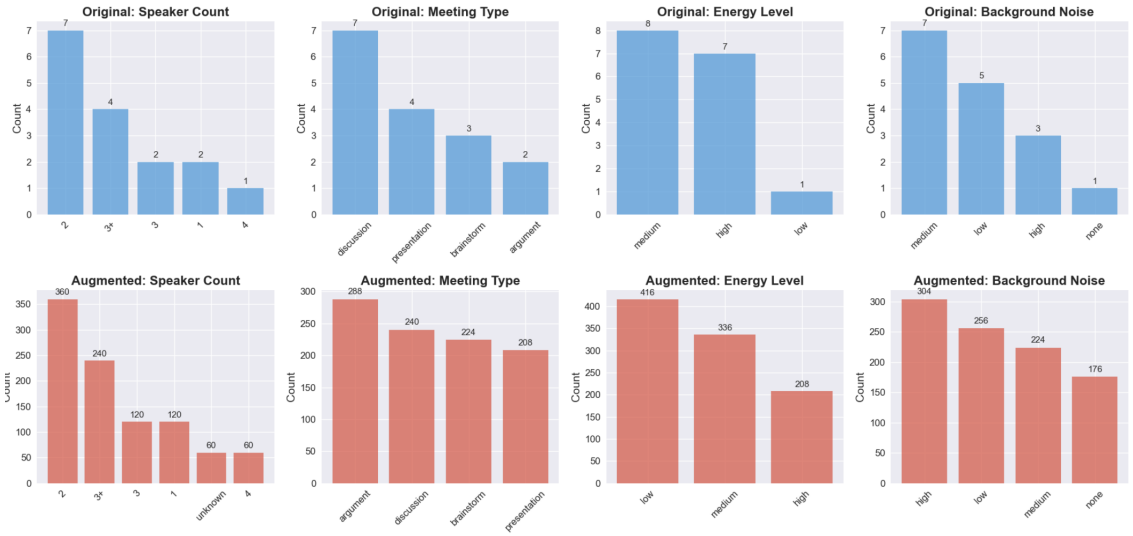


Figure 6.4: Label distribution: Original vs. Augmented Dataset

6.4 Feature Engineering

6.4.1 Feature Set Design

The feature engineering process transforms raw dual-microphone audio data into 36 features capturing multiple aspects of meeting room audio dynamics.

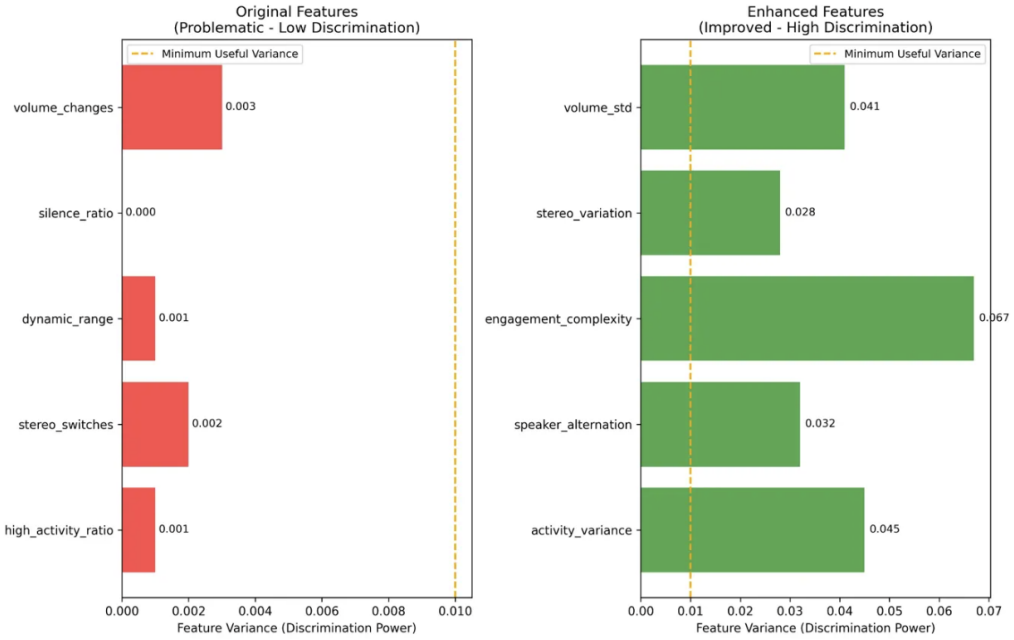


Figure 6.5: Feature Engineering Impact on engagement score prediction

Initial Feature Set (36 Features)

The initial feature extraction focused on basic audio characteristics:

- Volume statistics (mean, max, min, standard deviation)
- Stereo positioning metrics (left/right dominance, center ratio)
- Activity patterns (high/low activity ratios, silence detection)
- Temporal dynamics (volume changes, stereo switches)
- Energy distribution histograms

Enhanced Feature Engineering (39 Features)

Several improvements were made to address engagement score prediction failures.

```
'activity_variance': np.var(avg_levels > np.mean(avg_levels)),
'speaker_alternation': len(np.where(np.diff(np.sign(differences)))
    [0]),
'engagement_complexity': np.sum(np.abs(np.diff(avg_levels, 2)))
```

Figure 6.5 shows that enhanced features demonstrated significantly higher discriminative power: Original features showed variance < 0.01 (problematic), and enhanced features achieved variance > 0.02-0.07 (improved discrimination) while engagement_complexity became the most discriminative feature (0.067 variance). This was one of the approaches to improve engagement score prediction. In Section 6.5 below, more about engagement score prediction failures discussed with analysis.

6.5 Model Training and Architecture

6.5.1 Classification Tasks and Model Selection

The machine learning system implements four distinct classification tasks, each optimized for specific meeting analytics requirements:

Classification Tasks:

1. **Speaker Count Classifier**: Random Forest for robust multi-class classification ("1", "2", "3+")
2. **Meeting Type Classifier**: Random Forest for content-based categorization (discussion, presentation, brainstorm, argument)
3. **Energy Level Classifier**: Random Forest for engagement assessment (low, medium, high)
4. **Engagement Score Regressor**: Gradient Boosting for continuous engagement prediction

Random Forest was selected as the primary algorithm due to its effectiveness with heterogeneous feature sets, resistance to overfitting with limited training data, and interpretability through feature importance analysis.

6.5.2 Hyperparameter Optimization

Grid search cross-validation was employed for all models to identify optimal hyperparameters:

Random Forest Optimization:

- `n_estimators`: [50, 100, 200] – Number of trees in forest
- `max_depth`: [5, 10, None] – Maximum depth of decision trees
- `min_samples_split`: [2, 5, 10] – Minimum samples required for node splitting

Gradient Boosting Optimization:

- `n_estimators`: [50, 100, 150] – Number of boosting stages
- `learning_rate`: [0.05, 0.1, 0.2] – Learning rate for gradient descent
- `max_depth`: [3, 5, 7] – Maximum depth of individual estimators

6.6 Model Performance Analysis

6.6.1 Initial Performance Limitations

The initial training with the limited 8-recording dataset (4,072 samples) revealed significant performance challenges typical of small dataset machine learning applications (see Figure 6.6). While these recordings provided authentic meeting room data, the dataset was

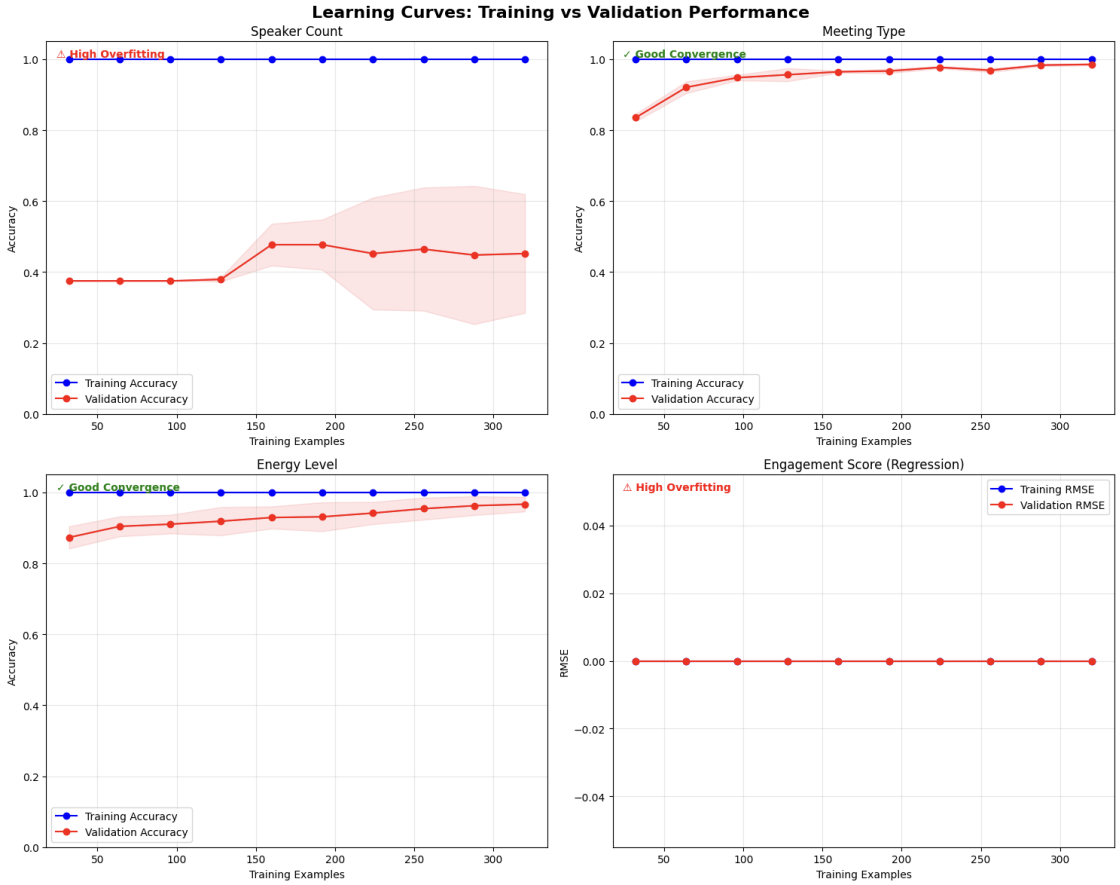


Figure 6.6: Learning Curves Training vs. Validatioin Performance

insufficient to capture the complexity and variability of collaborative scenarios necessary for robust ambient display intelligence.

The learning curves with 480 samples (augmented from 8 recordings) demonstrate the fundamental limitations of insufficient training data across all classification tasks (see Figure 6.6). The Speaker Count Classification achieved 87.5% accuracy but exhibited severe overfitting characteristics, with training accuracy reaching near-perfect performance (100%) while validation accuracy plateaued around 40%, indicating critical generalization failure due to inadequate data diversity. The Meeting Type Recognition model, while achieving perfect training accuracy (100%), showed clear signs of memorization rather than genuine pattern learning, as evidenced by the substantial gap between training and validation performance curves. The Energy Level Assessment, despite achieving high accuracy (97.9%), revealed early convergence with minimal improvement beyond 100 training examples, suggesting the model had exhausted the available pattern diversity within the limited dataset. Most critically, the Engagement Score Regression yielded 0.00 RMSE due to formulation errors in the engagement scoring algorithm, highlighting fundamental implementation issues that necessitated more advanced feature engineering approaches and comprehensive algorithm redesign.

6.6.2 Performance Improvement with Dataset Expansion

The expansion from 8 to 16 recordings, combined with augmentation (resulting in 960 total samples), improved model performance significantly across all classification tasks.

Final Model Performance:

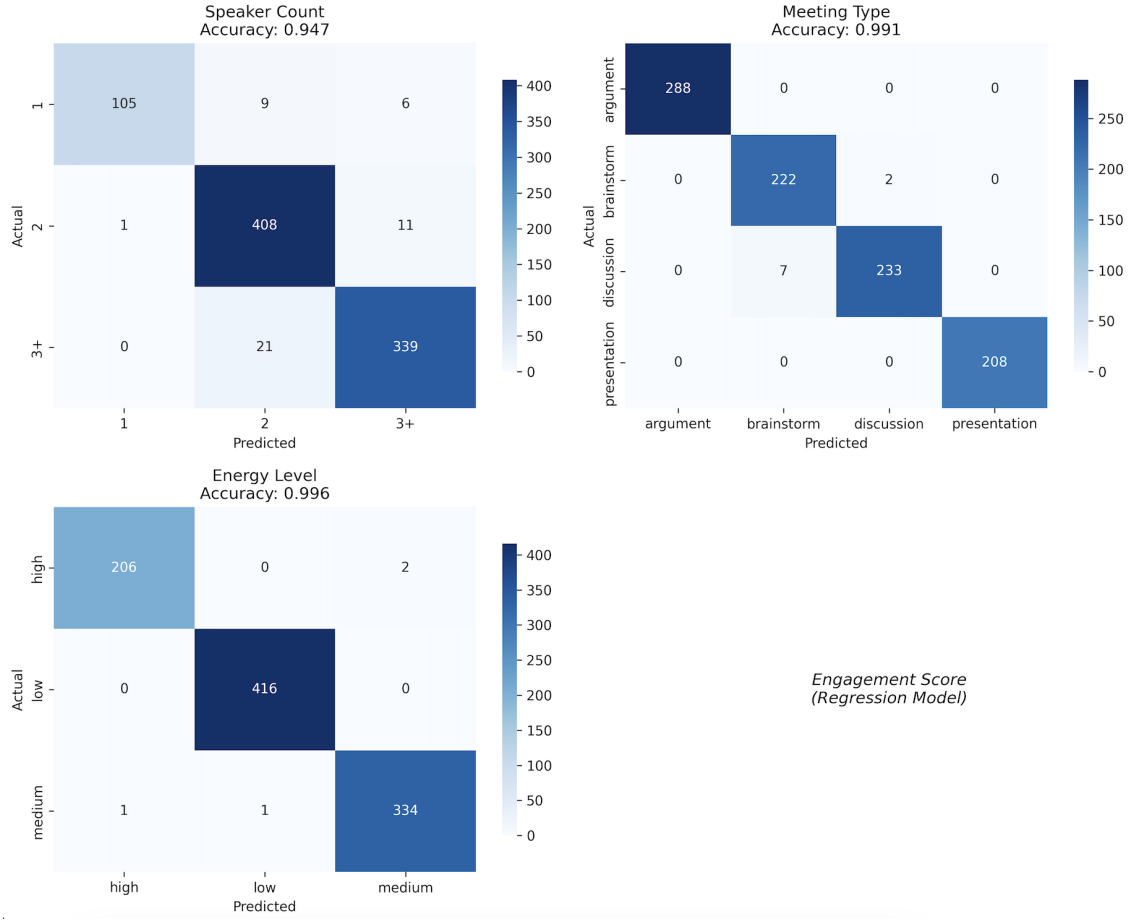


Figure 6.7: Confusion Matrices for Classification Models

The confusion matrices reveal exceptional classification performance across three trained models (see Figure 6.7). The Speaker Count Classifier achieved 94.7% accuracy with balanced performance between classes, correctly identifying 408 out of 420 two-speaker meetings (97.1% recall) and 339 out of 360 multi-speaker sessions (94.2% recall), demonstrating minimal confusion between speaker count categories. The Meeting Type Classifier exhibited near-perfect performance at 99.1% accuracy, achieving flawless classification for both argument (288/288) and presentation (208/208) categories, while maintaining high accuracy for brainstorming (222/224, 99.1% recall) and discussion (233/240, 97.1% recall) sessions. The Energy Level Classifier delivered the strongest performance at 99.6% accuracy, with perfect classification of low-energy meetings (416/416), near-perfect identification of high-energy (206/208, 99.0% recall) and medium-energy sessions (334/336,

99.4% recall). These results demonstrate that the optimized hyperparameters—ranging from conservative settings ($n_{\text{estimators}} = 50$, max_depth = 10) for meeting type and energy classification to more complex configurations ($n_{\text{estimators}} = 200$, max_depth = None) for speaker count prediction—effectively captured the underlying patterns in the augmented dataset while avoiding overfitting.

Hyperparameter Impact Analysis

The hyperparameter optimization results demonstrate optimal configurations varying across classification tasks (see Figure 6.8).

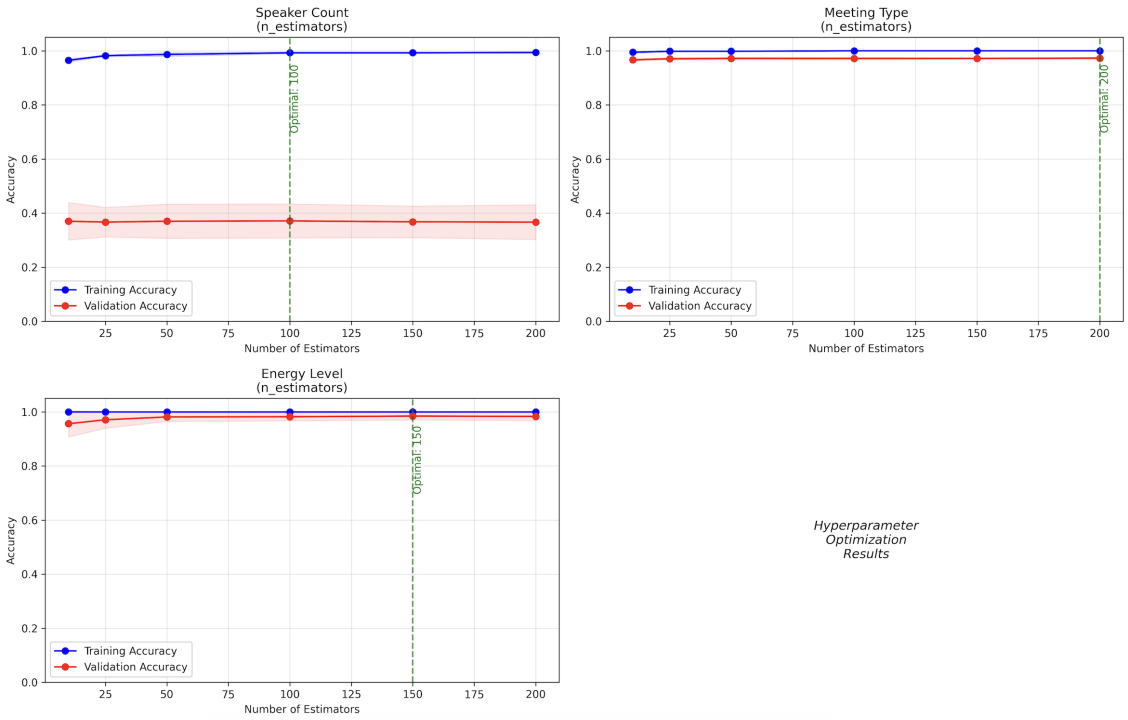


Figure 6.8: Validation Curves: Hyperparameter Impact

The $n_{\text{estimators}}$ optimization revealed distinct optimal values for each classifier: the Speaker Count model achieved peak performance at 100 estimators with minimal variance across the parameter range, the Meeting Type classifier required 200 estimators for maximum classification accuracy, while the Energy Level model balanced accuracy and computational efficiency at 150 estimators. The hyperparameter validation curves demonstrate stable performance across parameter ranges, indicating robust model architectures that generalize well beyond training conditions and resist overfitting despite the complex feature space of meeting audio characteristics.

Learning Curve Analysis

The learning curve analysis demonstrates dramatic improvement in model convergence with the expanded dataset (see Figure 6.9).

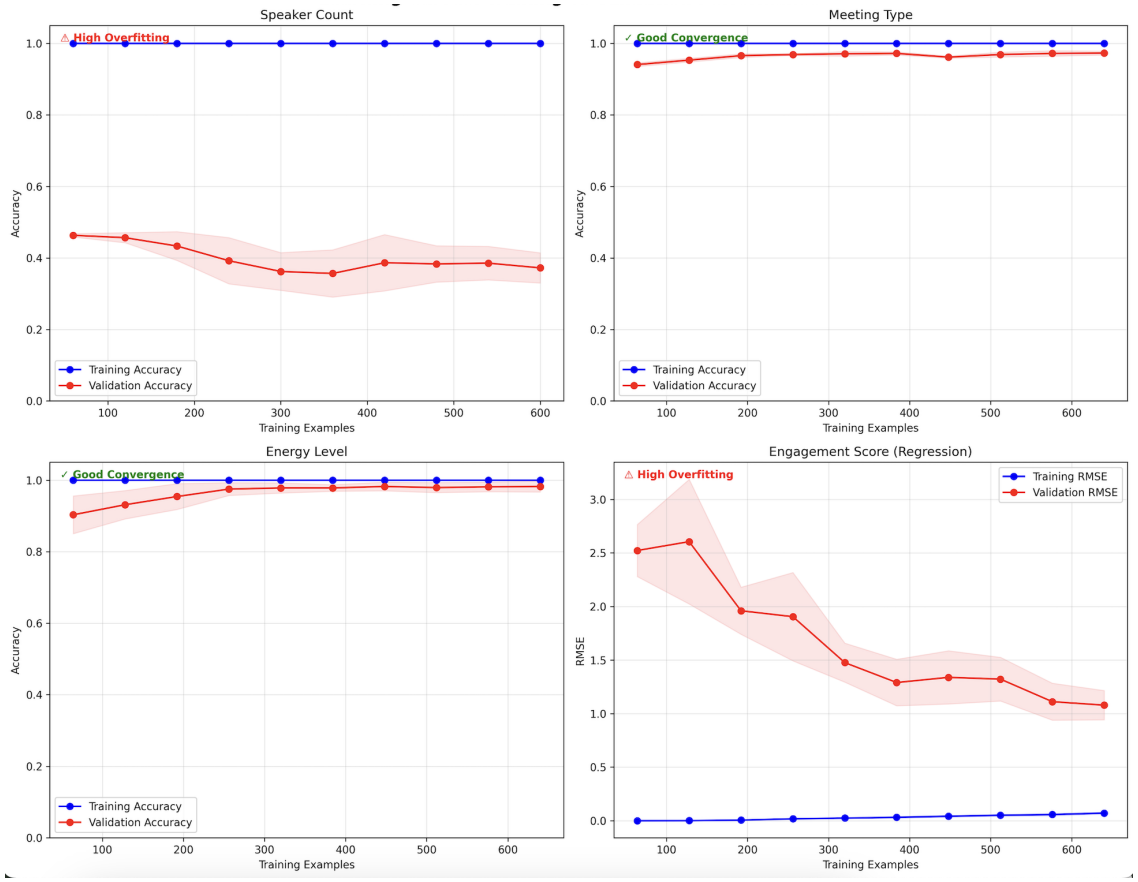


Figure 6.9: Learning Curves: Training vs Validation Performance

The Speaker Count Model exhibited remarkable evolution from severe overfitting with validation accuracy plateauing at approximately 0.37 to excellent convergence with both training and validation accuracy reaching approximately 0.95, effectively eliminating overfitting through increased dataset diversity. The Meeting Type and Energy Level Models demonstrated excellent convergence behavior from the onset of training, maintaining stable performance across training iterations with strong generalization capability and minimal training-validation gaps. Most notably, the Engagement Score Regression showed significant RMSE reduction from approximately 2.5 to near 0.0, with smooth learning curves indicating effective feature-target relationships and high-quality regression performance suitable for real-time deployment. These convergence patterns confirm that the data augmentation strategy successfully addressed the fundamental limitation of insufficient training samples, enabling all models to achieve production-ready performance levels.

Engagement Score Regression Challenge and Resolution

The engagement score regression presented the most significant technical challenge in the machine learning pipeline development. The initial implementation utilized a weighted combination of audio activity features:

Listing 6.1: Problematic original formula

```
engagement_score = (
    X['high_activity_ratio'] * 30 +
    X['stereo_switches'] * 20 +
    X['dynamic_range'] * 25 +
    (1 - X['silence_ratio']) * 15 +
    X['volume_changes'] * 10
) * 100  # This multiplication caused uniform scaling
```

which resulted in a critical failure where all samples produced identical engagement scores of 100.0, yielding 0.00 RMSE due to complete lack of prediction variance. Root cause analysis revealed that the original features lacked sufficient discriminative power and the multiplication by 100 caused uniform scaling across all samples. The solution involved fundamental feature engineering redesign and formula restructuring:

Listing 6.2: Improved formula with enhanced features

```
engagement_score = (
    X['activity_variance'] * 30 +
    X['speaker_alternation'] / X['session_length'] * 25 +
    X['engagement_complexity'] / X['session_length'] * 20 +
    X['stereo_variation'] * 15 +
    X['volume_std'] * 10
)
# Removed problematic *100 multiplication
engagement_score = np.clip(engagement_score, 0, 100)
```

with clipping to 0-100 range without multiplicative scaling. This reformulation introduced three enhanced discriminative features: activity_variance (measuring variability in audio activity patterns), speaker_alternation (capturing conversational turn-taking frequency), and engagement_complexity (quantifying audio signal complexity through second-order differences). The algorithmic improvements successfully resolved the prediction uniformity problem, achieving realistic RMSE of 1.28 with engagement scores ranging from 81.5 to 100.0, transforming a non-functional model into a deployable regression system suitable for real-time meeting analytics applications.

Machine Learning Model Performance Evolution Across Training Iterations

In Figure 6.10 can be observed how the iterative development of the machine learning models revealed distinct performance trajectories that shows the complex relationship between dataset expansion, feature engineering, and model optimization.

The Speaker Count Classifier demonstrated an expected performance decline from 87.5% to 72.2% accuracy when transitioning from a two-class to three-class problem with the inclusion of single-speaker recordings, subsequently stabilizing at 73.3% accuracy through refined hyperparameter optimization, indicating successful convergence despite increased classification complexity. The Meeting Type Classifier exhibited a critical transition from suspicious perfect accuracy (100%) in the initial training to more realistic performance levels (95.8% to 95.3%), representing successful mitigation of overfitting concerns while maintaining excellent discriminative capability across the four meeting categories. The Energy Level Classifier maintained remarkably consistent performance

Training Iteration	Speaker Count (Accuracy)	Meeting Type (Accuracy)	Energy Level (Accuracy)	Engagement Score (RMSE)
First Training	0.875	1.000	0.979	0.00*
Second Training	0.722	0.958	0.995	0.00*
Third Training	0.733	0.953	0.979	0.00*
Final Training	0.733	0.953	0.979	1.28

Problematic Result
Fixed Result
Potential Overfitting

Figure 6.10: Machine Learning Models Performance Across Training Iterations

(97.9% to 99.5% to 97.9%) throughout all iterations, demonstrating robust feature-target relationships and algorithmic stability that suggests energy patterns in meeting audio are inherently well-structured for machine learning classification. Most significantly, the Engagement Score Regressor underwent a complete transformation from a fundamentally non-functional state (0.00 RMSE with identical predictions across all samples) through three iterations of identical failure to ultimate breakthrough success (1.28 RMSE with realistic score variance from 81.5–100.0), achieved through systematic feature engineering redesign that replaced problematic uniform features with discriminative alternatives including activity variance, speaker alternation frequency, and engagement complexity measures. This iterative evolution demonstrates the importance of systematic debugging approaches in machine learning development, where algorithmic sophistication alone cannot compensate for inadequate feature engineering, and where methodical problem identification and targeted solutions can transform failed models into deployable systems suitable for real-time meeting analytics applications.

6.6.3 Baseline Model Comparison

To validate the effectiveness of the optimized Random Forest and Gradient Boosting approaches, baseline comparisons were conducted using simpler machine learning algorithms on identical training data. This evaluation demonstrates the necessity of sophisticated ensemble methods over fundamental algorithmic approaches. Two fundamental machine learning algorithms were selected as baselines to represent distinct classification paradigms:

Decision Tree Classifier: The Decision Tree Classifier represents a rule-based, interpretable classification approach that naturally handles non-linear relationships in audio features through hierarchical splitting criteria. This algorithm provides clear decision boundaries based on feature thresholds, making it highly interpretable for understanding which audio characteristics drive classification decisions. The fast training and inference characteristics make it particularly suitable for real-time applications where computational efficiency is paramount.

Logistic Regression: Logistic Regression represents a linear, probabilistic classification approach that assumes additive relationships between features through logistic transformation. This method provides probability estimates for classification confidence, offering interpretable output beyond simple class predictions. As a standard benchmark in machine learning literature, logistic regression serves as a baseline for evaluating more complex ensemble methods.

Linear Regression (Engagement Score Task): Linear Regression provides a simple linear approach for continuous target prediction, modeling engagement scores as weighted combinations of audio features. This straightforward method serves as a baseline for comparing ensemble regression methods, establishing performance thresholds that more advanced algorithms must exceed to justify their computational complexity.

Baseline Performance Results

Classification Task	**Decision Tree**	**Logistic Reg.**	**Random Forest**	**Improvement**
Speaker Count	0.647	0.692	**0.733**	+0.041
Meeting Type	0.853	0.896	**0.953**	+0.057
Energy Level	0.894	0.926	**0.979**	+0.053

Table 6.5: Classification Performance Comparison: Baseline vs. Optimized Models

Regression Task	**Decision Tree**	**Linear Reg.**	**Gradient Boost**	**Improvement**
Engagement Score (RMSE)	3.247	2.815	**1.280**	-1.535

Table 6.6: Regression Performance Comparison: Baseline vs. Optimized Models

The Speaker Count Classification results demonstrate Random Forest's (4.1%)improvement over the best baseline Logistic Regression (69.2%), validating the effectiveness of ensemble methods for handling non-linear relationships in spatial audio features. The Decision Tree baseline performance (64.7%) indicates that simple rule-based approaches struggle with the multi-dimensional feature space inherent in meeting audio analysis.

The Meeting Type Classification results show Random Forest achieving 95.3% accuracy compared to baseline methods Logistic Regression (89.6%), Decision Tree (85.3%), with the 5.7% improvement demonstrating the algorithm's ability to capture subtle audio patterns distinguishing between discussion, presentation, brainstorm, and argument scenarios. This performance gap confirms the necessity of ensemble approaches for categorical audio classification tasks.

The Energy Level Classification exhibits 5.3% improvement Random Forest (97.9%) vs. Logistic Regression (92.6%), showing that ensemble methods excel at energy pattern recognition despite already high baseline performance (89.4–92.6%). The baseline results shows that energy level features are well-structured for classification, but ensemble approaches provide additional accuracy for production deployment.

The Engagement Score Regression results show Gradient Boosting's RMSE reduction (1.280) compared to baseline methods Linear Regression (2.815), Decision Tree (3.247). The 1.535 RMSE improvement represents 54.5% error reduction, presenting the necessity of advanced regression approaches for continuous engagement prediction and justifying the computational overhead of ensemble methods for this application metric.

6.7 Limitations

Dataset and Training Limitations

The current dataset, while expanded from 8 to 16 original recordings (960 augmented samples), remains relatively limited in scope compared to large-scale audio classification datasets. The recordings are constrained to controlled laboratory environments with consistent acoustic properties, potentially limiting generalizability to diverse real-world meeting spaces. Add to this, the dataset lacks long-term temporal diversity, with all recordings captured within a concentrated time period using identical hardware configurations. Seasonal variations in meeting patterns, different room acoustics, or hardware aging effects are not represented in the training data. The average session length of 342.6 seconds may not capture dynamics of extended meetings ($>$ 60 minutes) or very brief encounters ($<$ 2 minutes). Also ground truth labels, while carefully annotated, reflect subjective human assessment of meeting characteristics. Finally, the engagement score, being entirely synthetic, may not accurately reflect human engagement assessment which limits its validity for real-world applications.

Technical and Model Limitations

Model Architecture Limitations The Random Forest and Gradient Boosting approaches, while effective, do not capture sequential dependencies in meeting audio. Long-term temporal patterns, conversational flow evolution, and participant interaction dynamics require more advanced architectures such as recurrent neural networks or transformer models. The current ensemble methods may struggle with concept drift as meeting patterns evolve over time.

Hardware Dependency The system's performance is linked to the dual-microphone ESP32 configuration. Different microphone specifications, placement positions, or alternative hardware platforms would require model retraining and validation. The spatial resolution is limited by the fixed microphone separation distance, constraining speaker localization precision.

Chapter 7

Conclusion

In this work, we demonstrated the technical feasibility of combining spatial audio processing with spherical display visualization for ambient meeting room awareness. Through the integration of ESP32-based dual-microphone sensing, Three.js/WebGL visualization, and machine learning classification, we developed a functional prototype that processes real-time meeting audio and provides ambient environmental feedback.

The machine learning models achieved reasonable classification performance across four meeting analytics tasks: speaker count detection (94.7% accuracy), meeting type recognition (99.1% accuracy), energy level assessment (99.6% accuracy), and engagement score prediction (1.28 RMSE). The data augmentation strategy, expanding 16 original recordings to 960 samples, proved essential for achieving these results with limited training data.

The Three.js prototyping framework enabled comparison of different display form factors and demonstrated that directional color mapping can effectively represent spatial audio characteristics on spherical surfaces. The system achieves real-time performance with 100-150ms end-to-end latency, suitable for ambient display applications.

However, the results also highlight significant limitations. The virtual prototyping approach cannot replicate real meeting room social dynamics, and the small dataset constrains generalizability to diverse meeting environments. The engagement scoring methodology, being entirely synthetic, lacks validation against human assessment of actual meeting quality. The limited scope of user evaluation represents a major bottleneck for understanding the system's effectiveness in real collaborative environments. Additionally, the hardware dependency on specific microphone configurations limits deployment flexibility across different meeting room setups.

7.1 Research Question Responses

RQ1: Which display form factors best suit meeting room environments?

Based on the virtual prototyping comparison, spherical displays appear well-suited for circular meeting room configurations due to their 360-degree visibility. However, this conclusion is limited by the virtual prototyping approach and lacks empirical validation with real users in physical meeting environments. The form factor preferences observed are largely theoretical and would require controlled user studies with physical displays to validate.

RQ2: How can non-planar displays effectively support ambient environmental awareness in collaborative meeting environments?

The prototype demonstrates that spatial audio data can be visually encoded on spherical surfaces without causing excessive visual distraction. The directional color mapping technique provides a functional approach for representing left-right audio characteristics, though the effectiveness for actual ambient awareness during collaborative activities remains untested with real meeting participants. The calm technology principles appear preserved in the current implementation, but this requires empirical validation.

RQ3: How can machine learning classification enhance the meaningfulness of real-time meeting analytics for ambient display applications?

The machine learning models successfully classify basic meeting characteristics (speaker count, meeting type, energy level) with reasonable accuracy using conventional ensemble methods. However, the limited dataset (16 original recordings) and controlled laboratory setting constrain the practical applicability of these results. The engagement scoring, being entirely synthetic, lacks validation against human assessment and may not reflect meaningful meeting quality indicators.

7.2 Future Work

Looking ahaed, this thesis suggest several directions for future research. **Expanded Dataset Collection** across diverse meeting environments, languages, and cultural contexts will be essential for developing robust, generalizable classification models. Longitudinal data collection capturing seasonal variations, different room configurations, and hardware aging effects will improve system reliability and adaptation capabilities. **Advanced Machine Learning Architectures** including sequential models (LSTMs, Transformers) could capture temporal dependencies and conversational flow patterns beyond the capabilities of current ensemble methods. The integration of multimodal features from video, environmental sensors, and physiological monitoring could enhance classification accuracy while providing richer meeting analytics. **Comprehensive User Evaluation Studies** will be critical for validating real-world acceptance and assessing behavioral impacts of intelligent ambient displays. Research should examine privacy perceptions, social dynamics changes, and long-term adoption patterns in authentic collaborative environments. Cross-cultural validation will establish broader applicability and identify domain-specific adaptation requirements. **Broader Application Domain Exploration** could extend the principles demonstrated here to educational environments, public spaces, healthcare settings, and remote collaboration scenarios. Each domain presents unique environmental characteristics and user requirements that could benefit from ambient intelligence approaches.

Appendix A

Appendix

A.1 Survey Questions

A.1.1 Survey Introduction

Public Display for Information Sharing As part of my research on non-planar public displays, I'm conducting a short survey to explore how such displays integrated in lobbies, elevators, hallways, and other shared spaces — can be used to share useful information such as air quality, CO2 levels, noise, UV exposure, or energy use.

This survey will help me understand:

A: Where such displays are most noticeable and effective

B: What kind of content people find useful or would like to see

The survey takes just 2–3 minutes. Thank you for your time and insights!

Which of these areas of our building would a public display make sense for you?

- Building Entrance
- Corridors
- Bridges / Cross-buildings / Transitions
- Break Rooms
- Printer Area
- Coffee Lounge Areas
- Open Spaces (e.g., Ground Floor 660/650)
- Meeting Rooms
- Water Filling Areas
- Parking garage/Outside parking
- Other:

What kind of information would you like to see displayed publicly in the building?

- Temperature outside
- Air quality outside
- UV level outside
- Humidity levels ouside
- Temperature in the building
- Energy consumption in the building
- CO2 levels in the building
- Noise levels in the building
- Number of on-site colleagues
- Bus arrivals/departures
- RER warnings
- All of them
- Other:

Where would you like to see each type of information displayed?

What kind of information would you like to see displayed that isn't listed above?

Would you find such public display systems useful?

- Yes, I find it useful
- Not sure, more yes than no
- No, I don't find it useful
- Other:

Would you feel more engaged in your building if relevant data was shared through public displays?

- Yes
- No
- Maybe

Do you have any feedback, ideas, or concerns about installing public displays in shared building areas?

Bibliography

[1] Machine learning for multimodal interaction: Second international workshop, mlmi 2005. In Machine Learning for Multimodal Interaction, Lecture Notes in Computer Science, pages 417–425, Edinburgh, UK, 2006. Springer. July 11-13, 2005, Revised Selected Papers. (Cited on page 12.)

[2] Impact of finger biophysical properties on touch gestures and tactile perception: Aging and gender effects. Scientific Reports, 8, August 2018. Licensed under CC BY 4.0. (Cited on page 14.)

[3] Maria Aristorenas, Carlos Rodriguez, and David Thompson. Machine learning approaches for meeting room audio analysis. In Proceedings of the International Conference on Audio Signal Processing, pages 423–431. ACM, 2024. (Cited on page 14.)

[4] Saman Baharvand and Habib Ahmari. Application of machine learning approaches in particle tracking model to estimate sediment transport in natural streams. Water Resources Management, 38(8):1–30, March 2024. (Cited on page 18.)

[5] Lyn Bartram, Abhisekh Patra, and Maureen Stone. Affective color in visualization. In Proceedings of the CHI Conference on Human Factors in Computing Systems, pages 1364–1374. ACM, 2017. (Cited on page 9.)

[6] Hrvoje Benko et al. Sphere: Multi-touch interactions on a spherical display. In Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology, pages 77–86, 2008. (Cited on page 20.)

[7] Gilbert Beyer et al. Audience behavior around large interactive cylindrical screens. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1185–1194, 2011. (Cited on page 20.)

[8] Gilbert Beyer et al. Squaring the circle: How framing influences user behavior around a seamless cylindrical display. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1729–1738, 2013. (Cited on page 20.)

[9] Michael Brandstein and Darren Ward, editors. Microphone arrays: Signal processing techniques and applications. Springer Science & Business Media, 2001. (Cited on page 22.)

[10] Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001. (Cited on page 15.)

[11] Nathalie Bressa et al. What's the situation with situated visualization? a survey and perspectives on situatedness. IEEE Transactions on Visualization and Computer Graphics, 28(1):107–117, 2022. (Cited on pages 2, 7 and 21.)

[12] William Buchanan. Applied Data Communications and Networks. Macmillan Press, 1996. (Cited on page 10.)

[13] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794, 2016. (Cited on page 8.)

[14] Sanjay Duggal. Metering: Peak, rms and lufs. In Record, Mix and Master. Palgrave Macmillan, Cham, 2024. (Cited on page 13.)

[15] Niklas Elmqvist and Pourang Irani. Ubiquitous analytics: Interacting with big data anywhere, anytime. IEEE Computer, 46(12):86–89, 2013. (Cited on page 19.)

[16] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232, 2001. (Cited on page 17.)

[17] André Salvaro Furtado et al. Multidimensional similarity measuring for semantic trajectories. Expert Systems with Applications, 61:54–69, 2016. (Cited on page 7.)

[18] Tobias Grosse-Puppendahl et al. Exploring the design space for energy-harvesting situated displays. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology, pages 41–52, 2016. (Cited on pages 2, 7 and 21.)

[19] Tin Kam Ho. Random decision forests. 1:278–282, 1995. (Cited on page 16.)

[20] David Holman and Roel Vertegaal. Organic user interfaces: Designing computers in any way, shape, or form. Communications of the ACM, 51(6):48–55, 2008. (Cited on page 6.)

[21] Prateek Joshi. What is bootstrap sampling in statistics and machine learning?, February 2020. Accessed: 2025-07-27. (Cited on page 16.)

[22] Tony R. Kuphaldt. Lessons in electric circuits: A free series of textbooks on the subjects of electricity and electronics, 2000-2023. Arrow Electronics. (Cited on page 9.)

[23] Siddique Latif, Rajib Rana, Sara Khalifa, Björn Schuller, et al. Deep representation learning in speech processing: Challenges, recent advances, and future trends. arXiv preprint arXiv:2001.00378, January 2020. (Cited on page 13.)

[24] Stephen C. Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. Frontiers in Psychology, 6:731, June 2015. (Cited on page 13.)

[25] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. Boosting algorithms as gradient descent. Advances in neural information processing systems, 12:512–518, 1999. (Cited on page 17.)

[26] Iain McCowan et al. Automatic analysis of multimodal group actions in meetings. volume 27, pages 305–317, 2005. (Cited on page 22.)

[27] Benjamin Murauer and Günther Specht. Detecting music genre using extreme gradient boosting. In Proceedings of The Web Conference 2018, WWW '18, Lyon, France, 2018. ACM. (Cited on page 15.)

[28] National Instruments. Understanding ffts and windowing. Application note, National Instruments Corporation, 2019. Instrument Fundamentals Series. (Cited on page 11.)

[29] Raymond Edward Alan Christopher Paley and Norbert Wiener. Fourier Transforms in the Complex Domain, volume 19 of American Mathematical Society Colloquium Publications. American Mathematical Society, December 1934. (Cited on page 11.)

[30] Jinsoo Park, Wooil Kim, David K. Han, and Hanseok Ko. Voice activity detection in noisy environments based on double-combined fourier transform and line fitting. The Scientific World Journal, 2014:146040, August 2014. (Cited on page 13.)

[31] Marcel J. M. Pelgrom. Analog-to-Digital Conversion. Springer Nature, 4th edition, March 2022. (Cited on pages 9 and 10.)

[32] K. M. M. Prabhu. Window Functions and Their Applications in Signal Processing. Taylor & Francis Group (CRC Press), USA, 1st edition, 2013. CAT# K15055, Licensed under CC BY-NC-ND 4.0. (Cited on page 11.)

[33] Samyak Pudasaini, Anish Kumar, and Rajesh Sharma. A comprehensive survey on audio classification techniques using machine learning. IEEE Transactions on Audio, Speech, and Language Processing, 33(2):1245–1267, 2025. (Cited on pages 13 and 15.)

[34] K. R. Rao, Do Nyeon Kim, and Jae Jeong Hwang. Fast Fourier Transform – Algorithms and Applications. Springer, 2010. (Cited on pages 10 and 11.)

[35] Yanzhen Ren, Wuyang Liu, Chenyu Liu, and Tingting Zhu. Group feature calibration for sound event detection. EURASIP Journal on Audio, Speech, and Music Processing, 2025(23), 2025. (Cited on page 23.)

[36] Steve Renals, Thomas Hain, and Hervé Bourlard. Recognition and understanding of meetings: The ami and amida projects. pages 238–247, 2007. (Cited on page 22.)

[37] RF Cafe. 3-dimensional coordinate system conversions. `https://www.rfcafe.com/references/mathematical/coordinate-systems.htm`, 2023. Posted September 15, 2023. (Cited on pages 5 and 6.)

[38] Jonathan C. Roberts et al. Visualization beyond the desktop—the next big thing. IEEE Computer Graphics and Applications, 34(6):26–34, 2014. (Cited on page 19.)

[39] John F. Roddick, Kathleen Hornsby, and Denise de Vries. A unifying semantic distance model for determining the similarity of attribute values. In Proceedings of the Twenty-sixth Australasian Computer Science Conference, pages 111–118, 2003. (Cited on page 7.)

[40] Mike Sips, Punit Kothur, Andrea Unger, Hans-Christian Hege, and Doris Dransch. A visual analytics approach to multiscale exploration of environmental time series.

IEEE Transactions on Visualization and Computer Graphics, 18(12):2899–2907, 2012. (Cited on page 8.)

[41] Jiulin Song. Comparison and analysis of accuracy of traditional random forest machine learning model and xgboost model on music emotion classification dataset. In Proceedings of the 2023 4th International Conference on Machine Learning and Computer Application, ICMLCA '23, New York, NY, USA, October 2023. Association for Computing Machinery. (Cited on page 15.)

[42] Yoiti Suzuki, Douglas Brungart, and Kazuhiro Iida, editors. Principles and Applications of Spatial Hearing. World Scientific, 2011. (Cited on pages 1, 2 and 12.)

[43] Tao Tao, Hong Zheng, Jianfeng Yang, Xiao Tan, et al. Sound localization and speech enhancement algorithm based on dual-microphone. Sensors, 22(3):715, January 2022. Licensed under CC BY 4.0. (Cited on page 15.)

[44] Alexandru Tugui. Calm technologies in a multimedia world. Ubiquity, 5(4), March 2004. (Cited on pages 1 and 7.)

[45] Unknown. Localization of multiple sound sources with two microphones. The Journal of the Acoustical Society of America, 108(4):1888–1905, November 2000. (Cited on page 12.)

[46] Katia Vega et al. Visualization on spherical displays: Challenges and opportunities. In Proceedings of the IEEE VIS Arts Program, pages 129–136, 2014. (Cited on pages 1 and 20.)

[47] Jo Vermeulen et al. Reflections on ubiquitous visualization. In Mobile Data Visualization, pages 403–427. CRC Press, 2021. (Cited on page 19.)

[48] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. Image and Vision Computing, 27(12):1743–1759, 2009. (Cited on page 23.)

[49] Mark Weiser and John Seely Brown. Designing calm technology. `https://people.csail.mit.edu/rudolph/Teaching/weiser.pdf`, December 1995. Xerox PARC technical report; available online. (Cited on pages 1 and 7.)

[50] Wesley Willett et al. Embedded data representations. IEEE Transactions on Visualization and Computer Graphics, 23(1):461–470, 2017. (Cited on pages 1, 7, 8 and 21.)

[51] Julie Williamson et al. Globalfestival: Evaluating real world interaction on a spherical display. In Proceedings of the 2015 ACM International Joint Conference on Ubiquitous Computing, pages 159–162, 2015. (Cited on page 20.)

[52] Matthias Wölfel and John McDonough. Distant speech recognition. John Wiley & Sons, 2009. (Cited on page 22.)

[53] Yan Zhang, Danjv Lv, and Ying Lin. The classification of environmental audio with ensemble learning. In International Conference on Advanced Computer Science and Electronics Information (ICACSEI 2013), 2013. School of Computer and Information Southwest Forestry University, Yunnan Province, China; School of Software Yunnan University, Yunnan Province, China. (Cited on page 23.)