

Bálint Turi

**DEEP LEARNING FOR HEAD ORIENTATION
ESTIMATION USING PHASE SPECTROGRAMS
FROM A SINGLE MICROPHONE ARRAY**

Master of Science Thesis
Faculty of Information Technology and Communication Sciences
October 2025

ABSTRACT

Bálint Turi: Deep Learning for Head Orientation Estimation Using Phase Spectrograms from a Single Microphone Array
Master of Science Thesis
Tampere University
Data Science
October 2025

Estimating a speaker's head orientation from audio provides valuable contextual information for applications such as smart environments, meeting analysis, and driver monitoring. This thesis presents a data-driven method that estimates head orientation from the audio signals captured by a single compact microphone array, using only phase information extracted from the short-time Fourier transform. The proposed deep neural network combines convolutional, recurrent, and self-attention layers to learn spatial, temporal, and contextual patterns directly from phase spectrograms.

In contrast to prior work relying on handcrafted, physics-based features or raw waveform inputs, the presented approach enables robust learning from both simulated and real data. A large-scale simulated dataset was generated using measured voice directivity patterns and room acoustic simulations, allowing the model to learn orientation-dependent cues under varied acoustic conditions. The network is then fine-tuned on real recordings, resulting in improved adaptation to real-world variability.

The proposed system achieves state-of-the-art accuracy, outperforming baseline models under both clean and noisy conditions. In reverberant environments, it attains a mean angular error of 26.0° , and further personalization to individual speakers and environments reduces this to 13.9° . Evaluation on a real dataset confirms the method's generalization capability, reaching 71.1% classification accuracy after fine-tuning. These results demonstrate that accurate head orientation estimation can be achieved with a single microphone array, without the need for cameras or dense array setups.

Keywords: Head orientation, phase spectrogram, speech processing

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

USE OF ARTIFICIAL INTELLIGENCE IN THIS WORK

Artificial intelligence (AI) has been used in generating this work:

- No
 Yes

I hereby declare, that the AI-based applications used in generating this work are as follows:

Application	Version
OpenAI ChatGPT	GPT-5

Purpose of the use of AI

A large language model was used to improve language and clarity during the editing process, as well as to support code development by accelerating coding tasks. The model was not used to generate complete paragraphs; rather, it was used to review and refine the text I had written. Furthermore, no AI-generated text or code was used without subsequent thorough verification of its correctness and accuracy.

Parts of this work, where AI was used

AI assistance was employed throughout the entire text of this work. While no complete sections, tables, or figures were generated solely by AI, the large language model had a hand in refining language, improving clarity, and supporting code development across most chapters, sections, and subsections.

Acknowledgement of risks

I hereby acknowledge, that as the author of this work, I am fully responsible for the contents presented in this thesis. This includes the parts that were generated by an AI, in part or in their entirety. I therefore also acknowledge my responsibility in the case, where use of AI has resulted in ethical guidelines being breached.

CONTENTS

1.	Introduction	1
1.1	Motivation and problem definition	1
1.2	Thesis structure	4
2.	Theoretical Background and Related Work	5
2.1	Theoretical Background.	5
2.1.1	Digital Signal Representation and Time–Frequency Analysis	5
2.1.2	Spatial Audio and Microphone Array Processing	7
2.1.3	Neural Networks for Acoustic Modeling	11
2.2	Related Work	14
3.	Methods	17
3.1	Voice activity detection	18
3.2	STFT feature extraction	19
3.3	Neural network architecture	21
4.	Evaluation	23
4.1	Training	24
4.1.1	Simulated Dataset Generation	24
4.1.2	Real Data	26
4.1.3	Training setup	27
4.2	Evaluation metrics	29
4.2.1	Mean Angular Error	29
4.2.2	Classification Accuracy.	29
4.3	Baselines.	30
4.3.1	Raw Audio + CNN.	30
4.3.2	ITD & ILD Features	30
4.3.3	Physics-Informed Features	31
4.4	Results.	32
4.4.1	Simulated data	32
4.4.2	Real data	37
4.5	Discussion	38
5.	Conclusions	39
	References.	42

GLOSSARY

CNN	Convolutional Neural Network
DFT	Discrete Fourier Transform
DoV	Direction of Voice
FFT	Fast Fourier Transform
GRU	Gated Recurrent Unit
MAE	Mean Angular Error
MHSA	Multi-Head Self-Attention
RIR	Room Impulse Response
STFT	Short-Time Fourier Transform
VAD	Voice Activity Detection
VDP	Voice Directivity Pattern

1. INTRODUCTION

1.1 Motivation and problem definition

Human–machine interaction is becoming an increasingly important part of everyday life, from voice assistants in our homes that allow people to control lights, appliances, and media with simple commands, to meeting rooms in offices where hybrid meetings require participants to be tracked and understood automatically, and to intelligent driver-assistance systems in cars designed to improve safety. For these technologies to feel natural and useful, machines need to understand more than just the words people say—they also need to interpret non-verbal cues that reveal where a person’s attention is directed and what their intent might be. One of the clearest and most informative of these signals is head orientation, as the direction a speaker is facing often indicates what or who they are engaged with. In smart homes, head orientation can help decide which device a user’s voice command is directed to when multiple smart appliances are present in the same room [1, 2, 3]. In meeting rooms, knowing who a speaker is facing makes it possible to identify the addressee and improve the quality of the transcription [4, 5]. In vehicles, monitoring where a driver is facing provides critical information about attention and distraction, making it an important part of road safety systems [6, 7].

Much of the existing research on estimating head orientation provides a vision-based solution, which has been studied extensively [8]. While these systems achieve high accuracy under controlled conditions, they show several limitations that restrict real-world deployment. They typically require calibrated and synchronized cameras, making setup costly and technically demanding. Their performance can drop significantly in real-world settings with occlusions, variable lighting conditions, or when the subject is out of view. Most importantly, continuous video monitoring raises privacy concerns that can make camera-based solutions unsuitable for many everyday environments, such as homes, workplaces, or healthcare facilities.

These drawbacks motivate interest in audio-based alternatives. Unlike cameras, microphones are already built into many of the devices we use daily, such as smart speakers, laptops, conferencing systems, and vehicles, making them an attractive and low-cost sensing option. Audio is also less sensitive to changes in lighting or line of sight and raises fewer privacy concerns compared to video recording.

The feasibility of audio-based orientation estimation arises from the directional nature of human speech production. When someone speaks, their whole body radiates sound unevenly, producing systematic variations in spectral coloration, inter-channel phase differences, and reflection patterns. By exploiting this directional property of speech, it becomes possible to infer the direction in which the speaker’s head is pointing. This makes audio-only systems a practical, hardware-efficient, and better privacy-preserving alternative to vision-based approaches to estimate head orientation.

Despite this promise, several open challenges remain for audio-based methods. Real environments are acoustically complex: reverberation introduces reflections that distort the direct signal, and background noise or interfering speakers often overlap with the target speech, further complicating the analysis. Traditional model-based approaches [9, 10, 11, 12, 13, 14, 15] typically required a large number of microphones distributed throughout a room, making them impractical and computationally expensive. More recent methods [16, 2] reduced hardware demands but still assumed prior knowledge of array locations and user positions, which is not realistic for everyday use. In contrast, approaches based on a single compact device with a small embedded array are far more practical and better aligned with real-world applications. To address these challenges, recent research has increasingly turned to data-driven methods [17, 18, 3, 1, 19]. However, the scarcity of large annotated datasets makes training deep models difficult, and robust generalization to unseen speakers and environments remains a significant challenge.

In this thesis, the objective is to estimate the head orientation of a single speaker in the azimuthal plane within a reverberant room environment, using only a single microphone array with specifications similar to those found in smart home audio devices. Restricting the estimation to the azimuthal plane simplifies the task while still addressing the most practically relevant aspect of head orientation, since the horizontal facing direction dominates in everyday interactions. In this setting, both the speaker and the array may occupy any position within the room at a fixed height, making the task more general and practical compared to methods that assume fixed or calibrated setups. The angle of orientation θ_{ori} is defined as the angle between the speaker’s facing direction and the line that connects the speaker’s mouth and the center of the microphone array, measured over the full 0° – 360° range, as illustrated in Figure 1.1. The system takes a single speech utterance as input and outputs a single orientation estimate, assuming a static head orientation.

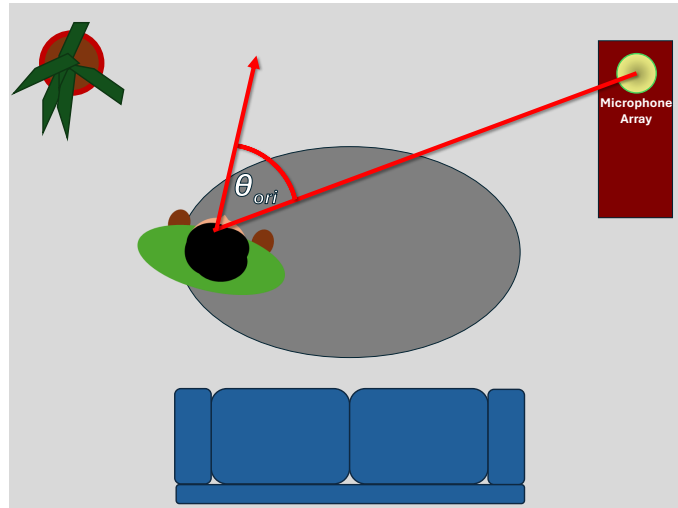


Figure 1.1. Top-down illustration of the problem setup showing the speaker, microphone array, and the orientation angle θ_{ori} defined between the speaker's facing direction and the line toward the array.

We propose a novel approach that uses the **phase component of the short-time Fourier transform (STFT)** as the primary input feature for a deep neural network that combines convolutional, recurrent, and self-attention layers. To address the challenge of training such a model with high-dimensional input features without access to large annotated datasets, we create a large-scale simulated dataset generated by leveraging voice directivity patterns (VDP). The model is first pre-trained on this simulated dataset and then fine-tuned on real recordings, achieving state-of-the-art accuracy and strong generalization across speakers and environments.

1.2 Thesis structure

The remainder of this thesis is organized as follows. Chapter 2 provides the theoretical background necessary to understand the proposed approach. Introduces the fundamentals of digital signal processing, including the representation of sound as a discrete one-dimensional signal and the short-time Fourier transform (STFT) for time–frequency analysis. The chapter also explains sound propagation and microphone array processing, emphasizing how spatial information is captured through inter-channel phase differences. In addition, it presents the basic principles of neural networks—covering the key layer types used in this work and their training process—to clarify how the proposed architecture models temporal and spatial acoustic patterns. Additionally, it reviews related work in both vision- and audio-based head orientation estimation, highlighting the limitations of existing methods and motivating the development of the proposed approach.

Chapter 3 describes the proposed system in detail. It outlines the overall processing pipeline, beginning with voice activity detection and feature extraction based on the phase component of the short-time Fourier transform, and proceeding to the design of the deep neural network architecture that combines convolutional, recurrent, and self-attention mechanisms for robust orientation estimation.

Chapter 4 presents the experimental setup and evaluation procedure. It explains how the simulated and real datasets were constructed or used, details the training strategy and optimization settings, and defines the metrics employed to assess model performance. The chapter then reports the results under different acoustic conditions, compares the proposed method with several state-of-the-art baselines, and investigates the effects of personalization through user- and environment-specific fine-tuning.

Chapter 5 concludes the thesis by summarizing the main findings and contributions. It discusses the implications of the results, outlines the current limitations of the approach, and suggests directions for future research toward more generalizable and adaptive audio-based head orientation estimation systems.

2. THEORETICAL BACKGROUND AND RELATED WORK

This chapter introduces the theoretical foundations required to understand the proposed approach for estimating speaker head orientation from multichannel audio. It is divided into two main parts. The first part (Section 2.1) presents the necessary background on digital signal processing, spatial audio, and neural modeling. The second part (Section 2.2) reviews previous research in head orientation estimation, summarizing the field’s development from model-based methods to data-driven approaches.

2.1 Theoretical Background

This section introduces the theoretical foundations required to understand the proposed approach for estimating speaker head orientation from multichannel audio. It is divided into three main parts. First, basic concepts of digital signal processing are introduced, including how a continuous sound wave is represented as a discrete signal and analyzed in the time–frequency domain using the Short-Time Fourier Transform (STFT). Next, the acoustic principles behind spatial audio are discussed, focusing on the directional properties of speech, the propagation of sound in a room, and how microphone arrays capture this spatial information. Finally, the chapter summarizes the neural network architectures used in this work and how they are applied to model spatial–temporal acoustic patterns.

2.1.1 Digital Signal Representation and Time–Frequency Analysis

One-Dimensional Discrete Signal

Speech signals are continuous in time, but for digital processing, they need to be represented in discrete form. A discrete signal x is defined as a finite sequence of N real- or complex-valued samples:

$$x = (x_1, x_2, \dots, x_N), \quad (2.1)$$

where each x_n represents the amplitude of the signal at discrete time n . Samples are

obtained from a continuous-time acoustic waveform $x(t)$ through uniform sampling at a rate f_s :

$$x_n = x(nT_s), \quad T_s = \frac{1}{f_s}. \quad (2.2)$$

According to the Nyquist–Shannon theorem, the highest representable frequency is

$$f_{\max} = \frac{f_s}{2}. \quad (2.3)$$

In this thesis, signals are sampled at $f_s = 48$ kHz, which sufficiently covers the frequency range of speech.

While the discrete-time signal $x[n]$ fully describes the waveform, it provides little direct insight into the underlying frequency content that characterizes speech. Speech is highly non-stationary: its spectral composition changes rapidly as phonemes and articulations evolve. To analyze these variations, it is therefore more informative to study the signal in the frequency domain.

Discrete Fourier Transform

The Discrete Fourier Transform (DFT) decomposes a signal into its frequency components:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N}, \quad k = 0, 1, \dots, N-1. \quad (2.4)$$

Each $X[k]$ is a complex coefficient representing the amplitude and phase at frequency $f_k = kf_s/N$. The transformation maps a time-domain sequence to its frequency-domain representation, revealing which frequencies are present and with what magnitude and phase.

In practice, the DFT is computed using the Fast Fourier Transform (FFT) algorithm [20], which efficiently exploits symmetries in the transform to reduce computational complexity from $O(N^2)$ to $O(N \log N)$.

The Fourier representation allows us to separate periodic structures and spectral content from the temporal evolution of the waveform. However, for speech and other signals that change over time, a global frequency analysis is not sufficient.

Windowing and Short-Time Fourier Transform

To capture how the frequency content evolves over time, the signal is analyzed in short, overlapping segments using the Short-Time Fourier Transform (STFT) [21]. For each frame, a window function $w[n]$ of length N_w is applied to emphasize the current segment and reduce discontinuities at its boundaries:

$$X[m, f] = \sum_{n=0}^{N_w-1} x[n + mH]w[n]e^{-j2\pi fn/N_w}, \quad (2.5)$$

where H is the hop size between consecutive frames and m is the discrete frame index. A commonly used window is the Hann window,

$$w[n] = \frac{1}{2} \left(1 - \cos \left(\frac{2\pi n}{N_w - 1} \right) \right), \quad (2.6)$$

which provides good spectral leakage suppression.

The STFT produces a two-dimensional complex-valued matrix $X[m, f]$, where m indexes time frames and f frequency bins. Each element can be written as

$$X[m, f] = A[m, f]e^{j\phi[m, f]}, \quad (2.7)$$

where $A[m, f]$ and $\phi[m, f]$ denote the local magnitude and phase, respectively.

Although most work in speech processing relies primarily on the magnitude spectrum, the phase term contains valuable information about the fine temporal and spatial structure of the signal. In this thesis, the phase information extracted from the STFT is the key feature used for orientation estimation. The phase difference across microphones encodes relative time delays and spatial cues, which are essential for capturing the directionality of the speech source.

2.1.2 Spatial Audio and Microphone Array Processing

Speech is inherently spatial: it originates from a physical source, propagates through an acoustic environment, and is captured by microphones placed at distinct positions. Understanding these spatial properties is essential for interpreting the directional cues used in head orientation estimation [22].

Sound Source Directivity

Human speech is not radiated uniformly in all directions. The vocal tract and head shape cause frequency-dependent variations in sound intensity, known as the *voice directivity pattern* (VDP). At low frequencies (below about 1 kHz), the radiation is nearly omni-

rectional, whereas higher frequencies become increasingly directional, with most of the acoustic energy radiated forward from the mouth region, as illustrated in Figure 2.1. As the speaker turns their head, the spectral coloration and phase patterns measured by a microphone array change systematically, providing cues about orientation [23].

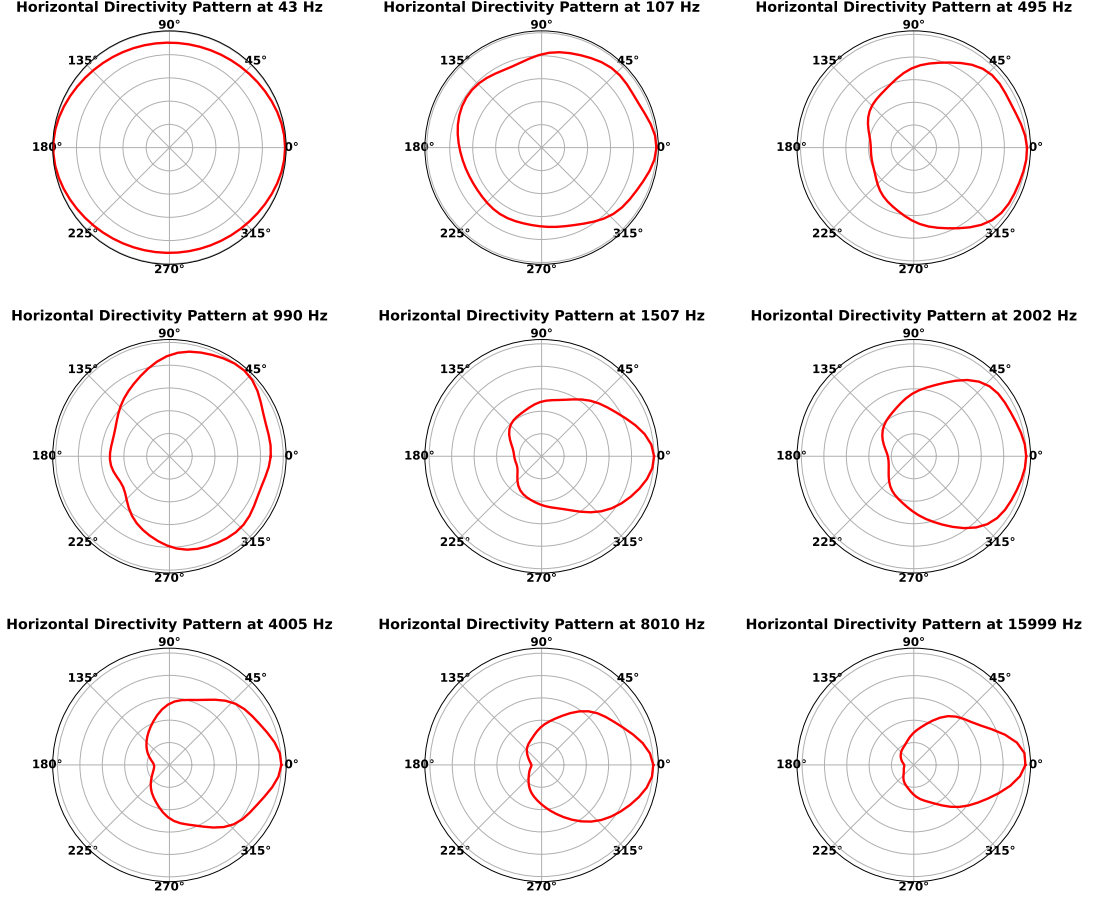


Figure 2.1. Horizontal-plane voice directivity pattern at different frequencies.

A directivity pattern $D(\theta, f)$ can be expressed as the normalized pressure amplitude measured at azimuth angle θ and frequency f :

$$D(\theta, f) = \frac{p(r, \theta, f)}{p(r, 0, f)}, \quad (2.8)$$

where $p(r, \theta, f)$ denotes the complex acoustic pressure of the radiated sound field measured at distance r and azimuth θ [23].

In this work, directivity is explicitly modeled when simulating the acoustic scenes. For each desired head orientation, the speech signal is weighted in the frequency domain by the corresponding directivity response before being propagated to the microphones.

Sound Propagation in Rooms

Once emitted, the sound wave propagates through the environment and interacts with surfaces such as walls, floor, and ceiling. The signal arriving at a microphone is therefore a combination of the direct path and multiple reflections, modeled by the room impulse response (RIR) $h_c(t)$ [24, 25]:

$$x_c(t) = s(t) * h_c(t) + n_c(t), \quad (2.9)$$

where $s(t)$ is the source signal, $*$ denotes convolution, and $n_c(t)$ is additive noise. Each RIR encodes the geometry and materials of the room and depends on both source and microphone positions. Each reflection introduces a delayed and attenuated version of the signal, leading to *reverberation*. The RIR can be approximated as

$$h_c(t) = \sum_i \alpha_i(f) \delta(t - \tau_i), \quad (2.10)$$

where $\alpha_i(f)$ and τ_i denote the frequency-dependent amplitude and delay of the i -th propagation path, respectively. The ratio between direct and reflected energy, together with the time constant of energy decay, determines the reverberation time T_{60} , which characterizes the acoustic liveness of the room [25].

In this work, room acoustics are simulated using the `Pyroomacoustics` library [26], which implements the image source method to generate RIRs for arbitrary room geometries. The resulting reverberant multichannel signals are obtained as

$$x_c(t) = \mathcal{F}^{-1}\{S(f)D(\theta, f)H_c(f)\}, \quad (2.11)$$

where $S(f)$ is the spectrum of the clean speech signal, $D(\theta, f)$ is the directivity pattern corresponding to the desired head orientation, $H_c(f)$ is the room transfer function for the c -th microphone, and $\mathcal{F}^{-1}\{\cdot\}$ denotes the inverse Fourier transform that maps the frequency-domain product back to the time domain. This process ensures that each simulated signal accurately reflects both the directionality of speech and the acoustic properties of the environment.

Reflections are often viewed as a source of distortion, but in this work, they provide useful secondary cues: the relative phase of reflections changes with the speaker's orientation, enriching the spatial information available to the model.

Microphone Array Capture

A microphone array consists of C microphones located at known positions $\mathbf{r}_c = [x_c, y_c, z_c]^\top$ in three-dimensional space. When a plane wave characterized by the wave vector \mathbf{k} reaches the array, the signal at microphone c can be expressed as

$$x_c(t) = s(t - \tau_c), \quad \tau_c = \frac{\mathbf{k} \cdot \mathbf{r}_c}{c_s}, \quad (2.12)$$

where \mathbf{k} points in the direction of propagation with magnitude $|\mathbf{k}| = 2\pi f/c_s$, and c_s is the speed of sound. The small time differences τ_c between microphones result in frequency-dependent phase shifts that encode the direction of arrival [27]. For speech, these phase shifts also depend on the head orientation relative to the array, especially when combined with directional radiation and room reflections.

The spatial information can be captured in the frequency domain by computing the cross-spectrum between microphones i and j :

$$C_{ij}(f) = X_i(f)X_j^*(f), \quad (2.13)$$

where $(\cdot)^*$ denotes complex conjugation. The phase of $C_{ij}(f)$ encodes the inter-channel time difference, forming the basis for many spatial features such as interaural phase difference (IPD) or generalized cross-correlation (GCC-PHAT) [28].

In practical systems, these inter-microphone phase differences, combined with direction-dependent spectral coloration, provide the main cues for estimating the orientation of the speaker relative to the array. The STFT-phase features used in this work capture exactly this type of information.

2.1.3 Neural Networks for Acoustic Modeling

Deep neural networks have become the dominant framework for modeling complex relationships in audio and speech processing. They can learn complex mappings from high-dimensional audio features to target variables such as head orientation. This section briefly summarizes the main concepts relevant to the architecture used in this thesis.

Feedforward and Multilayer Networks

The fundamental building block of most neural architectures is the feedforward network, or multilayer perceptron (MLP). Each neuron computes a weighted sum of its input feature vector \mathbf{x} followed by a non-linear activation function:

$$y = \sigma(\mathbf{w}^\top \mathbf{x} + b), \quad (2.14)$$

where \mathbf{x} contains the input features from the previous layer (or directly from the data), \mathbf{w} and b are the trainable weight vector and bias, and $\sigma(\cdot)$ is a non-linear activation such as the rectified linear unit (ReLU) or sigmoid function. When multiple layers of such units are stacked, the network can model highly non-linear mappings between input and output domains. Although MLPs are conceptually simple, they are limited in their ability to capture temporal or spatial structures in data.

Convolutional Neural Networks

Convolutional layers extend MLPs by applying small, shared kernels over local regions of the input, enabling efficient extraction of local time–frequency patterns. A kernel (or filter) is a compact weight matrix that slides over the input, computing local weighted sums that highlight relevant spatial or spectral features. Weight sharing across time and frequency reduces the number of parameters while preserving translation invariance [29], making them ideal for spectrogram-based features.

When applied to STFT representations of multi-channel audio, convolutional layers operate on two-dimensional feature maps where one dimension corresponds to time frames and the other to frequency bins. Each microphone channel can be treated as a separate input channel, analogous to color channels in images. The convolutional kernels slide across time and frequency, learning filters that capture local spatial–spectral correlations and inter-channel phase relationships. This allows the network to extract features that are sensitive to both spectral structure and spatial configuration.

Pooling Layers

Pooling layers are used to reduce the spatial or temporal resolution of feature maps while retaining the most informative activations. The most common form, max pooling, selects the maximum value within small non-overlapping regions. This operation introduces translation invariance, emphasizes the most important features, and reduces the computational complexity of subsequent layers.

Adaptive Pooling

While standard pooling operates with fixed kernel sizes, adaptive pooling adjusts the pooling regions dynamically to produce a predefined output size. This makes it particularly useful when processing input sequences of variable length, as it allows the network to generate fixed-size representations regardless of the input duration. In this work, adaptive max pooling is used in the aggregation stage to obtain a uniform embedding for orientation estimation.

Recurrent Layers and GRUs

Speech contains strong temporal dependencies that cannot be captured by static filters. Recurrent neural networks address this by maintaining a hidden state that evolves over time. Among their variants, the Gated Recurrent Unit (GRU) [30] provides a compact and effective mechanism to model long-term dependencies. GRUs use update and reset gates to control how much past information is retained, enabling stable training on sequences of varying lengths. In our architecture, bidirectional [31] GRU layers are used so that both past and future temporal context contribute to each frame's representation.

Self-Attention Mechanisms

The self-attention mechanism [32] computes relationships between all time frames simultaneously. Each output vector is a weighted combination of all input vectors, where the weights are determined by learned similarity scores. This enables the network to focus on the most relevant temporal segments or frequency regions in our case. Multi-head self-attention (MHSA) extends this by learning multiple independent attention patterns in parallel.

Training and Optimization

Neural networks are trained by minimizing a task-specific objective function through iterative optimization. The most common approach is gradient-based learning, in which the model parameters are updated according to the steepest downward direction in the

loss landscape. Gradients are computed efficiently using the backpropagation algorithm, which applies the chain rule of differentiation through all network layers.

Optimization is typically performed using variants of stochastic gradient descent (SGD) [33], where gradients are estimated from small random batches of data rather than the entire dataset. Among adaptive optimization methods, *Adam* [34] is one of the most widely used due to its stable and efficient convergence properties. Adam computes parameter-specific learning rates based on running estimates of the first and second moments of the gradients:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t, & v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, \\ \hat{m}_t &= \frac{m_t}{1 - \beta_1^t}, & \hat{v}_t &= \frac{v_t}{1 - \beta_2^t}, \\ w_t &= w_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}}, \end{aligned}$$

where g_t is the gradient at step t , β_1 and β_2 are decay rates (commonly set to 0.9 and 0.999), and η is the global learning rate. This adaptive behavior helps maintain fast convergence while remaining robust to noisy gradient estimates, which is especially important for complex tasks such as speech-based orientation estimation.

Dropout Regularization

Dropout is a widely used regularization technique that prevents overfitting by randomly deactivating a fraction of neuron outputs during training. This encourages the network to rely on distributed representations rather than specific neurons, leading to better generalization on unseen data.

In summary, the combination of convolutional, recurrent, and attention-based layers enables the model to capture local spatial–spectral cues, temporal dependencies, and long-range contextual relationships in a unified end-to-end framework for robust head orientation estimation.

2.2 Related Work

Early studies on acoustic estimation of head orientation primarily relied on model-based techniques that explicitly formulated the relationship between acoustic energy distribution and the speaker's facing direction. These approaches were typically designed for controlled environments and employed large microphone arrays to achieve spatial precision. One of the representative works in this category is by [9], who introduced the *Huge Microphone Array (HMA)* consisting of 448 microphones distributed around a laboratory environment. Their system estimated the talker's azimuthal orientation by analyzing spatial energy patterns derived from the array, exploiting the anisotropy of human speech radiation. Similar large-array systems were later explored in [10, 11], where energy-based or beamforming-based cues were used to infer the speaker's facing direction. In particular, an *orientation-extended amplitude beamforming (OE-ABF)* method using a 96-channel array was proposed in [11], achieving real-time estimation with an average angular error below 5° . While these systems demonstrated proof-of-concept feasibility, their reliance on dozens, hundreds of arrays at each wall around a room in order to cover all possible directions made them impractical for real-world scenarios.

Subsequent research shifted toward exploiting *inter-microphone correlations* rather than absolute energy fields to derive orientation cues. Many of these methods employed the *Generalized Cross-Correlation with Phase Transform (GCC-PHAT)* [28] as their core component. For example, [12] extended the conventional Steered Response Power (SRP-PHAT) localization algorithm by introducing orientation as an additional search parameter, weighting the contribution of each microphone pair according to the hypothesized head pose. This allowed estimating the joint position and orientation of the speaker. In [13], a two-step method based on GCC-PHAT was introduced in which the speaker position is first estimated using SRP-PHAT, and then the values of the cross-correlation peaks at the corresponding time delays are analyzed between microphone pairs to infer the most likely head orientation.

In parallel, other approaches [14, 15] focused on spectral energy cues, leveraging the frequency-dependent directivity of human speech radiation. These methods estimated orientation from the *High-to-Low Band Energy Ratio (HLBR)*, defined as the ratio between high- and low-frequency speech energy measured at distributed microphones. Since high frequencies are radiated more directionally than low frequencies, this ratio encodes orientation information. Moreover, the HLBR measure provides a form of normalization that mitigates calibration errors and propagation losses, enabling low-complexity orientation estimation when the speaker position is known beforehand.

These classical methods were appealing for their simplicity and lack of training requirements, yet they were inherently constrained by two main factors. First, their reliance on distributed microphone arrays required precise spatial calibration and synchronization,

since the estimation accuracy directly depended on the known geometry and fixed layout of the microphones. Second, energy- and correlation-based features were highly sensitive to reverberation, as reflected signals can create misleading peaks and distort the underlying time-delay structure.

In an effort to reduce the hardware requirements of dense microphone arrays, later studies investigated more compact array configurations and geometric simplifications. For example, [16] proposed estimating head orientation from multiple distributed but relatively small arrays. Their system employed six four-channel T-shaped arrays positioned along the room walls, each providing a local orientation estimate derived from the *Oriented Global Coherence Field (OGCF)* and *High-to-Low Band Energy Ratio (HLBR)* features. The final orientation was obtained by geometrically combining these local estimates. Similarly, [2] presented a model-based framework named *HOE*, designed for smart devices equipped with only two four-channel arrays. *HOE* models the speech radiation pattern as a frequency-dependent cardioid function and estimates orientation by matching measured energy distributions to the theoretical pattern. To address practical issues such as propagation attenuation, reverberation, and orientation ambiguity, the system incorporates distance- and direction-dependent energy compensation as well as a high-frequency energy ratio for disambiguation. Although these methods represented a clear step toward practical deployment, they continued to assume fixed array and speaker locations. Consequently, their usability remained limited to controlled rooms rather than the ad-hoc environments encountered in consumer devices.

To address these challenges, research has explored machine learning-based approaches. Early work by [17] introduced a single-channel method using Hidden Markov Models (HMMs) to estimate head orientation from reverberant speech. The system first separated the acoustic transfer function (ATF) of the speech signal using phoneme-based clean-speech HMMs, then classified the resulting ATF sequences with a Support Vector Machine (SVM) trained on data corresponding to known head orientations. Although this demonstrated that temporal patterns in the transfer function contain orientation information, its performance was limited compared to multichannel systems. A subsequent study by the same authors [18] extended this framework to a two-microphone configuration, introducing features derived from the full shape of the Cross-Power Spectrum Phase (CSP) coefficients rather than only their peak values. By exploiting reverberation-dependent variations in CSP patterns, this method achieved improved robustness and orientation discrimination in real environments, but was based on loudspeaker data and remained sensitive to real-world noise.

A significant step forward came with the introduction of the *Direction-of-Voice (DoV)* work [3], which formulated head orientation estimation as a classification problem over eight discrete azimuthal directions. The authors collected a large real-world dataset and trained a decision-tree classifier using a diverse set of features, including frequency band

power ratios, spectral polynomial coefficients, reverberation and autocorrelation measures, and cross-correlation and time-difference-of-arrival features. This work demonstrated that relatively simple learning architectures could outperform traditional models, particularly when trained on data representative of realistic acoustic environments.

The *Soundr* system [1] further extended this line of research by leveraging deep learning to infer head orientation directly from multichannel raw audio. Using over 700 minutes of recorded data, it used a convolutional–recurrent (CNN–LSTM) network trained on 16-channel waveforms to predict continuous angle of head orientation. While this end-to-end approach effectively learned orientation features, its generalization to unseen speakers and acoustic environments remained limited, achieving a mean angular error of approximately 57° . These findings highlighted the difficulty of purely data-driven estimation under limited real-world training data, where models tend to overfit to specific recording setups.

Overall, existing research has progressively evolved from model-based methods towards data-driven learning solutions. While this direction has improved flexibility and reduced dependence on large arrays or rigid calibration, current systems still face significant challenges in achieving robust generalization across diverse environments and speaker characteristics. These limitations highlight the need for approaches that can effectively leverage spatial and spectral cues while maintaining adaptability to real-world acoustic variability.

3. METHODS

This chapter presents the complete methodology used to estimate head orientation from multichannel speech recordings. The proposed system estimates the head orientation of a speaker from multichannel speech recordings captured by a compact microphone array. As shown in Figure 3.1, the pipeline consists of three main components: (i) voice activity detection (VAD), which filters out non-speech segments and ensures that only speech-related frames are processed; (ii) feature extraction, where the phase component of the short-time Fourier transform is computed from each microphone channel; and (iii) orientation estimation using a deep neural network that combines convolutional, recurrent, and self-attention mechanisms.

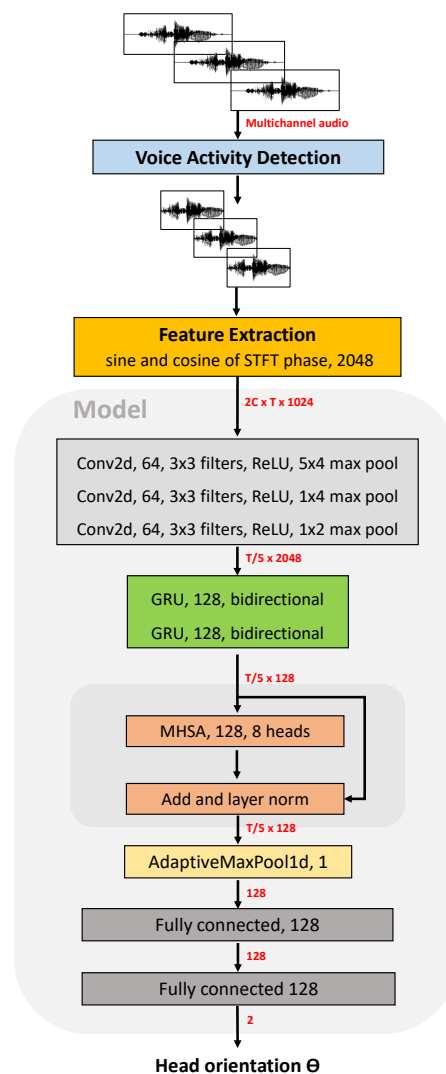


Figure 3.1. Head orientation estimation pipeline and model architecture overview.

3.1 Voice activity detection

The first stage of the pipeline is a voice activity detection module, which identifies and removes non-speech frames, such as silence or background noise, that do not carry useful directional information about the speaker’s head orientation and may introduce artifacts that degrade performance. To address this, we employ a state-of-the-art neural VAD model from `pyannote.audio` [35], based on a temporal convolutional neural network trained on large-scale datasets spanning diverse acoustic conditions, ensuring robustness to noise and different environments. The VAD operates in the time domain and is trained to predict a binary sequence indicating whether each frame is speech-active or not. During inference, it produces frame-level scores between 0 and 1 that represent the likelihood of speech activity, which are then binarized using two tunable thresholds: an *onset* threshold that determines when speech activity begins and an *offset* threshold that determines when speech activity ends (see Fig. 3.2). To avoid truncating relevant information, a small context margin is applied around the detected speech regions. As a result, detected leading and trailing non-speech frames are filtered out, and only the remaining speech segment is passed to the feature extraction stage, thereby improving the robustness and stability of the system.

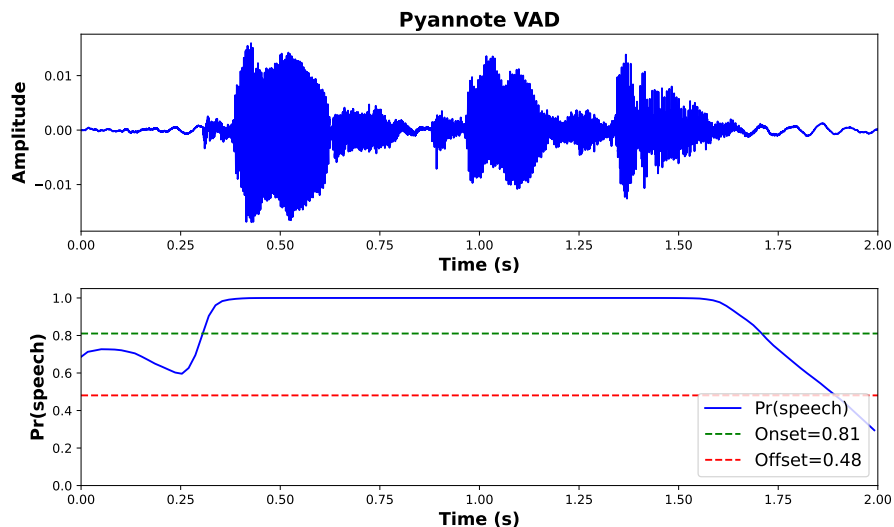


Figure 3.2. VAD output for a sample utterance. The top plot shows the speech waveform and the bottom plot shows the frame-level speech probability produced by the `pyannote.audio` model, along with the onset (green) and offset (red) thresholds used for binarization.

3.2 STFT feature extraction

The second stage of the pipeline is the extraction of time—frequency domain features from the speech signal.

For each microphone channel c , the short-time Fourier transform (STFT) is computed as

$$X_c(m, f) = \sum_{n=0}^{N_w-1} x_c[n + mH] w[n] e^{-j2\pi fn/N_{\text{fft}}},$$

where $x_c[n]$ is the discrete-time input signal, $w[n]$ is a Hann window of length N_w , H is the hop size, and m and f denote the time-frame and frequency-bin indices, respectively. In our implementation, we use a window size of 0.04 s, a hop size of 0.02 s, and an FFT size of $N_{\text{fft}} = 2048$.

The STFT is complex-valued and can be written in polar form as

$$X_c(m, f) = A_c(m, f) e^{j\phi_c(m, f)},$$

where $A_c(m, f) = |X_c(m, f)|$ is the magnitude and $\phi_c(m, f)$ is the phase. The phase is extracted using

$$\phi_c(m, f) = \text{atan2}(\Im\{X_c(m, f)\}, \Re\{X_c(m, f)\}),$$

which yields values wrapped to the interval $(-\pi, \pi]$.

Since phase values are circular, we represent each phase $\phi_c(m, f)$ by its sine and cosine components:

$$\phi_c(m, f) \mapsto [\sin(\phi_c(m, f)), \cos(\phi_c(m, f))],$$

which avoids discontinuities and preserves the circular structure of the phase. Following common practice in spatial feature modeling, this representation provides a numerically stable and rotation-invariant encoding of phase information [36, 37].

Because the FFT of a real-valued signal is conjugate-symmetric, only the first half of the spectrum contains unique information. Thus, when $N_{\text{fft}} = 2048$, the number of unique frequency bins is

$$F = \frac{N_{\text{fft}}}{2} = 1024.$$

For an array with C microphones, the resulting input tensor has size

$$2C \times T \times F,$$

where $2C$ accounts for sine and cosine values per channel, T is the number of time frames, and $F = 1024$ is the number of frequency bins. This high-dimensional representation captures rich directional information, which is then processed by the neural network.

3.3 Neural network architecture

The final stage of the system is a deep neural network that maps multichannel phase features to an orientation estimate. The model architecture combines previous work in spatial audio scene analysis [38] with self-attention feature learning [39].

Convolutional layers: Three 2D convolutional layers with 3×3 kernels, 64 filters, and ReLU activations are applied to the input features with a shape of $2C \times T \times 1024$. The three convolutional layers contain 7,104, 37,056, and 37,056 parameters, respectively. Each convolution is followed by max pooling with kernel sizes 5×4 , 1×4 , and 1×2 , respectively. Dropout is applied after each convolutional block to avoid overfitting. This progressively reduces temporal and frequency resolution while preserving spatial structure, resulting in a feature map of shape $64 \times T/5 \times 32$. After reshaping, the feature maps have size $T/5 \times 2048$.

Recurrent layers: To capture sequential dependencies across frames, the feature sequence is processed by two bidirectional GRU layers, each with 128 hidden units. The bidirectional design allows the network to capture both past and future contexts. Together, these layers account for 1,969,152 trainable parameters and produce embeddings of size $T/5 \times 128$.

Self-attention: To further improve the temporal representations, we apply two blocks of multi-head self-attention, each with 8 heads and an attention size of 128. Self-attention enables the model to dynamically weight frames based on their relevance, effectively focusing on the most informative time–frequency regions for orientation estimation. Residual connections and layer normalization are used after each block to ensure stable training and prevent gradient vanishing. Each attention block contains 66,048 parameters, totaling 132,096 across both layers.

Aggregation and output: To handle input of variable length, the temporal sequence is aggregated with adaptive max pooling, producing a fixed-size embedding of dimension 128. This embedding is then passed through a fully connected layer with 16,512 parameters and a final linear projection that outputs two values:

$$[\cos(\theta), \sin(\theta)],$$

where θ is the estimated orientation angle. By predicting sine and cosine instead of the angle directly, the network avoids discontinuities at the $0^\circ/360^\circ$ boundary, yielding smooth angle predictions. The final orientation in degrees can be recovered from the network output using

$$\hat{\theta} = \left(\text{atan2}(y, x) \times \frac{180}{\pi} \right) \bmod 360,$$

where (x, y) correspond to the predicted $\cos(\theta)$ and $\sin(\theta)$ components.

In total, the model contains approximately 2.2 million trainable parameters, making it compact compared to many modern architectures while still expressive enough for the orientation estimation task. A more detailed breakdown of the parameter counts per layer is provided in Table 3.1.

Table 3.1. *Number of trainable parameters per layer in the neural network. Only layers with parameters are included.*

Layer	Parameters
Convolutional layer 1	7,104
Convolutional layer 2	37,056
Convolutional layer 3	37,056
Bidirectional GRU (2 layers)	1,969,152
Multi-head attention block 1	66,048
Multi-head attention block 2	66,048
Fully connected layer	16,512
Output projection	258
Total	2,199,746

4. EVALUATION

We evaluated our proposed method on both simulated and real datasets. For comparison, we also re-implemented two state-of-the-art approaches [1, 19] under the simulated setup, and for real-world recordings, we directly compared against the reported results of a third approach [3].

The simulated setup serves two key purposes. First, it enables the generation of a large-scale annotated dataset for pre-training when real data are limited. Second, it provides a controlled environment for systematically testing performance under different acoustic conditions and baselines. Using this setup, we investigate the head orientation estimation under several scenarios.

We begin by comparing performance in anechoic versus reverberant environments to see how reverberation affects estimation accuracy: whether it helps the model by providing additional cues or instead introduces distortions that degrade performance. Next, we test robustness to background noise by training and evaluating models at different signal-to-noise ratios. This step is important because real-world environments are rarely noise-free, and reliable performance under such conditions is critical for practical deployment.

We then explore the effect of personalization through user- and room-specific fine-tuning. Since machine learning models typically perform best when adapted to their operating conditions—and users may be willing to provide a small number of calibration samples—this experiment gives insight into the realistic performance gains achievable through adaptation.

Finally, we evaluated our method on a real dataset to validate its effectiveness beyond simulation. This experiment also examines whether pre-training on simulated data generalizes to real-world recordings, a key question for deploying such systems when large annotated real datasets are unavailable.

4.1 Training

Training a model for head orientation estimation from multichannel speech requires careful consideration of both data generation and optimization. The diversity and realism of the training data play a crucial role in determining how well the model generalizes to real, unseen acoustic conditions. Since collecting large annotated datasets of real recordings with precise orientation labels is challenging, a realistic simulation becomes essential. This section describes the procedures used to construct the datasets and the training setup applied to ensure consistent model performance.

4.1.1 Simulated Dataset Generation

Since a large annotated dataset of real recordings with continuous head orientation labels does not exist, we created a simulated dataset by combining three key components: (i) a clean speech corpus to provide realistic content, (ii) voice directivity patterns (VDPs) to model the directional radiation of human speech, and (iii) room acoustic simulation with noise mixing to capture environmental conditions.

Speech corpus

Speech samples were taken from the VCTK corpus [40], which includes recordings from 110 English speakers with diverse accents. Each speaker reads approximately 400 sentences, totaling around 44,000 utterances. For computational efficiency and to reflect realistic interaction durations, we truncated each utterance to a maximum of 3 s. All recordings were originally captured using high-quality omni-directional microphones at 96 kHz and 24-bit resolution and were converted to 16-bit samples and downsampled to 48 kHz.

Voice directivity patterns

To capture the radiation of human voice, we combined three datasets of measured VDPs, totaling 22 patterns:

- **DirPat database** [41]: Eight measured singing-voice directivities from professional singers, recorded at a distance of 1.23 m. The measurements were taken on a spherical grid with 11.25° resolution in the horizontal plane, providing fine angular sampling.
- **Soprano_m database** [42]: Six directivities of soprano singers, measured with 36° resolution in the azimuthal plane.
- **Classical singers dataset** [43]: Eight directivities of professional opera singers measured under anechoic conditions, with microphones placed every 15° in the horizontal plane.

Since all VDPs are obtained through measurements at a finite set of discrete angles, they cannot directly represent continuous head orientations. To address this limitation, we used the method proposed by [44], which reconstructs continuous spherical harmonic representations of the measured VDPs from finite-distance measurements. This approach enables smooth interpolation of the directivity function over all azimuthal angles, allowing the simulated voice source to rotate freely in any direction rather than being restricted to the discrete measurement grid.

Room and array simulation

For each utterance, we randomly sampled a rectangular room with dimensions drawn uniformly from $[3, 12] \times [3, 12] \times [2, 6]$ meters, covering a wide range of acoustic environments. The microphone device was modeled as a compact circular array with six microphones evenly spaced at 60° intervals on a circle of radius 4.5 cm. Both the array and the speaker were placed at random positions at a fixed height within the room. The head orientation angle was then sampled uniformly from the full 0° – 360° range. The resulting signals are then simulated with the Pyroomacoustics toolkit [26].

Noise conditions

Finally, to simulate realistic acoustic interference, we added background noise from the WHAM dataset [45]. Since WHAM provides only monophonic signals, we generate a realistic multichannel version by phase-randomizing uncorrelated copies and mixing them to match the target inter-channel coherences under isotropic diffuse field assumptions, following [46, 47]. Three different noise level conditions were considered, measured in signal-to-noise ratio: clean (no noise), moderate noise (10–20 dB SNR), and high noise (0–10 dB SNR). We also included anechoic clean simulations to isolate the role of reverberation.

To test true generalization, we set aside eleven speakers and six VDPs for testing, resulting in 3,947 utterances. The training used 40,295 utterances from the remaining 99 speakers. The same speaker and VDP splits were used in all simulation conditions for consistency.

4.1.2 Real Data

To validate our method on real-world recordings, we used the Direction-of-Voice (DoV) dataset introduced in [3]. This dataset contains recordings from 10 participants (4 male, 6 female), each recorded across two independent sessions. Data were collected in two rooms with very different sizes and reverberation properties: a large classroom ($24.3 \times 9.1 \times 4.0$ m) and a smaller office ($13.7 \times 6.1 \times 3.6$ m). For each participant, utterances were recorded at three distances from the microphone array (1 m, 3 m, and 5 m), and at three lateral offsets (0° , 45° , 90°). Within each position, the participant was asked to face eight discrete azimuthal orientations: 0° , $\pm 45^\circ$, $\pm 90^\circ$, $\pm 135^\circ$, and 180° .

Two utterances were spoken in each trial: a short command phrase (“hey assistant”) and a longer sentence (“the quick brown fox jumped over the lazy sheep”), resulting in a total of 11,520 recordings (over 350 minutes of audio) captured with a 4-channel ReSpeaker USB microphone array.

Following the official evaluation protocol of the “Per Angle Classifier” method, we used the provided split to fine-tune the classifier on all session one samples, and to test it on session two samples, and vice versa, which ensures that the model is evaluated under slightly different conditions from those it was trained on.

Unlike our simulated dataset, the DoV dataset defines orientation as an eight-class classification task and all recordings. Accordingly, we adapted our model by replacing the continuous regression output with an 8-class linear classifier layer. Furthermore, while our simulations used a 6-channel circular microphone array, the DoV recordings were captured with a 4-channel array. To ensure compatibility, we created a new pre-trained model by re-simulating the moderate-noise, reverberant training conditions using a 4-channel circular array. This configuration enabled direct comparison with the baseline results reported in [3], providing a consistent and realistic validation of our method’s performance in real acoustic environments.

4.1.3 Training setup

The training of the proposed head orientation estimation model follows the general principles of supervised deep learning, but requires some task-specific adaptations due to the nature of variable-length utterances and angular prediction.

Batch Construction and Variable-Length Inputs

Each training step processes a *mini-batch* of B audio samples simultaneously, where $B = 16$ in our experiments. Mini-batch training allows efficient use of the GPU, while also stabilizing gradient estimates through averaging.

Utterances in the dataset vary in length from 0.5 to 3 seconds, which results in STFT representations with different numbers of time frames T . However, batching requires tensors of uniform dimensions. To achieve this, all signals are padded along the time axis to match the longest sequence within the current batch. Padding is applied with zeros in the phase feature representation, which corresponds to silence after feature transformation. To minimize the negative impact of excessive padding on training efficiency, utterances are sorted by length before batching, so that samples of similar duration are grouped together. This strategy reduces the proportion of padding and improves both computational efficiency and model convergence.

Thus, for each batch, we obtain a tensor of shape

$$X \in \mathbb{R}^{B \times 2C \times T_{\max} \times F},$$

where C is the number of microphones, T_{\max} is the maximum number of frames in the current batch, and $F = 1024$ is the frequency dimension. The corresponding ground-truth labels are the sine and cosine of the target orientation angles $\theta \in [0^\circ, 360^\circ)$.

Loss Function

The model outputs a two-dimensional vector (\hat{y}_1, \hat{y}_2) that is trained to approximate the cosine and sine of the orientation angle, i.e.

$$y = (\cos \theta, \sin \theta).$$

This representation avoids discontinuities at the $0^\circ/360^\circ$ boundary and allows smooth learning of angular relationships. The training loss is the *mean squared error (MSE)*

between predicted and target vectors:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{B} \sum_{i=1}^B [(\hat{y}_{1,i} - \cos \theta_i)^2 + (\hat{y}_{2,i} - \sin \theta_i)^2].$$

This loss penalizes deviations equally across all angles and directly encourages the network to approximate the circular geometry of orientations.

Optimizer

We train the model using the *Adam optimizer* [34], following the standard configuration with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. It was chosen for its stable convergence and robustness to noisy gradient estimates, which are important when training on variable-length and noisy speech signals.

Learning Rate and Scheduler

The initial learning rate is set to

$$\eta_0 = 4 \times 10^{-4}.$$

To ensure stable convergence we apply a multiplicative decay after each epoch using a scheduler:

$$\eta_e = \eta_0 \cdot (0.95)^e,$$

where η_e is the learning rate at epoch e . This schedule reduces the step size gradually, allowing the model to make large updates during early training and fine adjustments in later epochs.

Training Duration and Iterations

Each training run is conducted for up to *200,000 iterations*, with one iteration corresponding to a single forward and backward pass of a mini-batch. Model checkpoints are saved after each epoch, and the best-performing checkpoint on the validation set is selected for final evaluation.

4.2 Evaluation metrics

The evaluation of head orientation estimation depends on whether the task is formulated as a regression over continuous angles or as a classification problem over discrete orientation bins. For the simulated dataset, which allows continuous labels across the full 0° – 360° range, we use the *mean angular error (MAE)* as the primary metric. For the real dataset, which defines eight fixed orientations, we follow the dataset’s protocol and report the classification accuracy.

4.2.1 Mean Angular Error

The Mean Angular Error (MAE) measures the average difference between the predicted and ground-truth orientation angles across the evaluation set. Since angular variables are periodic, the error must account for the circular wrapping at $0^\circ/360^\circ$. Specifically, for each sample i with ground-truth orientation $\theta_{t,i}$ and predicted orientation $\theta_{p,i}$, the error is calculated as

$$e_i = \min(|\theta_{t,i} - \theta_{p,i}|, 360^\circ - |\theta_{t,i} - \theta_{p,i}|).$$

The mean angular error is then defined as

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N e_i,$$

where N is the total number of samples. This metric is well-suited to regression settings because it provides an interpretable measure in degrees, where smaller values indicate higher accuracy. A random predictor uniformly distributed over $[0^\circ, 360^\circ)$ would achieve a MAE of 90° , therefore, any meaningful method must substantially improve on this bound.

4.2.2 Classification Accuracy

In the Direction-of-Voice (DoV) dataset, the task is formulated as an 8-class classification problem, with classes corresponding to azimuthal orientations of 0° , $\pm 45^\circ$, $\pm 90^\circ$, $\pm 135^\circ$, and 180° . In this case, the evaluation follows the standard definition of classification accuracy.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \times 100\%.$$

This metric allows direct comparison with the published baseline of [3] while also reflecting the intended use case of discrete angle detection in that dataset. Random guessing on 8 classes yields a prior probability of $1/8 = 12.5\%$, serving as a lower bound against which the methods can be compared.

4.3 Baselines

To contextualize the performance of our approach, we compare against three representative baselines that reflect the main categories of prior work: models based on raw multichannel audio, models based on hand-crafted acoustic features, and physics-informed propagation features.

4.3.1 Raw Audio + CNN

The first baseline is Soundr [1], which processes raw 16-channel audio through a convolutional-recurrent architecture (CNN+LSTM). In its original form, Soundr was trained on more than 700 minutes of real recordings and reported an average orientation error of 57° on unseen speakers and rooms. With speaker and room personalization, the performance further improved to 40.0° on unseen speakers in seen rooms and 34.3° on seen speakers in seen rooms. For consistency with our simulation setup, we re-implemented their convolutional encoder while keeping our recurrent and attention-based layers intact. The convolutional encoder accepts raw multi-channel audio samples as input and processes them through a stack of convolutional and max-pooling layers with decreasing kernel sizes and sequence lengths, while progressively increasing the number of feature channels. This hierarchical processing captures increasingly abstract temporal representations of the waveform and serves as the baseline for evaluating whether directly learning from raw waveforms provides sufficient orientation cues compared to phase features of the STFT.

4.3.2 ITD & ILD Features

The second baseline is the method of [19], which exploits interaural time differences (ITD) and interaural level differences (ILD). Beyond the classical binaural cues, their system also separates ITD and ILD into low- and high-frequency bands to address aliasing effects (high-frequency ITD) and to leverage the stronger directional patterns of high-frequency ILD. In addition, they introduce an *inverse convolution feature*, defined as the convolution of high-frequency ILD with the reciprocal of low-frequency ILD, which emphasizes frequency-dependent asymmetries in vocal radiation.

In our implementation, we extracted ITD_{low} , ITD_{high} , ILD_{low} , ILD_{high} , and the inverse convolution feature for each microphone pair in our array and passed them as parallel feature channels into our network.

4.3.3 Physics-Informed Features

The third baseline comes from the Direction-of-Voice (DoV) study [3], which developed a feature set based on acoustic propagation models. Features include:

- **Spectral balance features:** low- and high-frequency power, power ratios, and polynomial regression coefficients of the spectrum.
- **Wavefront crispness measures:** autocorrelation ratios, statistical descriptors, and speech-to-reverberation modulation energy ratios.
- **Cross-microphone correlation features:** raw GCC-PHAT correlations, peak values, interchannel delays, area under the curve, and time-difference-of-arrival statistics.

Since the DoV dataset is publicly available, we did not re-implement their method in our simulated setup. Instead, we report our results directly on their dataset using the evaluation protocol described in Section 4.1.2, enabling direct comparison with their published numbers.

4.4 Results

The results are presented in two parts. First, we present a detailed analysis of performance on the simulated dataset, covering reverberant and anechoic conditions, robustness under varying noise levels, and the benefits of personalization through speaker and room-specific adaptation. Then, we report results on a real-world dataset to assess how well the model generalizes beyond simulation.

4.4.1 Simulated data

The simulated dataset allows us to investigate three key research questions: (i) how reverberation affects orientation estimation accuracy, (ii) how robust the model is to background noise, and (iii) how much personalization through user- or room-specific adaptation can improve performance.

Anechoic and Reverberant Conditions

We begin by analyzing the effect of reverberation in the absence of noise. Table 4.1 compares anechoic conditions to reverberant clean rooms. ITD and ILD features achieve the lowest error in the anechoic setup (27.3°), greatly outperforming our STFT-phase approach (40.4°). In contrast, when reverberation is present, our method benefits substantially, reducing the mean angular error to 26.0° , significantly outperforming both the ITD/ILD baseline (39.7°) and the raw waveform model (44.8°). This indicates that the network is able to exploit consistent orientation-dependent reflections captured in the phase spectrum, while handcrafted binaural cues degrade.

Table 4.1. Comparison of orientation estimation methods in anechoic vs. reverberant clean environments, measured in MAE (degrees).

Method	Anechoic	Reverberant
Raw audio	56.9°	44.8°
ITD & ILD	27.3°	39.7°
STFT phase	40.4°	26.0°

Angular error analysis. To gain a clearer understanding of the performance characteristics, we investigate the distribution of errors. First, we plot the error distribution for our STFT-phase approach. As seen in Figure 4.1, the histogram under anechoic conditions exhibits a long tail with frequent large errors. Reverberation mitigates this by introducing informative reflections, which reduces high-error instances and consequently sharpens the distribution curve.

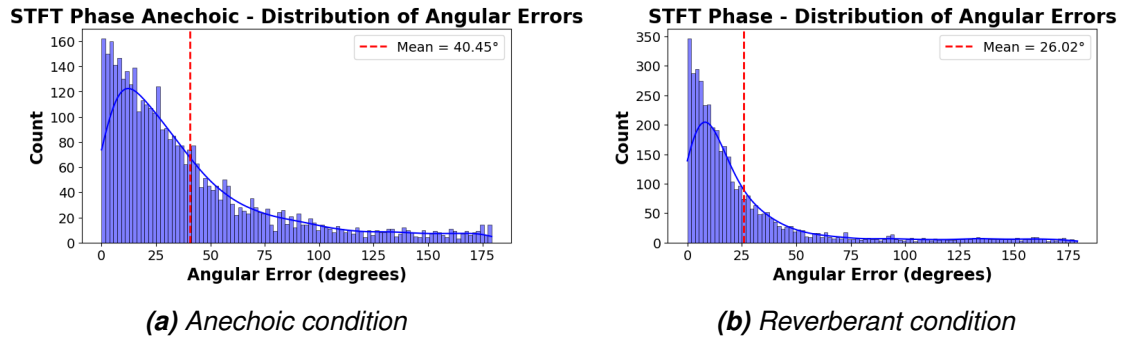


Figure 4.1. Comparison of angular error histograms for anechoic and reverberant conditions using the STFT-phase method.

Next, we investigate the error distribution across different angles to identify regions of highest confusion. Figure 4.2 shows the angular error aggregated over 10° bins with eight equal (45°) reference regions. The errors are smallest for orientations near 0° and 180° (front and back) and largest for side angles ($\pm 90^\circ$). This front–side asymmetry is particularly strong in the anechoic condition, where lateral cues are ambiguous. In reverberant rooms, the error distribution becomes more uniform, as reflections provide complementary cues for side orientations, though performance remains best at the front.

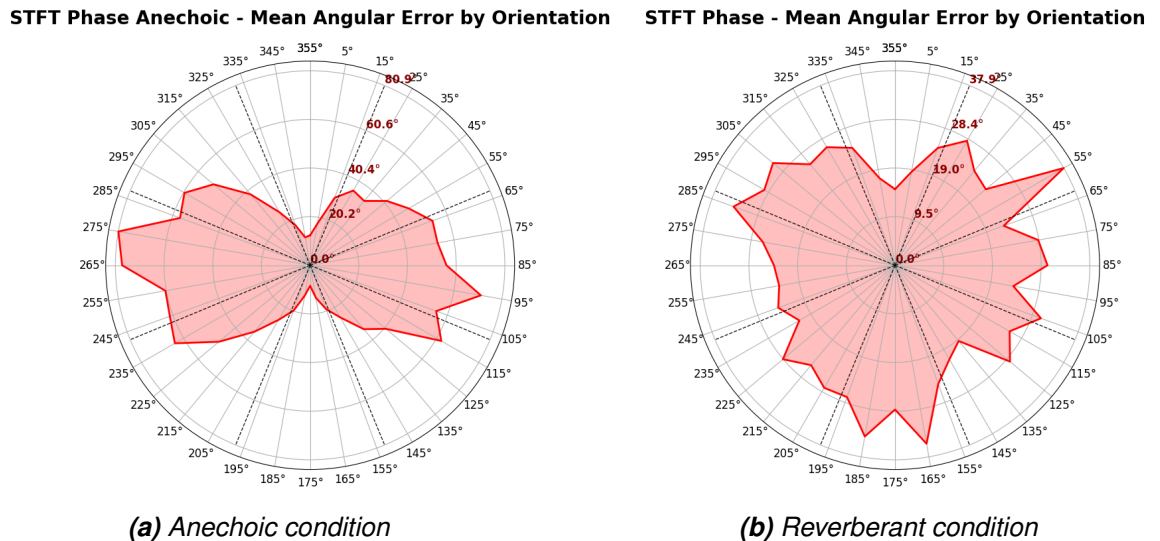


Figure 4.2. Spider charts of angular errors aggregated in 10° bins for anechoic and reverberant conditions using STFT-Phase.

Finally, we investigate the effect of distance in both anechoic and reverberant conditions for our STFT-Phase solution and the ITD&ILD approach. Figure 4.3 shows the angular error as a function of source-microphone distance; gray points represent individual localization errors, and the red line denotes the mean error trend computed in 0.5 m bins. In reverberant conditions, there is a clear positive correlation between angular error and distance: as the distance increases, the error also increases. In contrast, the angular error in anechoic conditions is more constant across distances, showing no such clear trend.

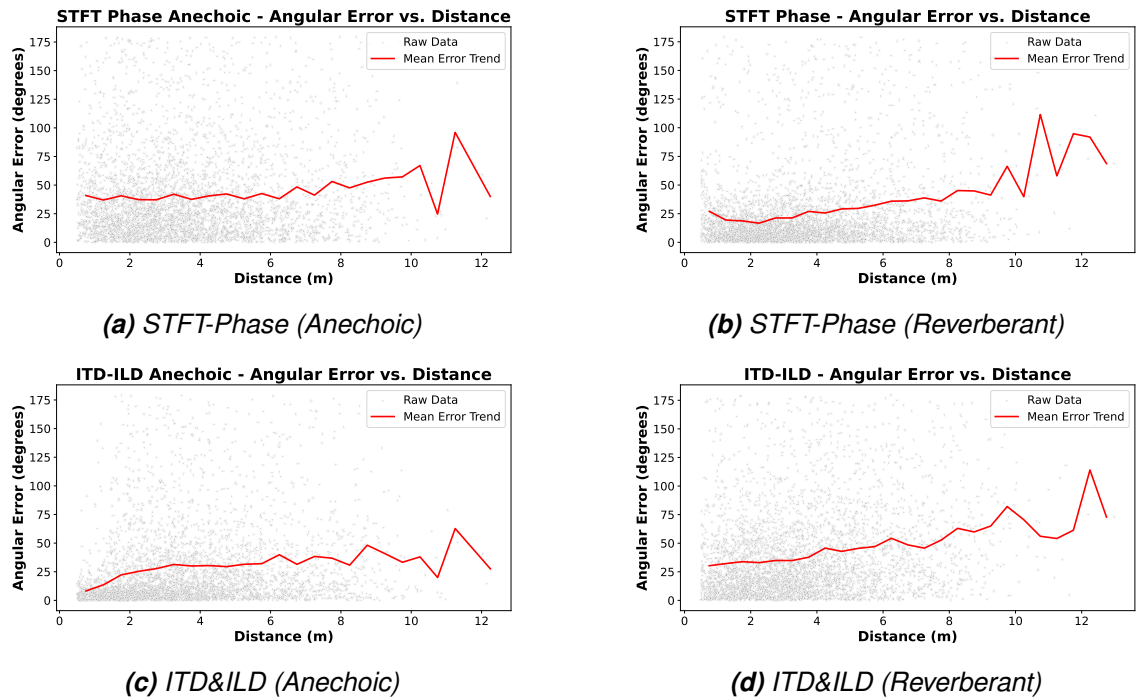


Figure 4.3. Angular error vs. distance for STFT-Phase and ITD&ILD methods under anechoic and reverberant conditions.

Noisy Conditions

We next evaluate robustness to background noise in reverberant rooms. Table 4.2 shows results under clean, moderate noise (10–20 dB SNR), and high noise (0–10 dB SNR) conditions. The signal-to-noise ratio (SNR) is defined as the ratio of signal power to noise power, expressed in decibels (dB), Figure 4.4 illustrates an example of a clean speech signal, background noise, and their mixture at 10 dB SNR.

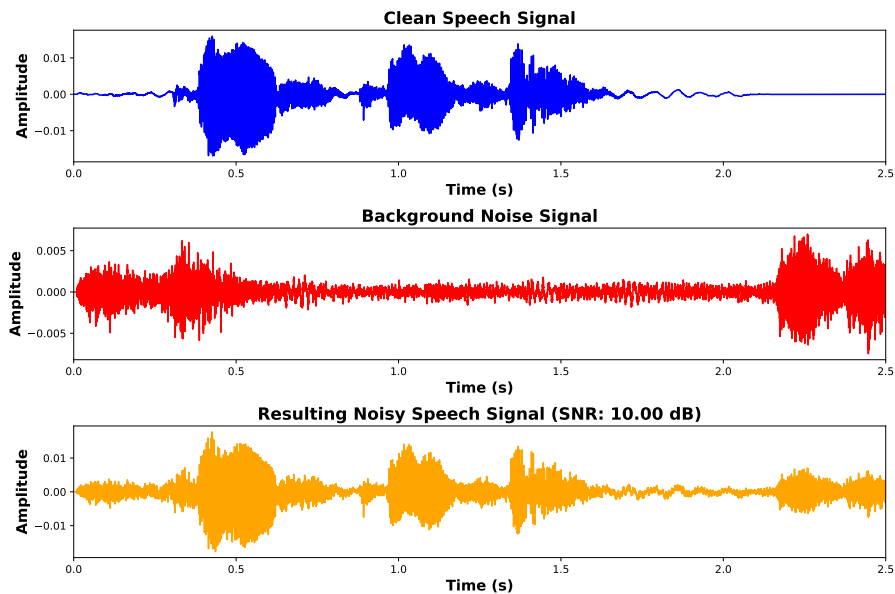


Figure 4.4. Example of a clean speech (top), background noise (middle), and noisy speech at 10 dB SNR (bottom).

Our STFT-phase method consistently outperforms the baselines in all conditions. Under moderate noise, the error increases only slightly to 35.4° , while the raw waveform approach degrades sharply to 63.7° . Even in the high-noise case, our method maintains a 48.5° error, well below the 90° random guessing level. In contrast, the raw audio method approaches near-random performance and the ITD/ILD also degrades substantially to 62.8° .

Table 4.2. Comparison of orientation estimation methods under reverberant noisy environments, measured in MAE (degrees). Noise levels indicate SNR: 10–20 dB (moderate) and 0–10 dB (high).

Method	Clean	Moderate noise (10–20 dB)	High noise (0–10 dB)
Raw audio	44.8°	63.7°	75.1°
ITD & ILD	39.7°	48.4°	62.8°
STFT phase	26.0°	35.4°	48.5°

Overall, these results highlight three key points. First, ITD/ILD features remain highly effective in idealized anechoic conditions, but degrade sharply once reverberation or noise is introduced. Second, raw waveform models are not good enough in either condition, indicating their inability to extract robust directional cues from limited training samples. Third, STFT phase features provide consistently strong performance in realistic environments, striking a balance between leveraging reflections and maintaining robustness to noise.

Personalization Experiments

In many real-world applications, users may be asked to provide a small number of samples to calibrate the model, since machine learning models are expected to perform better on matched data from the speakers and environments they were trained on. To investigate this effect, we evaluate our method under three personalization scenarios.

Experimental setup.

- **Room Only:** The model is fine-tuned using samples from the target environment but tested on unseen speakers. For this setup, we created ten fixed rooms and used 500 samples per room for fine-tuning, then evaluated on 400 samples per speaker from ten unseen speakers in these rooms.
- **Speaker Only:** The model is adapted using samples from the target speaker but tested in unseen rooms. To simulate this, we selected ten unseen speakers and used 150 samples per speaker for fine-tuning and 250 for testing.
- **Speaker + Room:** The model is fine-tuned using samples from both the target speaker and the target room, neither of which the model has encountered during pre-training. For this, we created a fixed room for each speaker and used 150 samples per speaker for fine-tuning and 250 for testing.

Fine-tuning was performed for 10 epochs at a reduced learning rate of 1×10^{-5} , updating all model weights. The remainder of the dataset was held out for testing.

This allows us to measure how much performance can be gained from personalization to the user and/or adaptation to the environment. As shown in Table 4.3, personalization consistently improves accuracy, with the largest gains observed when both user and room information are available. Notably, adapting to users provides a larger benefit than adapting to rooms, highlighting the strong variability in speech directivity between speakers. These findings demonstrate that even limited personalization data is sufficient to substantially improve accuracy, making adaptation a practical strategy for deployment.

Table 4.3. MAE across different personalization scenarios.

Prior knowledge	MAE
None (Baseline)	26.01°
Room Only	19.72°
Speaker Only	16.49°
Speaker + Room	13.91°

4.4.2 Real data

Finally, we validate our approach on the Direction-of-Voice (DoV) dataset [3]. Table 4.4 compares our approach against the DoV baseline. Training solely on real data yields weaker performance (59.8%), showing the lack of data to train such a complex network, but when pre-trained on large-scale simulated data and fine-tuned on real samples, our method achieves 71.1% accuracy, outperforming the reported DoV baseline (65.4%). This highlights the value of simulation-based pre-training, which provides a strong prior that transfers well to real environments despite differences in room acoustics, distances, and speaker variability.

Table 4.4. Classification accuracy on the DoV dataset (8 classes).

Method	Accuracy
DoV baseline [3]	65.4%
Ours (trained only on real data)	59.8%
Ours (pre-trained on simulated + fine-tuned on real)	71.1%

Figure 4.5 shows the confusion matrices for both [3] and our method. Similar to the baseline, most errors occur between adjacent orientation classes (e.g., 0° and 45°) rather than opposite ones (e.g., 90° and -90°), but our model exhibits a noticeably lower overall confusion rate. Over 97% of predictions fall within the correct or neighboring class, indicating that the model captures orientation relationships consistently and with minimal randomness.

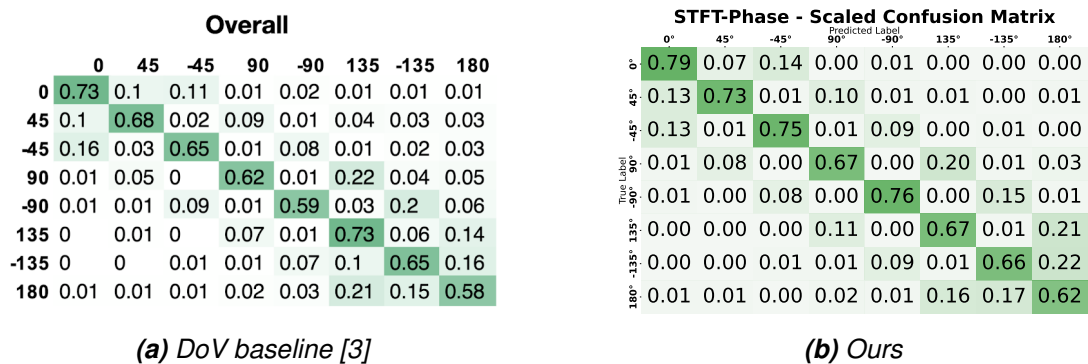


Figure 4.5. Confusion matrices on the DoV dataset [3].

The results confirm that while real data remains essential for adaptation, large-scale simulation provides a powerful foundation for robust orientation estimation. In practice, this combination allows developers to reduce reliance on costly annotated real-world recordings while still achieving great performance.

4.5 Discussion

The experimental results highlight several key insights.

First, the strong performance on both simulated and real datasets demonstrates that STFT phase features are superior to raw audio and hand-crafted inter-channel features, offering greater robustness to noise where other baseline methods degrade substantially. We hypothesize that the phase provides reliable cues for the relative rotation between the source and microphone array, including echoes from image sources, which the model can exploit. Similar prominence of phase features was observed in DNN-based speaker distance estimation with a single microphone [37], though further experiments are needed to validate these hypotheses.

Second, personalization experiments confirm that user-specific information contributes more to accuracy improvements than environment-specific information. Variations in vocal tract shape, articulation style, and speech radiation patterns differ significantly between individuals, and adapting to these variations through fine-tuning can greatly reduce estimation error. Similarly, acoustic properties of the recording environment, such as reverberation time and microphone placement, influence phase patterns and can benefit from environment-specific calibration.

Third, the evaluation on the real dataset from [3] further validates the generalization capability of our method and the effectiveness of pre-training on large-scale simulated data, providing a practical way to mitigate limited annotated real-world data.

Overall, these findings demonstrate that the use of STFT phase-based features within a deep neural architecture offers a promising and scalable path toward accurate and noise-robust speaker head orientation estimation using only a single compact microphone array, without requiring camera input or dense microphone configurations.

5. CONCLUSIONS

This thesis presented a novel approach to estimating speaker head orientation from multichannel audio recorded by a single microphone array. The proposed method departs from traditional handcrafted features based on sound source-to-receiver propagation models or raw audio-based approaches by leveraging phase information extracted from the short-time Fourier transform. Through extensive experiments on both simulated and real datasets, the work demonstrated that the phase domain provides a rich and robust representation for capturing spatial and directional cues inherent to speech, enabling accurate orientation estimation under a wide variety of acoustic conditions.

Key Contributions

The main contributions of this work can be summarized as follows:

- We introduced a new representation for audio-based orientation estimation that utilizes the sine and cosine of STFT phase components as model input. This representation preserves the circular structure of the phase while effectively encoding inter-channel relationships that are essential for spatial reasoning.
- We utilized a deep neural architecture combining convolutional, recurrent, and self-attention mechanisms to predict head orientation. The proposed network achieved strong performance while maintaining a compact model size.
- We constructed a large-scale simulated dataset that integrates measured voice directivity patterns, room acoustic simulations, and additive background noise. This dataset enabled training and pre-training of deep models without reliance on large annotated real recordings.
- We demonstrated that pre-training on simulated data, followed by fine-tuning on a small amount of real-world data, yields substantial improvements in generalization and accuracy. The combination of simulation-based pre-training and real-world adaptation provides a practical path for deploying orientation estimation systems in scenarios where labeled data are limited.
- We evaluated the proposed approach across a range of conditions, including different levels of reverberation, noise, and environments, and compared it to state-

of-the-art baselines based on raw waveform and interaural features. The results showed that the proposed phase-based model achieves lower mean angular errors and higher robustness to noise and reverberation than the other methods.

- Finally, we investigated user- and environment-specific fine-tuning and showed that even a small amount of adaptation data can significantly improve accuracy. This highlights the practical benefits of personalization and suggests that model calibration could be an effective strategy for real-world deployment.

Main Findings

The experimental results presented in this thesis allow several conclusions to be drawn about the proposed approach. The evaluations showed that the use of STFT phase features in combination with a deep neural architecture provides a robust and accurate framework for estimating speaker head orientation from audio. Across all conditions, the phase-based model consistently outperformed methods relying on raw waveform input or traditional interaural features. In reverberant environments, the proposed system achieved a mean angular error of 26.0° , compared to 39.7° for ITD/ILD-based features and 44.8° for raw audio. These results confirm that the phase domain encodes stable and informative spatial cues that remain reliable even when it is distorted by reflections and background noise. The model maintained strong performance under noisy conditions, with moderate degradation at low signal-to-noise ratios, indicating that it generalizes well to realistic acoustic scenarios.

The experiments further demonstrated the effectiveness of combining simulated and real data for training. Pre-training on a large simulated dataset followed by fine-tuning on a limited amount of real-world recordings significantly improved generalization and accuracy, increasing classification performance on the DoV dataset from 59.8% to 71.1%. This shows that simulation can effectively bridge the gap between controlled and real conditions, providing a practical solution when annotated data are scarce. In addition, user- and environment-specific fine-tuning further reduced the mean angular error to as low as 13.9° , highlighting that adaptation can yield significant improvements and that personalized calibration is a viable strategy for deployment.

Limitations and Future Work

Despite the promising results, several limitations remain. The proposed approach currently assumes static head orientation within each utterance, whereas in real scenarios, speakers may move dynamically while speaking. Extending the system to handle continuous orientation tracking over time would make it more suitable for real-time interactive applications.

Another limitation lies in the generalization to unseen users and environments. Although pre-training on simulated data provides a strong prior, the performance gap observed between generic and fine-tuned models indicates that the model still benefits significantly from adaptation. This suggests that larger and more diverse datasets are required to achieve robust zero-shot generalization.

Finally, the current method focuses on the azimuthal plane, assuming a fixed elevation. Extending the model to three-dimensional orientation estimation could provide a more complete spatial understanding.

Final Remarks

In conclusion, this thesis demonstrated that phase spectrogram features, when combined with a deep neural architecture, provide a powerful and noise-robust framework for estimating speaker head orientation from audio alone.

REFERENCES

- [1] J. Yang, G. Banerjee, V. Gupta, M. S. Lam, and J. A. Landay. “Soundr: Head position and orientation prediction using a microphone array”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–12.
- [2] Q. Yang and Y. Zheng. “Model-based head orientation estimation for smart devices”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.3 (2021), pp. 1–24.
- [3] K. Ahuja, A. Kong, M. Goel, and C. Harrison. “Direction-of-voice (dov) estimation for intuitive speech interaction with smart devices ecosystems”. In: *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 2020, pp. 1121–1131.
- [4] S. O. Ba and J.-M. Odobez. “A study on visual focus of attention recognition from head pose in a meeting room”. In: *International Workshop on Machine Learning for Multimodal Interaction*. Springer. 2006, pp. 75–87.
- [5] R. Stiefelhagen and J. Zhu. “Head orientation and gaze direction in meetings”. In: *CHI’02 Extended Abstracts on Human Factors in Computing Systems*. 2002, pp. 858–859.
- [6] Y. Zhao, L. Görne, I.-M. Yuen, D. Cao, M. Sullman, D. Auger, C. Lv, H. Wang, R. Matthias, L. Skrypchuk, and A. Mouzakitis. “An Orientation Sensor-Based Head Tracking System for Driver Behaviour Monitoring”. In: *Sensors* 17.11 (2017). ISSN: 1424-8220. DOI: 10.3390/s17112692.
- [7] S. Jha and C. Busso. “Estimation of driver’s gaze region from head position and orientation using probabilistic confidence regions”. In: *IEEE Transactions on Intelligent Vehicles* 8.1 (2022), pp. 59–72.
- [8] A. Asperti and D. Filippini. “Deep learning for head pose estimation: A survey”. In: *SN Computer Science* 4.4 (2023), p. 349.
- [9] J. M. Sachar and H. F. Silverman. “A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array”. In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 4. IEEE. 2004, pp. iv–iv.
- [10] A. Levi and H. Silverman. “A robust method to extract talker azimuth orientation using a large-aperture microphone array”. In: *IEEE transactions on audio, speech, and language processing* 18.2 (2009), pp. 277–285.
- [11] H. Nakajima, K. Kikuchi, T. Daigo, Y. Kaneda, K. Nakadai, and Y. Hasegawa. “Real-time sound source orientation estimation using a 96 channel microphone array”. In:

- 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2009, pp. 676–683.
- [12] A. Abad, C. Segura, D. Macho, J. Hernando, and C. Nadeu. “Audio person tracking in a smart-room environment”. In: *Interspeech*. 2006. DOI: 10.21437/Interspeech.2006-649.
- [13] C. Segura and J. Hernando. “GCC-PHAT based head orientation estimation”. In: *Interspeech*. 2012.
- [14] C. Segura, A. Abad, J. Hernando, and C. Nadeu. “Speaker orientation estimation based on hybridation of GCC-PHAT and HLBR”. In: *Interspeech*. 2008.
- [15] C. Segura, C. Canton-Ferrer, A. Abad, J. R. Casas, and J. Hernando. “Multimodal head orientation towards attention tracking in smartrooms”. In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP’07*. Vol. 2. IEEE. 2007, pp. II–681.
- [16] R. C. Felsheim, A. Brendel, P. A. Naylor, and W. Kellermann. “Head orientation estimation from multiple microphone arrays”. In: *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE. 2021, pp. 491–495.
- [17] R. Takashima, T. Takiguchi, and Y. Ariki. “Single-channel head orientation estimation based on discrimination of acoustic transfer function”. In: *Interspeech*. 2011. DOI: 10.21437/Interspeech.2011-147.
- [18] R. Takashima, T. Takiguchi, and Y. Ariki. “Estimation of talker’s head orientation based on discrimination of the shape of cross-power spectrum phase coefficients”. In: *Interspeech*. 2012. DOI: 10.21437/Interspeech.2012-403.
- [19] H. Takawale and N. Roy. “Learning speaker-listener mutual head orientation by leveraging hrtf and voice directivity on headphones”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 1171–1175.
- [20] J. W. Cooley and J. W. Tukey. “An algorithm for the machine calculation of complex Fourier series”. In: *Mathematics of Computation* 19.90 (1965), pp. 297–301. DOI: 10.1090/S0025-5718-1965-0178586-1.
- [21] J. Allen. “Short term spectral analysis, synthesis, and modification by discrete Fourier transform”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25.3 (1977), pp. 235–238. DOI: 10.1109/TASSP.1977.1162950.
- [22] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, Oct. 1996. ISBN: 9780262268684. DOI: 10.7551/mitpress/6391.001.0001.
- [23] B. B. Monson, E. J. Hunter, and B. H. Story. “Horizontal directivity of low- and high-frequency energy in speech and singing”. In: *The Journal of the Acoustical Society of America* 132.1 (July 2012), pp. 433–441. ISSN: 0001-4966. DOI: 10.1121/1.4725963. eprint: https://pubs.aip.org/asa/jasa/article-pdf/132/1/433/15298379/433_1_online.pdf.

- [24] J. B. Allen and D. A. Berkley. “Image method for efficiently simulating small-room acoustics”. In: *The Journal of the Acoustical Society of America* 65.4 (Apr. 1979), pp. 943–950. ISSN: 0001-4966. DOI: 10.1121/1.382599. eprint: https://pubs.aip.org/asa/jasa/article-pdf/65/4/943/11426543/943_1_online.pdf.
- [25] H. Kuttruff. *Room Acoustics*. 6th. Boca Raton: CRC Press, 2016, p. 322. ISBN: 9781315372150. DOI: 10.1201/9781315372150.
- [26] R. Scheibler, E. Bezzam, and I. Dokmanic. “Pyroomacoustics: A python package for audio room simulation and array processing algorithms”. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 351–355.
- [27] M. Brandstein and D. Ward. *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2001.
- [28] C. Knapp and G. Carter. “The generalized correlation method for estimation of time delay”. In: *IEEE transactions on acoustics, speech, and signal processing* 24.4 (2003), pp. 320–327.
- [29] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.
- [30] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179.
- [31] M. Schuster and K.K. Paliwal. “Bidirectional recurrent neural networks”. In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681. DOI: 10.1109/78.650093.
- [32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.
- [33] H. Robbins and S. Monro. “A Stochastic Approximation Method”. In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407. ISSN: 00034851.
- [34] D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In: *ICLR*. 2015.
- [35] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill. “Pyannote.Audio: Neural Building Blocks for Speaker Diarization”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020. DOI: 10.1109/ICASSP40776.2020.9052974.

- [36] T. Peer and T. Gerkmann. “Phase-aware deep speech enhancement: It’s all about the frame length”. In: *JASA Express Letters* 2.10 (2022).
- [37] M. Neri, A. Politis, D.A. Krause, M. Carli, and T. Virtanen. “Speaker distance estimation in enclosures from single-channel audio”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), pp. 2242–2254.
- [38] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen. “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks”. In: *IEEE Journal of Selected Topics in Signal Processing* 13.1 (2018), pp. 34–48.
- [39] P. Sudarsanam, A. Politis, and K. Drossos. “Assessment of Self-Attention on Learned Features For Sound Event Localization and Detection”. In: *Detection and Classification of Acoustic Scenes and Events (DCASE)*. 2021.
- [40] J. Yamagishi, C. Veaux, and K. MacDonald. *CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit*. Version 0.92. 2019. DOI: 10.7488/ds/2645.
- [41] M. Brandner, M. Frank, and D. Rudrich. “DirPat—Database and viewer of 2D/3D directivity patterns of sound sources and receivers”. In: *Audio Engineering Society Convention 144*. Audio Engineering Society. 2018.
- [42] N. R. Shabtai, G. Behler, M. Vorländer, and S. Weinzierl. “Generation and analysis of an acoustic radiation pattern database for forty-one musical instruments”. In: *The Journal of the Acoustical Society of America* 141.2 (2017), pp. 1246–1256.
- [43] D. Cabrera, P. J. Davis, and A. Connolly. “Long-Term Horizontal Vocal Directivity of Opera Singers: Effects of Singing Projection and Acoustic Environment”. In: *Journal of Voice* 25.6 (Nov. 2011), e291–e303. DOI: 10.1016/j.jvoice.2010.03.001.
- [44] Jens Ahrens. *Database of Spherical Harmonic Representations of Sound Source Directivities*. Version 2020-03-12. Zenodo, Mar. 2020. DOI: 10.5281/zenodo.3707708.
- [45] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux. “WHAM!: Extending Speech Separation to Noisy Environments”. In: *Interspeech*. Sept. 2019.
- [46] L. McCormack, A. Politis, and V. Pulkki. “Rendering of source spread for arbitrary playback setups based on spatial covariance matching”. In: *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2021.
- [47] D. Mirabilii, S. J. Schlecht, and E.A.P. Habets. “Generating coherence-constrained multisensor signals using balanced mixing and spectrally smooth filters”. In: *The Journal of the Acoustical Society of America* 149.3 (2021), pp. 1425–1433.