Reichman University

Efi Arazi School of Computer Science

M.Sc. Program – Research Track

Binaural sound source localization using a hybrid time and frequency domain model

Gil Geva

M.Sc. dissertation, submitted in partial fulfillment of the requirements for the M.Sc. degree, research track, School of Computer Science Reichman University (The Interdisciplinary Center, Herzliya)

December 2023

Supervision

I would like to acknowledge my supervisors, Prof. Yacov Hel-Or from the Efi Arazi School of Computer Science and Prof. Amir Amedi from the Baruch Ivcher Institute for Brain, Cognition & Technology, both at Reichman University, Prof. Shlomo Dubnov, Director of the Center for Research and Learning at UC San Diego, Olivier Warusfel, head researcher at the Acoustic and Cognitive Spaces, and Tammuz Dubnov, for their guidance and support throughout this research project.

Collaboration

This research was carried out in collaboration with Ircam (Institute for Research and Coordination of Acoustic Music) STMS Lab (Sciences and Technologies of Music and Sound), Music Representation Team, ERC REACH Project.

Acknowledgements

I would like to express my deep gratitude to PI Gérard Assayag for his help in conducting the research.

Personal note

I wish to express my sincere gratitude and great appreciation to my beloved partner, Noam Rahav. Her endless support, encouragement, and understanding have been essential throughout this academic journey, particularly during the research expedition to Paris. Furthermore, I am deeply thankful for her invaluable contributions in reviewing and refining this thesis, harnessing her academic experience and clarity.

Funding

This research was funded by the European Research Council under the Horizon 2020 programme (Grant 883313 ERC REACH).

Abstract

Sound source localization plays a foundational role in auditory perception, enabling both human and machines to determine the sound source location. Traditional sound localization methods often rely on manually crafted features and simplified conditions, which limit their applicability in real-world situations.

Accurate sound localization holds vital importance across diverse applications, spanning robotics, virtual reality, human-computer interactions, and medical devices. This significance is particularly amplified for individuals with cochlear implants (CI), who confront significant challenges in perceiving the direction of sound sources.

Previous research focused on extensive microphone arrays in the frontal plane, which exhibit accuracy and robustness limitations when employing small microphone arrays. These sound localization techniques are also impractical for CI users due to size and weight constraints, and the need for full-sphere localization capabilities.

This research introduces a new approach to sound source localization using head-related transfer function (HRTF) characteristics, from raw data, in both the time and frequency domains. Furthermore, it advances binaural sound localization by extending its capabilities from a 180-degree range to a full-sphere context.

The proposed approach introduces an end-to-end Deep-Learning (DL) hybrid model, that integrates spectrogram and temporal domain insights via parallel channels. The performance of our proposed hybrid model, surpasses the current state-of-the-art results. Specifically, it boasts an average angular error of 0.24° and an average Euclidean distance of 0.01 meters, while the known state-of-the-art gives average angular error of 19.07° and average Euclidean distance of 1.08 meters.

This level of accuracy is of paramount importance for a wide range of applications, including robotics, virtual reality, and aiding individuals with CI.

In conclusion, as the field of sound source localization continues to progress, this research contributes to a deeper understanding of auditory perception and offers practical applications within healthcare scenarios.

Contents

Chapter 1

Introduction

Sound localization, the ability to determine a sound's three- dimensional position, is vital in applications such as robotics, virtual reality, and human - computer interactions. It is especially crucial for CI users, who often struggle with accurate sound direction estimation compared to those with normal hearing.

While those with normal hearing exhibit direction estimation errors ranging between 4.1° and 17.8°, CI users encounter notably larger discrepancies. The average error for CI users stands at 43.5° on the CI side and 60° on the opposite side [19]. The precision in determining direction of arrival (DOA) holds significant implications for CI users, both in terms of safety considerations and their ability to navigate complex listening situations as in the cocktail party scenario.

1.1 Head related transfer function

Head related transfer function (HRTF) describes how sound signals are transformed by the human head and ears before it reaches the eardrums. HRTF is a frequency domain representation widely used in spatial sound perception. It refers to the intricate filtering effects imposed by the morphology of the human head and ears upon sound waves as they moves toward their final destination.

Integral to this, Head-Related Impulse Response (HRIR) constitutes a time-domain representation of the transformations a sound impulse undergoes when traveling from a designated direction in space to a listener's ears. The HRIR is typically derived through the inverse Laplace Transform of the transfer function, followed by convolving HRIRs with impulse signals. This time-domain response encapsulates both the filtering consequences and the initial signal modifications, influenced by the listener's head and ear structure. Ultimately, this process recreates the original signal as if it were played from a certain direction on the individual's head.

Binaural signals are generated as sound waves travel from a source to the ears, interacting with the head's geometry. Two key binaural signals, denoted as $Y_l(t)$ and $Y_r(t)$, correspond to the left and right ears, respectively. These signals are influenced by the source location and noise. The mathematical representations of these signals in the time domain are given by:

$$Y_l(t) = H_l((x_s, y_s, z_s), S(t)) + H_l(N_l(t))$$
$$Y_r(t) = H_r((x_s, y_s, z_s), S(t)) + H_l(N_r(t))$$

Here, H_l and H_r are the HRIR describing the interaction of the sound with the head for left and right ears, respectively. $S(t)$ represents the source signal, while $N_l(t)$ and $N_r(t)$ are noise signals for the left and right ears.

The spectrogram representations of these signals are derived using short-time Fourier transform (STFT) with frequency index k and time index n, creating $Y_{spectrogram,left}(k,n)$ and $Y_{spectrogram,right}(k,n)$.

$$Y_{spectrogram,left}(k,n) = H_l((x_s, y_s, z_s), S(k,n)) + H_l(k, N_l(n))$$
$$Y_{spectrogram,right}(k,n) = H_r((x_s, y_s, z_s), S(k,n)) + H_l(k, N_r(n)$$

1.2 Sound localization calculation

The main focus of binaural sound localization research is on the frontal horizontal plane. This area is preferred because of the fundamental challenges associated with crafting features for achieving a comprehensive 360-degree prediction, a task that is heavily dependant on individual variations in head size and shape.

The primary technique involves computing the interaural time difference (ITD) through the maximal cross-correlation between two recordings:

$$\tau = \underset{x}{\operatorname{argmax}}(corr(Y_l(t), Y_r(t+x))$$

The interaural phase difference (IPD) for each frequency can subsequently be deduced utilizing τ:

$$\phi_f = \Phi_f(t) - \Phi_f(t+\tau)$$

Where ϕ represent the phase difference and Φ represent the phase.

This approach empowers the accurate estimation of sound source location under frontal horizontal plane assumption. If, however, one is unable to make this assumption, we are left without information about sound source elevation and are unable to discern whether it is positioned in the front or rear. In reality, the sound source can emanate from any point within a cone centered along the line connecting the two sensors and extending vertically from this line (Figure 3.1). This phenomenon is known as the "cone of confusion" or "front-back confusion".

An additional source of information that can enhance the precision of sound source localization is the interaural level difference (ILD). ILD can be utilized in two different ways. The first, and simpler, approach involves dividing the magnitudes of the two recordings to obtain the same signal output:

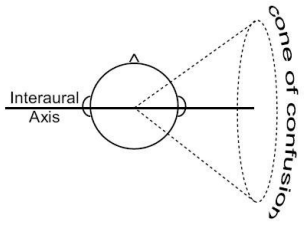


Figure 1.1: Cone of confusion

$$ILD(t) = \frac{Y_l(t)}{Y_r(t+\tau)}$$

The insight derived from this calculation is limited, as it only provides insight into the absolute direction of the sound (left or right), primarily determined by the time delay in most cases.

A more informative form of ILD considers the disparities at each frequency range, k:

$$ILD(k,n) = \frac{Y_{sl}(k,n)}{Y_{sr}(k,n+\tau)}$$

Predicting the location of a sound source using frequency-level ILD proves to be a challenging task without the application of ML techniques, because it relies heavily on specific configurations and the shape and size of the head.

1.3 Duplex theory

The duplex theory encapsulates a fundamental trade-off between the contributions of IPD and ILD to sound source localization across various frequencies. This trade-off is enabled because of the distinct characteristics of these cues and their relevance to different frequency ranges.

In lower frequencies, the spectral disparities between the ears tend to be minimal. This phenomenon is attributed to wave diffraction, which is influenced by the wavelength of sound waves. Considering room temperature, speed of sound of $c \sim 340\frac{m}{s}$, and a head width of 0.25 meters, sound waves with wavelengths greater than 0.25 meters should exhibit more pronounced diffraction. This corresponds to frequencies below 1,360 Hz.

Conversely, frequencies above this threshold are anticipated to showcase more significant differences in spectral ILD.

In contrast, the upper frequency limit at which two ears separated by 0.25 meters can detect phase differences is around 680 Hz. This is due to the fact that a phase difference of 180 degrees corresponds to half a wavelength. Beyond 180 degrees, ambiguity arises when trying to discern which ear's wave arrived first.

$$f_{max} = \frac{c}{2 * \text{distance between ears}} = \frac{340}{2 * 0.25} = \frac{340}{0.5} = 680Hz$$

This implies that if the frequency surpasses 680 Hz, it becomes practically impossible to determine which ear's phase delay corresponds to which or whether the delays stem from the same wave.

Hence, from a theoretical standpoint, IPD holds greater significance for lower frequencies, while spectral ILD proves more valuable for higher frequencies.

1.4 Research thesis

Traditional sound localization methods, such as beamforming and the Multiple Signal Classification (MUSIC) algorithm, face two key challenges when trying to help CI users. Firstly, there is a notable decrease in performance with small microphone arrays, and large arrays are not feasible for CI users. Secondly, while traditional approaches to binaural sound localization may perform well within the frontal sphere, their ability to accurately determine the localization of a sound source across the entire sphere, without front-back confusion, is limited.

DL has shown promise in improving the accuracy of sound localization by learning complex features from microphone signals. A significant portion of the research in this field has focused on using big arrays of microphones, leading to significant advancements. However, these advancements were primarily focused on large microphone arrays situated in the frontal sphere.

Harnessing the attributes of HRTF for sound localization holds potential for significantly enhancing complete sphere sound localization, even within the constraints of a compact microphone array. The intricate nature of modeling HRTF, influenced by the unique variations in individuals' head and ear anatomy, emphasizes its complexity. Hence, we propose to leverage DL to address this challenge.

This research seeks a way to integrate DL techniques into sound localization utilizing a two-microphone setup and the HRTF. This work proposes a model capable of accurately learning sound source localization from binaural recordings, especially for limited microphones or wearable devices such as hearing aids and CIs.

Chapter 2

Related work

2.1 Traditional sound localization and HRTF research

The phenomenon of sound localization has been extensively studied in the field of auditory perception. Traditional research on sound localization focused on the role of interaural time differences (ITDs), interaural level differences (ILDs), and interaural phase differences (IPDs) in sound localization. Pioneering works dating back to 1907 by Lord Rayleigh [1] and Wallach [2] established the foundational importance of these parameters in the context of sound localization.

Later investigations, as exemplified by Middlebrooks' work in 1991 [3], extended this understanding to encompass the significance of spectral cues in localization. This work highlighted the importance of the spectral shape of sound as it reaches each ear, particularly the modifying effects of the pinna, in accurately localizing the elevation of sound sources. These investigations laid the groundwork for considering both the anatomical and acoustic attributes of the outer ear in the realm of sound localization studies.

2.2 Machine Learning Methods for Sound Localization

In recent years, the field of sound localization has seen significant developments with a focus on the role of HRTF. Wightman and Kistler [4] used principal component analysis (PCA) to customize HRTF for individual listeners, based on their specific anatomical characteristics. Talagala and Thushara [5] introduced a novel approach by combining various localization cues, including both time and phase differences, with spectral interaural differences.

ML has played a crucial role in enhancing sound localization accuracy. Early algorithms such as Gaussian mixture models (GMM) and expectation–maximization (EM) were employed [6] [7]. However, the emergence of deep neural networks (DNNs) opened up many new options.

Tsuzuki et al [8] used DL for sound localization via multilayer-perceptrons to estimate sound source localization from time delay and amplitude differences. Takeda et al. [11]

exploited DNNs using phase information. Hirvonen et al. [9] were the first to use feature extraction from spectrograms.

Recent years have witnessed increased interest in leveraging DL for sound localization [20] [22], particularly in conjunction with large microphone arrays [14]. Efficient techniques for using limited microphone setups for a full sphere localization, however, are essential.

2.3 Prominent sound localization challenges

The field of sound localization has been enriched by some significant challenges, organized by the IEEE. These provide valuable platforms for driving and showcasing advancements in sound localization and its associated domains. Two of the notable challenges are:

1. Scoustic source LOCalization And TrAcking (LOCATA) Challenge [17]: The LOCATA Challenge was designed to address sound localization and tracking algorithms, encompassing a range of distinct challenges, each with its own specific objectives. This challenge involved the utilization of multiple microphone arrays, each catering to a different experimental setup. The arrays were:
    - A planar array equipped with 15 microphones, featuring linear uniform subarrays.
    - A spherical Eigenmike array with 32 microphones.
    - A pseudo-spherical array consisting of 12 microphones mounted on a humanoid robot.
    - Hearing aids integrated into a head-torso simulator.
2. DCASE Challenge [21]: The DCASE Challenge encompassed a diverse range of challenges, addressing distinct aspects of sound classification and analysis. For this particular challenge, a spherical Eigenmike array comprising 64 microphones was employed.

It is noteworthy that in the LOCATA challenge, only one microphone array configuration was binaural recordings. Interestingly, only one team chose to publish their work for this particular challenge [13].

2.4 Advancements in audio processing techniques

The concept of Audio-Visual Correspondence, as exemplified by "Look, Listen, and Learn" [12], showcases the potential of combining audio and visual information to acquire semantic knowledge. In this study, the authors effectively processed each data modality individually, employing convolutional techniques, and subsequently combining the embeddings using dense layers (Figure 2.1). This approach had a distinct impact on audio processing research.

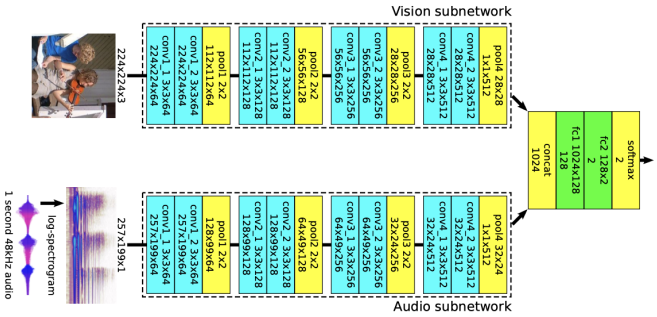


Figure 2.1: Look-Listen-Learn architecture

The Hybrid Spectrogram and Waveform Source Separation method, introduced by Facebook AI Research [18], extends the U-Net model to handle both time and frequency domains in separating sources from hybrid waveform and spectrogram data (Figure 2.2).

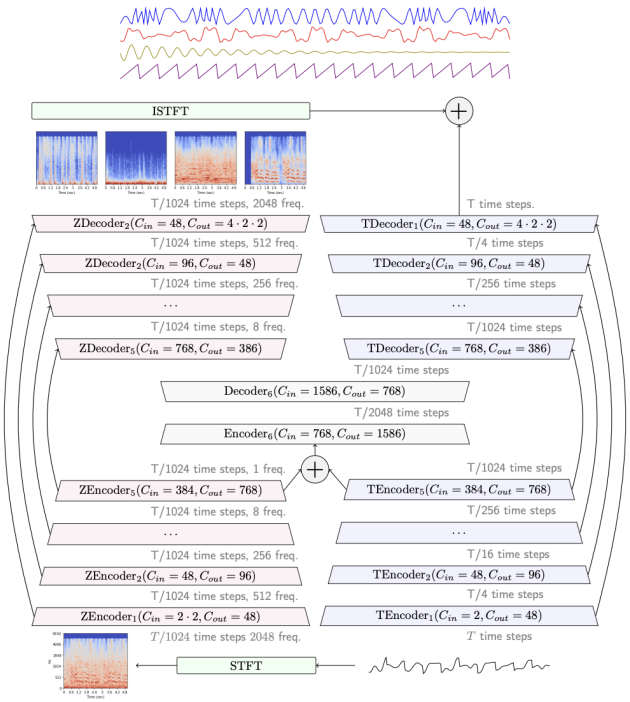


Figure 2.2: Demucs architecture

2.5 Benchmark model comparison

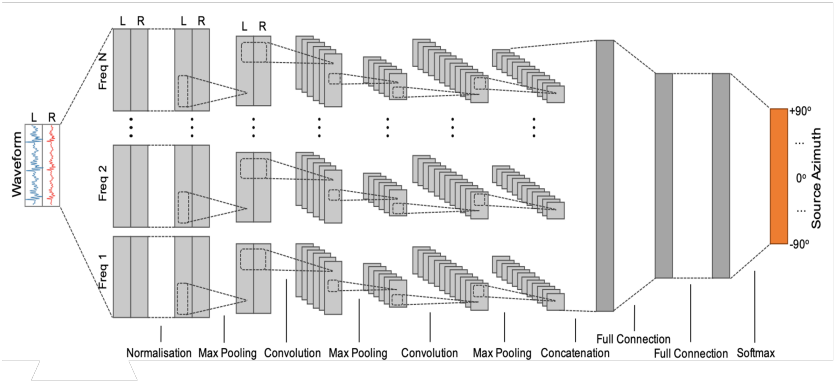


Figure 2.3: Benchmark model architecture

Vecchiotti et al. [16] introduced an influential method using Convolutional neural network (CNN) to analyze waveform data directly for sound localization, diverging from traditional feature engineering. This innovative approach marks a change from the conventional practice of relying on manually engineered features. We use this work as a benchmark model for our comparison. Their model excelled in classifying sound sources locations within a frontal 180-degree range among 37 speakers. We retrained it using our dataset and evaluated it using the same metrics.

In their study, the authors presented two distinctive models: one model is inspired by the auditory mechanisms and incorporates a gammatone filtering layer, while the other model is data-centric, utilizing trainable convolutional layers to analyze frequency patterns. The benchmark's model architecture includes a convolutional layer with linear activation function for frequency analysis, applied separately on both recordings. Subsequent layers process both recordings, followed by two fully-connected hidden layers. The output layer uses softmax activation for sound localization classification, with training guided by the root mean square error (RMSE) loss function.

Chapter 3

Methodology

3.1 Approach to Addressing the Challenge

We adopted a hybrid approach that combines time and frequency domain data. Waveform data inherently contain temporal, spectral, and phase details. Incorporating spectrogram information, however, has been shown to improve performance. Recent advances in DL, particularly in audio processing, have emphasized the utility of end-to-end systems for diverse applications. These frameworks excel when dealing with tasks requiring the integration of multiple elements. Hence, we embraced an end-to-end hybrid model that leverages the strengths of both waveform and spectrogram representations. We believe that our model maximizes the benefits of recent advancements.

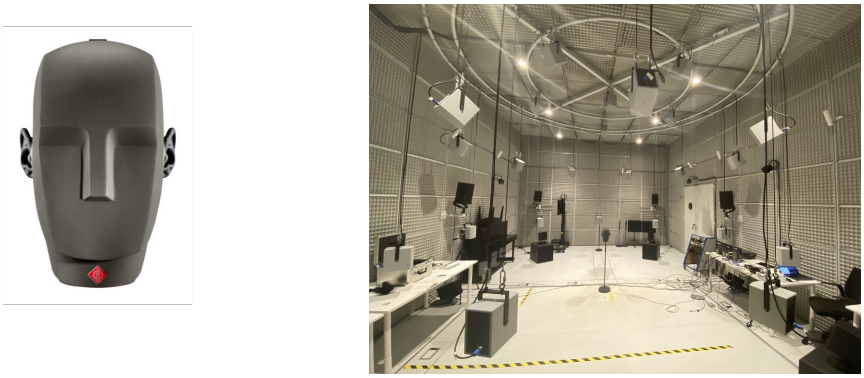
3.2 Data collection and modeling

Data collection was conducted using two different setups. The initial one employed 3DIO binaural microphones with a custom-designed pinna for capturing audio. Recordings were conducted in a studio at Reichman University, where an arrangement of 72 speakers spanned across three elevation levels. The data acquisition included playing a range of everyday sounds and "Sweeps" covering frequencies from 0 to 20,000 Hz, emanating from various speakers in each instance. Subsequently, the obtained stereo recordings were divided into short segments.

Figure 3.1: Reichman recording room and 3DIO FS XLR microphone

The acquired data yielded favorable outcomes for frequencies exceeding 1 KHz. Nevertheless, the model faced limitations in its ability to predict lower frequencies, mainly due to the "ear's" undersized structure impeding sound wave propagation. Additionally, the substantial differences between our experimental setup and the configuration of an actual human head were evident. Given our aspiration for this research to yield practical implications, it became crucial to minimize the disparities and closely emulate real-world situations.

Hence, the second and final data collecting setup occurred at the IRACM studio in Paris, using the KU 100 dummy head microphone system, which provided an augmented sense of realism in contrast to the pinna microphones. This dummy head system significantly improved predictions of sound localization, especially at lower frequencies—a challenge that the singular pinna approach encountered. A precise arrangement of 24 strategically positioned speakers spanned three dimensions, with meticulous alignment using laser pointers for precision.



(a) KU100 Dummy head (b) Ircam – Studio 1

Figure 3.2: KU100 Dummy head and IRCAM recording studio

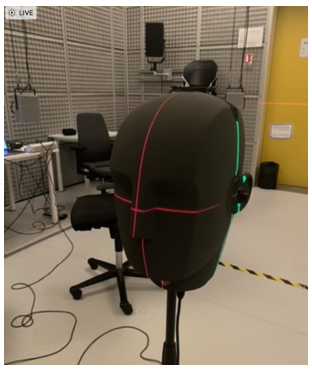


Figure 3.3: Head alignment

"Sweeps" across the frequency spectrum from 0 to 20,000 Hz, were emitted from various speakers, and recorded. These recordings were used to generate HRIRs for all directions and speakers.

Once HRIRs were established for each ear and sound source location, the recordings were convolved with these HRIRs. This convolution process allowed for the simulation of playback as if the recordings were emanating from specific spatial positions.

MUSDB18 dataset was used for training and testing the model. The dataset encompassed diverse musical genres, featuring isolated tracks for drums, bass, vocals, and other instruments, all in stereo format at 44.1kHz.

We attached a visual representation of the frequency distribution of all testing data, with each color denoting a unique data slice.

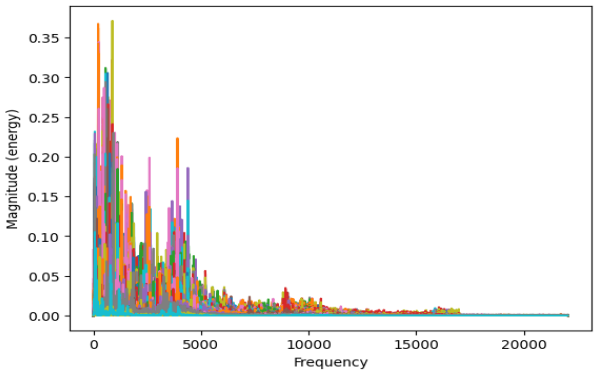


Figure 3.4: Frequency span of the testing data, colour represent different sound slices

From the MUSDB18 dataset, 10 songs were randomly selected and convolutions were applied across 24 directions for both ears to create two datasets:

1. Raw Waveforms: This dataset consists of paired recordings from the left and right ears, concatenated along the channel dimension.
2. Spectrogram Dataset: We applied STFT on the waveform, using a 256-sample STFT window with a hop length of 64.

The labels for these datasets denote the speaker's three-dimensional coordinates, with the head's location designated as the origin.

To prevent data contamination, we applied these processes to different five songs from the MUSDB18 test dataset for the test validation.

For GPU compatibility, we randomly selected 30% of the training data and 15% of the test data while keeping the dataset size at 24 GB for training and evaluation.

3.3 Segment length variation

The evaluation included two distinct data slice lengths. The initial slice length was set at 25 ms. To address potential overfitting, training was also conducted with shorter 5 ms slices. Slices shorter than 5 ms were avoided due to lost phase information. A 5-ms slice yields a maximum detectable delay of 5 ms, equivalent to a frequency of 200 Hz.

While the most optimal performances of the hybrid model were attained using 25-ms segments, the performance of the 5-ms segments did not significantly impair them. Certain models in the ablation studies exhibited noteworthy improvements and interesting

distinctions when employing the 5-ms slices. We provide a more in-depth analysis when discussing the results of each model separately.

Furthermore, a motivation for utilizing shorter time samples is rooted in the context of noise cancellation. CI users often complement their hearing with lip reading, and other hearing aids do not eliminate real-world sound but rather amplify it. Consequently, in the realm of noise cancellation, the time range of 9-10 ms is recognized as an informal threshold consciously avoided by hearing aid engineers [10].

3.4 Architecture

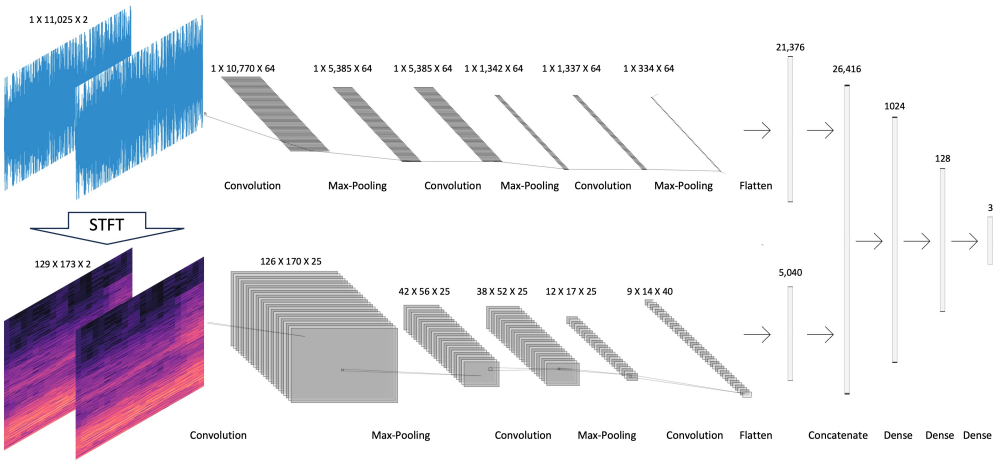


Figure 3.5: Hybrid time and frequency domain model architecture

The sound localization model architecture we propose is presented in Figure 3.4. This end-to-end approach is designed to process two variations of the same audio sample: waveform and spectrogram data formats. Each sample consists of two channels—representing the left ear and the right ear recordings. The architecture comprises three key components: a spectrogram unit, a waveform unit and a hybrid element that processes the concatenation of both former components through fully connected (FC) layers. The architecture is summarized in Table 3.1.

Spectrogram	Waveform	Hybrid
conv 4x4, 25 chan, Relu	conv 63, 75 chan, Relu	FC 1024 units, Relu
max-pool 3x3, Relu	conv 59, 91 chan, tanh	Dropout 0.5
conv 5x5, 25 chan, Relu	conv 58, 96 chan, tanh	FC 128 units, Relu
2 max-pool 3x3, Relu	max-pool 10	Dropout 0.5
conv 4x4, 40 chan, Relu	flatten	FC 3 units, Linear

Table 3.1: Model's architecture

3.5 Loss Function and Experimental Configuration

Previous studies [15] have shown that using Cartesian coordinates produces better results than polar coordinates. Accordingly, we selected this representation. We initially used the MSE loss function but observed suboptimal convergence during the learning phase.

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(y_i - y_i')^2$$

Here, y_i represents the actual sound source location, while y_i' denotes the model's predicted location.

To achieve stable convergence and prioritize directional accuracy over distance accuracy, we used a sum of the Euclidean distance and the angular error.

$$Loss = \frac{1}{n}\left(\sum_{i=1}^{n}||\mathbf{x}_i - \mathbf{x}_i'||_2 + \lambda \arccos(\sum_{i=1}^{n}\hat{\mathbf{x}}_i \cdot \hat{\mathbf{x}_i'})\right)$$

where n is the number of instances, $\mathbf{x}$ is the sound source location, $\mathbf{x}'$ is the model prediction and $\hat{\mathbf{x}}$ is the normalized vector location. Lambda was set to be 180/pi, approximately 57.3. We adjusted this value to ensure the model's loss function processes errors in degrees instead of radians, promoting convergence even with small errors. The model's performance showed small differences around this value, but deviating significantly higher or lower resulted in suboptimal outcomes.

The model was trained using the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 128 samples. The training process spanned 100 epochs, although models typically reached saturation between 40 to 70 epochs. A dynamic learning rate schedule was employed to enhance training.

3.6 Assessment Metrics

The evaluation metrics employed to measure model performance were as follows:

$$\text{Mean Euclidean Distance} = \frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_i - \mathbf{x}_i'||_2$$

$$\text{Mean Angular Error} = \frac{1}{n}\arccos(\sum_{i=1}^{n}\hat{\mathbf{x}}_i \cdot \hat{\mathbf{x}}_i') \cdot \frac{180}{\pi}$$

Chapter 4

Results

This section presents the results achieved by the developed models, accompanied by a comprehensive comparative analysis. To assess the contribution of different elements, ablation studies were conducted. These tests involved the removal of specific components from the hybrid model, resulting in the creation of two separate models. One is a spectrogram-only model, preserving the architecture without the waveform part. The second is a waveform-only model, mirroring the hybrid model architecture but excluding the spectrogram part. Additionally, another hybrid model was constructed, maintaining the same architecture but incorporating only the right ear information. This was done to investigate the information gained from binaural settings compared to single ear recording. All models were evaluated for both 25 ms and 5 ms segments.

A subsequent comparison was made between the developed hybrid model and the benchmark model. The performance of the benchmark model was assessed using the same dataset and evaluation framework as the hybrid model.

Various visual representations were used to analyze the results:

1. Prediction error for each sound source location

   We examined the mean prediction error for each speaker associated with a specific direction, measured in Euclidean distance and angular error between prediction and ground truth. The angular error is measured between the plain created between the two point. The numbering order of the speakers is shown in Figure 4.1.

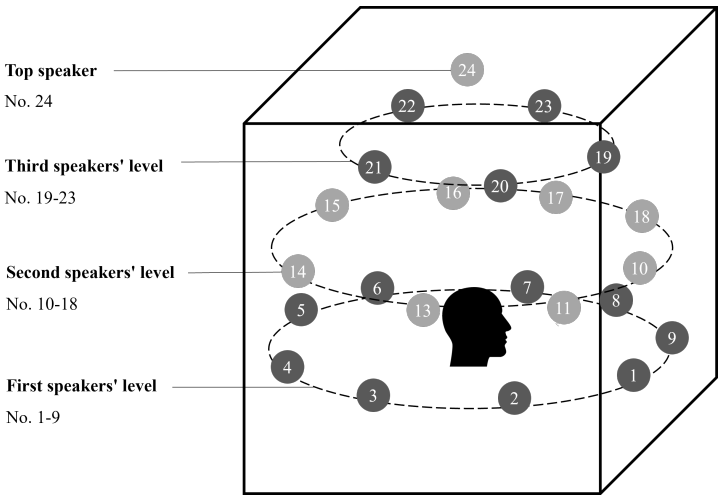


Figure 4.1: Schematic illustration of speakers layout in relation to the dummy head. The initial nine speakers are positioned on the lowest level, speakers 10 through 18 occupy the second level, speakers 19 to 23 are situated on the third level, and the 24th speaker is located precisely at the top of the spherical arrangement. Within each level, speakers are sequentially numbered starting from the speaker right to the front and proceeding in a clockwise direction.

2. Prediction error for each frequency range

   We examined the mean prediction error for each frequency range determined by isolating the frequency component with the highest magnitude in each data sample. The error is measured in both Euclidean distance and angular error.

3. Random sample of prediction

   Ten data records were randomly selected, showing the ground truth location of the recording and the model's prediction for each.

4. 3D angular error presentations

   The graph contains the same data shown in the average angular error for speakers but presented in a 3D visualization, showing the angular error for each speaker.

In each graph, the number at the top of each bar indicates the number of segments in that category.

4.1 Hybrid model results

In this section, we present the comprehensive results of the hybrid model, which surpassed all other tested models, including the benchmark model.

The hybrid model achieved an average angular error of 0.24° and average Euclidean distance of 0.01 meters. This compared to the benchmark results of 19.07° and average Euclidean distance of 1.08 meters, that we will elaborate on in the continuation of this chapter.

Notably, there is no visible consistent trend in the quality of results concerning either sound source location. There are two instances of directional outliers observed on the mid-level front side that repeated in all of out tests. Nevertheless, it is noteworthy that even the outliers demonstrate superior performance compared to other models.

The mean angular error and average Euclidean distance corresponding to each speaker are illustrated in Figure 4.2, following the numbering order presented above.

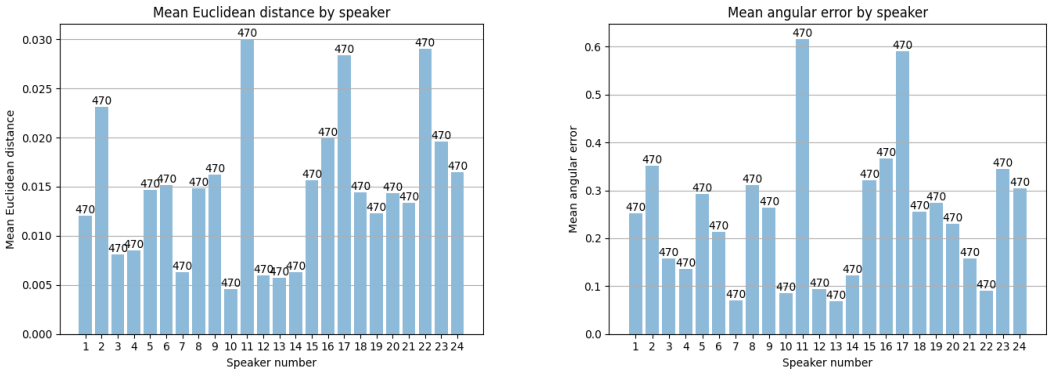


Figure 4.2: Mean angular error in degrees and mean Euclidean distance in meters for each speaker. The model's average angular error is 0.24° while the benchmark's average angular error is 19.07°.

Additionally, we explored the prediction error across frequency ranges. Although no discernible pattern emerged, a notable outlier around the 3,000 Hz range displayed a significantly larger angular error compared to other frequencies. Figure 4.3 presents the prediction error by mean Euclidean distance and mean angular error according to frequency spans.

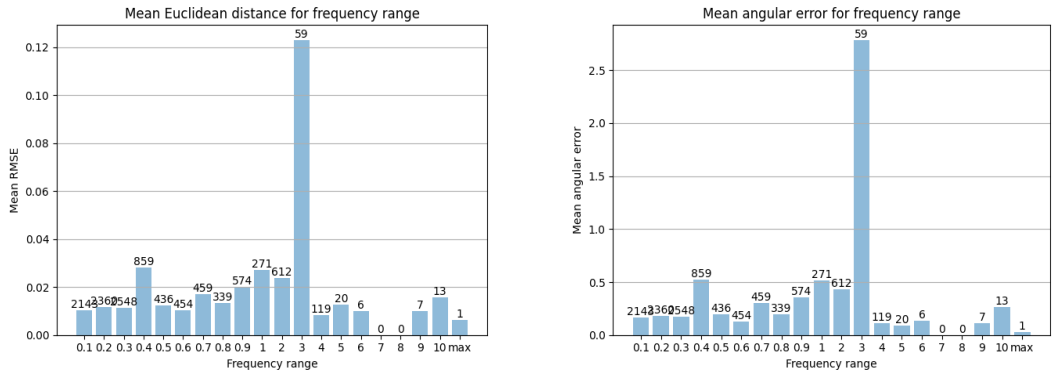


Figure 4.3: Mean angular error in degrees and mean Euclidean distance for each frequency range in kHz

In addition to quantitative analysis, we examined the 3D spatial distribution through two evaluations. The first utilized the same data of mean angular error by speaker, presented in a more intuitive 3D visualization.

The second involved a random sample of 10 recordings, displaying the true speaker location alongside the model's prediction. Notably, the model's predictions closely matched the ground truth for all samples.

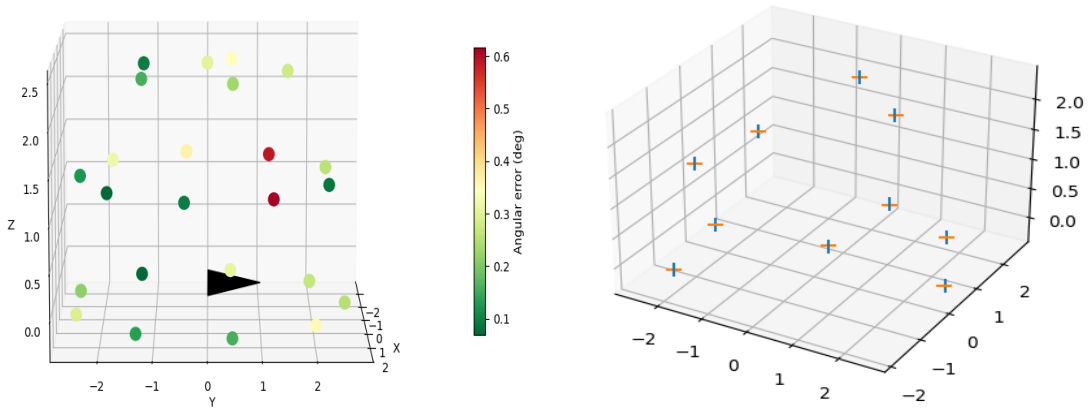


Figure 4.4: The left graph shows the mean angular error for each speaker. The right graph presents an example of the source's location (blue vertical line) and the model prediction (orange horizontal line).

4.1.1 5 ms slices

A comparison was conducted between the performance of the hybrid model trained on 25-ms slices and 5-ms slices. The 5-ms model achieved slightly lower results than the 25 ms model, yet its performance is significantly better than the other models and the benchmark model. The angular error measured 0.79°, with an average Euclidean distance of 0.08 meters. The decline in performance was more pronounced in frequencies below 200 Hz, as anticipated. This suggests that while the model does utilize phase information, the latter's impact is limited.

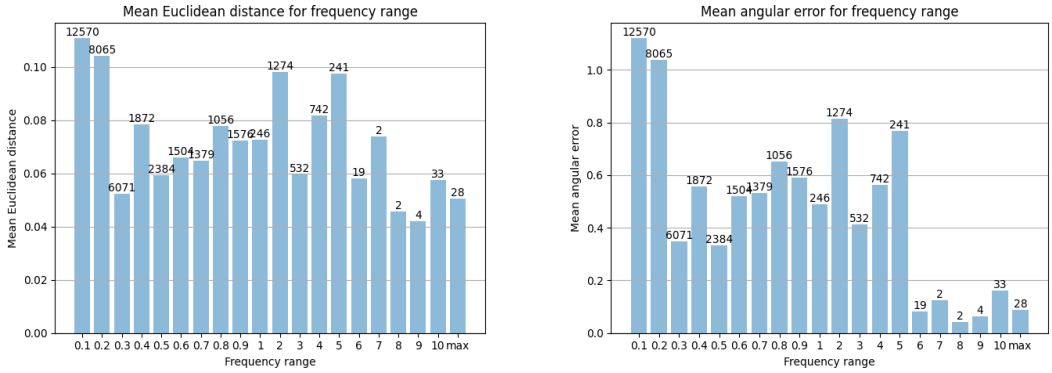


Figure 4.5: Mean angular error in degrees and mean Euclidean distance for each frequency range in kHz. The average angular error is 0.79° while in the 25-ms model it was 0.24°. We can see the performances decrease under 200 Hz derived from loss of phase information.

4.2 Spectrogram-only model results

The spectrogram-only model achieved the second-best results in our evaluations. This model shares the same architecture as the hybrid model but excludes the concatenation with waveform data.

This model achieved an average angular error of 1.3° and an average Euclidean distance of 0.1 meters. Clear differences between directions are not evident. A decline in performance is observed on the lower right side of the head, attributed to the presence of computer fans and other sound-emitting components in that region. This model demonstrated the highest sensitivity to this noise compared to all other models.

The average angular error and Euclidean distance associated eith each speaker are depicted in Figure 4.6

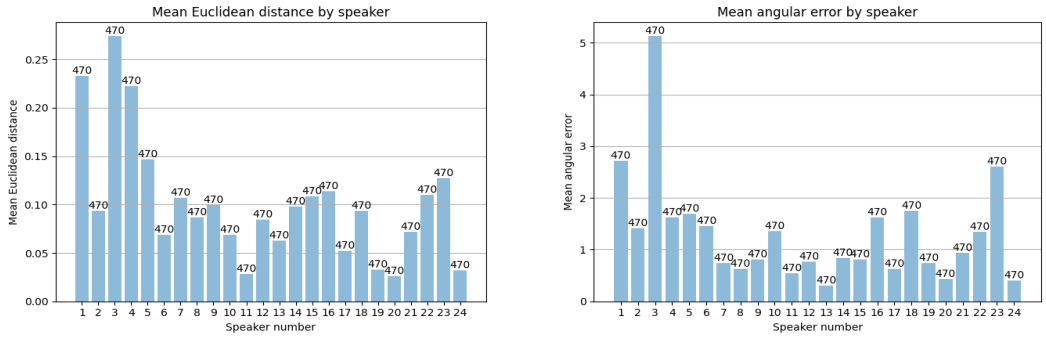


Figure 4.6: Mean angular error in degrees and mean Euclidean distance in meters for each speaker. This model's average angular error is 1.3° while the hybrid model's angular error is 0.24°

Moreover, we explored the prediction error across frequency ranges. Conversely, a noticeable trend is observable in the frequency range performances. As anticipated, the model performs more effectively as the frequency increases. This trend becomes more pronounced when considering the smaller number of samples for higher frequencies. Figure 4.7 presents the prediction according to frequency spans.

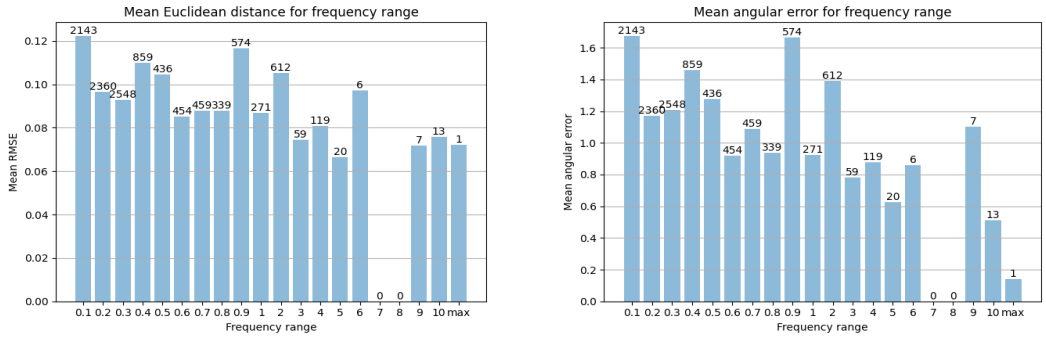


Figure 4.7: Mean angular error in degrees and mean Euclidean distance for each frequency range in kHz. As expected, the model's effectiveness improves with higher frequencies, even though it was trained with a small number of samples than the hybrid model.

In the spatial evaluation, we can see a consistent angular error excluding the noise sensitivity, and decent prediction accuracy although distinct decrease from the hybrid model. We can see this evaluation in Figure 4.8.

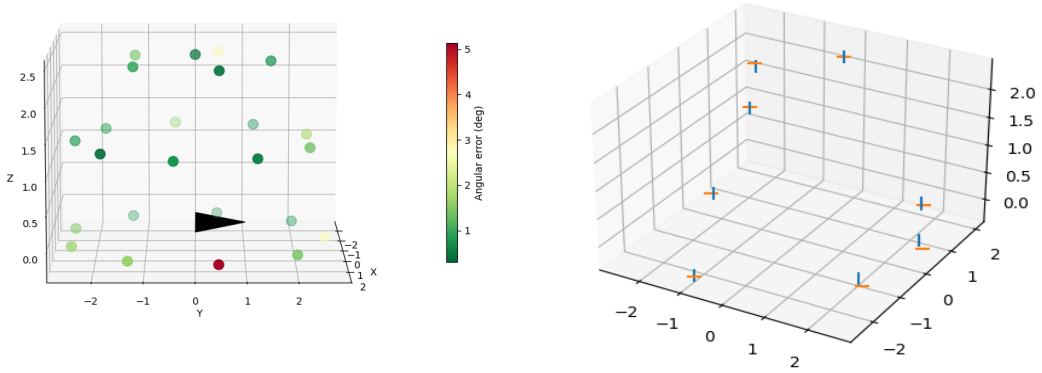


Figure 4.8: The left graph shows the mean angular error for each speaker. The right graph present example of source location (blue vertical line) and the model's prediction (orange horizontal line).

4.2.1 5 ms slices

The 5-ms model exhibited a slight decrease in performance, yet remained notable with an angular error of 1.29° and an average Euclidean distance of 0.24 meters. Graphs for these results are omitted due to their similarity with the 25-ms results.

4.3 Waveform-only model results

The waveform-only model, sharing the same architecture as the hybrid model but excluding the concatenation with spectrogram data, exhibited significantly lower performance compared to the hybrid and spectrogram models.

The model's performance was significantly lower than the performance of the hybrid and spectrogram models. The model achieved an average angular error of 7.39° and an average Euclidean distance of 0.37 meters. Notably, the speakers positioned on the right rear side of the head exhibited relatively less favorable outcomes.

Figure 4.9 illustrates the average angular error and the mean Euclidean distance associated with each speaker.

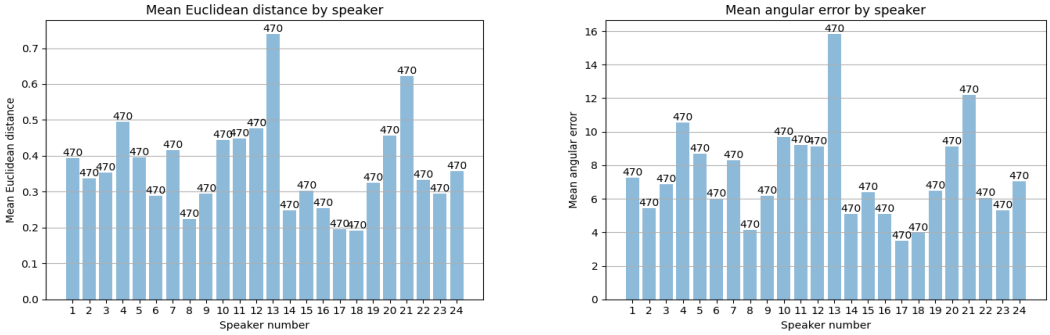


Figure 4.9: Mean angular error in degrees and mean Euclidean distance in meters for each speaker. The model's average angular error is 7.39° while the hybrid model's angular error is 0.24°.

A slight reduction in accuracy was observed with increasing frequencies until 3 kHz, however, this decrease is not noteworthy. This was somewhat surprising and against our expectations from duplex theory.

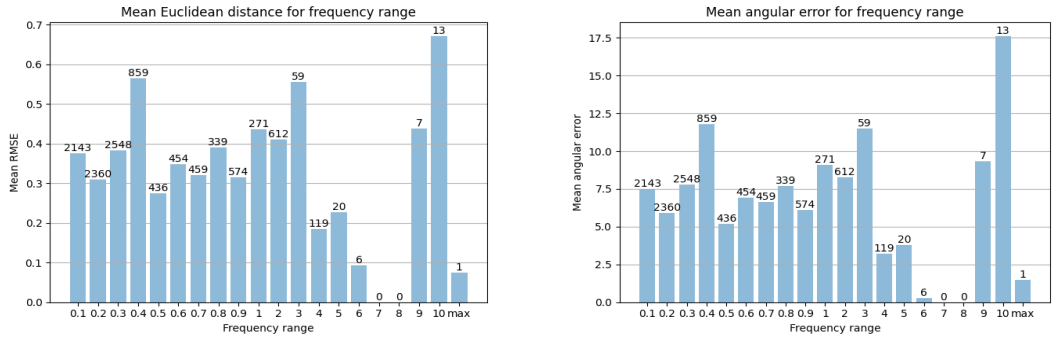


Figure 4.10: Mean angular error in degrees and mean Euclidean distance for each frequency range in kHz. Accuracy decreased slightly as frequencies increase, as expected from duplex theory, but this decline is not significant.

In the spatial evaluation, we can see the accuracy decrease in the speakers located on the right rear side of the head. In adition we can see the decrease in results at the random samples relative to the previous models. We can see this evaluation in Figure 4.11.

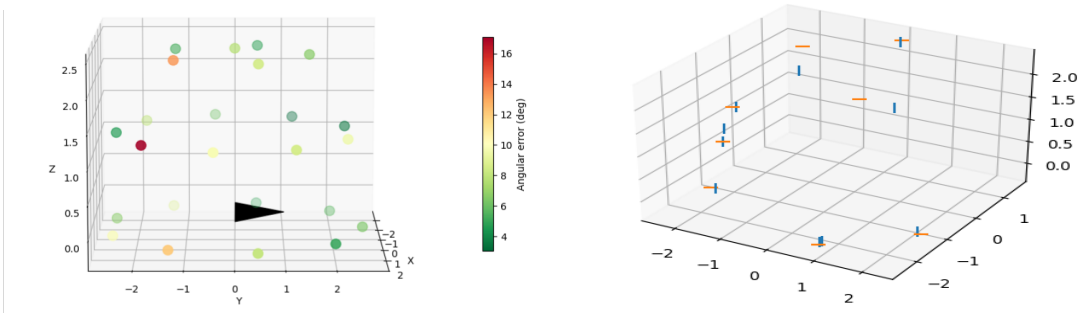


Figure 4.11: The left graph shows the mean angular error for each speaker. The right graph presents an example of source location (blue vertical line) and model prediction (orange horizontal line).

4.3.1 5 ms slices

Intriguingly, the waveform-only model's results improved markedly when working with shortened sample slices. The 5-ms model achieved an average angular error of 2.78° and an average Euclidean distance of 0.13 meters. This represents approximately a three-fold enhancement. It is important to note that this improvement does not carry over to the hybrid model. The prediction error is demonstrated in Figure 4.12

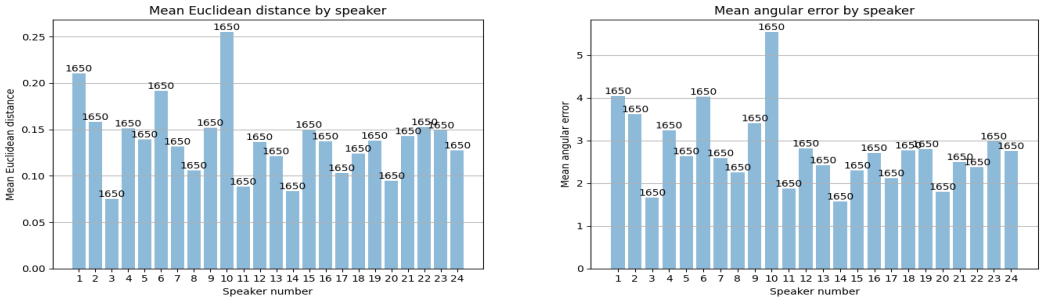


Figure 4.12: Mean angular error in degrees and mean Euclidean distance in meters for each speaker. The model's average angular error is 2.78° while the 25-ms model's angular error is 7.39°

The model's least favorable results are observed in slices dominated by frequencies below 200 Hz, primarily due to the absence of phase information. Beyond these frequency ranges, a marginal decrease in accuracy is evident with increasing frequencies, up to 5000 Hz. However, frequencies higher than 5000 Hz are disregarded due to the limited number of samples.

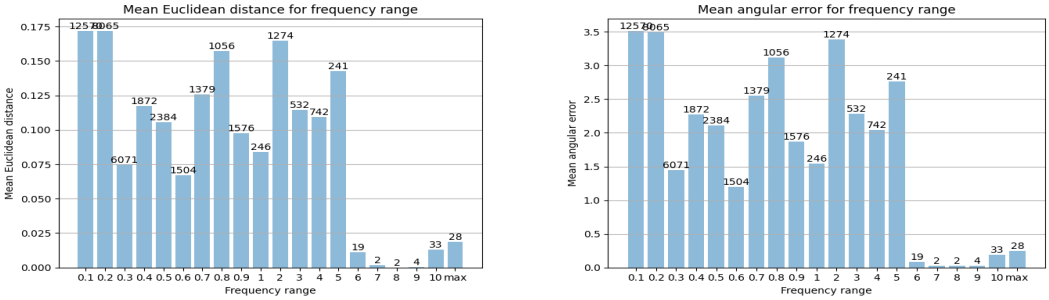


Figure 4.13: Mean angular error in degrees and mean Euclidean distance for each frequency range in kHz.

4.4 One-ear only model results

In addition to the investigated binaural models, we explored whether sound source localization could be achieved using a single microphone. In this scenario, binaural information is entirely absent, and only resonance and spectral behavior can be leveraged.

Results indicate a significant degradation in performance compared to the binaural models. The one-ear only model achieved an average angular error of 19.5° and an average Euclidean distance of 1.03 meters. Pronounced differences between the sources on the recording ear and the opposite side are evident, reflecting the substantial impact of the head on sound waves arriving from the opposite side. This behavior aligns closely with human sound localization ability using only one ear. The noticeable contrast between the sound source on the recording ear side and the opposite side clearly visualized in Figure 4.16.

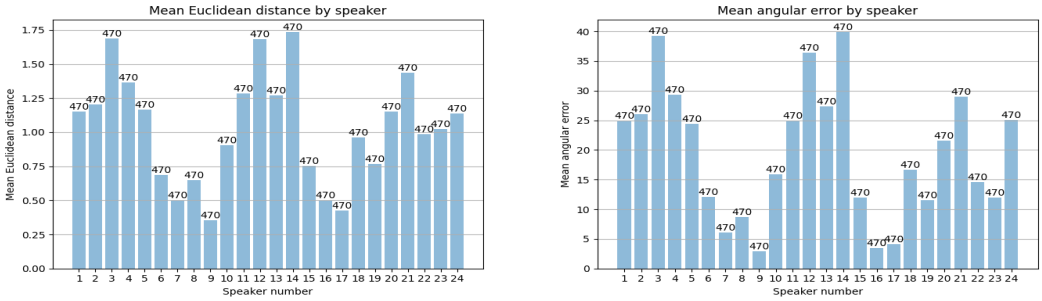


Figure 4.14: Mean angular error in degrees and mean Euclidean distance in meters for speaker. The model's average angular error is 19.5° while the hybrid model's is 0.24°

Moreover, a clear improvement in performance with increasing frequency is observable. This trend shows that spectral differences become clearer in higher frequencies, as expected from the duplex theory.

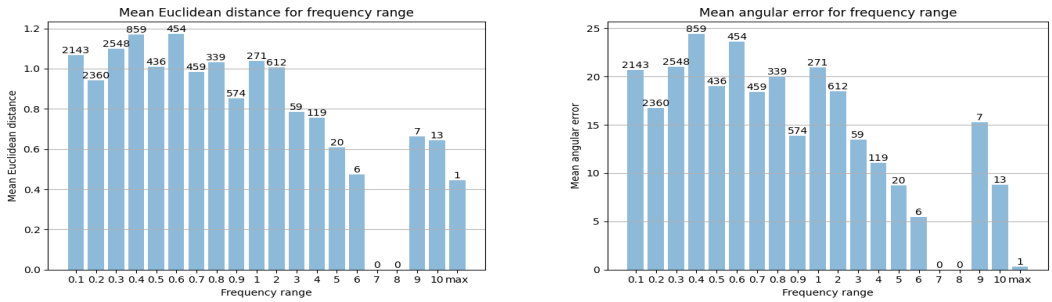


Figure 4.15: Mean angular error in degrees and mean Euclidean distance for each frequency range. Higher frequencies result in noticeable performance improvement.

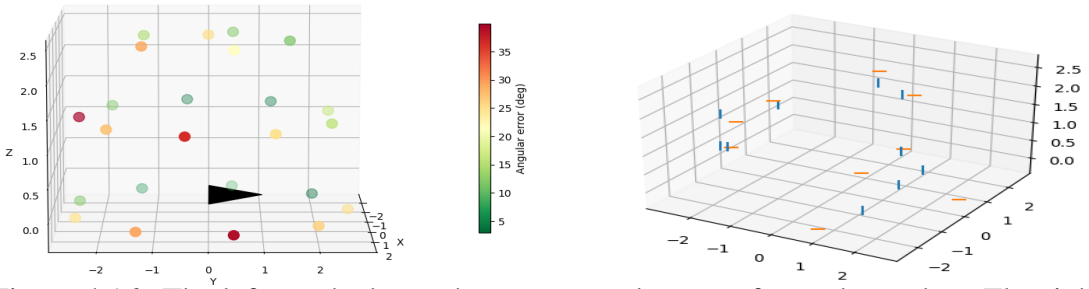


Figure 4.16: The left graph shows the mean angular error for each speaker. The right graph presents an example of source location and model prediction.

4.4.1 5 ms slices

Notably, the 5-ms model outperformed the 25-ms model in this scenario, achieving results approaching those of the binaural models. The 5-ms model demonstrated enhanced performance, with an average angular error of 10.15° and an average Euclidean distance of 0.68 meters. Additionally, the directional impact exhibited different behavior: the 5-ms model's worst performance was on the face center plane and speakers closest to the functional ear, while the range between the center and the ear displayed better results.

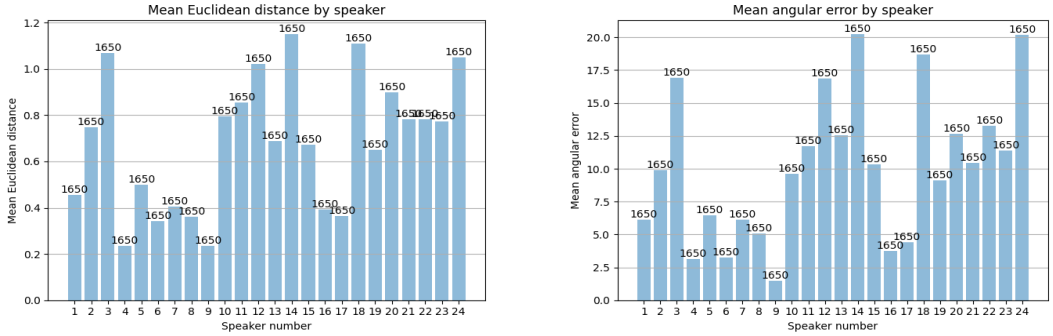


Figure 4.17: Mean angular error in degrees and mean Euclidean distance in meters for each speaker. The model's average angular error is 10.15° while the 25-ms model's angular error is 19.5°

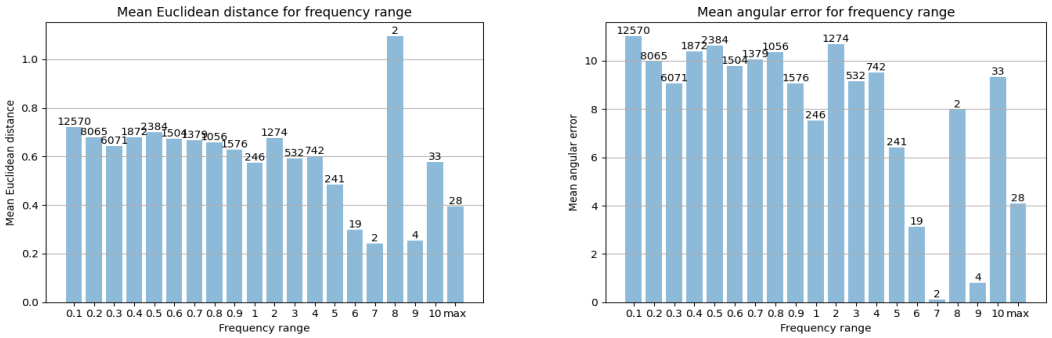


Figure 4.18: Mean angular error in degrees and mean Euclidean distance for each frequency range in kH

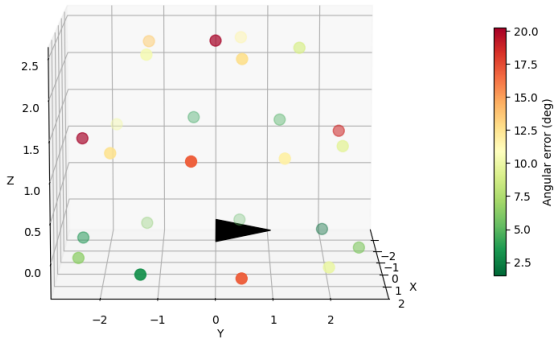


Figure 4.19: Mean angular error for each speaker.

4.5 Comparison to Benchmark

A comparison was made between our models and the benchmark model, which was presented in chapter 'Related work'. Key differences exist between our hybrid model and the benchmark model. Our approach incorporates a waveform-spectrogram hybrid methodology as opposed to the benchmark's waveform-only approach. In the frequency analysis stage, the benchmark model employs separate one-dimensional convolutions on each recording, while we employ convolutions on both recordings simultaneously along the depth dimension. Moreover, the benchmark model utilizes a linear activation function for frequency analysis, whereas our approach integrates activation functions across all layers. Lastly, our model employs a loss function tailored to the key metrics, while the benchmark utilizes the RMSE loss function.

The 25-ms model yielded superior results for the benchmark. The results show an average angular error of 19.07° and an average Euclidean distance of 1.08 meters. Table 4.1 shows the comparison between our model and the benchmark model.

	Benchmark model	Hybrid model
Angular error	19.07°	0.24°
Euclidean distance	1.08 m	0.01 m

Table 4.1: Hybrid model and benchmark model results

Besides performance differences, the benchmark model shows distinctions in predicting frontal, rear, and above locations. Figure 4.20 shows the average angular error and the mean Euclidean distance associated with each speaker. Figure 4.21 illustrate this in a 3D visualization.

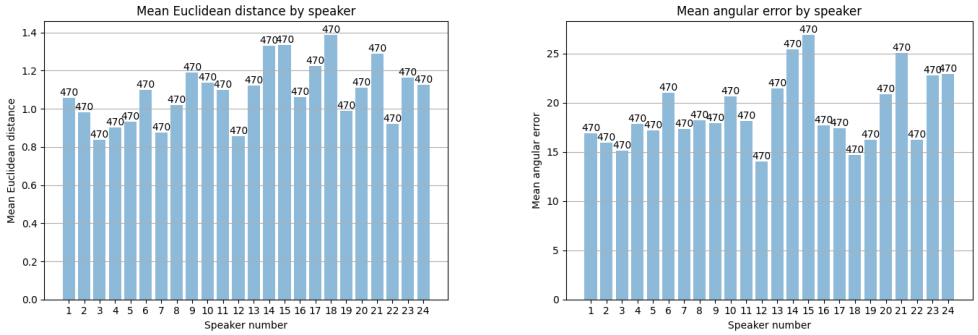


Figure 4.20: Mean angular error in degrees and mean Euclidean distance in meters for each speaker. The benchmark's average angular error is 19.07° while the hybrid's angular error is 0.24°.

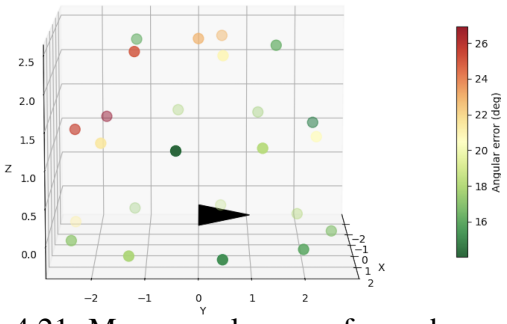


Figure 4.21: Mean angular error for each speaker.

Figure 4.22 shows both, the benchmark model (right) and the proposed hybrid mode (left). Clearly, the hybrid model performs an order of magnitude better than the benchmark model (note the difference in scales between the two plots). Note also that the performance of the benchmark model declines as the horizontal angle increases, further supporting the effectiveness of hybrid time and frequency domain model's.

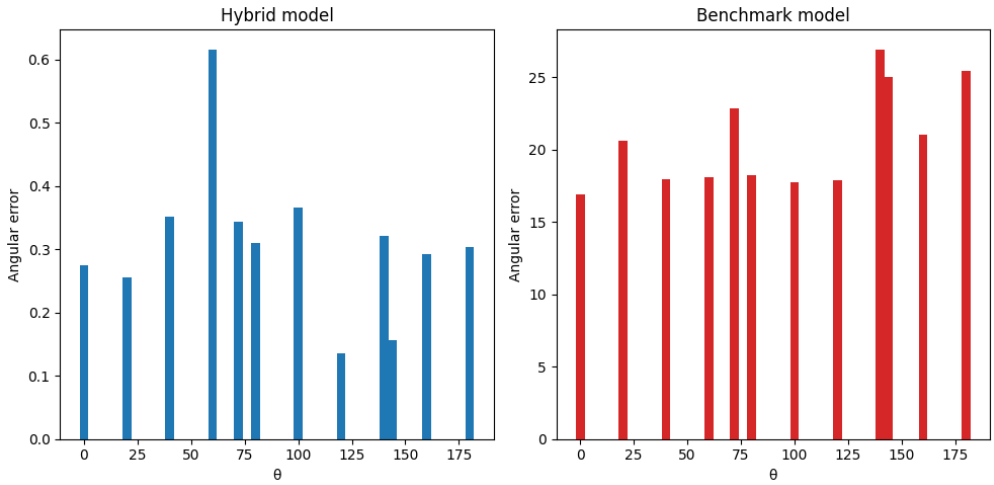


Figure 4.22: Angular error by horizontal angle θ for hybrid and benchmark models. Note both the magnitude of the angular error and how it trends concerning θ

Additionally, we investigated the prediction error across frequency ranges. As expected, the model exhibits more effective performance at low frequencies. Figure 4.23 illustrates the prediction error according to frequency spans.

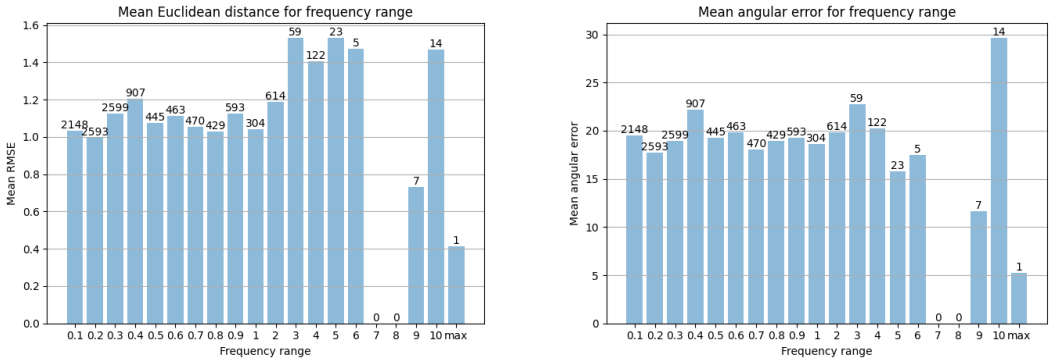


Figure 4.23: Mean angular error in degrees and mean Euclidean distance for each frequency range in kHz. The Euclidean distance grows with increasing frequency.

Chapter 5

Conclusions and Discussion

This study has introduced a novel approach to sound source localization. We have, moreover, expanded the scope of binaural sound localization from a limited 180-degree span to a comprehensive full-sphere scenario. In contrast to traditional machine localization systems that rely heavily on handcrafted features such as ITD and ILD, our approach adopts an end-to-end hybrid model that integrates both spectrogram and temporal domain information. Our results surpass by a wide margin the current state of the art in sound source localization.

The evaluation of our models has provided valuable insights into the effectiveness of different architectures and approaches. The hybrid model, incorporating end-to-end learning from raw data encompassing both waveform and spectrogram information, emerged as the most successful. This success underscores the importance of leveraging both time and frequency domain features for enhanced performance in sound source localization tasks.

Ablation studies, where specific components were selectively removed from the hybrid model to create spectrogram-only and waveform-only models, highlighted the individual contributions of these elements. The spectrogram model demonstrated robust performance, affirming the significance of frequency information in accurate localization. In contrast, the waveform model exhibited lower accuracy, emphasizing the complementary role of spectrogram data. Nevertheless, integrating waveform information into the model led to a significant 5-fold improvement in accuracy.

Exploring one-ear-only models, representing scenarios with a single microphone, emphasized the challenges posed by the absence of binaural information. The significant degradation in performance compared to binaural models underscores the crucial role of interaural differences in sound localization. However, the model demonstrated unexpectedly decent accuracy, particularly on the ear's opposite side of the head. Furthermore, the 5 ms version of the single-ear-only model exhibited improved results, approaching the performance level of the binaural waveform-only model.

Moreover, the influence of the Duplex-Theory, specifically the heightened accuracy within high frequencies observed in the spectrogram-only and one-ear-only models, was noted. This observation provides further support for our comprehension of the frequency analysis conducted by the DL model.

Regarding the difference between the 25 and 5 ms tests, various models exhibited interesting behavior. Although the short time-frame decreased the performances of the hybrid model, the impact on other models was more complex. In the spectrogram-only model, the angular error improved slightly, but the distance prediction decreased, leading to higher Euclidean distance prediction error. The waveform-only model showed a significant improvement despite the clear harm to low-frequency prediction caused by the loss of phase information. For the one-ear-only model, notable improvement was also observed with a shorter time-frame. This suggests a higher tendency for overfitting in these models, capturing more specific details with a shorter time-frame. Additionally, the frequency range analysis revealed that the hybrid and waveform-only models rely to some extent on phase information, while the spectrogram and one-ear models don't use this information as much, being more frequency-oriented or missing the information completely.

Comparisons with the benchmark model, which utilized a waveform-only approach and distinct architectural choices, further underscored the superiority of the hybrid model. The hybrid model consistently outperformed the benchmark across various metrics, demonstrating the effectiveness of the chosen hybrid time and frequency domain methodology.

These findings contribute to the broader understanding of sound source localization models and their practical applications. The success of the hybrid model suggests its potential utility in diverse fields. The nuanced insights gained from the ablation studies and comparisons with alternative models pave the way for further refinements and optimizations in future iterations of sound source localization systems. Overall, this research lays the groundwork for more accurate models, bringing us closer to achieving robust and reliable sound source localization in real-world scenarios.

Future advancements in sound source localization research should explore additional research avenues that could generate substantial contributions to the field. Three major ones are:

1. Developing an agnostic Model for different heads and environments:

   A promising avenue for future research involves transforming our current model into a head-agnostic one and adaptable to various settings. This adaptation could eliminate the need for user-specific training, ensuring accurate predictions across a diverse range of users and scenarios. By creating a model that generalizes well to different users and configurations, we aim to enhance the practicality and accessibility of sound source localization technologies.

2. Building a Unified Model for Localization and Noise Cancelling:

   An intriguing possibility is the development of an integrated model that addresses not only sound localization but also related tasks such as noise cancellation or source separation. Combining these functionalities within a unified model could lead to mutually reinforcing enhancements in both localization accuracy and noise reduction. Such comprehensive solution may be vital in environments where background noise is a significant factor, contributing to improved user experiences in various settings.

3. Optimizing one-microphone based localization:

   A novel direction for future exploration involves optimizing the design of the microphone cover. Our study highlights the potential for achieving credible sound localization using a single microphone, through a ear's natural structural features. By refining the design, materials, and geometry of the microphone cover, there is a possibility of achieving reliable sound localization even with a lone microphone. This innovation could have widespread implications, including cost reduction, increased user comfort, and application in compact devices.

In conclusion, as the field of sound source localization advances, this research deepen our understanding of auditory perception and presents real life applications, particularly within healthcare settings. We anticipate that the implementation of this research could improve the quality of life for CI users and may even play a crucial role in advancement of other hearing aids and various contemporary technologies.

References

[1] Lord Rayleigh. “On the dynamical theory of gratings”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 79.532 (1907), pp. 399–416.

[2] Hans Wallach. “The role of head movements and vestibular and visual cues in sound localization.” In: *Journal of Experimental Psychology* 27.4 (1940), p. 339.

[3] John C Middlebrooks and David M Green. “Sound localization by human listeners”. In: *Annual review of psychology* 42.1 (1991), pp. 135–159.

[4] Frederic L Wightman and Doris J Kistler. “The dominant role of low-frequency interaural time differences in sound localization”. In: *The Journal of the Acoustical Society of America* 91.3 (1992), pp. 1648–1661.

[5] Dumidu S Talagala and Thushara D Abhayapala. “HRTF aided broadband DOA estimation using two microphones”. In: *2012 International Symposium on Communications and Information Technologies (ISCIT)*. IEEE. 2012, pp. 1133–1138.

[6] John Woodruff and DeLiang Wang. “Binaural localization of multiple sources in reverberant and noisy environments”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.5 (2012), pp. 1503–1512.

[7] Ofer Schwartz and Sharon Gannot. “Speaker tracking using recursive EM algorithms”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.2 (2013), pp. 392–402.

[8] Hirofumi Tsuzuki et al. “An approach for sound source localization by complex-valued neural network”. In: *IEICE TRANSACTIONS on Information and Systems* 96.10 (2013), pp. 2257–2265.

[9] Toni Hirvonen. “Classification of spatial audio location and content using convolutional neural networks”. In: *Audio Engineering Society Convention 138*. Audio Engineering Society. 2015.

[10] Joshua Alexander et al. “Hearing aid delay and current drain in modern digital devices”. In: *Canadian Audiologist* 3.4 (2016).

[11] Ryu Takeda and Kazunori Komatani. “Sound source localization based on deep neural networks with directional activate function exploiting phase information”. In: *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2016, pp. 405–409.

[12] Relja Arandjelovic and Andrew Zisserman. “Look, listen and learn”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 609–617.

[13] Semih Ağcaer and Rainer Martin. “Binaural source localization based on modulation-domain features and decision pooling”. In: *arXiv preprint arXiv:1812.02399* (2018).

[14] Soumitro Chakrabarty and Emanuël AP Habets. “Multi-speaker DOA estimation using deep convolutional networks trained with noise signals”. In: *IEEE Journal of Selected Topics in Signal Processing* 13.1 (2019), pp. 8–21.

[15] Zhenyu Tang et al. “Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks”. In: *arXiv preprint arXiv:1904.08452* (2019).

[16] Paolo Vecchiotti et al. “End-to-end binaural sound localisation from the raw waveform”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 451–455.

[17] Christine Evers et al. “The LOCATA challenge: Acoustic source localization and tracking”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 1620–1643.

[18] Alexandre Défossez. “Hybrid spectrogram and waveform source separation”. In: *arXiv preprint arXiv:2111.03600* (2021).

[19] Alexandra Annemarie Ludwig et al. “Sound localization in single-sided deaf participants provided with a cochlear implant”. In: *Frontiers in psychology* 12 (2021), p. 753339.

[20] Bing Yang, Hong Liu, and Xiaofei Li. “Learning deep direct-path relative transfer function for binaural sound source localization”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3491–3503.

[21] Irene Martın-Morató et al. “Low-complexity acoustic scene classification in dcase 2022 challenge”. In: *arXiv preprint arXiv:2206.03835* (2022).

[22] Qi Hu, Ning Ma, and Guy J Brown. “Robust Binaural Sound Localisation with Temporal Attention”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

תקציר

איכון של מקור הקול ממלא תפקיד בסיסי בתפיסה השמיעתית, ומאפשר לאדם ולמכונות לקבוע את מיקום מקור הקול. שיטות איכון קול מסורתיות מסתמכות לרוב על תכונות שנוצרו באופן ידני ותנאים פשוטים, המגבילים את הישימות שלהן במצבים אמיתיים.

כלים המאפשרים איכון מדויק של צליל הינם חיוניים במגוון יישומים, הכוללים רובוטיקה, מציאות מדומה, אינטראקציות בין אדם למחשב ומכשירים רפואיים. שיפור יכולת זו חשוב במיוחד עבור אנשים עם שתלי שבלול ,אשר מתמודדים עם אתגרים משמעותיים בתפיסת הכיוון של מקורות הקול.

מחקרים קודמים שעסקו באיכון צלילים, התמקדו בפיתוחים מבוססי מערכי מיקרופונים נרחבים במישור הקדמי בלבד. מודלים אלו בעלי מגבלות דיוק והתמודדות עם רעש בעת שימוש במערכי מיקרופונים קטנים. טכניקות איכון אלו אינן מעשיות עבור משתמשי שתל שבלול גם בשל אילוצי גודל ומשקל, והמגבלה המרחבית.

מחקר זה מציג גישה חדשה לאיכון של מקור קול באמצעות מאפייני פונקציית תמסורת ראש , מנתונים גולמיים, הן בתחום הזמן והן בתחום התדר. יתר על כן, הוא מקדם איכון קול על בסיס דו־סנסורי, ומרחיב את טווח הגילוי מ081 מעלות במישור הקדמי לכלל המרחב.

הגישה המוצעת מציגה מודל היברידי של למידה עמוקה, המשלב קלטי ספקטוגרמה וגלים באמצעות ערוצים מקבילים. המודל ההיברידי מאפשר איכון ברמת דיוק גבוהה יותר מהמודלים הקיימים בתחום, פי 08 יותר מדויק באומדן השגיאה הזוויתיות ופי 801 יותר מדוייק בהערכת המרחק האוקלידי.

מחקר זה תורם להבנה של תפיסה שמיעתית ומציע יישומים מעשיים בעולם הבריאות.

הנחייה

ברצוני להודות למנחיי, פרופ' יעקב הל־אור מבית ספר אפי ארזי למדעי המחשב, ופרופסור אמיר עמדי ממכון ברוך איבצ'ר למוח, קוגניציה וטכנולוגיה באוניברסיטת רייכמן. תודה גם לשלמה דובנוב, מנהל המרכז למחקר ולמידה באוניברסיטת קליפורניה סן דייגו, אוליבייה וורוסספל, חוקר ראשי במרכז המחקר IRCAM בפריז, ולתמוז דובנוב, על ההדרכה והתמיכה שסיפקו לטובת מחקר זה.

שיתוף פעולה

מחקר זה בוצע בשיתוף פעולה עם אוניברסיטת קליפורניה סן דייגו ומרכז המחקר IRCAM. ברצוני להודות לראש המעבדה גררד אסייג ושלמה דבנוב על תרומתם לשיתוף פעולה זה.במחקר.

תודה אישית

אני מבקש להביע את תודתי והערכתי הרבה לזוגתי האהובה נעם רהב. התמיכה האינסופית, העידוד וההבנה שלה היו חיוניים לאורך מסע אקדמי זה, במיוחד במהלך משלחת המחקר לפריז. בנוסף, אני מודה מאוד על תרומתה החשובה ביותר בקריאה חוזרת וסיוע בחידוד כתיבת תזה זו, תוך רתימת ניסיונה האקדמי וחדות כתיבתה.

מימון

מחקר זה מומן על ידי מועצת המחקר האירופית במסגרת אופק 0202, מענק מספר 313388.

אוניברסיטת רייכמן

בית־ספר אפי ארזי למדעי המחשב

התכנית לתואר שני (M.Sc.)־ מסלול מחקרי

איכון מיקום מקור צליל בתנאי שמיעה אנושיים תוך שימוש במודל משולב תחום זמן ותדר

גיל גבע

עבודת תזה המוגשת כחלק מהדרישות לשם קבלת תואר מוסמך M.Sc. במסלול המחקרי בבית ספר אפי ארזי למדעי המחשב, אוניברסיטת רייכמן

דצמבר 2023