Artificial-Pinna Acoustic Encoding for Monaural Sound Source Localization

Meghan Kret, Tiffany Shum, Grace Tseng, and Lani Wang
The Cooper Union
New York, NY, USA

***Abstract*—This project investigates whether the geometry of an artificial pinna can give a single microphone useful directional hearing. Instead of relying on multiple microphones, the system uses a four-wall artificial outer ear to passively reshape incoming sound before it reaches the sensor. The front, right, back, and left walls create direction-dependent reflections, shadowing, and spectral notches that act as acoustic fingerprints for the detection of sound sources. These fingerprints are characterized through an anechoic-chamber measurement process in which the pinna-microphone assembly is mounted on a turntable, exposed to a known loudspeaker signal from many directions, and converted into a grid of pinna-related impulse responses (PRIRs) and pinna-related transfer functions (PRTFs). A convolutional neural network is then used as a decoder to test whether the measured monaural cues are strong enough to recover source direction. On the controlled 632-direction PRTF grid, the system achieves a mean great-circle angular error of 8.61°, a median error of 3.70°, and 86.7% accuracy within 10°. Frequency-domain analysis shows that performance is strongest in the mid-to-high frequency range, where pinna-induced spectral cues are expected to be most informative.**

***Index Terms*—monaural localization, artificial pinna, HRTF, anechoic chamber, turntable measurement, spatial audio, convolutional neural network**

I. Introduction

Sound source localization is usually treated as a multi-sensor problem. Human listeners compare signals between two ears, while engineered systems often use two or more microphones to estimate time differences or level differences. [1], [2]. These cues are powerful because they compare the same sound after it arrives at different spatial positions.

Multiple microphones require physical spacing, additional electronics, synchronization, and calibration, which can be difficult in small robots, wearables, hearing-assistive devices, or compact acoustic sensors. A single-microphone design could reduce hardware complexity, costs, and space where needed.

However, a single microphone alone does not have access to time and level differences. With only one recorded waveform, there is no second signal that can provide an arrival-time difference or a level difference. If a monaural device is to estimate direction, directional information must be introduced before the signal reaches the microphone. Previous work has shown that monaural localization can be learned when a direction-dependent spectral structure is available [3]. In this project, that information comes from the sensor's environment.

The central idea is that an artificial pinna can act as a passive acoustic encoder. As sound travels around the walls and through the cavities of the pinna, different source directions create different reflection paths, shadowing effects, and interference patterns. These effects appear in the recorded signal as direction-dependent peaks and notches in the frequency spectrum, consistent with the known role of spectral detail in localization [1], [4]. The microphone still records only one channel, but that channel has already been shaped by the geometry of the artificial outer ear.

A neural network is used as a decoder rather than as a replacement for acoustic design. Strong localization performance is evidence that the pinna geometry converted source direction into measurable spectral structure. Similar learning-based approaches have been used to study sound localization from acoustic cues, including single-microphone localization and neural models of spatial hearing [3], [5].

This paper makes two contributions. One is that a pinna structure was created and validated through simulations and measurements that would create unique transfer functions in each direction. The other is that a neural network was created to predict the direction of incident sound.

II. Background and Related Work

A. Human Pinna and Monaural Spectral Cues

Human spatial hearing relies on three primary cue classes: interaural time differences (ITDs), interaural level differences (ILDs), and spectral shape cues introduced by the outer ear [1]. The pinna filters incoming sound in a direction-dependent way, creating peaks and notches in the frequency response that allow listeners to resolve front-back ambiguity and estimate elevation, which are tasks that are impossible with ITD and ILD alone [1], [4]. Kulkarni and Colburn demonstrated that the fine spectral structure in the range of roughly 4–16 kHz is particularly important for elevation perception [4]. This project uses the same principle artificially: an engineered outer ear creates the spectral cues rather than the biological one.

A head-related transfer function (HRTF) captures the direction-dependent filtering applied to sound as it travels from a source to the eardrum [1]. The UC Davis CIPIC database, one of the most widely used HRTF resources, measured responses for 45 subjects at 1250 directions and has enabled significant research in spatial audio rendering, hearing-aid processing, and localization modeling [6]. In this paper, we use the terms pinna-related transfer function (PRTF) and pinna-related impulse response (PRIR) to refer to the frequency- and time-domain responses of the pinna structure, respectively.

These terms are analogous to HRTF and HRIR, but describe directional filtering caused by the pinna alone rather than by the combined head, torso, and outer ear.

B. Monaural Sound Source Localization

Monaural sound localization has been studied in both human perception and engineered systems. The first artificial monaural localizer used a single microphone with a shaped reflecting structure and analog Very-Large-Scale Integration (VLSI) echo-time processing to estimate elevation [7]. Saxena and Ng later demonstrated that a single microphone can estimate azimuth over a full 360° range when pinna-induced spectral structure is decoded by a machine learning model, achieving a best-case average error of 13.5° for their best pinna design on a variety of natural sounds [3]. Deleforge et al. showed that speech sources can be localized using simple scattering structures, such as irregular cubes surrounding an omnidirectional sensor. These structures alter incoming sounds in direction-dependent ways, and non-negative matrix factorization (NMF) can be used to learn the resulting acoustic patterns and infer the source location [8]. Sun et al. extended this idea using a 3D metamaterial enclosure, which provides strong direction-dependent spectral signatures and enables both localization and source separation from a single microphone via compressive sensing [9]. More recent work by Francl and McDermott showed that deep neural networks trained on binaural signals can reproduce many perceptual localization phenomena, suggesting that the underlying acoustic cues can be learned from data [5]. Our work differs from all of these in that it targets full 3D localization (azimuth and elevation jointly) across 632 measured directions using an anechoic-chamber and a CNN decoder that operates on monaural spectrograms.

C. Learning-Based Spatial Hearing

Residual networks [10] and squeeze-and-excitation attention [11] have become standard components for deep audio classification tasks. SpecAugment [12] is a widely used spectrogram-domain augmentation method that improves robustness by masking time and frequency bands during training. AdamW [13] provides stable optimization with decoupled weight decay. We draw on these methods to build the localization decoder, but the primary novelty of this work lies in the acoustic design of the artificial pinna and live measurement pipeline, rather than in the network architecture itself.

III. Artificial Pinna Design

A. Four-Wall Geometry

The 63.5 mm x 139.7 mm x 130 mm artificial pinna (see Fig. 1) was 3D-printed in PLA plastic and divided into four asymmetrical walls: front, right, back, and left, which are connected together via a spider-like component that branches into each structure and has a hole in the middle for the microphone to fit. The goal of this geometry is not to amplify sound uniformly, but to instead make the recorded spectrum depend on the direction of arrival. Each wall presents a different orientation and surface path to incoming sound, so each wall can contribute different reflections, delays, shadowing effects, and interference patterns.

If the structure were highly symmetric, several source directions could produce similar filters and become difficult to distinguish. The four-wall design intentionally breaks that symmetry. Sounds from the front, side, and rear encounter different acoustic boundaries before reaching the microphone. As a result, two directions that might otherwise produce similar monaural recordings can produce different spectral notches once the pinna is added.

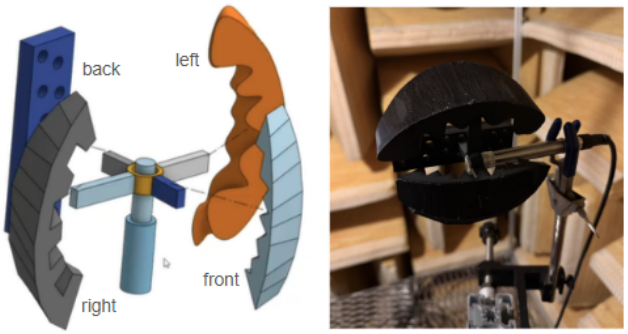

Fig. 1. 3D model of the pinna, and the printed model of the pinna used as the monaural acoustic sensor enclosure respectively. The microphone is positioned and held at the center of the pinna, which is connect to its external structures on the paths of the dashed lines, so that incoming sound is filtered by the pinna geometry before it reaches the sensor.

B. Acoustic Role of Each Wall

Table I summarizes the intended role of each wall. These roles are qualitative because the response at the microphone is produced by the combined pinna geometry, not by one wall in isolation. Still, the wall-based description makes the design easier to explain and gives physical meaning to the later wall-specific analysis.

TABLE I
Intended acoustic role of each artificial-pinna wall.

Wall	Intended acoustic role
Front (Light Blue)	Composed of jagged, pyramid-like peaks with flat sides.
Right (Grey)	Has varying sawtooth bumps of different heights and rectangular cavities.
Back (Blue)	A brick-like rectangular prism with a random assortment of 6.35 mm wide holes.
Left (Orange)	A wedge with smooth and organic inner curves.

The key design principle is that small geometric differences create varied and unique audio reflections. At low frequencies, wavelengths are large compared with the pinna features, so the structure has limited ability to create direction-specific notches. At higher frequencies, wavelengths are short enough for the wall geometry to produce stronger reflections and cancellations. This frequency dependence is consistent with the role of pinna-related spectral cues in spatial hearing [1],

[4]. This is why the frequency-domain analysis in Section X is central to validating the design.

C. Simulation and Design Validation

Before the final pinna model was printed, Mesh2HRTF simulations were used to evaluate whether the design produced useful direction-dependent acoustic features. The pinna geometry was first created as a 3D model, then remeshed in Meshmixer to improve mesh quality and make it suitable for acoustic simulation. This step helped reduce issues such as irregular triangles, nonuniform mesh density, and geometry artifacts that could affect the simulation results.

The cleaned mesh was then imported into Mesh2HRTF, which simulated the acoustic transfer function of the pinna structure for different source directions to estimate how the pinna would filter incoming sound before reaching the microphone. The resulting simulated PRTFs were then used as inputs to our localization model to test whether the design created distinguishable acoustic cues across direction.

By comparing model performance across the simulated responses, we were able to check whether each wall of the pinna prototype contributed useful directional information and identify any geometry that needed adjustment. The selection of the final design was based on both the simulated localization performance and the practicality of printing and mounting the physical structure.

IV. Anechoic-Chamber Measurement Process

A. Measurement Goal

The purpose of the measurement process is to characterize how the artificial pinna filters sound from each direction. This direction-specific filtering is represented as a PRIR in the time domain or as a PRTF in the frequency domain [1], [6]. For this project, the PRTF is the measured acoustic fingerprint of the pinna-microphone system.

Because the goal is to measure the pinna rather than the room, the setup uses an anechoic chamber. The chamber suppresses reflections from walls, floor, and ceiling, so the recorded signal is dominated by the direct sound path and the acoustic modifications caused by the pinna itself. This makes the resulting PRTFs easier to interpret and more suitable for controlled localization experiments.

B. Turntable-Based Direction Sampling

The pinna-microphone assembly is placed at a fixed distance from a loudspeaker and mounted on a turntable. For each measurement direction, the loudspeaker plays an excitation signal consisting of a logarithmic sine sweep from 20 Hz to 20 kHz. The microphone records the response after the sound has interacted with the artificial pinna. The turntable then rotates the assembly to the next elevation angle, once a full rotation of elevation angles is completed, the pinna is manually rotated to the next azimuth angle, and the process is repeated over the desired set of azimuth and elevation directions. In the set analyzed by the machine learning model in Section VI, measurements were taken every 5° for elevation angles and in 20° increments for azimuth. Each measured direction produces its own impulse response. Once the full grid has been measured, the dataset describes how the artificial pinna transforms incoming sound across space. This step is the bridge between the physical design and the localization experiment: it turns the geometry of the pinna into direction-labeled acoustic data.

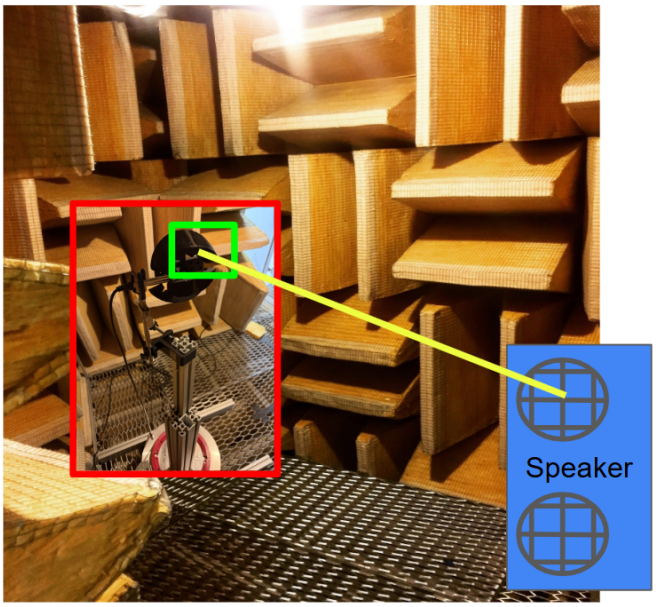

Fig. 2. Anechoic-chamber measurement setup for the artificial pinna. The pinna-microphone assembly is mounted on a turntable and measured relative to a loudspeaker in a reflection-suppressed environment. The turntable changes the source direction relative to the pinna, while the microphone records the signal after it has been filtered by the pinna geometry. This measurement process produces the direction-specific impulse responses that are later stored as an PRTF grid.

C. From Recorded Responses to PRTFs

At each direction (θ, ϕ), the acoustic response was calculated from the recorded signals in the frequency domain. The recorded output signal was transformed via FFT to obtain $Y(\theta, \phi)$, while the input signal was transformed to obtain X. For each direction, five repeated recordings were taken, and the output spectra were averaged to reduce noise. The transfer function was then calculated as $H(\theta, \phi) = \frac{\overline{Y}(\theta,\phi)}{X}$, where $\overline{Y}(\theta, \phi)$ is the average recorded output spectrum. This gives the frequency-domain response of the pinna and microphone system for that direction. The corresponding impulse response can then be obtained by taking the inverse FFT of $H(\theta, \phi)$ [1], [14].

The final dataset is stored as a grid of direction-labeled responses. In the experiments below, the grid contains 632 directions. The responses are sampled at 48 kHz and contain 480 samples per impulse response. Azimuth spans $0°$ to $340°$, and elevation spans $-90°$ to $90°$. Only the single-microphone response is used, so the localization task is strictly monaural.

V. Spectral Analysis of PRTFs

The measured PRTFs provide a direct way to inspect whether the pinna creates useful directional structure. Figure 3 shows the spatial sampling grid, example horizontal-plane spectra, a direction-versus-frequency heat map, and example impulse responses. The most important observation is that the PRTFs are not identical across direction. Their spectral peaks and notches shift as the source direction changes. For a monaural system, these moving spectral notches are the main cue: the model must learn how a particular spectral pattern corresponds to a particular source direction.

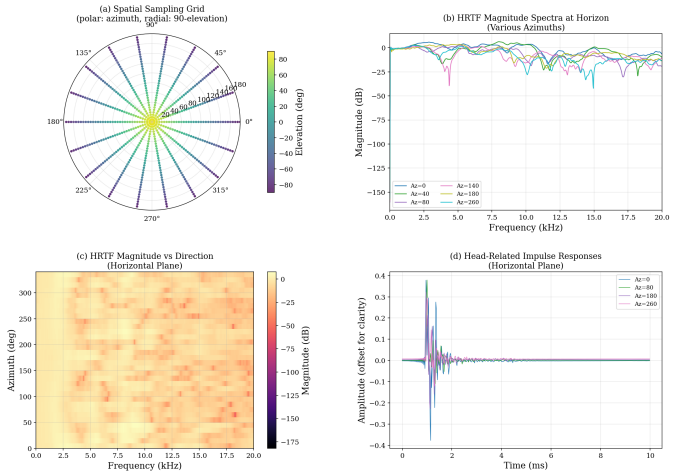

Fig. 3. Directional structure in the artificial-pinna PRTFs. The measured direction grid defines the localization task, while the spectra and heat map show that different source directions produce distinct frequency-domain patterns. For a one-microphone system, these direction-dependent spectral patterns are the key localization cue. The heat map illustrates how the artificial pinna introduces direction-dependent spectral notches and peaks across frequency and azimuth, which provide monaural localization information. The corresponding time-domain PRTFs for multiple directions reveals the impulse responses that give rise to these characteristic directional filters.

VI. Dataset and Signal Generation

The localization experiments use a SOFA-format PRTF grid with 632 source directions. SOFA provides a standardized format for storing spatial acoustic data such as PRTFs [14]. Each direction has an associated impulse response for the artificial pinna-microphone system. Because the experiment uses only one channel, each example is treated as a monaural recording.

For each of the 632 directions, source signals are generated by convolving with the direction-specific impulse response. The training set contains 1000 examples per direction, giving 6320 training examples. The validation set contains 8 examples per direction, giving 5056 validation examples.

The source signals are randomized broadband signals drawn from three families: white noise, pinkish noise, and Gaussian band-limited noise. After convolution with the PRTF, each waveform is normalized to a target RMS level of -20 dBFS, where dBFS denotes decibels relative to full scale. Thus, the target RMS amplitude is 20 dB below the maximum representable digital amplitude. During training, the waveform was augmented with random gain jitter of ± 3 dB and additive white noise with a randomly sampled SNR. Specifically, the SNR was drawn uniformly from a range of 20–40 dB for each training example, rather than being fixed at 30 dB. These variations prevent the model from memorizing a single source waveform and encourage it to focus on direction-dependent filtering.

The validation protocol uses the same direction grid as training but different source realizations. This means the experiment tests whether the decoder can generalize across different input sounds for known measured directions. It does not yet test interpolation to unseen directions or robustness in real rooms.

VII. Localization Decoder

The neural network is used as a decoder for measured acoustic fingerprints. The goal is not to claim that the network alone solves the localization. Instead, the model tests whether the artificial pinna has produced monaural patterns that are consistent enough to decode.

Each waveform is converted into a four-channel time-frequency tensor of size $4 \times 64 \times 256$. The raw training signals were generally 6.0 s long, although they were not required to have exactly identical durations before preprocessing. After convolution with the HRIRs and feature extraction, all examples were standardized to a fixed feature length of $T_{\text{target}} = 256$ time frames. With an STFT hop size of 256 samples, this corresponds to an effective analyzed duration of approximately 1.37 s per feature tensor. This standardization ensured that every input to the model had the same size.

The short-time Fourier transform used $N_{\text{FFT}} = 1024$ with a hop size of 256 samples. The resulting spectra were then projected onto 64 mel-frequency bins, which convert the linear frequency axis in Hertz to a perceptually motivated, approximately logarithmic frequency scale. The four feature channels are

$$\mathbf{F} = [\log |M|,\ \Delta \log |M|,\ \cos \angle M,\ \sin \angle M]\,, \qquad (1)$$

where M is the complex mel spectrogram. The log-magnitude channels describe spectral shape, while the phase channels preserve circular phase information without creating a discontinuity at $\pm\pi$.

The decoder, called *PinnaResNet*, is a residual convolutional neural network with anisotropic kernels, squeeze-and-excitation attention, and SpecAugment-style masking during training. Residual convolutional networks and squeeze-and-excitation attention have been widely used to improve deep feature learning, while SpecAugment is a common spectrogram-domain augmentation method [10]–[12]. Rather than predicting azimuth and elevation directly, the model predicts a 3-D unit vector on the sphere. This avoids the wraparound problem at $0^\circ/360^\circ$ and makes the training objective correspond directly to angular similarity. The model is trained using cosine distance between the predicted unit vector and the ground-truth direction vector.

Training uses AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 32, gradient clipping at 5.0, and cosine-annealing warm restarts. AdamW decouples weight decay from the adaptive gradient update, making it a common optimizer choice for deep neural networks [13]. Early stopping selects the best checkpoint, which occurs at epoch 342.

VIII. Evaluation Metrics

The primary metric is great-circle mean angular error (GC-MAE), which measures the shortest angular distance between the predicted and true directions on the sphere. We also report median angular error, 95th-percentile error, azimuth mean absolute error, elevation mean absolute error, and the percentage of examples within 10°.

Because the PRTFs lie on a discrete measurement grid, the predictions are also evaluated as bucket classifications. Azimuth predictions are assigned to the nearest measured azimuth bucket, and elevation predictions are assigned to the nearest measured elevation bucket. Azimuth bucket accuracy counts whether the predicted azimuth bucket matches the true azimuth bucket. The Az–El bucket accuracy counts a prediction as correct only when both the azimuth and elevation buckets match the true measured direction.

IX. Results

A. Overall Localization Performance

Table II gives the main validation result. The system achieves 8.79° mean great-circle error and 3.44° median error. The gap between the mean and median indicates that most predictions are close to the true direction, while a smaller number of difficult cases create a long error tail. The 95th-percentile error of 43.4° measures that tail directly.

The model also performs far above chance on the discrete grid. Az–El bucket accuracy is 48.7% over 632 possible measured directions, while uniform chance would be approximately 0.15%. The azimuth bucket accuracy is 75.1%, showing that the decoder often recovers the correct measured azimuth bucket even when the full azimuth-elevation direction is not exactly matched.

TABLE II
PRIMARY VALIDATION METRICS. THE VALIDATION SPLIT USES THE SAME 632 DIRECTIONS AS TRAINING BUT NEW SOURCE REALIZATIONS, SO THE TEST MEASURES SOURCE INVARIANCE ON A FIXED SPATIAL GRID.

Set	GC-MAE	Median	P95	Az MAE	El MAE	$< 10^\circ$	Az bucket	Az–El bucket
Validation	8.61	3.70	43.4	13.14	5.11	86.7%	75.1%	48.7%

B. Local Error Structure

Average error alone does not show whether failures are random or local. Figure 4 and Figure 5 show that predictions are concentrated near the diagonal. This means that when the model misses the exact bucket, it often predicts a neighboring or nearby direction.

Elevation errors are especially local. The strongest elevation confusions are mostly one-bin slips between adjacent 5° bins. Azimuth errors are broader, which matches the larger azimuth MAE reported in Table II. This suggests that the pinna and decoder capture vertical spectral structure more reliably than horizontal structure in the current setup.

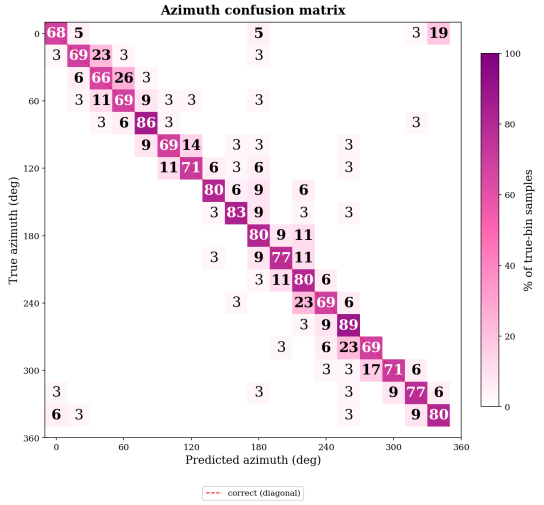

Fig. 4. Azimuth bucket confusion matrix. Rows correspond to true azimuth bins and columns correspond to predicted azimuth bins. Cell color indicates the percentage of samples from each true azimuth bin assigned to each predicted bin, with darker magenta indicating a larger percentage. The dashed red diagonal marks correct bin-level predictions.

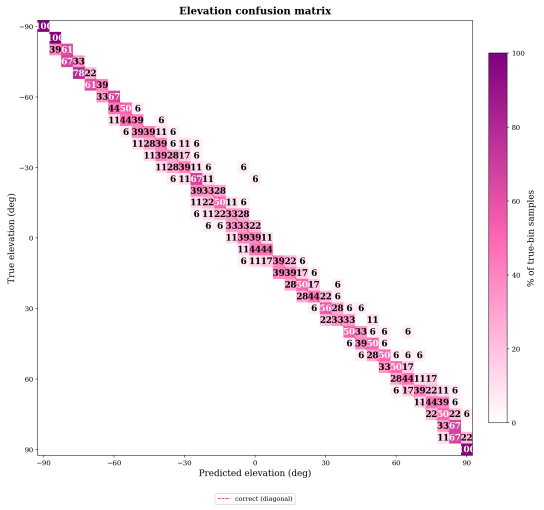

Fig. 5. Elevation bucket confusion matrix. Elevation predictions are tightly concentrated near the diagonal, indicating that most elevation mistakes are small neighboring-bin errors rather than large directional failures.

C. Spatial Distribution of Error

Figure 6 shows that the remaining errors are not uniformly distributed across space. The easiest elevation band is $[-90^\circ, -60^\circ)$, where GC-MAE falls to 6.66°. The hardest elevation band is $[30^\circ, 60^\circ)$, where GC-MAE rises to 10.77°. The per-sample map shows spatial pockets where localization is more difficult.

This pattern matters because it points back to the acoustic design. If errors were random, the spatial map would be

relatively unstructured. Instead, the difficult regions suggest directions where the PRTF cues are more ambiguous, less stable, or harder for the decoder to separate. These regions provide useful targets for future pinna redesign.

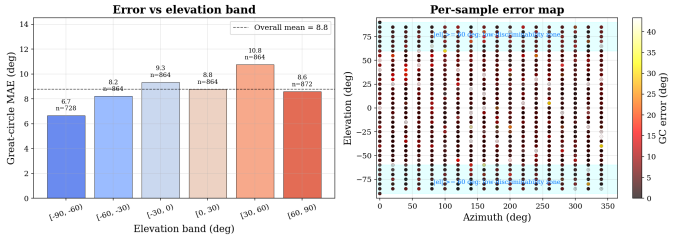

Fig. 6. Spatial distribution of localization error. Error varies by elevation band and by direction, showing that difficult cases cluster in specific spatial regions rather than being uniformly distributed over the sphere.

D. Four-Wall Analysis

The wall-specific results connect the localization performance back to the pinna geometry. The right sector is easiest at $8.0°$ GC-MAE, while the left sector is hardest at $9.2°$. However, the differences are modest. Azimuth bucket accuracy remains in a narrow range from 73.9% to 76.6%, and the Az–El bucket accuracy, which counts a prediction as correct only when both the azimuth and elevation buckets match the true measured direction, remains between 47.3% and 50.5%.

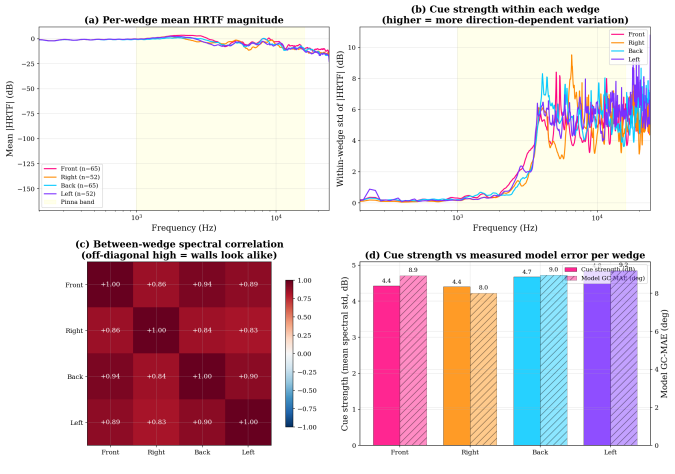

Fig. 7. Wall-sector PRTF analysis for the four artificial-pinna regions. The mean PRTF curves compare the average spectral response of the front, right, back, and left sectors. The cue-strength plot shows how much the PRTF varies within each sector as a function of frequency, with stronger variation indicating more direction-dependent acoustic information. The correlation matrix shows that the four sectors remain spectrally similar at a coarse level, while the final panel compares cue strength with model error. This is a results figure because it analyzes measured PRTF data and model behavior; it is not the physical pinna design image.

This suggests that the four-wall design creates directional asymmetry, but no single wall alone explains the system performance. Localization depends on the full combined acoustic field produced by the pinna. The wall analysis is still useful because it shows where future geometry changes might be tested.

TABLE III
FOUR-WALL VALIDATION METRICS. EACH VALIDATION EXAMPLE IS ASSIGNED TO THE FRONT, RIGHT, BACK, OR LEFT SECTOR OF THE ARTIFICIAL PINNA.

Wall	n	Az bucket	Az–El bucket	GC-MAE
Front	1416	76.6%	47.3%	8.9
Right	1120	75.7%	47.6%	8.0
Back	1400	73.9%	50.5%	9.0
Left	1120	74.3%	49.4%	9.2

X. Frequency-Domain Evidence

The strongest evidence that the decoder is using meaningful pinna cues comes from the frequency ablation in Figure 8. If the model were relying on an unrelated artifact of the dataset, there would be no clear reason for performance to follow the expected frequency behavior of pinna acoustics. Instead, performance is poor at low frequencies and much stronger at mid and high frequencies.

At roughly 500 Hz, the system is essentially unusable: GC-MAE is $90.3°$, median error is $90.2°$, and only 1.5% of samples fall within $10°$. At 1.5 kHz, performance is still poor at $48.8°$. Once the band reaches about 3 kHz, error drops sharply to $15.2°$, then to $9.3°$ near 6 kHz. A 10 kHz band remains informative at $15.7°$, and broadband mixtures perform best overall at $8.5°$.

This trend matches the physical interpretation of the pinna. At low frequencies, wavelengths are too large for the small wall geometry to create strong direction-dependent notches. At higher frequencies, wavelengths are short enough to interact with the pinna shape, producing spectral features that vary with direction. This is consistent with prior work showing that spectral detail is important for sound-source localization [2], [4]. The decoder performs best where those features are strongest.

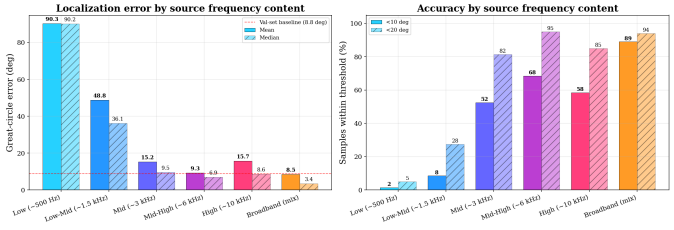

Fig. 8. Localization performance by source frequency content. Performance is weak at low frequencies, improves sharply around 3 kHz, is strongest near 6 kHz, and is best overall for broadband mixtures. This pattern supports the interpretation that the model is using pinna-driven spectral cues.

XI. Discussion

The main result is not simply that a neural network achieved low angular error. The full system demonstrates a plausible monaural localization mechanism: the artificial pinna introduces direction-dependent acoustic structure, the measurement process captures it as an PRTF grid, and the decoder recovers direction from new monaural examples.

Several findings support this interpretation. First, performance is far above chance on a dense 632-direction grid.

Second, the median error is much smaller than the mean, and the confusion matrices show mostly local mistakes. Third, the frequency ablation aligns with the expected physics of pinna cues: localization fails at low frequencies and improves in the mid-to-high-frequency range. Together, these observations suggest that the model is decoding the intended acoustic cue rather than memorizing arbitrary labels.

The results also reveal useful design limitations. Azimuth is harder than elevation, with an azimuth MAE of 13.14° compared with an elevation MAE of 5.11°. This may mean that the current pinna geometry creates stronger vertical spectral variation than horizontal variation. The spatial error map and wall-specific results point to regions where the geometry could be redesigned to create more distinctive cues.

XII. Ethical Considerations

This project is fundamentally a sensing technology, and like all sensing technologies it has both beneficial and potentially harmful applications that deserve explicit consideration.

Beneficial applications. The primary motivation for this work is to enable spatial hearing in devices where microphone arrays are impractical. Hearing-assistive devices and cochlear implant processors could use a single-microphone directional front end to help users localize speech in noisy environments, an area where many current devices perform poorly. Small robots and drones could use monaural direction estimation without the additional weight, power, and calibration complexity of multi-microphone arrays. Accessibility tools for people with single-sided deafness represent a particularly compelling use case, since a compact directional sensor could partially restore the monaural spatial cues that a biological pinna normally provides.

Privacy and surveillance risks. A compact, single-microphone direction-finding system could also be misused for covert acoustic surveillance. A device that can estimate the direction of a sound source from a single small sensor is, by design, harder to detect and deploy than a visible microphone array. The research itself does not create this risk, the underlying physics of pinna-shaped directional filtering is well known, but deployment decisions should consider the potential for misuse in monitoring or tracking applications without the knowledge of those being recorded.

Limitations and fairness. The current system was measured and evaluated in a controlled anechoic environment. Performance in real rooms with reverberation, background noise, and non-stationary sound sources is unknown. Any deployment in assistive technology or robotics should be preceded by testing across diverse acoustic environments and user contexts. Additionally, if the system is extended to person-tracking applications, careful attention to consent, transparency, and data governance would be required.

XIII. Limitations and Future Work

The current evaluation is controlled and should not be overstated. The validation set uses the same 632 directions as training, with different source realizations. This tests source invariance on a fixed spatial grid, but it does not prove interpolation to unseen continuous directions. Future experiments should hold out directions during training and evaluate whether the decoder can generalize between measured positions.

The current experiments also rely on a fixed SOFA-format response set. A complete real-world demonstration should include physical recordings from the fabricated pinna, measured hardware variation, background noise, and room reverberation. Real rooms add reflections that may interfere with the pinna cues measured in the anechoic chamber. Robust deployment would require testing in multiple rooms and with realistic sound classes such as speech, impacts, or alarm operating noises.

Future work should also use the wall-specific and frequency-specific results to redesign the pinna. The left and back sectors could be modified to create stronger or more stable directional notches. Additional ridges, cavities, or depth changes could be introduced to increase azimuth separability. The design process could then become iterative: fabricate a geometry, measure its PRTF grid, evaluate localization, identify weak regions, and redesign the acoustic structure.

XIV. Conclusion

This work shows that artificial pinna geometry can provide useful directional information to a single microphone. The four-wall pinna acts as a passive acoustic encoder: sounds from different directions interact with the walls differently, creating direction-dependent spectral patterns. By measuring these patterns in an anechoic chamber with a turntable-based PRTF process, the system obtains a controlled dataset that captures the spatial filtering created by the pinna.

The neural network results show that these measured cues are learnable. On the controlled 632-direction grid, the system achieves 8.61° mean great-circle error, 3.70° median error, and 86.7% accuracy within 10°. The frequency analysis further supports the physical interpretation, since performance is strongest in the mid-to-high-frequency range where pinna geometry is expected to produce distinctive spectral notches.

Future work should extend this design–measure–decode pipeline to unseen directions, richer sound classes, real recordings, and realistic acoustic environments.

References

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1997, revised edition.

[2] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–2200, 1990.

[3] A. Saxena and A. Y. Ng, "Learning sound location from a single microphone," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2009, pp. 1737–1742.

[4] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, no. 6713, pp. 747–749, 1998.

[5] A. Francl and J. H. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments," *Nature Human Behaviour*, vol. 6, pp. 111–133, 2022.

[6] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001, pp. 99–102.

[7] J. G. Harris, C.-J. Pu, and J. C. Principe, "A monaural cue sound localizer," *Analog Integrated Circuits and Signal Processing*, vol. 23, pp. 149–163, 2000.

[8] A. Deleforge, R. Horaud, Y. Schechner, and L. Girin, "Direction of arrival with one microphone, a few LEGOs, and non-negative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[9] X. Sun, H. Jia, Z. Zhang, Y. Yang, Z. Sun, and J. Yang, "Sound localization and separation in 3D space using a single microphone with a metamaterial enclosure," *Advanced Science*, vol. 7, no. 3, p. 1902271, 2020.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[12] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of Interspeech*, 2019, pp. 2613–2617.

[13] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[14] P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, H. Ziegelwanger, and M. Noisternig, "Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions," in *Proceedings of the AES Convention*, 2013.