EarCase: Sound Source Localization Leveraging Mini Acoustic Structure Equipped Phone Cases for Hearing-challenged People

Xin Li*, Yilin Yang*, Zhengkun Ye†, Yan Wang†, Yingying Chen*
*Rutgers University, †Temple University

ABSTRACT

Sound source localization is vital for daily tasks such as communication or navigating environments. However, millions of adults struggle with hearing impairment, which limits their ability to identify the direction and distance of sound sources. Traditional methods for sound spatial sensing, such as microphone arrays, are not suitable for resource-constrained IoT devices like smartphones due to power consumption or hardware complexity. To overcome these limitations, this paper proposes EarCase, an alternative scheme that utilizes commercial smartphones with only two microphones to recognize 3D acoustic spatial information. EarCase draws inspiration from the human auditory system, where two ears amplify minute differences in acoustic signals to help pinpoint sound sources. This ability can be regarded as a response function trained through a large amount of sound source information, which can be used to extract spectral cues from a sound source position to the ears drums. We imitate this effect by designing a smartphone case with perforated mini-structures covering the microphones to help the smartphone infer the location of the sound source. Sound waves that pass through the mini-structure will undergo unique changes in diffraction at the hole, amplifying directional information similar to ears. Our scheme uses the top and bottom microphones to eliminate noises and multi-path effects, making the design robust to different sound sources in varying environments. By using only built-in microphones and low-cost phone cases, EarCase provides an accessible tool to enhance the quality of life for hearing impaired individuals. Extensive experimental results show that EarCase achieves high accuracy in localizing sounds, with a mean error of 3.7° at a distance of 200cm and 96% accuracy for real-world sounds (e.g., car horns).

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing**.

KEYWORDS

Sound Source Localization, Acoustic Sensing, Smartphone Cases

ACM Reference Format:
Xin Li*, Yilin Yang*, Zhengkun Ye†, Yan Wang†, Yingying Chen*. 2023. EarCase: Sound Source Localization Leveraging Mini Acoustic Structure Equipped Phone Cases for Hearing-challenged People. In *The Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '23), October 23–26, 2023, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3565287.3610270

MobiHoc '23, October 23–26, 2023, Washington, DC, USA

ACM ISBN 978-1-4503-9926-5/23/10…$15.00
https://doi.org/10.1145/3565287.3610270

1 INTRODUCTION

The NIH estimates approximately 15% of American adults (37.5 million) aged 18 and over report some trouble hearing [5], causing unique challenges in their daily lives, such as avoiding dangers (e.g., honking car) or finding misplaced items (e.g., IoT devices). These people usually use hearing aid devices to help them perceive sound. However, existing hearing aid devices do not provide spatial information due to constrained form factors and computational power, resulting in "I can hear you, but I can't find you". We envision that a low-cost sound source localization solution can address this limitation, providing an affordable and accessible option for people who are hard of hearing. Such technology can significantly enhance the quality and safety of users' lives. In this work, we seek to explore the potential of using commercial-off-the-shelf (COTS) smartphones for sound source localization and how this technology can benefit users who are hard of hearing in their daily lives.

Existing Work and Limitations. Sound source localization has been used in many applications, such as speech localization [14], indoor ranging [25], and authentication [36]. These approaches rely on time difference of arrival (TDOA) and the direction of arrival (DOA) calculation using distributed networks of external microphones, which incurs high energy costs and complex system design [12]. Traditional microphone array based applications also use TDOAs of multiple microphones to determine the sound source. For instance, two microphone linear arrays are used to locate the sound source in an indoor environment by using the generalized cross-correlation algorithm to calculate the TDOA [11]. Recently, researchers proposed leveraging multiple smartphone microphones to find objects (e.g., earbuds, keys, or wallets) attached with sound-emitting tags within a short distance [37]. However, the existing methods can only localize sound sources at a close distance on a horizontal plane (2D), which restricts the applications in complex environments (e.g., a room with multiple obstacles). All these findings motivate us to develop a new approach that does not compromise on distance, accuracy, or hardware complexity.

Acoustics of Smartphone Case Mini-structures. We are inspired by the Head-Related Transfer Functions (HRTF) of the human auditory system [9], where fine-grained differences in spectral cues are distinguishable due to the structural shapes of the ear. We thus consider the possibility of enhancing the acoustic sensitivity of microphones by imitating HRTF on smartphones using well-designed acoustic structures. Toward this end, we observe that the use of protective cases for smartphones is common in daily

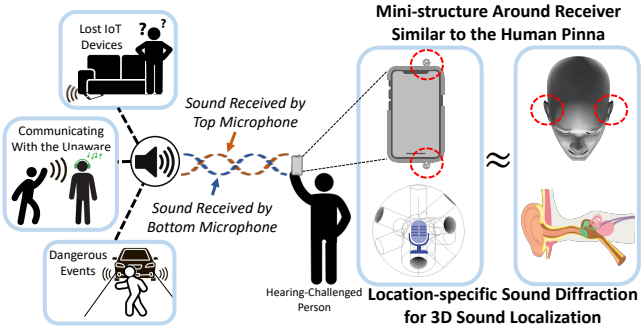


Figure 1: EarCase provides 3D sound source localization for hearing-challenged people leveraging smartphone cases with mini-structures similar to the human pinna.

life [18]. These cases can be highly varied in shape, making them easy to adopt stencils shielding the microphones. By embedding *mini-structures* with holes or tubes to filter sounds into these stencils, we induce diffraction and effectively treat the holes as new sound sources. Consequentially, the sound received in the stencils has rich multi-path effects that amplify the sound and create unique patterns correlated to the sound location and frequency, which can be leveraged to localize the sound with high accuracy. We noticed that existing work [15, 21, 32] has tried to design structures (e.g., holes, tubes, blockages, and cavities) to enhance acoustic sensing for sound angle inference, user interfaces, and acoustic filters. However, none of them can enable 3D sound source localization by using the smartphone with only two microphones.

In this paper, we present a novel acoustic sound inference system that introduces a smartphone case with mini-structures around the two microphones of the smartphone to embed directional cues, named EarCase. Similar to the binaural localization method of human ears, EarCase adds special physical structures with holes and tubes around two microphones to imitate the HRTF and change the frequency response of sound waves reaching the smartphone, as depicted in Figure 1. Such wave conduction is called diffraction, which is common in our daily life. Based on these observations, we explore various 3D-printed mini-structures to generate unique acoustic patterns to smartphone microphones when sound comes from different locations. We design distinct mini-structures (e.g., distribution of the holes) in smartphone cases for each microphone to allow signal separation and facilitate unique patterns for different locations. Such acoustic patterns are robust to variations in sound source signals and highly accurate at localizing sound sources in various environments. Moreover, our mini-structures are low-cost as they do not require additional sensing devices (e.g., microphone arrays), needing only a smartphone case producible at ~$5 USD.

The major challenge faced by EarCase is designing a case that can consistently generate distinct patterns for the sound from different spatial locations regardless of environments and types of sound. In other words, the system should work by design, and the user does not need to calibrate or train the system in practice. In addition, the reflected sound by the environment will cause the multi-path effect, which will also impact the frequency response. For instance, in a room-scale sensing area, the acoustic wave will interact with others physical structures (e.g., reflected by the wall) so that even the same sound source from different locations shows differences. To mitigate the multi-path effect caused by different room layouts, EarCase leverages the top and bottom microphones as each other's reference, which effectively captures similar multi-path effects due to their close distance, and cancels the multi-path effect during pattern generation. Although we can eliminate the impact of the environment, it's still hard to collect comprehensive samples in different environments. Thus, we conduct rigorous acoustic analyses to generate fine-grained location-related acoustic patterns for training effective localization models without the heavy burden of manual data collection.

This paper introduces novel acoustic mini-structures around two microphones as a pair of ears, which provides fine-grained sound spatial information in various applications for ubiquitous sensing, e.g., helping people who are hard of hearing localize acoustic alerts to avoid danger. We studied the feasibility of achieving such sound source localization and made the following contributions:

- We design a novel acoustic sound inference system, EarCase, that localizes the sound source by introducing mini-structures in a smartphone case instead of microphone arrays. Sound propagation through the mini-structures embeds information that is specific to the sound source location, allowing us to identify the sound source location without traditional DoA measurements.
- We extensively study mini-structures and their unique effects on the frequency response that can facilitate sound source localization. We develop a method based on the Rayleigh-Sommerfeld equation to simulate the diffraction and capillary effects in small physical metamaterial structures and help us design the mini-structures in smartphone cases.
- We design an interference removal method to eliminate environmental interference using 3D-printed mini-structures based on the controlled diffraction in small physical metamaterial structures to ensure EarCase remains robust.
- We conduct extensive experiments with multiple smartphone case designs and smartphones. Results demonstrate that EarCase can achieve a mean error of 3.7° at a distance of 200cm, while it is robust under varying real-world factors, including background noises and room layouts.

2 RELATED WORK

Sound Source Localization Applications. Sound source localization (SSL) systems focus on finding the direction of a sound source. SSL has multitudinous practical applications, for instance, in source separation [8], automatic speech recognition (ASR) [22], speech enhancement [34], human-robot interaction [24], noise control [10], and room acoustic analysis [2]. In addition, SSL has been utilized to aid people with mild-to-moderate hearing loss to segregate multiple talkers and understand speech in a noisy environment [3].

Direction of Arrival Estimation. Stefanakis *et al.* [29] present a modification in the propagation model and used it for 2D direction-of-arrival (DOA) estimation in the half-space. Badawy *et al.* [13] acquire DoA results with one microphone and non-negative matrix factorization. Cagli *et al.* [6] address the DOA estimation problem for speech signals via sparsity models using MEMS microphone arrays. As the most closely related work, Garg *et al.* propose Owlet, an approach for acoustic DoA estimation and source localization using 3D-printed metamaterial structure that covers the microphone.

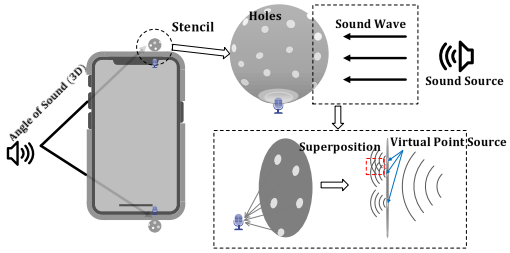


Figure 2: Passive directional filtering using acoustic mini-structures built into smartphone cases. The stencil envelops the microphone to embed direction-specific information in the frequency response.

Acoustic Sensing Using Physical Structures. The impact of structures on the sound field has been broadly studied, and acoustic metamaterials have recently gained attention for their ability to attenuate sound. Casarini *et al.* [7] present subwavelength small-scale 3D printed acoustic metamaterials based on Helmholtz resonators capable of generating stop bands where the sound is attenuated. Li *et al.* [23] present a computational approach that automates the design of acoustic filters comprised of a set of parameterized shape primitives to satisfy target acoustic properties. Priyantha *et al.* [27] use coded ultrasound pulses to locate and identify users in an instrumented room. Harrison *et al.* [16] exhibit acoustic barcodes, an identifying tag that uses notches to produce sound when dragged across. Although acoustic structures are a long-standing and widely researched topic, it remains a very challenging problem to date, and many works demonstrate feasibility of using structures, but none have been applied to localization on mobile devices.

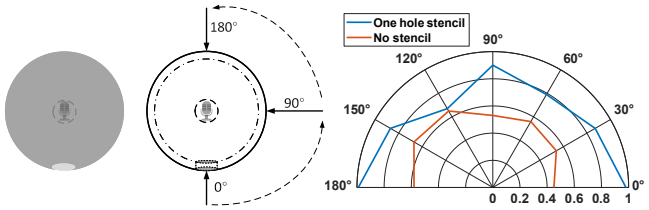
3 BACKGROUND AND PRELIMINARY

3.1 Analogue to Head-related Transfer Function

Our ears have a very special shape, with complex folds and protrusions that allow the original sound to be superimposed on the sound reflected by the pinna. Such sound reflections increase the sound propagation time, causing a difference in the phase of the signal. As a result, some sound frequencies will be enhanced while others will be weakened. The influence of pinna on the frequency characteristics of the sound can be described by a function named Head-related Transfer Function (HRTF) [35]. HRTF is a function of sound location. Sounds from different places will have different frequency response characteristics after passing through the HRTF. Human brains gradually learn the HRTF for each ear, which enables us to recognize the locations of sounds. Inspired by the function of the pinna, we propose to develop mini-structures around the microphones in smartphone cases, generating unique frequency characteristics of the sound from different locations as an analogue to HRTF so that the smartphone can recognize sound locations using two built-in microphones as human ears.

3.2 Acoustics of 3D-Printed Mini-structures

Acoustic mini-structures [21] can impact and regulate sound waves along the sound propagation path. These mini-structures rely on their specially crafted designs, rather than the characteristics of the underlying 3D-printed materials, to achieve their sound controls. Through the use of miniature holes, tubes, and blocks, these



(a) Top-down view of a 1-hole stencil (b) Sound amplitude at different angles from the direction of sound source

Figure 3: Feasibility study of the diffraction of wave near the mini-structure.

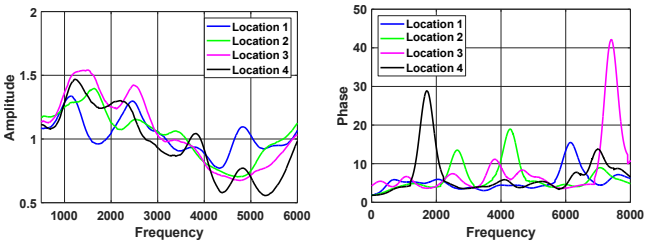
structures possess unique properties that allow them to manipulate sound waves. When sound interacts with the mini-structures, the signal response can be either amplified or attenuated. Similar phenomena (i.e., frequency variations) can be observed from multi-path reflections due to constructive and destructive interference on a large scale (e.g., room size). In order to achieve fine-grained sound source localization using smartphones, we design smartphone cases with mini-structures based on the principles according to two phenomena: *diffraction* and *capillary effects*.

Diffraction. When a sound wave encounters a small opening (hole) with an aperture smaller or comparable to the wavelength of the sound, it diffracts at the hole and acts as a virtual sound source [4]. Assuming the sound passes through a barrier of multiple holes, each diffracted sound becomes a new sound source. The sound received on the other side of the barrier combines all the diffracted sounds, creating a unique pattern of constructive and destructive interference similar to the multi-path effect, which is dominated by the receiver's location and the sound's frequency. Figure 2 shows the details about the virtual point sound sources and the superposition of each new sound source. To facilitate sound source localization using mini-structure-enabled smartphone cases, we employ diverse patterns of small holes around the designated mini-structures, where the diffraction of the sound coming from different directions has distinct multi-path effects.

Capillary Effect. Acoustic impedance is known to change when sound propagates through a small capillary tube [26]. Furthermore, the length and cross-sectional area of the tubes affects the speed of sound transmission. Therefore, to increase the diversity of the frequency spectrum of the diffracted sound through the mini-structure-enabled smartphone case, we design capillary tubes connecting the open holes in the mini-structure. With such designs, the transmission time of the sound coming from different directions is distinct despite the small tube length, leading to unique frequency shifts and enhanced diversity [31].

3.3 Feasibility Study

Diffraction of Wave Near the Mini-structure. We verify the diffraction by using a sphere shape stencil design that has only one hole in one direction, shown in Figure 3(a). The top and bottom microphone stencils have their holes facing opposite directions. We place the sound source 200cm away from the smartphone and play a sound with 7kHz fixed frequency at 0° to 90° for every 30°. We measured the sound pressure received by two microphones of the smartphone, where the sampling rate is 16kHz. Figure 3(b) demonstrates that there will be a substantial increase in sound



(a) Amplitude of different locations (b) Phase of different locations

Figure 4: Diversity in frequency responses (amplitude and phase) of 3D-printed mini-structure.

pressure even when the hole is over 90 degrees away from the direction of the sound source, indicating that the sound waves are bending. Moreover, it also shows that there is a high level of sound pressure directly opposite the incoming sound. This is a result of the sound fields from both sides of the sphere merging.

Sound Location Diversity Through the Mini-structure. In order to verify multiple holes and tubes can provide diversity in the sound frequency responses, we use a 3D-printed smartphone case with multiple embedded mini-structures. We place the sound source at four different locations far away from the smartphone and play a sound with frequency band from 1Hz to 8kHz. We revive the sound signal and obtain the frequency and phase domain of signal. The results show the sound from the different locations can provide distinct frequencies responses, as shown in Figure 4.

4 SYSTEM OVERVIEW

Our system is designed to recognize subtle differences in distances, angles, and heights of the sound source without the aid of external sensors in the environment or intensive calibration setups. We consider example sound sources such as signals emitted by devices (e.g., smartphones, earbuds, car horns) as well as human utterances (e.g., "Excuse me!"), which may be difficult for people to notice due to hearing loss or low attentiveness caused by music or earmuffs. To provide accessible and accurate sound localization information to the user, we develop an acoustic signal processing architecture for smartphones and smartphone cases, shown in Figure 5.

Sound Data Pre-processing. Microphones convert sound pressure to electrical signals, represented in the time-domain as oscillations in voltage. We apply smoothing and segmentation algorithms to mitigate interference in the recordings. Presence of the sound source is detected using a peak-energy calculation across a fixed time window computed one second. The specific window length can be varied depending on the target sound source and sampling rate supported by the microphones.

Location-specific Pattern Extraction. Once detected, the sound is then filtered for patterns can be used to reveal the sound source location in 3D space. To ensure training is a one-time process only, our data must be location-specific without being sound-specific or environment-specific. By using two microphones, we always detect the sound source at one side of the smartphone before the other. We therefore leverage the stereophonic properties of our recordings and filter out any information observed by both microphones at the same point in time. This process guarantees the filtered output consists only of direction-based location information. We then

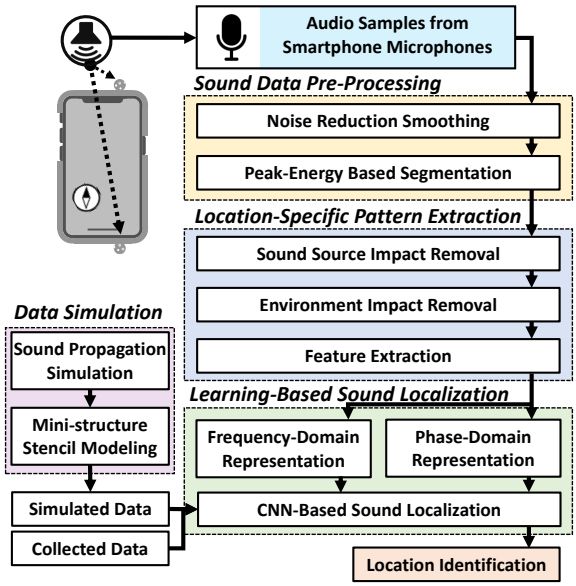


Figure 5: EarCase Overview.

convert the output to the frequency-domain and phase-domain to obtain directional frequency responses.

Learning-Based Sound Localization. EarCase uses a one-time training process at the developer-side to learn the relationship between the frequency responses produced by the two smartphone microphones. No effort is required from the user. Performance is dependent on the granularity of training data collected, with fine-grained samples (i.e., $1°$, $1cm$) yielding the best precision. This data can be sourced from real-world microphone recordings or simulations of sound propagation. We build a neural network model and provide our extracted location-specific features for training. We supply this data to a CNN neural network for training. The output labels correspond to the predicted angle of the sound. This trained model can later be downloaded by users (e.g., mobile app store) to localize new sounds in their own local environments.

5 SOUND SOURCE LOCALIZATION

5.1 Mini-structure Design

5.1.1 Sound Diffraction Caused by the Designed Mini-structures. The diffraction pattern for a collective aperture of multiple holes can be accurately derived by solving the Helmholtz wave equation. Less computationally expensive formulations with Fresnel or Fraunhofer approximations often assume far-field separation between the source and observation-plane. However, our application requires a smaller arrangement where the distance between the planes are comparable to one wavelength of the lowest frequency. Our experiments show poor performance for the far-field or planar wavefront approximation in this setup. Therefore, we assume spherical wavefronts for the sounds from the virtual sources and adapt the following Rayleigh-Sommerfeld equation [33] to calculate the resultant sound field on the observation plane.

$$U(x,y) = \iint_{\Sigma} U_1(x',y') h\left(x-x', y-y'\right)\, \mathrm{d}x'\mathrm{d}y', \tag{1}$$

where $h(x,y) = \frac{D}{j\lambda r^2} e^{\frac{j2\pi r}{\lambda}}$ represents the response on the diffraction screen at unit amplitude, called the impulse response or point

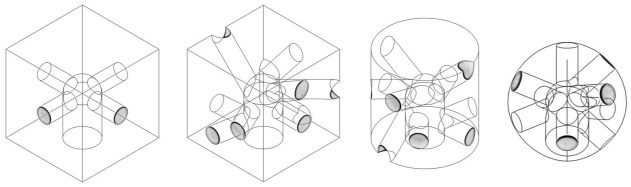


Figure 6: Iterative prototyping of EarCase designs from initial concepts (left) to finalized design (right).

source spread function and $r = \sqrt{x^2 + y^2 + D^2}$. The diffraction pattern is a function of the frequency. A microphone placed at a specific location on the observation plane will experience phase and amplitude of a particular frequency as per the diffraction pattern corresponding to that frequency. This means, a microphone separated by a stencil of holes will see different complex gains for a set of frequencies present in the incoming sound. We change parameters of Eq. (1) as follows to formulate the gain-pattern of the frequency spectrum for a given hole-pattern:

$$H(\lambda) = \frac{D}{j\lambda} \sum_n \frac{e^{\frac{j2\pi r_n}{\lambda}}}{r_n^2}. \tag{2}$$

This formulation will guide us to simulate data in the Section 6.2.

5.1.2 Mini-structure Design for Stencils of Smartphone Case. We influence the acoustic patterns received by the device by controlling sound propagation through a mini-structure consisting of a perforated stencil surrounding the microphone. Conventional localization via TDOA calculations approximates the acoustic signal as a single wave arriving at the microphone, which cannot sufficiently capture fine-grained height, distance, or angles. We therefore diversify the sound patterns received at microphones by blocking direct propagation paths with a stencil barrier and strategically distributing holes or entryways to simulate the arrival of multiple pseudo-sound-sources. The staggered arrival of these sounds and their altered diffraction and/or capillary effects within the stencil provides rich acoustic signatures indicative of the direction and position of the actual sound source.

To understand the constraints and ideal design qualities, we conduct iterative tests by prototyping different mini-structures with varied parameters. This data-driven trial-and-error approach progressively optimized our mini-structure into the final design shown in Figure 6. Our initial concept utilizes a cubical structure with 4 holes evenly distributed leading to an internal tube connected to the microphone. By increasing the number of holes, we theoretically intensify the capillary effect. However, we find the flat surfaces of the cube to introduce minimal changes as holes of the same plane are too similar in position and angle. Therefore, we experiment in rounding the surface into a cylinder before settling on a spherical design, which allows every hole to have a unique position and arrival time of the sound wave. We set the diameter of our holes at 2mm to allow as many holes as possible without being too small to manufacture.

5.2 Location-specific Pattern Extraction

In order to understand how EarCase infers the sound location, we use simple terms to model this process, shown in Figure 7. First, we assume that the sound source signal without environmental effects

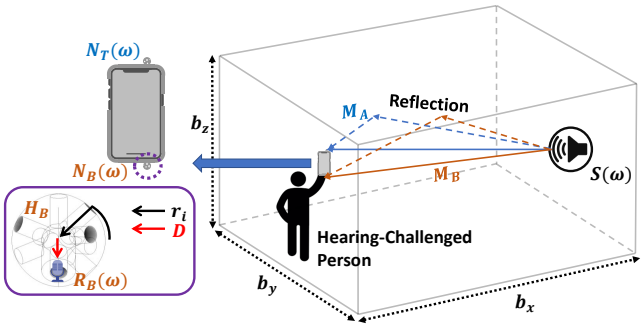


Figure 7: Sound modeling in a room with $6m \times 5m \times 3m$ based on the idea of Rayleigh-Sommerfeld diffraction model.

is $S(\omega)$ where ω is the frequency. To generate location-specific patterns P_l of the mini-structure for location l, we send signals $S(\omega)$ with the wide frequency-band from various locations and use a fixed-location smartphone with two microphones covered by mini-structures to record signals. The sound received by the top and bottom microphone at location l can be define as $R_T^l(\omega)$ and $R_B^l(\omega)$. The set of location-specific patterns $\mathbb{P}$ of the mini-structure for different locations derived from $R_T(\omega)$ and $R_B(\omega)$ can be used for later sound source localization. We process the received signal to extract the unknown specific pattern P' and map this pattern to the pre-collected specific patterns $\mathbb{P}$ to obtain location information. However, in reality we can only obtain the complex signal that contains the original sound source and additional signal superposition introduced by the environments. Thus, if we want to pre-train a model to predict the sound location, we need to extract the location-specific pattern P introduced by the mini-structure from the complex signal.

5.2.1 Sound Source Independence. In most scenarios, the sound source signal is unknown, so we cannot extract the special pattern caused by the stencil mini-structures from the received signal. Ideally, no matter what sound source is used, we can always find a special pattern generated by the stencil. When the sound propagates from the source to the two microphones, the received signal will be impacted by the multi-path effect from nearby objects called channel frequency responses (CFR). We assume the CFR at the top and bottom microphone are M_T and M_B. At the two channels, there will exist independent noise defined as $N_T(\omega)$ and $N_B(\omega)$ at the frequency ω. The $R_T(\omega)$ and $R_B(\omega)$ can be expressed in the frequency domain as the following:

$$\begin{aligned} R_T(\omega) &= S(\omega)M_T H_T + N_T(\omega), \\ R_B(\omega) &= S(\omega)M_B H_B + N_B(\omega), \end{aligned} \tag{3}$$

where H_T and H_B is the frequency response caused by the interaction with the top and bottom stencil. Our goal is to use the received signal $R_T(\omega)$ and $R_B(\omega)$ to derive the location-specific pattern H_T and H_B. However, sound source localization is used in many complex environments, e.g., localize a car honking on the road or finding misplaced beeping IoT devices for people who are hard of hearing. In such scenarios, the sound source is unknown and background noise is also hard to detect. Luckily, we have the top and bottom microphone, which give us a chance to derive the special pattern.

In our scenario, the distance between two microphones is small and fixed. We can assume $N_T(\omega) \approx N_B(\omega)$. Then, if we want to

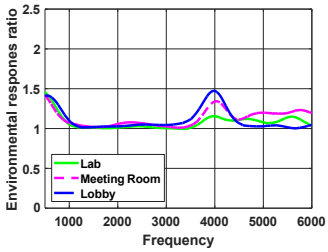


Figure 8: Environment ratio under different environments.

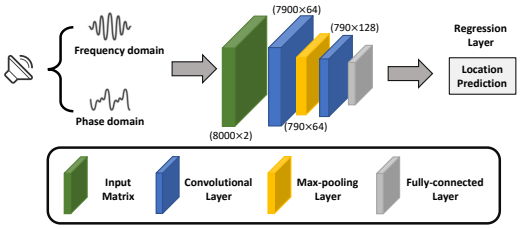


Figure 9: Sound localization on a CNN-based model.

eliminate them, we can consider two cases: 1) If the independent noise $N_T(\omega)$ is considerable to $S(\omega)M_T H_T$, it means that there is a strong noise near the smartphone microphone, meaning we can adopt processing techniques to reduce this noise. 2) Then, we can assume $N(\omega) \ll S(\omega)M_T H_T$. If we divide $R_T(\omega)$ and $R_B(\omega)$, it will remove the independent noise by using the theorem $\frac{A+\varepsilon}{B+\varepsilon} \approx \frac{A}{B}, \varepsilon \ll A$ (or B) and eliminate the sound source $S(\omega)$. What remains are:

$$\frac{R_T(\omega)}{R_B(\omega)} \approx \frac{M_T(\omega)}{M_B(\omega)} \cdot \frac{H_T(\omega)}{H_B(\omega)}. \tag{4}$$

CFRs caused by the environment still remain in the term $\frac{M_T(\omega)}{M_B(\omega)}$.

5.2.2 Environment Independence. In our scenario, M_T and M_B vary with the environment. However, we only need a one-time calibration of the stencil in lab and do not require collecting any data at the target environment. Because microphones of the smartphone are close (13cm), the path difference between the two microphones may not be very sensitive to changes in the environment. In other words, although $M_T(\omega)$ and $M_B(\omega)$ will be different in various environments, the ratio $\frac{M_T(\omega)}{M_B(\omega)}$ should be bounded within a range. As we can see, when sound waves transmit from the source and propagate through the air, they reflect off objects, causing different path lengths. These reflected components add together with the no reflection path waves at the microphone, resulting in a unique environmental response. That is, even if two microphones record the same signal, they may detect different responses because of the varying path lengths of the reflections. However, if the microphones are located close to each other, the differences in path lengths of the reflections are limited. In the extreme case where two microphones are collocated, they will detect the same environmental response. Therefore, the ratio $\frac{M_T(\omega)}{M_B(\omega)}$ should have a bounded value within a range. Note that users can still contribute new training data collected in their environments, if desired. In extreme circumstances where environmental variations significantly affect performance, EarCase can update the pre-calibrated model with new acoustic data to cater to individual preferences, resulting in improved accuracy and robustness, even in noisy environments.

In order to prove this thought, we conduct experiments to explore the ratio range. We use the smartphone without a phone case to collect data from three different environments: lab, meeting room and lobby. In each experimental setup, we fixed the relative positions of the smartphone and the sound source, where the sound source is a chirp signal with frequency band from 1Hz to 8kHz. Then, after we divide the received signal of the top and bottom microphones, we obtain the equation similar to Eq. (4) without the location-specific pattern: $\frac{R_T(\omega)}{R_B(\omega)} \approx \frac{M_T(\omega)}{M_B(\omega)}$. As shown in Figure 8, the environment response ratio has a similar distribution under different environments, consistent with previous analysis.

5.3 Deep Learning for Sound Localization

Once the environmental response ratio $M_{ratio} = \frac{M_T(\omega)}{M_B(\omega)}$ is bounded within a range, Eq. (4) can be further derived as $\frac{R_T(\omega)}{R_B(\omega)} \approx \frac{H_T(\omega)}{H_B(\omega)} \cdot k$, where k is a constant value. Rather than further extracting the location-specific pattern H_T and H_B separately, we prefer to use ratio $H_{ratio} = \frac{H_T(\omega)}{H_B(\omega)}$ to represent the special location pattern. That is, the $\frac{R_T(\omega)}{R_B(\omega)}$ is only related to H_{ratio}. Thus, we can collect the data from different locations and use the ratio of the received signals from top and bottom microphones to train a deep learning model. We leverage a convolutional neural network, which has superior performance in signal processing [19]. The illustration of our CNN model is shown in Figure 9. To optimize the method for resource-limited smartphones, we employ a lightweight one-dimensional model with two convolutional layers and one max-pooling layer followed by a fully-connected layer and an output regression layer. The input matrix is with size of 8001×2, where 8001 is the separated frequencies between 0-8kHz and 2 is the frequency and phase domain. The convolution layers have 64 and 128 filter maps, respectively. In this model, we add a max-pooling layer between the convolution layers to reduce the matrix size and speed up the training process. The regression layer computes the half-mean-squared-error loss for location estimation. We set the learning rate at 0.01 with 500 epochs when train the model. This design ensures compatibility with mobile devices while still capturing the necessary subtle cues from mini-structures.

6 IMPLEMENTATION

6.1 Signal Design and Data Collection

A multiple frequency-band chirp signal is generated in Matlab, which will be used as the training sound source. To make sure the sound wave interacts (e.g., diffraction or reflection) when it passes through the mini-structure, the wavelength of the signal should be comparable to the size of the smartphone and mini-structures. On the other hand, according to the Nyquist-Shannon sampling theorem [20], the upper frequency limit is restricted by the frequency sampling (fs) rate. Due to hardware limitations, the microphone of commercial smartphones usually has a sampling rate fs as 16kHz, restricting the maximum transmittable frequency to 8kHz. Even though some smartphones can reach high sampling rates of 44.1kHz, most common sounds in daily life are lower than 8kHz. Given these factors, our system adopts a 100ms length multiple frequency-band signal with chirp sweeps using the 1Hz to 8kHz bandwidth. Such

signal is played by a off-the-shelf speaker (e.g., smartphones, loud speakers) and the smartphone with our EarCase prototype is held by a stand to record the sound. We note that the smartphone should enable true stereo recording in order to record the sound signal with two microphones separately.

6.2 Simulation Data

Sound propagation simulation can generate a wide variety of sound signal types and physical setups. Although we can eliminate the impact of the environment, it is still a heavy burden to collect comprehensive samples in different environments manually. Thus, we conduct rigorous acoustic analyses to generate fine-grained location-related acoustic patterns for training effective localization models. We initiate the simulation process by placing two virtual microphones 14cm apart, which is the average distance for smartphones. Additionally, a randomly chosen background noise is included in the scene. We then simulate room impulse responses (RIRs) for a randomly sized room using the image source method implemented in the pyroomacoustics library [1, 28]. We randomly generate rooms with lengths 6-10m, widths 5-9m, heights 3-5m and RT60 (time it takes for the RIR to decay by 60 dB) 0.3-0.5s. All signals undergo convolution with the RIR and are then rendered to the two-channel microphone array. The volumes of the background sounds are randomly selected, ensuring that the input signal-to-distortion ratio falls within the range of approximately -5 to 5 dB. This process simulates the signals before they interact with the two stencils, shown in Figure 7. Next, we simulate the signal interaction with the stencil. As mentioned in Section 5.1.1, Eq. (1) guides the interaction simulation process to obtain the parameter H. Once we determine the exact stencil, we can compute the hole distances and obtain the parameter r_n and D. Then, we can easily derive the interaction parameters H. Finally, we use the received signals as the simulation data.

7 PERFORMANCE EVALUATION

7.1 Experimental Methodology

Experimental Setup. We developed a prototype app of EarCase for Android platforms. Our smartphone cases were designed using Autodesk Inventor to fit the *Nexus 5* and *Galaxy Note 5* devices, shown in Figure 11. Note that EarCase is designed to be compatible with any smartphone featuring two microphones as the case structure plays a dominant role in capturing acoustic signal patterns for sound source localization. Physical copies were produced using Prusa I3MK3S printers and Nylon filament. We conduct our experiments by setting the microphone sampling rate of the smartphone at 16kHz and utilize an external speaker to act as the sound source. We record audio samples of the sound source at angles from 0° to 180° in increments of 1° and distances from 100cm to 250cm in increments of 50cm. We assume the smartphone and sound source are at equal elevation levels but control the precise height of the sound source using an adjustable stand to study 3D localization. Our hardware setup is shown in Figure 10. Our sound sources include multi-frequency wideband signals, human speeches, random sounds, and car honking. For reproducibility, we use a looping wideband signal maintained at 40 dB SPL as our default sound source and

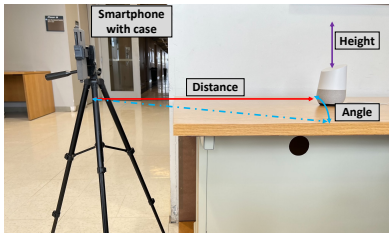


Figure 10: Experimental setup in the lab environment. An external speaker acts as the sound source. Distance, height, and angle are controllable.

(a) Nexus 5 (b) Galaxy Note 5

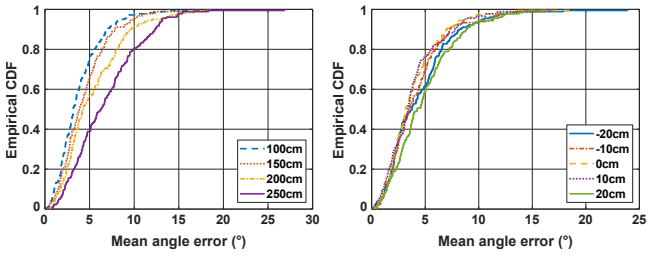
Figure 11: EarCase prototypes for different smartphones.

localize using our *Nexus 5* smartphone unless mentioned otherwise. We collect 20 times at each test location.

Evaluation Metrics. We quantify the performance of EarCase through widely used metrics including Mean Error (angles and distances), Location Accuracy and DL Model Accuracy. Mean error of one location is computed by $ME = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$, where x_i is the predicted location, x is the ground truth location and n is the testing times. When we get the mean errors of all the testing locations, we use the cumulative distribution function (CDF) to show the distribution of all locations' mean errors. Meanwhile, the location accuracy shows the ratio of identifying location successfully, computed by $LA = \frac{n'}{n} \times 100\%$ (n is the testing time and n' is the times of identifying location successfully). Finally, the DL Model Accuracy indicates how the model performs across all datasets.

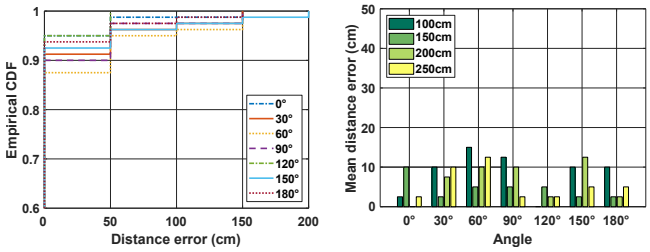
7.2 Performance Analysis

7.2.1 How well is EarCase in three location dimensions? We measure the performance while the sound source is placed at various locations from the smartphone. Figure 12(a) shows the mean error for angle estimation when we set the distance at 100cm to 250cm with intervals of 50cm. EarCase can achieve mean angle errors as low as 3.7° at 100cm distances. When distance is more than doubled to 250cm, we still support low errors of 6.3°. This error is mainly dominated by the change in signal-to-noise ratio at the receiver due to increasing distance. The volume intensity of the sound source was kept constant despite any changes in location of the source. We also control 3D position by adjusting height of the sound source relative to height of the receiver. We observe in Figure 12(b) that height has negligible impact on performance, yielding an average 4.2° error across all tested deviations. Moreover, we keep same angle while change the distance. The results illustrated in Figure 13(a) shows that nearly 90% locations can be correctly identified. Meanwhile, we evaluate the location accuracy of EarCase at the area $1m \times 1m \times 0.6m$. As shown in Figure 14, all



(a) Mean angle error vs. Distance (b) Mean angle error vs. Height

Figure 12: Empirical CDF of mean angle errors when the sound source is at varying distances and heights.



(a) Distance error vs. Angle (b) Angle vs. Mean distance error

Figure 13: Distance error for sounds at different angles.

144 sound source locations can achieve over 90% accuracy. This demonstrates EarCase is effective at localizing sounds in 3D space.

7.2.2 How effective is EarCase in real application scenarios? We verify EarCase can be deployed in multiple application scenarios by experimenting in different real-world environments and sound sources. We consider in Figure 15(a) five locations (L1-L5) across two environments (E1-E2). E1 is an indoor lab with three sound locations whereas E2 is an outdoor parking lot with two sound locations. We use a smartwatch device beeping to act as the sound source in E1, replicating the conditions of locating a lost mobile device, and a car horn honking as the sound source in E2, safely simulating the acoustics of a dangerous event such as cars approaching pedestrians. We also use live human speech as an example sound source that could be found in both environments. The locations of the sound sources are circled in red. In each test round, we ask one participant to wear headphones with loud music to simulate someone who is hard of hearing. We use the localization accuracy as the metric to see if the participant can localize the sound source. Our findings showed in Figure 15(b) all sound sources can be reliable localized with over 80% accuracy. Specifically, we observed smartwatch beeps and car honking can be consistently recognized with over 90% accuracy while human speaking may be slightly more difficult. This is because the frequency and sound pressure level of human speaking are always changing.

7.2.3 How stable is EarCase when under various impact factors? **Different Mini-structure Case Designs.** We confirm our initial observations from Section 5.1 by conducting large-scale experiments using the many case designs developed during the prototyping process from Figure 6. In addition to different structural shapes (e.g., cubes, cylinders, spheres), we also consider how the precise locations of the holes bored into the mini-structure can

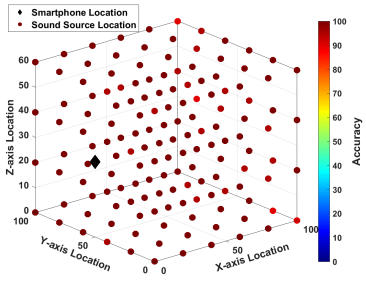
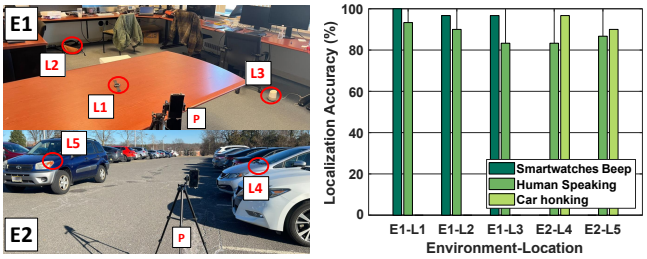
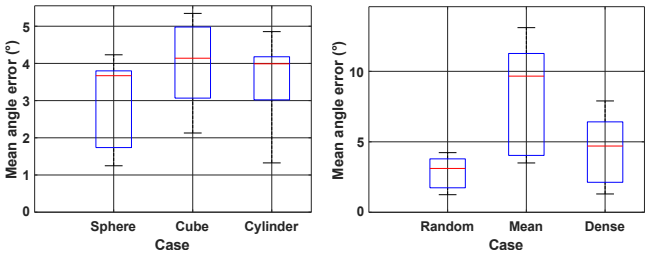


Figure 14: Comparison of localization performance using different testing sound sources.



(a) The environments and locations (b) The localization accuracy at different locations with real world sound source

Figure 15: Experimental setup and system performance in simulating real application scenarios.



(a) Mean Angle Error vs. Case Type (b) Mean Angle Error vs. Holes Distribution

Figure 16: Mean angle errors using different mini-structure case designs.

influence localization performance. Figure 16(a) reinforces our earlier observations, showing spherical structures can produce the lowest mean angle error. Meanwhile, we study in Figure 16(b) the impacts of having the mini-structure holes distributed throughout a spherical structure randomly, equidistant apart, and densely concentrated in a single area. We find randomly placed holes to be the best performing, achieving 3.7° error, likely due to the highly varied distance referred to as r of Eq.(2). The equidistantly positioned holes performed the worst, likely because the distribution of holes is symmetric (i.e., sound propagation path is the same). Meanwhile, the densely packed holes can achieve low error only for the direction they are oriented towards.

Different Smartphone Models. To ensure EarCase can be generalized to other devices, we adapt our spherical structure with randomly placed holes, the best performing design from Figure 16 for the Galaxy Note 5 smartphone model. Figure 17 shows the

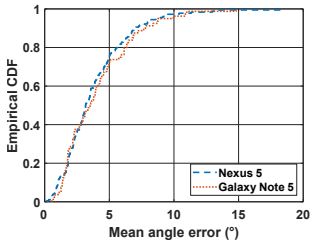


Figure 17: Mean angle errors using Nexus 5 and Galaxy note 5 smartphone models.

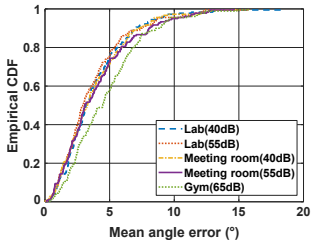


Figure 18: Mean angle errors under several environments with different sound noises.

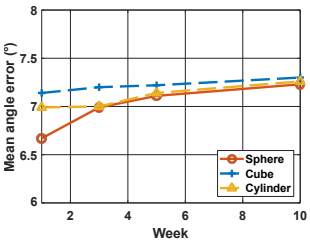


Figure 19: Mean angle errors of testing data collected in different weeks after training the system at the first week.

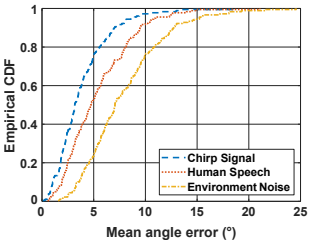


Figure 20: Performance using different testing sound sources when the system is trained by chirp signal sound.

performance differences across our devices. We find the Galaxy Note 5 to be equally effective at localizing sound sources as the Nexus 5, achieving 4.2° mean angle error which is a difference of only 0.5°. This suggests EarCase is robust to hardware variables such as microphone quality and positions.

Different Noise Levels. We confirm EarCase can continue operating even with persistent background noise. We control the background noise level by playing the random noise at lab and meeting room, and the background noise of the Gym is 65dB. Figure 18 shows our results when localizing the sound source with different levels of background noise. Our data shows most noise sources have minimal impact, resulting in average mean angle errors of 4.0°. Our highest tested noise level, 65dB, was the only setting to show noticeable changes in performance, achieving 4.8° error which is still acceptable for reliable localization.

Stability Over Time. We repeat our localization experiments regularly over several months using the same initial training data to determine whether our performance is time invariant. Figure 19 shows the results for three variations of our stencils across a 10 week period. Our spherical design is the most stable, supporting our findings from Figure 16. All cases were observed to have minor performance degradation as time progressed, converging around 7.2° after 5 weeks, which is less than 1° increase. This error can be reduced through optional re-sampling and training.

Different Sound Sources. We consider different real-world sound sources. In addition to our existing wideband signal, we transmit samples of human speech and random noise from our sound source to replicate the conditions people with difficulty hearing may find themselves in during daily routines. We show the differences in performance between these real-world sounds in Figure 20. We find that EarCase can also localize human speech and random noise with high accuracy, achieving 5.2° and 7.8° errors,

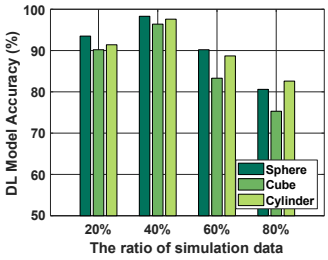


Figure 21: DL model accuracy of using different ratio of simulation data as training data.

respectively. These errors are only slightly higher than our results using controlled chirp signals, likely due to frequency and sound pressure level changing. Overall, our findings indicate EarCase is agnostic to different target sounds for localization.

Simulated Data vs. Real-World Data. We explore the feasibility of increasing training data volume by using modeled acoustic behavior in virtual environments as inputs. Figure 21 shows our real-world training data progressively supplemented with simulant data, eventually reaching a 1:5 ratio or 80% of training data consisting of simulants. Our findings suggest overall performance may worsen if simulation data comprises the majority of training data. Small amounts of simulation data may still improve accuracy, by introducing modest diversity to the model without overfitting.

8 DISCUSSION

Multiple Sound Sources. We primarily consider applications where one target sound source is active at a time with secondary sounds being regarded as interference only. Based on our peak-energy detection approach, EarCase will normally fixate on the loudest sound as the target. We believe multi-sound source support is also possible by designating the second or third strongest peaks as sound sources, for example. Our system can use existing sound source separation methods (e.g., deep cocktail party [17], independent component analysis [30]) to isolate the sound peaks and localize each individual sound source using the same method, respectively. The user can indicate either beforehand or in real-time whether they are seeking multiple sounds.

Optimal Stencil Design. Our experiments test several prototype stencil designs and achieved high resolution localization in many practical scenarios. However, the resolution of EarCase is also heavily influenced by design parameters such as the number and distribution of holes and tubes in the stencils, the diameter of the holes, and shape and size of the stencils. We provide an initial exploration of how different combinations of these factors can amplify the location-specific information embedded in acoustic sound propagation through our mini-structures. Future work will research how these design parameters can be further optimized and study the trade-offs between different design choices.

Phone Case Material. We manufacture EarCase prototypes using Nylon filament due to its tensile strength and low sound absorption. We also considered other candidate materials during our prototyping process, including TPU and PLA plastic resins. Plastic-based materials are popular in commercial products like phone cases because of their flexibility. However, we find this flexibility to be detrimental for building our small stencil designs, making them

brittle and prone to collapse. The ideal material should have good acoustic properties to ensure that sound can reflect through case effectively. Additionally, the material should be durable enough to protect the phone from everyday wear and tear, while also being visually appealing and affordable for consumers. We leave the exploration of additional case materials for future work.

9 CONCLUSION

We proposed EarCase, a novel sound localization technique for smartphones that utilizes a smartphone case equipped with mini-structures to enhance the acoustic sensitivity of built-in microphones. The core idea of our work is to surround the microphones with well-designed stencils to alter the frequency responses of sound waves such that precise location-based information can be derived, similar to how the unique shape of human ears enables general localization. We develop location-specific pattern extraction techniques using only the audio recordings provided by built-in smartphone microphones and manufacture physical prototypes of our smartphone cases with mini-structures for commercial-off-the-shelf devices. We demonstrate EarCase can localize the origin of multiple sound source types in real-world conditions with 96% accuracy. We believe our findings have the potential to greatly enhance the quality-of-life of people with hearing loss.

ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation Grants CCF1909963, CCF2211163, CNS2120396, and CNS2145389.

REFERENCES

[1] Jont B Allen and David A Berkley. 1979. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America* 65, 4 (1979), 943–950.
[2] Sebastià V Amengual Garí, Winfried Lachenmayr, and Eckard Mommertz. 2017. Spatial analysis and auralization of room acoustics using a tetrahedral microphone. *The Journal of the Acoustical Society of America* 141, 4 (2017), EL369–EL374.
[3] Eugena Au, Shirley Xiao, CT Justine Hui, Yusuke Hioka, Hinako Masuda, and Catherine I Watson. 2021. Speech intelligibility in noise with varying spatial acoustics under Ambisonics-based sound reproduction system. *Applied Acoustics* 174 (2021), 107707.
[4] G Bekefi. 1953. Diffraction of sound waves by a circular aperture. *The Journal of the Acoustical Society of America* 25, 2 (1953), 205–211.
[5] Debra L Blackwell, Jacqueline W Lucas, and Tainya C Clarke. 2014. Summary health statistics for US adults: national health interview survey, 2012. *Vital and health statistics. Series 10, Data from the National Health Survey* 260 (2014), 1–161.
[6] Eleonora Cagli, Diego Carrera, Giacomo Aletti, Giovanni Naldi, and Beatrice Rossi. 2013. Robust DOA estimation of speech signals via sparsity models using microphone arrays. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 1–4.
[7] Cecilia Casarini, Benjamin Tiller, Carmelo Mineo, Charles N MacLeod, James FC Windmill, and Joseph C Jackson. 2018. Enhancing the sound absorption of small-scale 3-D printed acoustic metamaterials based on Helmholtz resonators. *IEEE Sensors Journal* 18, 19 (2018), 7949–7955.
[8] Shlomo E Chazan, Hodaya Hammer, Gershon Hazan, Jacob Goldberger, and Sharon Gannot. 2019. Multi-microphone speaker separation based on deep DOA estimation. In *27th European Signal Processing Conference*. IEEE, 1–5.
[9] Corey I Cheng and Gregory H Wakefield. 1999. Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space. In *Audio Engineering Society Convention 107*. Audio Engineering Society.
[10] Paolo Chiariotti, Milena Martarelli, and Paolo Castellini. 2019. Acoustic beamforming for noise source localization–Reviews, methodology and applications. *Mechanical Systems and Signal Processing* 120 (2019), 422–448.
[11] Ming-An Chung, Hung-Chi Chou, and Chia-Wei Lin. 2022. Sound Localization Based on Acoustic Source Using Multiple Microphone Array in an Indoor Environment. *Electronics* 11, 6 (2022), 890.
[12] Maximo Cobos, Fabio Antonacci, Anastasios Alexandridis, Athanasios Mouchtaris, and Bowon Lee. 2017. A survey of sound source localization methods in wireless acoustic sensor networks. *Wireless Communications and Mobile Computing* 2017 (2017).
[13] Dalia El Badawy and Ivan Dokmanić. 2018. Direction of arrival with one microphone, a few legos, and non-negative matrix factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 12 (2018), 2436–2446.
[14] Anshuman Ganguly, Chandan Reddy, Yiya Hao, and Issa Panahi. 2016. Improving sound localization for hearing aid devices using smartphone assisted technology. In *IEEE International Workshop on Signal Processing Systems*. IEEE, 165–170.
[15] Nakul Garg, Yang Bai, and Nirupam Roy. 2021. Owlet: Enabling spatial information in ubiquitous acoustic devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 255–268.
[16] Chris Harrison, Robert Xiao, and Scott Hudson. 2012. Acoustic barcodes: passive, durable and inexpensive notched identification tags. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 563–568.
[17] Simon Haykin and Zhe Chen. 2005. The cocktail party problem. *Neural computation* 17, 9 (2005), 1875–1902.
[18] Andy Kiersz. 2014. REVEALED: Here's Who Uses iPhone Cases And Why. https://www.businessinsider.com/iphone-case-survey-2014-7.
[19] Serkan Kiranyaz, Turker Ince, Osama Abdeljaber, Onur Avci, and Moncef Gabbouj. 2019. 1-D convolutional neural networks for signal processing applications. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 8360–8364.
[20] HJ Landau. 1967. Sampling, data transmission, and the Nyquist rate. *Proc. IEEE* 55, 10 (1967), 1701–1706.
[21] Gierad Laput, Eric Brockmeyer, Scott E Hudson, and Chris Harrison. 2015. Acoustruments: Passive, acoustically-driven, interactive controls for handheld devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2161–2170.
[22] Ho-Yong Lee, Ji-Won Cho, Minook Kim, and Hyung-Min Park. 2016. DNN-based feature enhancement using DOA-constrained ICA for robust speech recognition. *IEEE Signal Processing Letters* 23, 8 (2016), 1091–1095.
[23] Dingzeyu Li, David IW Levin, Wojciech Matusik, and Changxi Zheng. 2016. Acoustic voxels: Computational optimization of modular acoustic filters. *ACM Transactions on Graphics* 35, 4 (2016), 1–12.
[24] Xiaofei Li, Laurent Girin, Fabien Badeig, and Radu Horaud. 2016. Reverberant sound localization with a robot head based on direct-path relative transfer function. In *International Conference on Intelligent Robots and Systems*. IEEE, 2819–2826.
[25] Kaikai Liu, Xinxin Liu, Lulu Xie, and Xiaolin Li. 2013. Towards accurate acoustic localization on a smartphone. In *2013 Proceedings IEEE INFOCOM*. IEEE, 495–499.
[26] KS Peat. 1994. A first aproximation to the effects of mean flow on sound propagation through cylindrical capillary tubes. *Journal of sound and vibration* 175, 4 (1994), 475–489.
[27] Nissanka B Priyantha, Anit Chakraborty, and Hari Balakrishnan. 2000. The cricket location-support system. In *Proceedings of the 6th annual international conference on Mobile computing and networking*. 32–43.
[28] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. 2018. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *IEEE international conference on acoustics, speech and signal processing*. IEEE, 351–355.
[29] Nikolaos Stefanakis and Athanasios Mouchtaris. 2016. Direction of arrival estimation in front of a reflective plane using a circular microphone array. In *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 622–626.
[30] James V Stone. 2002. Independent component analysis: an introduction. *Trends in cognitive sciences* 6, 2 (2002), 59–64.
[31] H Tijdeman. 1975. On the propagation of sound waves in cylindrical tubes. *Journal of Sound and Vibration* 39, 1 (1975), 1–33.
[32] Aaron Visschedijk, Hyunyoung Kim, Carlos Tejada, and Daniel Ashbrook. 2022. ClipWidgets: 3D-printed Modular Tangible UI Extensions for Smartphones. In *Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction*. 1–11.
[33] Francis M Wiener. 1947. Sound diffraction by rigid spheres and circular cylinders. *The Journal of the Acoustical Society of America* 19, 3 (1947), 444–451.
[34] Angeliki Xenaki, Jesper Bünsow Boldt, and Mads Græsbøll Christensen. 2018. Sound source localization and speech enhancement with sparse Bayesian learning beamforming. *The Journal of the Acoustical Society of America* 143, 6 (2018), 3912–3921.
[35] Bosun Xie. 2013. *Head-related transfer function and virtual auditory display*. J. Ross Publishing.
[36] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 1080–1091.
[37] Hongzi Zhu, Yuxiao Zhang, Zifan Liu, Xiao Wang, Shan Chang, and Yingying Chen. 2021. Localizing acoustic objects on a single phone. *IEEE/ACM Transactions on Networking* 29, 5 (2021), 2170–2183.