

SonicSieve: Bringing Directional Speech Extraction to Smartphones Using Acoustic Microstructures

Kuang Yuan*

Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
kuangy@andrew.cmu.edu

Yifeng Wang*

Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
yifengw3@andrew.cmu.edu

Xiyuxing Zhang*

Department of Computer Science and
Technology
Tsinghua University
Beijing, China
zxyx22@mails.tsinghua.edu.cn

Chengyi Shen

Human-Computer Interaction
Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
chengyis@andrew.cmu.edu

Swarun Kumar

Department of Electrical and
Computer Engineering
CMU
Pittsburgh, Pennsylvania, USA
swarun@cmu.edu

Justin Chan

Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
justinchan@cmu.edu

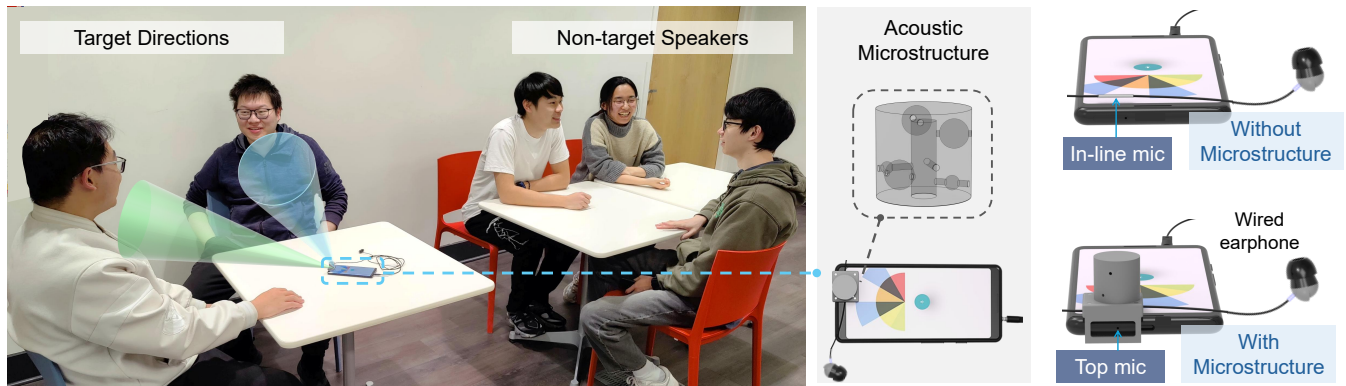


Figure 1: SonicSieve enables directional speech extraction on smartphones using a lightweight, passive acoustic microstructure. (Left) Our system leverages the distinct spatial cues created by the 3D-printed microstructure with a real-time neural network to intelligently amplify speech from target directions while attenuating others. (Right) Our design attaches to the in-line microphone of low-cost earphones which can be plugged into a smartphone. The system records sound mixtures from the in-line and top microphone which are used by the neural network to generalize across different sound sources and environments.

Abstract

Imagine placing your smartphone on a table in a noisy restaurant and clearly capturing the voices of friends seated around you, or recording a lecturer’s voice with clarity in a reverberant auditorium. We introduce *SonicSieve*, the first intelligent directional speech extraction system for smartphones using a bio-inspired acoustic microstructure. Our passive design embeds directional cues onto incoming speech without any additional electronics. It attaches to the in-line mic of low-cost wired earphones which can be attached

to smartphones. We present an end-to-end neural network that processes the raw audio mixtures in real-time on mobile devices. Our results show that *SonicSieve* achieves a signal quality improvement of 5.0 dB when focusing on a 30° angular region. Additionally, the performance of our system based on only two microphones exceeds that of conventional 5-microphone arrays.

CCS Concepts

• Human-centered computing → Sound-based input / output; Ubiquitous and mobile computing systems and tools; • Computing methodologies → Machine learning.

Keywords

Directional speech extraction, Acoustic microstructures, Spatial sensing, Audio interfaces, Machine Learning

*Co-primary authors



This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '26, Barcelona, Spain

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
<https://doi.org/10.1145/3772318.3790376>

ACM Reference Format:

Kuang Yuan, Yifeng Wang, Xiyuxing Zhang, Chengyi Shen, Swarun Kumar, and Justin Chan. 2026. SonicSieve: Bringing Directional Speech Extraction to Smartphones Using Acoustic Microstructures. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3772318.3790376>

1 INTRODUCTION

In this paper, we ask the question: *Can we enable directional speech extraction on smartphones?* A positive answer can enable smartphones to intelligently amplify target speakers and suppress unwanted interference. Imagine having a meeting in a noisy restaurant, and you only want to capture the speech of colleagues at your table for clear meeting notes and to remotely stream to other colleagues back at the office (Fig. 1). Or imagine being a student in a classroom and you want to record the sound of the lecturer without needing a wearable microphone or a dedicated setup.

While rich prior work [11, 22, 28, 37, 44, 48, 55, 56, 61, 67] on intelligent sensing for audio and speech using spatial information exist, they usually rely on arrays of microphones and estimate spatial cues of sounds by analyzing differences in arrival time and sound intensity across the microphones. Although microphone arrays have proliferated across various computing devices from smart speakers and smart glasses to AR/VR headsets, smartphones often contain only two microphones and remain largely unaware of the spatial richness of sound. Additionally, smartphones do not natively support external microphone arrays. While directional microphones exist, they only filter sound from a single, fixed direction, and don't provide multi-directional separation.

We present *SonicSieve*, the first intelligent directional speech extraction system on smartphones that uses a bio-inspired acoustic microstructure to embed distinct directional cues into incoming speech sounds from different angles. The design of the acoustic microstructure is inspired by directional hearing mechanisms found in nature. While human ears function like a two-microphone array, they are able to achieve directional hearing by leveraging spatial cues from the interaction of sound with the structure of the head and pinna [8]. Perhaps closest to *SonicSieve* in prior art is Owlet [19], a lightweight, miniature 3D-printed cylinder patterned with coded holes to encode spatial information into the signals arriving from different directions, enabling Direction-of-Arrival (DoA) estimation for a single sound source. Inspired by Owlet, *SonicSieve* address a real-world user-centered challenge: going beyond localizing *where* a sound signal originates, *SonicSieve* tackles directional speech extraction that determines *what* to extract from an acoustic mixture in a noisy multi-speaker environment.

To this end, we co-design both novel hardware and software solutions for *SonicSieve*. *First*, we optimize the dimensions and the coded pattern of the microstructure specifically for speech frequencies. We then create a practical design that attaches the microstructure to the in-line microphone of low-cost wired earbuds, which can be plugged into a smartphone. The system captures audio mixtures using both the in-line microphone and the smartphone's co-located built-in top microphone, which serves as a reference (Fig. 1) and enables the system to function reliably across diverse sound sources. *Second*, we develop an end-to-end neural network that runs in

real-time on the smartphone, processing the audio from both microphones to interpret the rich spatial cues and separate the speech from the target directions. We also design a smartphone user interface that lets users divide the surrounding space into a semicircular region with six 30° sectors. Users can select one or multiple sectors simultaneously, enabling functionality in multi-speaker scenarios (Fig. 16) such as transcribing presentations with multiple speakers or conducting remote meetings in noisy environments like a park.

We build our end-to-end system on a smartphone with a low-cost wired earphone and a 3D-printed microstructure. This prototype therefore targets smartphones that can physically connect to a wired earphone and currently relies on an app-guided alignment procedure between the microstructure and the phone's top microphone. We revisit these design trade-offs in our discussion section. Our results are as follows:

- Our system achieves a 5.0 dB Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) improvement when focusing on a 30° sector for directional speech extraction, which significantly outperforms the baseline system without microstructure.
- Our system demonstrates generalizable performance across 9 locations in 3 rooms.
- Our average model inference time to process an 8 ms audio chunk is 7.12 ms and 4.46 ms on the Motorola Edge and Google Pixel 7 respectively, demonstrating real-time processing capability.
- Our user study with 20 participants rating 720 audio clips in total shows our system achieve a higher mean opinion score than a system based on a 5-channel microphone array.

We will open source our microstructure design, code, and datasets, which can democratize directional speech extraction capabilities and make it available to the public.

2 RELATED WORK

To the best of our knowledge, no prior work has explored the use of acoustic microstructures to enable directional speech extraction on smartphones. Below we describe work related to acoustic microstructures, spatial sensing, and AI-enabled acoustic systems.

Acoustic sensing using microstructures. Acoustic microstructures have been designed for a variety of applications, including spatial sensing [6, 19], sound absorption [5, 17], acoustic filtering [32], and biometric identification [66]. Owlet [19] designed a 3D-printed microstructure that embedded directional cues into incoming sounds, enabling direction-of-arrival (DoA) estimation and sound localization. SPiDR [6] extended this design by using microstructures to project and capture spatially coded signals to generate depth maps of nearby objects. EarCase [34] adapted microstructures for smartphones, embedding them into a custom case for DoA estimation. However, its design required a new custom case for each device, limiting its ability to scale across multiple phone models and microphone layouts. In addition, case-based designs fix the relative placement between the microphones, which can make it harder to optimize for close microphone spacing that improves robustness across environments. Our work differs in two key ways. First, our microstructure is designed for low-cost wired earphones, and is compatible with different smartphones. This form factor allows the in-line microphone to be positioned close to the phone's built-in microphone, which empirically improves reliability under

Reference	Mics	Single device	Directional speech extraction	Smartphone compatible	Real-world evaluation	Real-time
Conventional mic array	4–8+	✓	✓	✗	N/A	N/A
Dia (2014) [46]	2–8	✗	✗	✓	✓	✗
Owlet (2021) [19]	2	✓	✗	✗	✓	✓
DSENet (2023) [28]	3	✓	✓	✓	✗	✓
EarCase (2023) [34]	2	✓	✗	✓	✓	DNS
Pandey et al. (2024) [37]	8	✓	✓	✗	✗	✓
SonicSieve (ours)	2	✓	✓	✓	✓	✓

Table 1: Comparison of *SonicSieve* with related systems. *SonicSieve* enables directional speech extraction across six 30° sectors on smartphones in real-time, with evaluation in real-world environments. (DNS = did not specify)

changing room acoustics (Sec. 3.3). Second, our work is focused on directional speech extraction by leveraging spatial cues to selectively amplify and suppress speech based on their direction. This enables a broader set of spatial audio interactions beyond what is achievable through DoA alone.

Directional speech extraction. Directional speech extraction is the task of isolating speech from a spatial region that is fixed or adjustable [37]. Prior work [37] explicitly relies on DoA information derived from 8-channel microphone arrays, and does not focus on smartphones. Related approaches [28, 67] explore two- or three-microphone configurations found on some smartphones. However, the spatial cues provided by these configurations are fundamentally limited by both the small number of microphones and their suboptimal placement for spatial sensing tasks. In contrast to these works, we present an acoustic microstructure design which is optimized to enhance the diversity of spatial cues for speech signals. Our system does not require explicit DoA estimation, instead it takes the raw audio mixture as input, and uses an end-to-end network to implicitly infer spatial and acoustic cues, enabling real-time directional speech extraction on smartphones.

Beamforming using statistical algorithms. Traditional beamforming techniques are typically based on statistical algorithms, including non-adaptive approaches such as Bartlett (delay-and-sum) and superdirective beamformers, as well as adaptive methods like minimum-variant distortionless-response (MVDR), and linearly constrained minimum variance (LCMV). However, these algorithms are designed for conventional microphone arrays with fixed geometries, and do not account for the complex direction-dependent transformations introduced by acoustic microstructures. As a result, they cannot be directly applied to microstructure-enabled systems like ours, where spatial cues arise not only from microphone placement but also from the physical design of the structure. Moreover, statistical beamformers are limited to capturing spatial cues, and lack the ability to model acoustic cues as neural beamformers do.

Neural beamformers. Prior work on neural beamformers have leveraged LSTM networks [12, 21], but are non-causal and are intended to support full-length audio inputs. Subsequent works [25, 41] introduced causal models that support online processing, but are not real-time. More recent work [53, 54] has focused on designing causal real-time neural beamformers focused on directional hearing for AR headsets, smart glasses, and custom headphones equipped with four or more microphones that provide spatial cues. In contrast, we present a real-time neural network that can process the spatial

and acoustic cues of audio that has been filtered by a microstructure, and use this for the task of directional speech extraction.

AI-enabled spatial-aware audio processing. These systems have focused on speech enhancement [10, 39], sound localization [36, 45, 71], sound classification [58, 70, 73, 74] and speech separation [26, 59]. Recent works have leveraged spatial information to enable new capabilities on smart devices. Cone of Silence [26] extracts speech from multiple target speakers at different locations. ReZero [22] extracts all sounds within a user-defined spatial region—angular, spherical, or conical. However, most of these systems either require large microphone arrays with at least six elements, or are not designed for causal, real-time use. ClearBuds [10] and Sound Bubble [11] achieve causal, real-time performance using wearable devices. ClearBuds leverages wireless earbuds [9] for real-time speech enhancement in telephony, while Sound Bubble isolates audio within a defined radius. However, these systems are designed to enhance the wearer’s hearing in human-human communication. In contrast, our work targets smartphone-based directional speech extraction for general audio recording and streaming applications, such as meeting transcription and voice assistants. Smartphones are well-suited for ad-hoc application scenarios as they are mobile devices that people carry throughout the day.

Spatial sensing on smartphones. Prior work has focused on estimating DoA [29, 49, 75], and beamforming [20] on smartphones using statistical algorithms and neural networks. However, the performance of these systems is fundamentally limited by the number of microphones. Ad hoc microphone arrays composed of smartphones have also been proposed for spatial sensing [46]. These systems emulate larger arrays by synchronizing multiple devices. However, they rely on complex synchronization protocols such as aligning smartphone CPU clocks using WiFi beacon timestamps, synchronizing audio I/O clocks through a backend server, or using time-of-arrival techniques [33, 40], which can make deployment in real-world scenarios challenging.

3 SONIC SIEVE

This section outlines the technical approach of *SonicSieve*. Our goal is to let users *pick a direction* on a phone interface and have the phone *amplify that speaker* in real time, using a low-cost add-on rather than specialized microphone arrays. We start with a high-level overview of how our system uses an acoustic microstructure for directional speech extraction. Then, we will detail the techniques for optimizing this microstructure, integrating it with a smartphone,

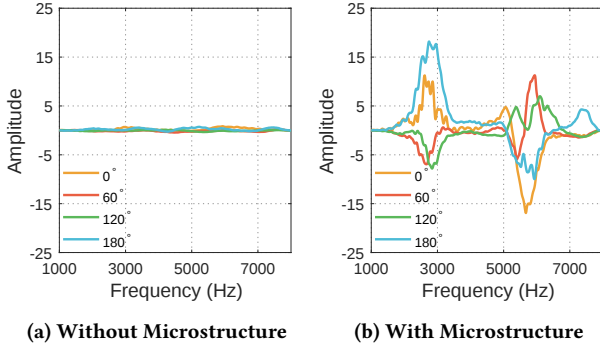


Figure 2: Effect of microstructure on incoming sound signals across different angles of arrival. The microstructure introduces larger variations in the frequency response $M_\theta(f)$ across different angles, providing enhanced spatial cues that can be leveraged for directional speech extraction.

using a neural network to extract speech in real-time, and real-world dataset collections.

3.1 System Overview

Traditional directional speech extraction relies on microphone arrays that use beamforming to isolate sound from specific directions. These arrays are common in dedicated devices (e.g., smart speakers), but are less accessible on everyday smartphones due to hardware constraints. These systems work by analyzing timing and amplitude differences across multiple microphones—when sound arrives from a particular direction θ , each microphone in the array receives the signal at slightly different times and intensities due to their spatial arrangement. The system can be modeled as:

$$\mathbf{x}(f, t) = \mathbf{a}(f, \theta) s_\theta(f, t) + \mathbf{u}(f, t) \quad (1)$$

where $\mathbf{a}(f, \theta)$ encodes the directional information based on known geometric delays between microphone channels, and $\mathbf{u}(f, t)$ represents interference and noise. Intuitively, $\mathbf{a}(f, \theta)$ describes how a sound from direction θ “shows up” across microphones, enabling spatial filtering. Beamforming algorithms then apply spatial filtering to enhance signals from the target direction while suppressing others.

While various beamforming techniques work reliably for directional audio enhancement [2, 3], they typically require a microphone array with multiple channels (usually 4-8) to handle the spatial ambiguity [15, 18] and achieve satisfactory performance, which does not exist on smartphones.

Instead of encoding the spatial information through microphone array geometry, *SonicSieve* leverages a passive acoustic microstructure that directly embeds direction-dependent spatial cues into the signal. Such a design is inspired by nature’s directional hearing system. For example, humans can localize and separated sound effectively with just two ears, which is due to the geometry of the head and ears creating distinctive reflecting and scattering effects that vary with sound angle [50, 65]. Similarly, owls achieve exceptional sound localization through asymmetric ear structures [13]. In practice, our microstructure acts like an acoustic lens placed over

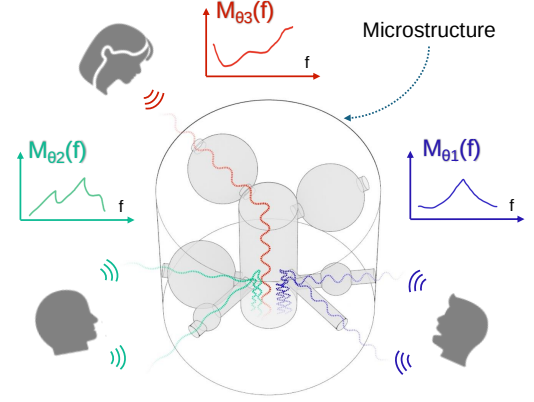


Figure 3: Acoustic microstructure: principle of operation. The structural elements of the microstructure (holes, tubes, and resonators) form a complex multipath environment that creates variations to incoming acoustic signals based on their direction of arrival.

a microphone: it reshapes incoming sound differently depending on where it comes from (Fig. 3).

Specifically, our acoustic microstructure applies a direction-specific filter $M_\theta(f)$ to a sound source arriving at the microstructure from a relative angle θ and propagating through the internal structure. We present an example in Fig. 2 of the direction-specific filter $M_\theta(f)$ at four different angles without and with the microstructure. The microstructure introduces substantially greater variation across different directions of arrival—without it, the signals look almost identical across angles; with it, each angle leaves a more distinctive spectral signature that a model can learn to exploit.

The received signal at the microphone inside the microstructure can be denoted as:

$$\mathbf{x}(f, t) = M_\theta(f) s(f, t) + \mathbf{u}(f, t) \quad (2)$$

However, with only single-channel audio, we can not reliably isolate the spatial filter $M_\theta(f)$ in real-world settings. The challenge is that different sound sources have different intrinsic spectra, and room reflections further modify the signal. With one channel, the model cannot distinguish whether a spectral pattern comes from the microstructure’s directional filtering, the source itself, or the room. A single-mic system might work for fixed sources in controlled environments, but cannot generalize across diverse real-world conditions.

Thus, we incorporate a secondary reference microphone placed outside the microstructure. This reference gives the model a “before/after” view: one channel is shaped by the microstructure, and one is not. The reference microphone records the unmodified sound as:

$$x_{ref}(f, t) = s'(f, t) + u_{ref}(f, t) \quad (3)$$

We note that the received signal at the reference microphone from the target source $s'(f, t)$ differs slightly from the one at the inside microphone $s(f, t)$, since the two microphones are not positioned at the exact same location, such that the arriving signals are propagated through slightly different environmental reflections. To minimize this mismatch and enable robust estimation of spatial cues

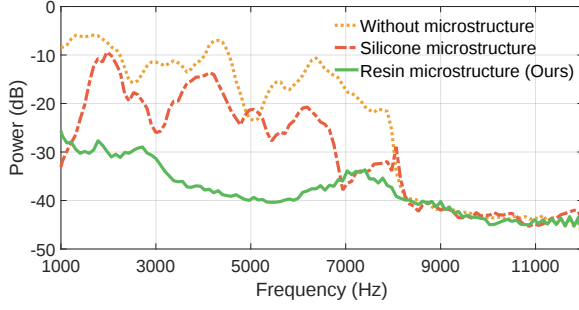


Figure 4: Effect of microstructure material (without holes) on acoustic attenuation. The resin material effectively attenuates sound and ensures that it primarily travels through the microstructure’s holes, rather than its walls.

across different environments, we position the two microphones as close as practically feasible.

Finally, we develop a deep neural network that processes the two-channel input to extract directional speech from the desired direction. Specifically, our directional speech extraction network h estimates the target speech from the selected directions θ :

$$\hat{s}_\theta(f, t) = h(x(f, t), x_{ref}(f, t), \theta) \quad (4)$$

In other words, given the two recordings and the user-selected sector(s), the network outputs only the speech coming from those directions.

To summarize, our system has four main components, which we describe below and present the details in the following sections:

- **Acoustic microstructure optimized for speech.** We present a compact acoustic microstructure that encodes spatial cues into the signal detectable by microphones. Building on prior microstructure work [19], we further optimize the design specifically for spatial diversity at speech frequencies.
- **Integration of acoustic microstructure to smartphones.** We present our design that can attach the acoustic structure to the in-line microphone of low-cost wired earbuds, which can be plugged into smartphones. This approach offers two key benefits: the in-line mic can be positioned close to the built-in mic to minimize signal differences, and the design can potentially be generalized to different smartphone models.
- **Real-time directional speech extraction network.** While conventional beamformers work with known array geometries, they cannot handle the complex propagation patterns created by microstructures. We design a real-time neural network that learns the spatial and acoustic cues from our microstructure to extract speech from multiple selected directions simultaneously, enabling multi-speaker transcription for interviews and group discussions.
- **Real-world dataset.** Acoustic and spatial cues signatures learned in one environment may not generalize well to others, as variations in acoustic properties, such as reverberation [64], can distort the spatial cues encoded by the microstructure. To enable robust performance across different environments, we construct a diverse training dataset that captures the variability encountered in real-world acoustic environments. This dataset is designed to enable

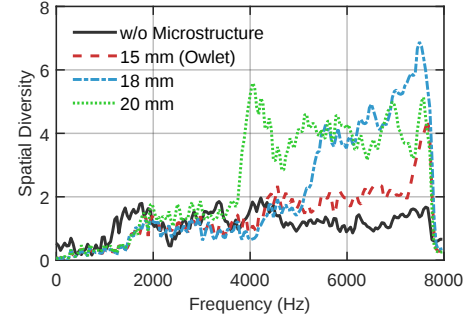


Figure 5: Effect of microstructure diameter on spatial diversity. The microstructure with a larger diameter (20 mm) provides an overall higher spatial diversity across the speech frequencies.

generalization across different spaces, speaker configurations, and background noise conditions.

3.2 Speech-aware acoustic microstructure

We build upon Owlet’s microstructure design [19] which consists of a hollowed-out cylinder that encases a microphone. The design incorporates three key structural elements: surface holes that act as virtual sound sources, capillary tubes with varying lengths that modify sound wave propagation, and Helmholtz resonators that selectively amplify different frequency bands. Together, these elements create unique acoustic signatures for sounds arriving from different angles. As visualized in Fig. 3, when sound arrives from angle θ , the microstructure applies a direction-specific filter $M_\theta(f)$. This directional fingerprint enables our system to distinguish between sounds from different locations.

In this section, we present our approach to optimize the acoustic microstructure and detail the key design factors that impact the overall system performance below:

Material Selection for Acoustic Attenuation. Our material selection aims to maximize sound attenuation, ensuring audio travels through the designed holes rather than leaking through the walls. We tested three configurations using a 1–8 kHz chirp at 70 dBA (normal conversation level): no microstructure, silicone microstructure, and resin microstructure.

As shown in Fig. 4, the resin microstructure achieves 33 dB average sound reduction across tested frequencies, significantly outperforming silicone (17 dB reduction). This superior attenuation ensures sound primarily diffracts through the holes as intended, making resin our material of choice.

Spatial Diversity Optimization. To optimize performance for directional speech extraction, we maximize spatial diversity—the variation in how the structure responds to speech from different directions. Mathematically, we seek to maximize the variance $\mathbb{V}_\theta[|M_\theta(f)|]$, ensuring speech signals from different angles produce maximally different frequency responses. Speech signals are wideband acoustic signals with frequency components mainly under 8 kHz, and the components that are most critical for intelligibility are in 1–4 kHz [30]. We conduct experiments using wideband

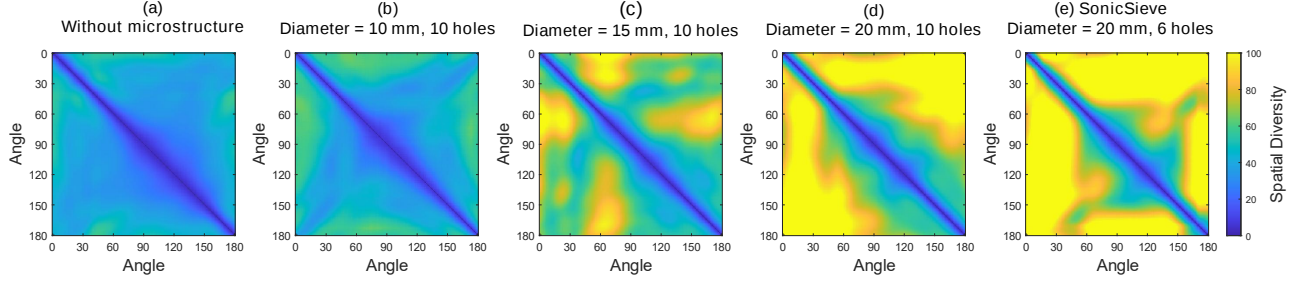


Figure 6: A comparison of spatial diversity across different microstructure designs of varied diameter and hole number. Spatial diversity is computed across angles in a semicircle.

audio under 8 kHz in a constrained environment (anechoic chamber) on different microstructure designs.

To measure spatial diversity, we record speech signals from 0° to 180° at 1° resolution using speakers and microphones inside and outside the microstructure. Since the measurements are taken in an anechoic chamber, we can assume the environmental reflections and background noise are negligible (i.e. $s'(f, t) = s(f, t)$ and $u(f, t) = 0$). From Eq. 2 and Eq. 3, the spatial diversity can then be calculated by:

$$\mathbb{V}_\theta [|M_\theta(f)|] = \mathbb{V}_\theta \left[\frac{|x(f)|}{|x_{ref}(f)|} \right] \quad (5)$$

where $x(f)$ and $x_{ref}(f)$ are signals from the internal and reference microphones. Based on the measurement setup, we optimize the following two key factors to improve the spatial diversity.

- **Microstructure Diameter.** Fig. 5 shows that Owllet’s original 15 mm design achieves spatial diversity mainly at 7–8 kHz, with poor performance below 7 kHz, which can be potentially problematic for speech signals. Based on the principle that microstructure size correlates with resonating wavelengths, we conduct measurements on larger diameters. Both 18 mm and 20 mm diameters show improved diversity below 8 kHz, with 20 mm providing a better performance across speech frequencies.

We further visualize spatial diversity across different angles by calculating the pairwise spectral distance (i.e. $\|M_{\theta_1}(f) - M_{\theta_2}(f)\|_2$) for the microstructure with diameters of 10 mm, 15 mm, and 20 mm. As shown in Fig. 6 (a-d), the microstructure with a higher diameter achieves higher spatial diversity across a wider range of angles. We use 20 mm as our final design. We note that we did not further increase the diameter as a larger microstructure may obstruct the smartphone screen and hinder usability.

- **Number of Surface Holes.** As we observed in Fig. 6(d), the spatial diversity across the angles is not evenly distributed—angles $> 60^\circ$ show lower diversity than $0-60^\circ$. To address this, we further optimized the surface hole pattern. We empirically find that, too few holes limit directional sensitivity (sounds from non-hole directions get attenuated), while too many holes eliminate spatial diversity entirely. Thus, we selectively reduce the original ten-hole design to six holes by removing four holes in the $60-180^\circ$ range. Fig. 6(e) demonstrates that this six-hole configuration



Figure 7: Illustration of the process to integrate acoustic microstructure to the smartphone. (a) UI on the smartphone app instructs the user to place the microstructure in the proper location. (b) Place the inline microphone of a wired earphone on the bottom holder of the microstructure. (c) The top part housing the optimized microstructure is attached to the holder using tiny screws. Proper alignment is achieved with the built-in top microphone of the smartphone.

outperforms the ten-hole design in overall spatial diversity. Therefore, our final design uses a 20 mm diameter cylinder with six strategically placed surface holes.

3.3 Integration of acoustic microstructure to smartphones

Microstructure Fabrication. We implement our optimized microstructure design using computer-aided design and 3D printing. The microstructure geometry is modeled in Rhino 8¹ and fabricated using stereolithography 3D printing. We use ELEGOO Standard Photopolymer Grey Resin² printed on a Formlabs Form 3 printer³, followed by post-processing with Form Wash for cleaning and Form Cure for curing. This fabrication process ensures precise

¹<https://www.rhino3d.com/>

²<https://us.elegoo.com/products/elegoo-standard-resin>

³<https://formlabs.com/3d-printers/form-3/>

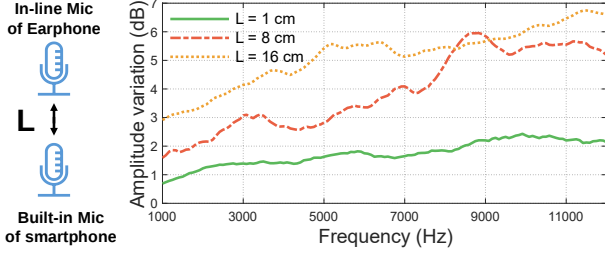


Figure 8: Effect of separation between earphone’s in-line mic and phone’s top mic on spectral variation. We compute the ratio of the signals at the two microphones ($s_{ratio}(f)$) and estimate the variation across environments (lower is better). The variations are lower when the mics are closely spaced.

hole dimensions and smooth internal surfaces critical for acoustic performance.

Assembly Design. Our integration approach attaches the microstructure to the in-line microphone of standard wired earphones, enabling compatibility with any smartphone equipped with an audio jack. As illustrated in Fig. 7, the system consists of two main components: a bottom holder that secures the earphone’s in-line microphone, and a top housing containing the optimized microstructure. The smartphone app guides users through proper placement, ensuring correct alignment between the microstructure and the device’s built-in top microphone.

The final assembly process involves three steps: (1) positioning the earphone’s in-line microphone in the bottom holder, (2) attaching the microstructure housing to the holder using precision-fitted connections, and (3) aligning the complete assembly with the smartphone’s top microphone. A silicone washer at the interface between components minimizes acoustic leakage and ensures consistent positioning.

Design Benefits. This integration approach offers two key advantages over device-specific solutions. First, it enables close microphone spacing—our design positions the in-line microphone within 1 cm of the smartphone’s built-in microphone, which is critical for minimizing environmental variation effects as established in Section 3.1. Second, the design generalizes across different smartphone models since it relies on standard audio jacks and commonly positioned top microphones rather than custom cases for specific devices.

Microphone Separation Validation. To validate the importance of close microphone spacing, we empirically measure how separation distance affects signal consistency across environments. Specifically, we measure the signal ratio of two microphones without attaching the microstructure, defined by:

$$s_{ratio}(f) = \frac{s(f)}{s_{ref}(f)} = \frac{x(f)}{x_{ref}(f)} \quad (6)$$

where $x(f)$ and $x_{ref}(f)$ are signals from the inline and reference microphones.

We conducted experiments across 160 locations in 4 different environments, measuring the magnitude of $s_{ratio}(f)$ at separation distances of 1 cm, 8 cm, and 16 cm. Fig. 8 demonstrates that

amplitude variation increases dramatically with microphone separation. At 1 cm separation, variations remain below 2.2 dB across all frequencies, while at 16 cm separation, variations exceed 6 dB at higher frequencies. This validates our design requirement for close microphone positioning and confirms that our 1 cm separation distance minimizes environmental artifacts that could interfere with directional cue extraction.

3.4 Directional speech extraction network

As described in Section 3.1, our microstructure encodes direction-specific spatial cues into the recorded audio signal. However, real-world acoustic environments contain multiple sound sources, environmental reflections, and noise that complicate the extraction of these spatial cues. We develop a deep neural network that processes the two-channel recordings (from the microstructure and reference microphones) to perform directional speech separation in real-time.

Our network design adapts a state-of-the-art real-time sound separation architecture TF-GridNet [11, 59] for angular-based directional speech extraction. The system processes short audio chunks (8 ms) to enable real-time applications such as live audio streaming and meeting transcription on smartphones.

Pre-processing Module. Our system processes 2-channel audio input $\mathbf{x} = [x_1, x_2]^T \in \mathbb{R}^{2 \times T}$, where x_1 and x_2 correspond to the reference microphone and microstructure microphone signals respectively, with T representing the audio sample length. To accommodate model lookahead requirements, we apply zero padding to create $\mathbf{x} \in \mathbb{R}^{2 \times (T+\sigma)}$, where σ is the lookahead sample count.

We implement a learnable Short-Time Fourier Transform (STFT) encoder with sufficiently large window size to enhance model stability across different reverberation environments [14]. After applying the STFT encoder with discrete Fourier transform length L and step size δ , each microphone channel i produces $X_i[f, t] \in \mathbb{R}^{F \times \frac{(T+\sigma-L)}{\delta}}$, where F represents the number of frequency bins.

Feature Extraction. The core challenge in our approach is extracting the spatial cues embedded by the microstructure while remaining robust to environmental variations and multiple sound sources. While we cannot directly recover the microstructure’s direction-specific filter $M_\theta(f)$ due to its entanglement with unknown source signals and environmental effects, we can compute inter-channel difference features that preserve the essential directional information.

We extract interchannel phase differences (IPD) and interchannel level differences (ILD)—well-established spatial cues in binaural hearing research [7, 22]:

$$\begin{aligned} \text{IPD}[f, t] &= \angle X_2[f, t] - \angle X_1[f, t] \\ \text{ILD}[f, t] &= 20 \log \{|X_2[f, t]|/|X_1[f, t]|\} \end{aligned} \quad (7)$$

The IPD captures phase relationships between the two microphones, which encode timing differences created by the microstructure’s directional filtering. The ILD captures magnitude differences that reflect the microstructure’s direction-dependent attenuation and amplification patterns. Together, these features encode the spatial signatures that enable directional discrimination.

To facilitate neural network processing, we convert the circular IPD values into continuous representations using trigonometric functions:

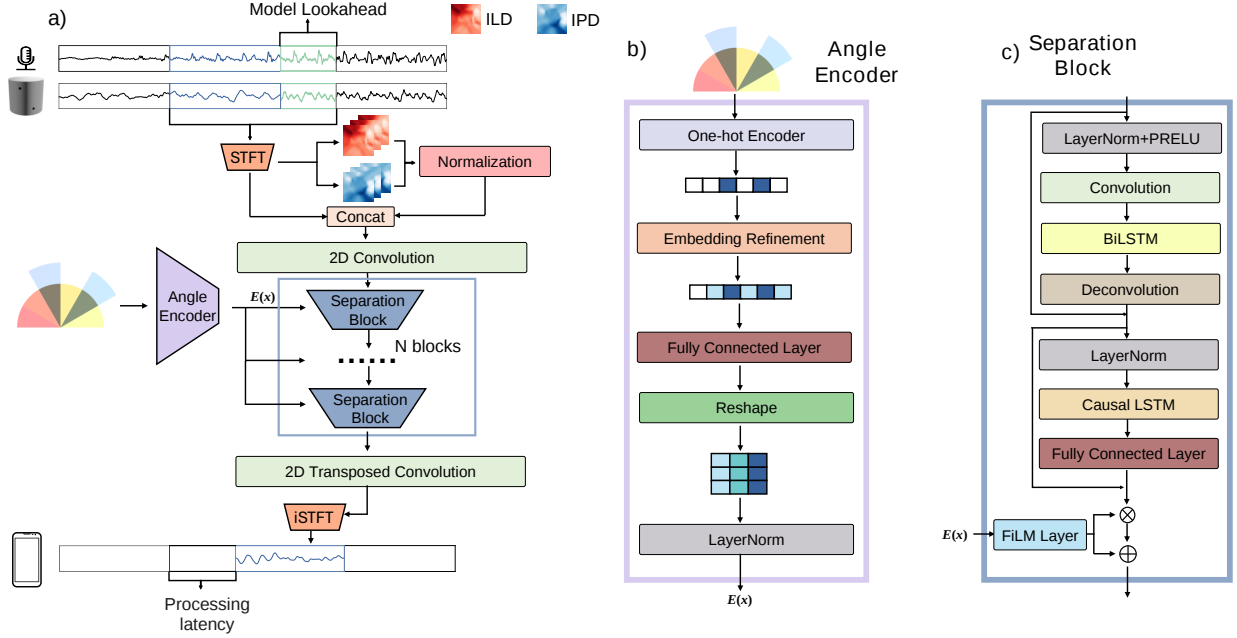


Figure 9: Real-time directional speech extraction network architecture. (a) Audio from the reference mic and the microstructure mic is converted to a spectrogram, ILD/IPD features are extracted, and directional separation is performed based on a user-selected spatial sector. The transformed spectrogram is then converted back to the time domain. (b) The angle encoder converts the selected sector into a normalized embedding for conditioning the separation blocks. (c) Separation blocks use this angle-conditioned embedding to apply directional masking on the spectrogram.

$$X_{feature}[f, t] = \{\cos(\text{IPD}[f, t]), \sin(\text{IPD}[f, t]), \text{ILD}[f, t]\} \quad (8)$$

We normalize each feature across frequency bands to ensure consistent scaling and improve training stability. For ILD, per-frequency normalization to zero mean and unit variance reduces the impact of frequency-dependent variations inherent in the microstructure design. The cosine and sine IPD components receive similar normalization to balance their influence across frequencies where phase relationships have varying reliability.

The complete feature set combines the real and imaginary STFT components from both microphone channels (4 channels) with our three spatial features, yielding 7 feature channels total. A 2D convolutional block encodes these into the representation $X \in \mathbb{R}^{c \times F \times \frac{(T+\sigma-L)}{\delta}}$, where c is the expanded feature dimension.

Angle Encoder. We represent user directional selections through a 6-sector encoding system. Selected sectors receive value 1 in a one-hot vector, while adjacent sectors receive 0.25 to provide smooth spatial context. This vector undergoes embedding refinement, fully connected transformations, and reshaping to create directional embeddings $E(x)$ matching our feature dimensions. After LayerNorm, these embeddings condition the separation blocks through Feature-wise Linear Modulation (FiLM).

Real-time Separation Blocks. Our processing pipeline consists of N sequential LSTM-based blocks optimized for real-time performance. Each block applies LayerNorm and PReLU activation, followed by convolutional downsampling to reduce computational

cost for the bidirectional LSTM. After frequency-domain processing, deconvolution with skip connections restores original dimensions.

A causal LSTM with fully connected layer processes temporal relationships while maintaining real-time constraints. The FiLM layer integrates directional information by modulating features according to $E(x)$:

$$E_{sep} = X' \odot \text{Conv1D}(E(x)) + \text{Conv1D}(E(x)) \quad (9)$$

where X' represents processed features and $\odot$ denotes element-wise multiplication.

The final output undergoes transposed convolution and inverse STFT to produce time-domain audio. For continuous processing, we output T samples for playback while retaining σ samples for subsequent chunk processing, enabling seamless real-time operation.

3.5 Real-world dataset

Motivation and Overview. Acoustic microstructures encode spatial cues through complex interactions between incoming sound waves and their geometric features. However, these spatial signatures can be significantly altered by environmental acoustic properties such as reverberation, room geometry, and background noise [64]. A model trained solely on clean, simulated data may fail to generalize across the diverse acoustic environments encountered in real-world deployments.

To address this challenge, we construct a comprehensive real-world dataset that captures the acoustic variability our system will

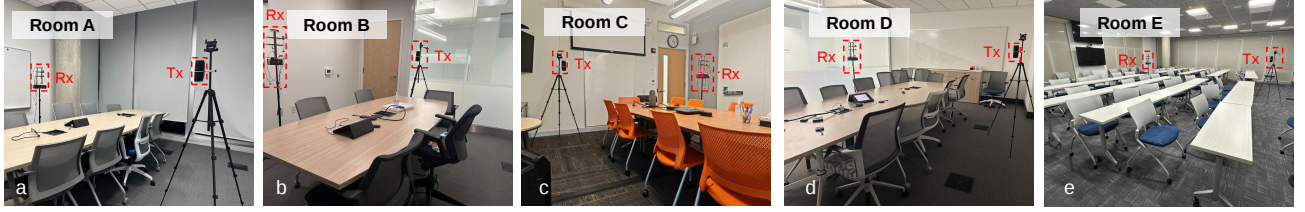


Figure 10: Real-world evaluation setup. (a–e) Five different test environments used for evaluating our system. The room dimensions are as follows (in meters): length $\times$ width $\times$ height. (a) Room A, $6.1 \times 4.6 \times 3.7$; (b) Room B, $4.9 \times 3.6 \times 3.5$; (c) Room C, $6.1 \times 4.6 \times 3.7$; (d) Room D, $7.0 \times 4.9 \times 2.7$; (e) Room E, $11.5 \times 9.1 \times 2.9$. In each room we play sound samples from an external speaker (model name: JBL Flip 5), and record it with multiple devices on our data capture rig.

encounter in practice. Rather than relying on synthetic room impulse responses or simulated acoustics, we physically deploy our hardware in multiple real environments and record actual speech playback. This approach ensures our training data reflects the true complexity of how our microstructure interacts with real acoustic spaces, including reflections, ambient noise, and other environmental factors that synthetic approaches may not accurately model.

Our data collection process involves five distinct rooms with varying acoustic properties, multiple speaker-receiver configurations, and systematic coverage of all directional sectors. This methodology produces a diverse dataset that enables robust performance across different meeting rooms, conference spaces, and classrooms. **Data Sources.** We use speech content from the VCTK [51] and LibriTTS [72] datasets, which provide high-quality recordings of human speech from multiple speakers. All audio clips are resampled to 24 kHz to balance audio quality with real-time processing requirements on mobile devices.

Collection Setup and Procedure. Our data collection setup consists of a JBL Flip 5 speaker⁴ for audio playback and a receiver rig equipped with both a MiniDSP microphone array⁵ and our *SonicSieve* system (Motorola Edge smartphone with microstructure). The receiver rig includes six LED lasers that divide the surrounding space into 30° sectors, enabling precise speaker placement.

We systematically collect data across five different rooms (meeting rooms, conference rooms, and classrooms) as shown in Fig. 10. For each room, we place the receiver rig at 3 randomly selected positions. At each receiver position, we place the speaker at 5 different locations distributed across all 6 directional sectors. For each speaker-receiver configuration, we play 6 random 5-second speech clips, resulting in 2,700 individual recordings (6 clips $\times$ 6 sectors $\times$ 5 speaker positions $\times$ 3 receiver positions $\times$ 5 rooms).

Mixture Generation and Dataset Statistics. To create training data for different directional extraction scenarios, we generate synthetic mixtures in software by combining recordings from different sectors. We create three types of mixtures: single-target (6 possible sector selections), two-target (15 combinations), and three-target (20 combinations) scenarios.

For each room and receiver position, this produces 1,200 single-target mixtures, 900 two-target mixtures, and 900 three-target mixtures. We control the Signal-to-Noise Ratio (SNR) between target

and interfering sectors within -5 to 5 dB by scaling the audio appropriately. The ground truth for each mixture is the clean combination of target sector signals as recorded by the reference microphone.

Our complete dataset comprises 45,000 mixtures across 5 rooms. We employ leave-one-room-out validation [63] to ensure generalization, reserving one room’s data (9,000 mixtures) for testing while using the remaining rooms for training. We additionally reserve 1,200 samples from the training data for validation.

Training Methodology. We process audio in 8 ms chunks (192 samples at 24 kHz) with 4 ms lookahead for real-time performance. The STFT uses a window size of 288 samples with 192-sample shifts, and we apply 96-sample zero padding at signal boundaries. We optimize using a Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) loss from the Asteroid package [38]:

$$\mathcal{L}_{\text{SI-SDR}} = \begin{cases} \lambda \|\hat{s} - s\|_1 & \text{no speakers in target area} \\ -10 \log \left[\frac{\|\alpha \cdot s\|^2}{\|\alpha \cdot s - \hat{s}\|^2} \right] & \text{otherwise} \end{cases} \quad (10)$$

where s is the target clean sound, $\hat{s}$ is the network output, $\lambda = 50$ is a weighting factor, and $\alpha = \frac{\hat{s}^T s}{\|\hat{s}\|^2}$ provides scale invariance.

We train for 150 epochs using the Adam optimizer [27]. The learning rate starts at 5×10^{-4} , increases to 5×10^{-3} over 10 epochs, remains constant for 20 epochs, then decays by 0.95 every 2 epochs. We apply data augmentation (Time Shift, Time Stretch, Frequency Mask, and Gain Perturbation) with 30% probability for each method [62].

4 RESULTS

4.1 Model results

For our evaluation, we began by using the real-world datasets (Fig. 10) collected in rooms B-E for training and validation, while testing on data from room A. We further present cross-room evaluation results at the end of the section. In all our evaluations, we use data from the in-line mic version of our system.

System performance across different sectors. We measured the SI-SDR across each of the six sectors in two conditions: before applying the acoustic microstructure and after applying it, with only a single sector selected in both cases. As shown in Fig. 11a, applying the microstructure consistently achieve higher SI-SDR improvement (SI-SDRi) across all sectors (range: 3.6–6.8 dB; mean: 4.5 dB). Interestingly, we found that performance varies across

⁴<https://www.jbl.com/bluetooth-speakers/JBL+FLIP+5-.html>

⁵<https://www.minidsp.com/products/usb-audio-interface/uma-8-microphone-array>

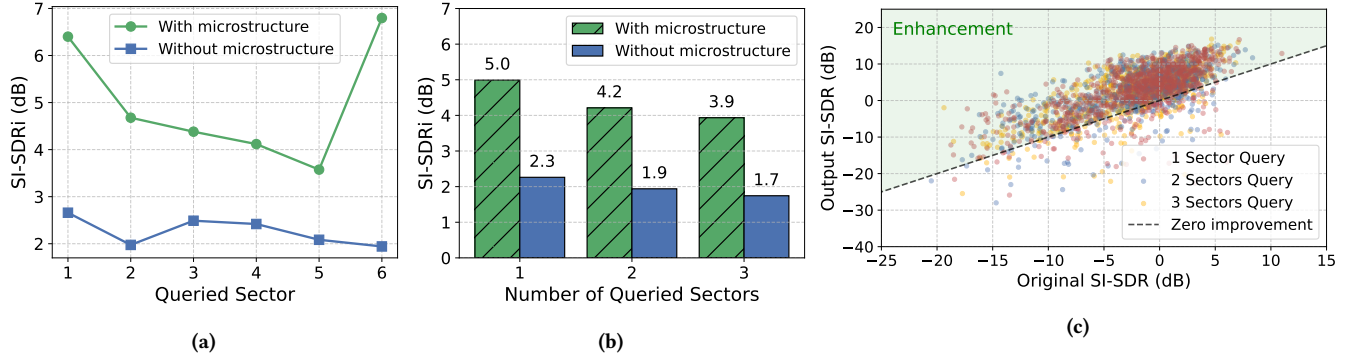


Figure 11: Overall System performance. (a) Effect of selected sector index on SI-SDR improvement. (b) SI-SDR improvement as a function of the number of simultaneously selected sectors. (c) Input vs. output SI-SDR of our model across all samples, grouped by 1-sector, 2-sector, and 3-sector selections. Points above the dashed line indicate cases where SI-SDR improved.

sectors due to the non-uniform distribution of holes in the microstructure. Specifically, sector 1 and 6 achieve an SI-SDRi of ~ 6 -7 dB, while sector 4 and 5 achieve around ~ 4 dB. This variation is consistent with the characterization of the microstructure’s spatial diversity for different angles as shown in Fig. 6, which shows that sectors 1 and 6 correspond to angles with higher spatial diversity, while sectors 3 to 5 exhibit reduced diversity. This characteristic can be leveraged in real-world use where users can rotate the device such that the target speaker is contained by sector 1 or 6 to achieve better directional filtering performance.

Effect of number of selected sectors. As our model is designed to support simultaneous selection of up to three sectors, we conducted a comprehensive evaluation of system performance across all possible sector combinations. This includes evaluating all individual sector (1-sector selection), all possible pairs of sectors (2-sector combinations), and all possible triplets of sectors (3-sector combinations). Each configuration was evaluated both with and without the acoustic microstructure. Our results in Fig. 11b reveal two key trends. First, the microstructure-enabled system consistently outperforms the baseline without the microstructure, regardless of selected number. When a single sector is selected, our system achieves a SI-SDRi of 5.0 dB, compared to 2.3 dB for the baseline. With two sectors selected, performance drops slightly to 4.2 dB, but still maintaining an advantage over the baseline of 1.9 dB. When three sectors are selected, the SI-SDRi with the microstructure is 3.9 dB which is greater than the baseline’s 1.7 dB. The second trend is that SI-SDRi decreases as the number of selected sectors increases for both systems. This is expected as extracting speech from multiple directions becomes more challenging when the spatial geometry of the selected region becomes more complex.

Fig. 11c shows the relationship between the input SI-SDR of the speech samples in our dataset and the output SI-SDR after processing with our model when 1, 2, or 3 sectors are selected. The diagonal dashed line represents zero improvement, where the output SI-SDR equals the input. Points above this line indicate enhanced speech quality after processing. As evident from the figure, the vast majority of our test samples (86.6%) demonstrate positive enhancement, with output SI-SDR values consistently higher than their original values. Notably, the one-sector selection configuration (represented

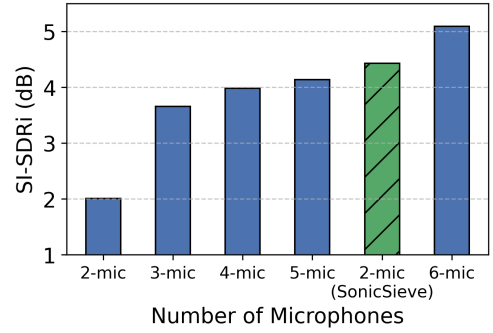


Figure 12: Comparison of SI-SDR improvement across microphone arrays of varying sizes versus our *SonicSieve* system.

by red dots) shows the fewest samples (9.9%) falling below the zero-improvement line compared to two- and three-sector selections (15.2% and 16.3%, respectively), confirming our earlier observation that the task becomes more challenging as the number of selected sectors increases.

Comparison against microphone array and traditional beamforming. We compared the end-to-end performance of *SonicSieve* against a conventional microphone array (UMA-8 USB mic array V2.0) as a baseline. Our network design is adapted from TF-GridNet [59], a state-of-the-art real-time sound separation architecture that supports arbitrary numbers of input channels, enabling fair comparison across different microphone configurations. For this baseline, we used 2 to 6 channels from the array and trained separate neural networks for each configuration. All baseline models were trained on the same dataset with the same architecture as *SonicSieve*, except for minor modifications to accommodate the varying number of input channels.

Fig. 12 shows the SI-SDRi, computed as the difference in SI-SDR between the sound mixture at the reference microphone and the enhanced speech output from the model. As the number of microphone channels increases from 2 to 6, SI-SDRi improves from 2.1 dB to 5.1 dB, due to the increased spatial diversity. *SonicSieve* achieves

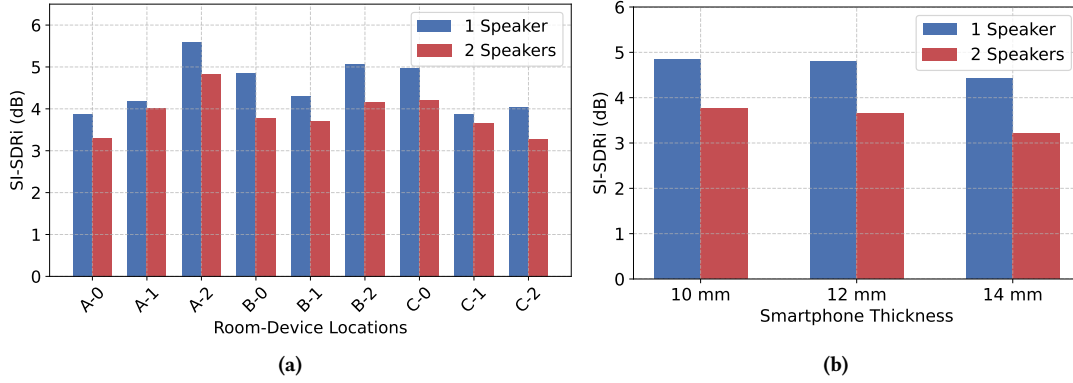


Figure 13: Stratified analysis by system variables. (a) Average SI-SDR improvement across the 3 rooms (Room A, B, C) in our dataset, computed from 3 random locations (location 0, 1, 2) per room. Results are categorized by scenarios with 1 or 2 target speakers and averaged over 3000 samples per location. (b) Average SI-SDR improvement with different smartphone thickness. The experiment is conducted at the B-0 location with three prototypes.

an SI-SDRi of 4.4 dB, outperforming the 5-microphone configuration (4.1 dB), and approaching the performance of the 6-microphone configuration (5.1 dB). Notably, the 2-microphone neural baseline achieves only 2.1 dB SI-SDRi, demonstrating that *SonicSieve*'s 4.4 dB improvement stems from the microstructure's ability to fundamentally enrich spatial cues beyond what microphone spacing alone provides.

We also evaluated the classical MVDR (Minimum Variance Distortionless Response) beamforming algorithm on the microphone array dataset. MVDR was configured to beamform toward the center of each selected sector (e.g., 15° for sector 1 spanning 0° to 30°). However, MVDR achieved less than 2 dB SI-SDRi even when using all 6 microphone channels. This poor performance stems from several fundamental limitations: (1) the coarse 30° sector resolution cannot provide the precise direction-of-arrival information that MVDR requires for effective spatial filtering, (2) MVDR's linear formulation struggles to handle complex multipath propagation and reverberation, which diffuse the spatial cues, (3) MVDR cannot effectively separate speech from environmental noise when they originate from similar directions, as it relies solely on spatial cues rather than the richer acoustic signatures that neural models can exploit.

Generalization across real-world environments. The previous results were developed using datasets from Rooms B–E and evaluated in Room A. To further evaluate generalizability of *SonicSieve* across environments, we conducted leave-one-room-out cross-validation for Rooms B and C. Rooms D and E were excluded from this analysis due to data collection inconsistencies in the data collection procedure (with portions of data missing). Fig. 13a presents the SI-SDRi performance across various device locations in different rooms, with each label indicating the test room (A, B, or C) and device position (0, 1, or 2) within that room. For single-speaker scenarios (blue bars), our system achieves consistent performance across all test conditions, with SI-SDRi values ranging from 3.9 to 5.7 dB. The highest performance is observed at location A-2 (5.7 dB), while the lowest occurs at C-1 (3.9 dB). Similarly, for two-speaker

scenarios (red hatched bars), our system maintains stable performance across all environments, with SI-SDRi values between 3.3 and 4.8 dB. The highest performance is observed at location A-2 (4.8 dB), while the lowest occurs at C-2 (3.3 dB).

Impact of smartphone thickness. We evaluated how smartphone thickness affects system performance since our system is designed to generalize to different smartphone geometries. Using our default prototype with the Motorola Edge 2024 (10 mm thickness), we modeled the effects of thicker devices by adding foam tape between the phone screen and microstructure to simulate 12 mm and 14 mm thickness. Fig. 13b shows that *SonicSieve* maintains robust performance across this thickness range. For single-speaker scenarios, SI-SDRi decreases slightly from 4.84 dB at 10 mm to 4.79 dB at 12 mm and 4.43 dB at 14 mm. Similarly, for two-speaker scenarios, performance drops from 3.77 dB at 10 mm to 3.65 dB at 12 mm and 3.21 dB at 14 mm. While thicker devices show modest performance reduction, the degradation is relatively small (less than 0.6 dB across the tested range). Since most off-the-shelf smartphones have thickness between 8–12 mm, our system demonstrates good compatibility with existing device form factors.

Effect of sector resolution. Although our model is mainly trained in 6 sector cases (30 degree resolution), we conducted evaluation on how the resolution affects model performance. We compared performance between 6 sectors and 9 sectors using simulation in Pyroomacoustics [43]. With one speaker in the target sectors, *SonicSieve* achieves 7.4 dB improvement in the 9 sector case and 7.9 dB improvement in the 6 sector case. With two speakers in the target sectors, *SonicSieve* achieves 5.0 dB improvement in the 9 sector case and 5.6 dB improvement in the 6 sector case. This demonstrates that even with more sectors and a more challenging task, *SonicSieve* still achieves good performance.

Model Latency. We evaluated the latency of our directional sound extraction neural network on a variety of different smartphones models: Samsung Galaxy S21, Google Pixel 7, and Motorola Edge 2024. To do this, we exported our PyTorch model to an ONNX model that could be run on these mobile devices. We used an audio chunk size of 8 ms, which we used in our main evaluations. We

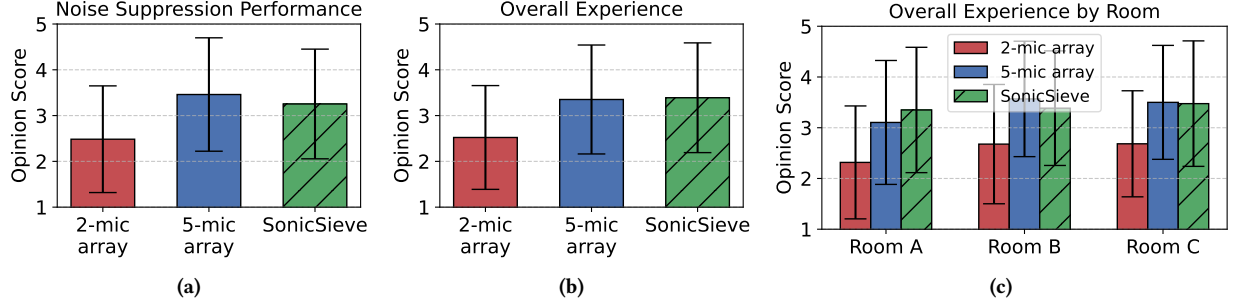


Figure 14: User study measuring listening experience via mean opinion scores: (a) audibility of interfering speakers and background noise, (b) overall listening quality, (c) overall listening quality across the audio samples collected in different rooms

ran the model for 1000 iterations and obtained mean and standard deviation latencies ranging from 4.5 ± 0.2 to 7.2 ± 0.2 ms across the phone models which is less than the 8 ms requirements needed for real-time execution.

4.2 Subjective user evaluation of audio quality

While objective metrics like SI-SDR provide valuable quantitative assessment, they cannot fully capture the perceptual quality of speech enhancement systems in real-world scenarios. To evaluate the actual listening experience, we conduct a user study to complement our objective measurements. This study was approved by our Institutional Review Board (STUDY2025_00000146). All studies complied with relevant ethical regulations.

We randomly select 25 generated audio samples from our test set collected from three different rooms (Room A-C). The selected audio contains one target speech and 1-2 interfering sources (undesired speech or other sound). Each sample is processed by three systems: the 2-channel microphone array (baseline), 5-channel microphone array, and our proposed *SonicSieve*. 44 participants in total (31 males and 13 females) with an average age of 25 years were recruited for the listening test. The evaluation protocol consisted of three steps: (1) listening to the noisy audio mixture, (2) listening to the clean target speaker audio as reference, and (3) listening to the three processed outputs presented in random order to avoid position bias. Participants were asked to rate each processed audio sample on two criteria using 5-point scales. We note that two separate questions [42] are used to decouple the assessment of noise removal from the preservation of target speech quality, as aggressive noise suppression can often introduce undesirable artifacts that a single question would fail to capture.

- (1) Noise Suppression Performance: *How intrusive/noticeable were the interfering speakers and background noises? (1 – Very intrusive to 5 – Not noticeable)*
- (2) Overall Experience: *If the goal is to focus on the target speaker, how was your overall experience? (1 – Very intrusive to 5 – Not noticeable)*

As shown in Fig. 14, both *SonicSieve* and the 5-channel microphone array significantly outperform the 2-channel microphone baseline in subjective evaluations. Across the 25 samples drawn from three rooms, for noise suppression performance (Fig. 14a), *SonicSieve* achieved a Mean Opinion Score (MOS) of 3.25 ± 1.20 , compared to the 5-microphone array's 3.46 ± 1.24 and the 2-channel

baseline's 2.48 ± 1.17 . For overall listening experience (Fig. 14c), *SonicSieve* achieved 3.39 ± 1.20 , comparable to the 5-microphone array (3.35 ± 1.19) and higher than the 2-channel baseline (2.52 ± 1.13). Fig. 14(c) further breaks down overall experience by room, showing consistent gains over the 2-channel baseline across Room A–C, with *SonicSieve* remaining comparable to the 5-channel array in each environment.

4.3 User Interface Evaluation

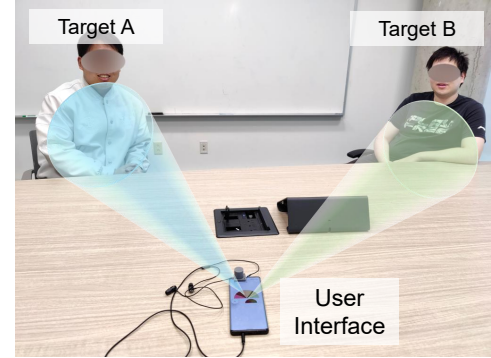


Figure 15: User interface evaluation setup for directional speech extraction, where the participants are asked to select one or both speakers on the smartphone UI with different configurations in different environments.

While our objective metrics demonstrate technical performance, understanding how users interact with directional selection in real-world scenarios is critical for practical deployment. We conduct a user study evaluating the usability of our interface across different sector resolution configurations, examining both selection accuracy and user confidence when targeting speakers in multi-person environments.

We recruit 10 participants to evaluate two configurations: our default 6-sector design (30° resolution) and a 9-sector design (20° resolution). As shown in Fig. 15, participants were positioned in three conference rooms with two experimenters as target speakers. For each configuration, participants completed three tasks: selecting target A, target B, and both targets simultaneously. Experimenter positions remained consistent across configurations while task order was randomized to minimize learning effects.

Table 2: Comparison of system performance and usability across different sector resolutions

Config	Resolution	Preference	Accuracy	Confidence (Single)	Confidence (Multiple)
6 Sectors	30°	0.96	0.90	0.82	1.00
9 Sectors	20°	0.88	0.87	0.70	0.86

Participants were instructed to select single or multiple sectors they believed necessary to capture the target speaker(s). We evaluated interface usability through three metrics: (1) *Selection Accuracy*, measured as the probability that selected sectors correctly included the target speakers. (2) *User Confidence*, assessed on a 5-point scale (5=very confident, 1=not confident at all) for both single and multiple speaker selections. (3) *Overall Preference*, rated on a 5-point scale (5=strongly prefer to use, 1=strongly disprefer to use) comparing the two configurations.

Table 2 presents our results by normalizing the scores and accuracy to 0–1. The results reveal that the 6-sector configuration consistently outperforms the 9-sector design across all metrics, achieving higher selection accuracy (0.90 vs. 0.87), stronger user preference (0.96 vs. 0.88), and greater confidence for both single-sector (0.82 vs. 0.70) and multiple-sector (1.00 vs. 0.86) scenarios. These findings suggest that finer angular resolution does not necessarily improve usability—instead, the increased cognitive load of managing more sectors can reduce both accuracy and confidence.

We observed two key factors driving this preference. First, the coarser 6-sector resolution provides sufficient spatial granularity for typical meeting and lecture scenarios where speakers are distributed around a table or room, while remaining simple enough for rapid selection without requiring precise angular estimation. Second, participants reported that the 9-sector interface felt "overwhelming" and made them second-guess their selections, particularly when attempting to capture multiple speakers. The perfect confidence score (1.00) for multiple-sector selection with 6 sectors suggests that users found it intuitive to select adjacent sectors to create broader capture regions when needed.

These results validate our design choice of 6 sectors as providing an effective balance between spatial resolution and interaction simplicity. For future work, adaptive interfaces that automatically adjust sector granularity based on speaker density could potentially combine the benefits of both approaches.

5 LIMITATIONS AND DISCUSSION

Dual-microphone design rationale. *SonicSieve* requires two closely spaced microphones for robust directional speech extraction. While the microstructure embeds direction-dependent spectral cues into incoming sound, these cues become entangled with the source’s intrinsic spectrum and environmental reflections in single-channel recordings. A dual-microphone design provides a "before/after" comparison—one channel shaped by the microstructure, one unmodified—that allows the neural network to factor out source/room differences and isolate the directional signature. Our choice of an in-line microphone as the second channel offers two key advantages: (1) It can be positioned very close to the phone’s built-in top microphone (within 1 cm in our prototype), which our ablations (Fig. 8) show is critical for cross-environment robustness;

(2) It is device-agnostic and can work with smartphones that have a standard audio jack or USB-C port.

Smartphone generalizability. Our system design prioritizes compatibility across different smartphone models through several key considerations:

- *Microphone layout adaptation.* Our smartphone application adapts to different device configurations by providing model-specific placement instructions. The UI guides users to position the microstructure optimally based on their device’s microphone locations, ensuring proper alignment between the in-line microphone and the smartphone’s top microphone (Fig. 7).
- *Generalization across smartphone thicknesses.* Our evaluation across devices ranging from 10 to 14 mm in thickness (Fig. 13b) shows robust performance with only modest degradation for thicker smartphones. These results indicate that the system can generalize across different form factors.
- *Generalization across microphone frequency responses.* To enable generalization across different smartphones microphones, device-specific calibration can be performed by measuring the frequency response of the built-in and in-line microphones, and deriving equalization weights to normalize for their effects. This can allow our system to robustly perform directional speech extraction across a range of devices. A complementary approach is the use of established data-driven approaches such as training on diverse device recordings [4, 31] and targeted data augmentation techniques [35, 68] which may enable zero-shot generalization without calibration for unseen devices.

Pathways to device-native integration. While our current prototype uses a microstructure attached to an external wired earphones, the same core concept can be integrated directly into devices with two closely located microphones. Several integration pathways are feasible: (1) Smartphones with a top and a rear microphone near the camera (e.g., Google Pixel 7) often have <2 cm spacing. A microstructure could be embedded near the rear mic while using the top mic as reference (Fig. 16a). (2) Standalone voice recorders with dual front-facing microphones could house the microstructure in a compact form factor (Fig. 16b). (3) Future smartphone designs could incorporate a miniaturized microstructure as part of the device chassis, with one microphone inside the structure and one outside.

Source separation within a sector. Our system currently performs spatial filtering along predefined directional sectors and produces a mixture of all sounds it. This can be undesirable when multiple speakers are close to each other and fall within the same sector. Future extensions of this work could address this limitation in two ways: (1) A low-latency speech separation framework [23, 24, 69] can be applied after directional speech extraction, using permutation-invariant training to separate multiple speakers. (2)



Figure 16: Applications of SonicSieve. (a–d) Our system can be physically connected to various computing devices with a microphone jack including smartphones, handheld digital voice recorders, tablets, and laptops. (e–h) *SonicSieve* can be used in various application scenarios including (e) Transcribing a presentation with speaker attribution. (f) Conducting a remote meeting by a noisy park. (g) Recording a podcast without specialized soundproof booths. (h) Interacting with an AI voice assistant in a noisy restaurant.

More fine-grained separation can be performed in the spatial domain by building on techniques for region-customizable sound extraction [22] to spatially isolate the target speaker.

Speaker mobility. Our current system is designed primarily for scenarios where the speaker remains relatively stationary such as in meetings. If the speaker moves, the user is expected to manually update the target sector. However, this is not a fundamental limitation of our neural network approach. Future extensions of the model could be explicitly trained to estimate the direction of arrival (DoA) for multiple speakers [16, 57, 60], and enable dynamic conditioning of the directional speech extraction network.

Audio playback. While our current system is focused on speech transcription and remote streaming, our model’s processing latency is low enough for real-time, on-device playback of the extracted audio to users. With a smartphone-native design (Fig. 16a), playback can be achieved using noise-canceling headphones, similar to prior AI-enabled acoustic systems [11, 52], which leverage both passive and active noise cancellation to suppress ambient sounds and deliver clear playback of the extracted sounds.

A limitation of the current in-line microphone setup is the close proximity of the mic to the earphones, resulting in insufficient wire length to comfortably reach a user’s ears when the smartphone is placed on a table. This can be addressed in two ways. First, a custom but low-cost headphone could be designed with the in-line mic positioned farther from the speakers, allowing the phone to remain on a surface while maintaining sufficient wire length. Alternatively, wireless headphones could be used for playback by integrating the acoustic microstructure onto one of the outward-facing microphones typically used for noise cancellation. This approach is likely possible as modern headphones often include dual outward-facing microphones for noise cancellation, and low-power GPUs are becoming integrated into wearable devices [1, 47].

Human-centered interaction and deployment considerations. While our user study (Sec. 4.2 and 4.3) demonstrates the system

performance and UI usability in controlled settings, broader human-centered aspects require deeper investigation. Real-world conversational dynamics, including natural turn-taking, situational awareness of non-target speakers, and adaptation to speaker movement, present challenges beyond spatial selection. Accessibility considerations, such as how users with hearing challenges interact with directional audio interfaces, require dedicated study. Similarly, longitudinal in-the-wild deployment across diverse acoustic environments would reveal how interaction patterns evolve over time. Our current work establishes the technical foundation and demonstrates basic usability. These human-centered dimensions represent valuable directions for future research toward broader real-world adoption.

References

- [1] 2025. GPU-WEAR, Ultra-low power heterogeneous Graphics Processing Units for Wearable/IoT devices. <https://cordis.europa.eu/project/id/717850>
- [2] Hidri Adel, Meddeb Souad, Abdulqadir Alaqeeli, and Amiri Hamid. 2012. Beam-forming techniques for multichannel audio signal separation. *arXiv preprint arXiv:1212.6080* (2012).
- [3] Shoko Araki, Hiroshi Sawada, and Shoji Makino. 2007. Blind speech separation in a meeting situation with maximum SNR beamformers. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 1. IEEE, 1–41.
- [4] Rosana Ardila, Megan Branson, Kelly Davis, et al. 2019. Common voice: A massively-multilingual speech corpus. *Proceedings of the 12th Language Resources and Evaluation Conference* (2019).
- [5] Arun Arjunan, Ahmad Baroutaji, John Robinson, Aaron Vance, and Abul Arafat. 2024. Acoustic metamaterials for sound absorption and insulation in buildings. *Building and Environment* 251 (2024), 111250.
- [6] Yang Bai, Nakul Garg, and Nirupam Roy. 2022. Spidr: Ultra-low-power acoustic spatial sensing for micro-robot navigation. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 99–113.
- [7] Jens Blauert. 1997. *Spatial hearing: the psychophysics of human sound localization*. MIT press.
- [8] Andrew Brughera, Jason Mikiel-Hunter, Mathias Dietz, and David McAlpine. 2019. Brainstem biophysics contribute to sound-source localisation in reverberant scenes. *bioRxiv* (2019). doi:10.1101/694356
- [9] Gaoshuai Cao, Kuang Yuan, Jie Xiong, Panlong Yang, Yubo Yan, Hao Zhou, and Xiang-Yang Li. 2020. Earphonetrack: involving earphones into the ecosystem of acoustic motion tracking. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 95–108.

- [10] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven M Seitz. 2022. ClearBuds: wireless binaural earbuds for learning-based speech enhancement. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 384–396.
- [11] Tuochao Chen, Malek Itani, Sefik Emre Eskimez, Takuya Yoshioka, and Shyamnath Gollakota. 2024. Hearable devices with sound bubbles. *Nature Electronics* (2024), 1–12.
- [12] Zhuo Chen, Xiong Xiao, Takuya Yoshioka, Hakan Erdogan, Jinyu Li, and Yifan Gong. 2018. Multi-channel overlapped speech recognition with location guided speech extraction network. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 558–565.
- [13] Roger B Coles and Anna Guppy. 1988. Directional hearing in the barn owl (Tyto alba). *Journal of Comparative Physiology A* 163, 1 (1988), 117–133.
- [14] Tobias Cord-Landwehr, Christoph Boeddeker, Thilo Von Neumann, Cătălin Zorilă, Rama Doddipatla, and Reinhold Haeb-Umbach. 2022. Monaural source separation: From anechoic to reverberant environments. In *2022 international workshop on acoustic signal enhancement (IWAENC)*. IEEE, 1–5.
- [15] Dhvani Desai and Ninad Mehendale. 2022. A Review on Sound Source Localization Systems. *Archives of Computational Methods in Engineering* 29 (05 2022). doi:10.1007/s11831-022-09747-2
- [16] Nilanjan Dey and Amira S Ashour. 2018. *Direction of arrival estimation and localization of multi-speech sources*. Springer.
- [17] Thomas Dupont, Philippe Leclaire, Raymond Panneton, and Olga Umnova. 2018. A microstructure material design for low frequency sound absorption. *Applied Acoustics* 136 (2018), 86–93.
- [18] Andrew Franc and Josh McDermott. 2022. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behaviour* 6 (01 2022), 111–133. doi:10.1038/s41562-021-01244-z
- [19] Nakul Garg, Yang Bai, and Nirupam Roy. 2021. Owllet: Enabling spatial information in ubiquitous acoustic devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 255–268.
- [20] Nikolay D Gaubitch, Jorge Martinez, W Bastiaan Kleijn, and Richard Heusdens. 2014. On near-field beamforming with smartphone-based ad-hoc microphone arrays. In *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 94–98.
- [21] Rongzhi Gu, Lianwu Chen, Shi-Xiong Zhang, Jimeng Zheng, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, and Dong Yu. 2019. Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information.. In *Interspeech*. 4290–4294.
- [22] Rongzhi Gu and Yi Luo. 2024. Rezero: Region-customizable sound extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).
- [23] Rongzhi Gu, Shi-Xiong Zhang, Meng Yu, and Dong Yu. 2021. 3d spatial features for multi-channel target speech separation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 996–1002.
- [24] Rongzhi Gu, Shi-Xiong Zhang, Yuexian Zou, and Dong Yu. 2022. Towards unified all-neural beamforming for time and frequency domain speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), 849–862.
- [25] Rongzhi Gu and Yuexian Zou. 2020. Temporal-spatial neural filter: Direction informed end-to-end multi-channel target speech separation. *arXiv preprint arXiv:2001.00391* (2020).
- [26] Teerapat Jenrungrot, Vivek Jayaram, Steve Seitz, and Ira Kemelmacher-Shlizerman. 2020. The cone of silence: Speech separation by localization. *Advances in Neural Information Processing Systems* 33 (2020), 20925–20938.
- [27] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [28] Anton Kovalyov, Kashyap Patel, and Issa Panahi. 2023. DSEnet: Directional signal extraction network for hearing improvement on edge devices. *IEEE Access* 11 (2023), 4350–4358.
- [29] Abdullah Küçük, Anshuman Ganguly, Yiya Hao, and Issa MS Panahi. 2019. Real-time convolutional neural network-based speech source localization on smartphone. *IEEE Access* 7 (2019), 169969–169978.
- [30] Milind N Kunchur. 2023. The human auditory system and audio. *Applied Acoustics* 211 (2023), 109507.
- [31] Bo Li, Khe Chai Sim, Tara N Sainath, Ron J Weiss, Kevin W Wilson, et al. 2017. Acoustic modeling for Google home. *Proc. Interspeech* (2017).
- [32] Dingzeyu Li, David IW Levin, Wojciech Matusik, and Changxi Zheng. 2016. Acoustic voxels: Computational optimization of modular acoustic filters. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–12.
- [33] Guinan Li, Xue Bao, and Zhi Wang. 2017. The design and implementation of a smartphone-based acoustic array system for DOA estimation. In *2017 36th Chinese Control Conference (CCC)*. IEEE, 5416–5423.
- [34] Xin Li, Yilin Yang, Zhengkun Ye, Yan Wang, and Yingying Chen. 2023. Earcase: Sound source localization leveraging mini acoustic structure equipped phone cases for hearing-challenged people. In *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*. 240–249.
- [35] Tobias Morocutti, Florian Schmid, Khaled Koutini, and Gerhard Widmer. 2023. Device-robust acoustic scene classification via impulse response augmentation. In *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 176–180.
- [36] Thi Ngoc Tho Nguyen, Karn N Watcharasupat, Ngoc Khanh Nguyen, Douglas L Jones, and Woon-Seng Gan. 2022. Salsa: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 1749–1762.
- [37] Ashutosh Pandey, Sanha Lee, Juan Azcarreta, Daniel Wong, and Buye Xu. 2024. All Neural Low-latency Directional Speech Extraction. *arXiv preprint arXiv:2407.04879* (2024).
- [38] Manuel Pariente, Samuele Cornell, Joris Cosentino, Sunit Sivasankaran, Efthymios Tzinis, Jens Heitkaemper, Michel Olvera, Fabian-Robert Stöter, Mathieu Hu, Juan M Martín-Doñas, et al. 2020. Asteroid: the PyTorch-based audio source separation toolkit for researchers. *arXiv preprint arXiv:2005.04132* (2020).
- [39] Santiago Pascual, Antonio Bonafonte, and Joan Serra. 2017. SEGAN: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452* (2017).
- [40] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. 2007. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *Proceedings of the 5th international conference on Embedded networked sensor systems*. 1–14.
- [41] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Dinei Florencio, and Mark Hasegawa-Johnson. 2018. Deep learning based speech beamforming. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5389–5393.
- [42] ITUT Recommendation. 2003. Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. *ITU-T recommendation* 835 (2003).
- [43] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. 2018. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 351–355.
- [44] Sheng Shen, Daguang Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury. 2020. Voice localization using nearby wall reflections. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.
- [45] Kazuki Shimada, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, and Yuki Mitsufuji. 2021. ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 915–919.
- [46] Sanjib Sur, Teng Wei, and Xinyu Zhang. 2014. Autodirective audio capturing through a synchronized smartphone array. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. 28–41.
- [47] Thierry Tambe, En-Yu Yang, Glenn G Ko, Yuji Chai, Coleman Hooper, Marco Donato, Paul N Whatmough, Alexander M Rush, David Brooks, and Gu-Yeon Wei. 2022. A 16-nm soc for noise-robust speech and nlp edge ai inference with bayesian sound source separation and attention-based dnns. *IEEE Journal of Solid-State Circuits* 58, 2 (2022), 569–581.
- [48] Kristina Tesch and Timo Gerkmann. 2022. Insights into deep non-linear filters for improved multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), 563–575.
- [49] Serkan Tokgöz, Anton Kovalyov, and Issa Panahi. 2020. Real-time estimation of direction of arrival of speech source using three microphones. In *2020 IEEE Workshop on Signal Processing Systems (SiPS)*. IEEE, 1–5.
- [50] John Van Opstal. 2016. *The auditory system and human sound-localization behavior*. Academic Press.
- [51] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2017. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)* 6 (2017), 15.
- [52] Bandhav Veluri, Malek Itani, Justin Chan, Takuya Yoshioka, and Shyamnath Gollakota. 2023. Semantic hearing: Programming acoustic scenes with binaural hearables. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–15.
- [53] Bandhav Veluri, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota. 2024. Look Once to Hear: Target Speech Hearing with Noisy Examples. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.
- [54] Anran Wang, Maruchi Kim, Hao Zhang, and Shyamnath Gollakota. 2022. Hybrid neural networks for on-device directional hearing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11421–11430.
- [55] Anran Wang, Dan Nguyen, Arun R Sridhar, and Shyamnath Gollakota. 2021. Using smart speakers to contactlessly monitor heart rhythms. *Communications biology* 4, 1 (2021), 319.
- [56] Anran Wang, Jacob E Sunshine, and Shyamnath Gollakota. 2019. Contactless infant monitoring using white noise. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.
- [57] Rui Wang, Li Li, and Tomoki Toda. 2024. Dual-Channel Target Speaker Extraction Based on Conditional Variational Autoencoder and Directional Information.

- IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), 1968–1979. doi:10.1109/TASLP.2024.3376154
- [58] Zeyu Wang, Xiyuxing Zhang, Ruotong Yu, Yuntao Wang, Kenneth Christofferson, Jingru Zhang, Alex Mariakakis, and Yuanchun Shi. 2024. DreamCatcher: A Wearer-aware Multi-modal Sleep Event Dataset Based on Earables in Non-restrictive Environments. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 85155–85178. https://proceedings.neurips.cc/paper_files/paper/2024/file/9ab8bb568825d49ce31aa87b7e2f4ad7-Paper-Datasets_and_Benchmarks_Track.pdf
- [59] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. 2023. TF-GridNet: Integrating full- and sub-band modeling for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 3221–3236.
- [60] Zhong-Qiu Wang, Xuiliang Zhang, and DeLiang Wang. 2019. Robust Speaker Localization Guided by Deep Learning-Based Time-Frequency Masking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 1 (2019), 178–188. doi:10.1109/TASLP.2018.2876169
- [61] Julian Wechsler, Srikanth Raj Chetupalli, Wolfgang Mack, and Emanuel A. P. Habets. 2023. Multi-Microphone Speaker Separation by Spatial Regions. arXiv:2303.07143 [eess.AS] <https://arxiv.org/abs/2303.07143>
- [62] Shengyun Wei, Shun Zou, Feifan Liao, and weimin lang. 2020. A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification. *Journal of Physics: Conference Series* 1453, 1 (jan 2020), 012085. doi:10.1088/1742-6596/1453/1/012085
- [63] Tzu-Tsung Wong. 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern recognition* 48, 9 (2015), 2839–2846.
- [64] Bo Wu, Kehuang Li, Minglei Yang, and Chin-Hui Lee. 2016. A reverberation-time-aware approach to speech dereverberation based on deep neural networks. *IEEE/ACM transactions on audio, speech, and language processing* 25, 1 (2016), 102–111.
- [65] Bosun Xie. 2013. *Head-related transfer function and virtual auditory display*. J. Ross Publishing.
- [66] Yilin Yang, Xin Li, Zhengkun Ye, Yan Wang, and Yingying Chen. 2023. BioCase: Privacy Protection via Acoustic Sensing of Finger Touches on Smartphone Case Mini-Structures. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 397–409.
- [67] Yang Yang, George Sung, Shao-Fu Shih, Hakan Erdogan, Chehung Lee, and Matthias Grundmann. 2024. Binaural angular separation network. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1201–1205.
- [68] Kuang Yuan, Yang Gao, Xilin Li, Xinhao Mei, Syavosh Zadissa, Tarun Pruthi, and Saeed Bagheri Sereshki. 2025. Lightweight and Generalizable Acoustic Scene Representations via Contrastive Fine-Tuning and Distillation. *arXiv preprint arXiv:2510.03728* (2025).
- [69] Kuang Yuan, Shuo Han, Swarun Kumar, and Bhiksha Raj. 2024. DeWinder: Single-Channel Wind Noise Reduction using Ultrasound Sensing. In *Proc. Interspeech 2024*. 627–631.
- [70] Kuang Yuan, Mohamed Ibrahim, Yiwen Song, Guoxiang Deng, Robert A Nerone, Suvendra Vijayan, Akshay Gadre, and Swarun Kumar. 2024. ToMoBrush: Exploring Dental Health Sensing Using a Sonic Toothbrush. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–27.
- [71] Kuang Yuan, Dong Li, Hao Zhou, Zhehao Li, Lili Qiu, Swarun Kumar, and Jie Xiong. 2025. WindDancer: Understanding Acoustic Sensing under Ambient Airflow. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 2 (2025).
- [72] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882* (2019).
- [73] Xiyuxing Zhang, Yuntao Wang, Yuxuan Han, Chen Liang, Ishan Chatterjee, Jiankai Tang, Xin Yi, Shwetak Patel, and Yuanchun Shi. 2024. The EarSAVAS Dataset: Enabling Subject-Aware Vocal Activity Sensing on Earables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 83 (May 2024), 26 pages. doi:10.1145/3659616
- [74] Xiyuxing Zhang, Yuntao Wang, Jingru Zhang, Yaqing Yang, Shwetak Patel, and Yuanchun Shi. 2023. EarCough: Enabling Continuous Subject Cough Event Detection on Hearables. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI EA '23). Association for Computing Machinery, New York, NY, USA, Article 94, 6 pages. doi:10.1145/3544549.3585903
- [75] Hao Zhou, Kuang Yuan, Mahanth Gowda, Lili Qiu, and Jie Xiong. 2024. Rethinking Orientation Estimation with Smartphone-equipped Ultra-wideband Chips. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 1045–1059.