Pushing the Limits of Acoustic Spatial Perception via Incident Angle Encoding

YONGJIAN FU, Central South University, China
YONGZHAO ZHANG, University of Electronic Science and Technology of China, China
HAO PAN, Shanghai Jiao Tong University, China
YU LU, Shanghai Jiao Tong University, China
XINYI LI, Tsinghua University, China
LILI CHEN, Tsinghua University, China
JU REN*, Tsinghua University, China and Zhongguancun Laboratory, China
XIONG LI, University of Electronic Science and Technology of China, China
XIAOSONG ZHANG, University of Electronic Science and Technology of China, China
YAOXUE ZHANG, Tsinghua University, China and Zhongguancun Laboratory, China

With the growing popularity of smart speakers, numerous novel acoustic sensing applications have been proposed for low-frequency human speech and high-frequency inaudible sounds. Spatial information plays a crucial role in these acoustic applications, enabling various location-based services. However, typically commercial microphone arrays face limitations in spatial perception of inaudible sounds due to their sparse array geometries optimized for low-frequency speech. In this paper, we introduce MetaAng, a system designed to augment microphone arrays by enabling wideband spatial perception across both speech signals and inaudible sounds by leveraging the spatial encoding capabilities of acoustic metasurfaces. Our design is grounded in the fact that, while sensitive to high-frequency signals, acoustic metasurfaces are almost non-responsive to low-frequency speech due to significant wavelength discrepancy. This observation allows us to integrate acoustic metasurfaces with sparse array geometry, simultaneously enhancing the spatial perception of high-frequency and low-frequency acoustic signals. To achieve this, we first utilize acoustic metasurfaces and a configuration optimization algorithm to encode the unique features for each incident angle. Then, we propose an unrolling soft thresholding network that employs neural-enhanced priors and compressive sensing for high-accuracy, high-resolution multi-source angle estimation. We implement a prototype, and experimental results demonstrate that MetaAng maintains robustness across various scenarios, facilitating multiple applications, including localization and tracking.

*Corresponding author.

Authors' addresses: Yongjian Fu, School of Computer Science and Engineering, Central South University, Changsha, China, fuyongjian@csu.edu.cn; Yongzhao Zhang, University of Electronic Science and Technology of China, Chengdu, China, zhangyongzhao@uestc.edu.cn; Hao Pan, Shanghai Jiao Tong University, Shanghai, China, panh09@sjtu.edu.cn; Yu Lu, Shanghai Jiao Tong University, Shanghai, China, yulu01@sjtu.edu.cn; Xinyi Li, Department of Computer Science and Technology, Tsinghua University, Beijing, China, xinyili@tsinghua.edu.cn; Lili Chen, Department of Computer Science and Technology, Tsinghua University, Beijing, China, lilichen@tsinghua.edu.cn; Ju Ren, Department of Computer Science and Technology, Tsinghua University, China and Zhongguancun Laboratory, Beijing, China, renju@tsinghua.edu.cn; Xiong Li, University of Electronic Science and Technology of China, Chengdu, China, lixiong@uestc.edu.cn; Xiaosong Zhang, University of Electronic Science and Technology of China, Chengdu, China, johnsonzxs@uestc.edu.cn; Yaoxue Zhang, Department of Computer Science and Technology, Tsinghua University, Beijing, China and Zhongguancun Laboratory, Beijing, China, zhangyx@tsinghua.edu.cn.

2474-9567/2024/5-ART52 $15.00
https://doi.org/10.1145/3659583

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Acoustic sensing, metasurface, angle estimation, compressive sensing

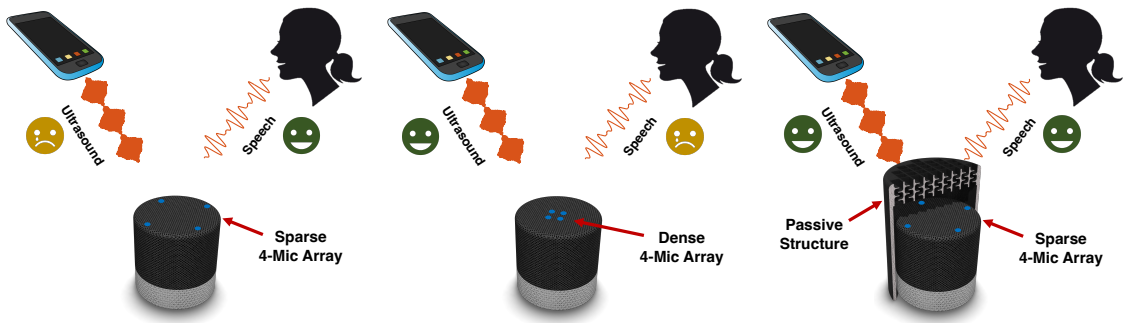
ACM Reference Format:
Yongjian Fu, Yongzhao Zhang, Hao Pan, Yu Lu, Xinyi Li, Lili Chen, Ju Ren, Xiong Li, Xiaosong Zhang, and Yaoxue Zhang. 2024. Pushing the Limits of Acoustic Spatial Perception via Incident Angle Encoding. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 52 (June 2024), 28 pages. https://doi.org/10.1145/3659583

1 INTRODUCTION

In recent years, the evolution of smart speakers has marked a significant milestone in the realm of consumer technology. With the adoption of microphone arrays, smart speakers can be integrated with capabilities like audio source separation [10, 58], speech enhancement [68], and voice localization [18, 53]. These features are underpinned by the speech interface (SI) that processes and responds to human speech in the audible band. Moreover, researchers also extend smart speaker capabilities beyond audible sound interactions, exploring the use of inaudible sounds for some novel applications, such as motion-based interaction [23, 34], health monitoring [52, 66], elderly fall detection [4, 26], and indoor localization [25]. By exploring the potential of the internal microphone array, smart speakers have a broad range of applications to improve everyday efficiency and interconnectivity.

Spatial perception, i.e., Angles-of-Arrival (AoAs) estimation, is crucial to enable many smart speaker applications involving spatial information (*e.g.*, localization and tracking), which leverages the internal microphone array to capture the spatial diversity of incoming sounds. The efficacy of spatial perception largely depends on the separation between two adjacent microphones (called inter-element separation). Generally, the inter-element separation should be smaller than the half wavelength of the operating frequency [31] (refers to a dense array), while larger separations may cause ambiguous angles [50, 52] (refers to a sparse array), *i.e.*, several possible solutions for incident angles. This phenomenon creates a dilemma for the design of microphone arrays to support both audible and inaudible sounds, as it is difficult to balance the need for inter-element separation between high-frequency and low-frequency signals. Specifically, the frequency range of speech signals is generally between 300 and 3400Hz [11], with a half wavelength larger than around 5cm (given the sound speed is 340m/s). While inaudible sound for acoustic sensing often operates around 20kHz [44, 56, 65], with a half wavelength of about 0.85cm, which is almost 6 times smaller than the minimum required inter-element separation for human speech. Therefore, a dense microphone array for speech signals is relatively sparse for inaudible sounds due to their wavelength discrepancy.

This dilemma is more challenging for Commercial-Off-The-Shelf (COTS) smart speakers, because they have only a few microphones evenly spaced on the outside of the body in a circular array. For example, the most recently announced Apple HomePod 2nd [1] (2023) and has 4 microphones for around 10cm inter-element separation, while the compact product Google Nest Mini [2] (2019) has 3 microphones for around 6cm inter-element separation. It implies that COTS smart speakers are mainly designed for speech interactions and are not ready for sensing inaudible sounds, as illustrated in Fig. 1(a). Instead, if we adopt a dense microphone array to support inaudible sounds, the performance of speech interfaces will be significantly affected since the inter-element spacing is far less than the half wavelength of speech, as shown in Fig. 1(b). Prior works [34, 54] adopt a non-uniform array to overcome the wavelength discrepancy, which, however, is not applicable to COTS smart speakers because they have only a few mics and adopt a circular array for omnidirectional sensing. The most recent works [15, 24] explore the acoustic stencils for extracting spatial information with only two microphones. However, these efforts primarily focus on low-frequency sound, and their designs face challenges in extending to multi-source scenarios. In this paper, we ask the following question: *Can we empower COTS sparse microphone arrays to estimate the*



(a) Commercial smart speaker with sparse microphone array

(b) Adopt dense microphone array

(c) Combine a passive shell with sparse microphone array

Fig. 1. Microphone arrays can be used to interact with human speech and also sense inaudible sounds. (a) Commercial smart speakers have only a few microphones sparsely distributed on the outside of the body in a circular array due to the large wavelength of human speech. Thus, commercial smart speakers have ambiguities when interacting with inaudible sounds. (b) If we adopt a dense microphone array according to the short wavelength of inaudible sounds, the performance of speech interfaces will be significantly affected. (c) MetaAng combines a passive shell with a sparse microphone array to enable spatial perception for inaudible sounds without affecting the performance of speech interfaces.

fine-grained AoAs of inaudible sounds? In other words, the system should also be able to (i) work properly with multiple sources and (ii) achieve high accuracy and angular resolution.

This paper introduces MetaAng, a system designed to augment microphone arrays by enabling wideband spatial perception capabilities across both speech signals and inaudible sounds. As illustrated in Fig. 1(c), MetaAng employs a passive structure (also known as acoustic metasurface) that envelops a microphone array with sufficiently large inter-element separation for human speech. The passive structure is a 3D-printed shell with sub-wavelength internal structures for inaudible sounds to manipulate incident waves, which can, by carefully designing, encode different spatial features for different incident angles. Such distinct spatial features can be used to resolve the angular ambiguity. Meanwhile, the passive structure barely impacts the processing of human speech, owing to the significant difference in wavelength. Therefore, MetaAng solves the microphone array design dilemma caused by the wavelength discrepancy by transforming such a discrepancy into a beneficial feature via introducing a passive structure. Moreover, since the passive structure is a 3D-printed low-cost shell, MetaAng can be easily integrated into existing microphone arrays in COTS smart speakers.

Though the idea of introducing a passive structure to resolve the wavelength discrepancy is promising, it further poses three new challenges for spatial perception using inaudible sounds:

First, how to estimate AoA with the presence of the passive structure and support multiple sources? On the one hand, AoA estimation is, conventionally, achieved by leveraging the phase difference of each microphone channel, which changes accordingly with the incident angle. However, the presence of the passive structure will alter the phase difference, making the conventional algorithms fail. On the other hand, the passive structure introduces new spatial features that can be used to classify the incident angle [15, 24]. Nevertheless, the intuitive classification approach fails to support tracking multiple sources since there are a mass of combinations of possible angles. To address this challenge, we propose a compressive sensing based angle estimation model to estimate the AoA with the presence of the passive structure. Specifically, we formulate the passive structure aided microphone array as a linear model, and we can solve this ill-posed inverse problem with compressive sensing by leveraging

the sparsity of the incident angle, under the assumption that there is only a tiny portion of incident angles are occupied by sound sources.

Second, how to find the optimal design of spatial encoding for better sensing accuracy? The sensing accuracy of AoA estimation with compressive sensing is highly related to the design of a combination of the passive structure and the microphone array. There are many possible designs for the passive structure, and each will encode different spatial features with a given form factor of a microphone array, thus affecting the performance of AoA estimation. Therefore, the optimal design of the passive structure should maximize the difference of spatial encoding features at different frequencies, which can be used to improve the sensing accuracy. To achieve this, we build an optimization model to minimize the average similarity of encoding features across all possible angles and the maximum similarity of encoding features between different angles to search for the optimal design of the passive structure with the given form factor of microphone arrays.

Third, how to improve the angular resolution? The conventional compressive sensing algorithm, such as the Iterative Soft Thresholding Algorithm (ISTA) [42] can produce sharp peaks with one sound source in the angle profile (*i.e.*, the AoA estimation results), which is expected to manifest satisfactory angular resolution. However, when two sound sources are present and getting closer (*e.g.*, 5° apart), the two sharp peaks will be merged to one broader peak, limiting the angular resolution for multiple-source estimation. This is due to the fact that closer incidence angles will tend to produce more similar spatial features even with the optimization of the passive structure. The received signal from two closely located sound sources may be easily confused with the linear combination of many nearby angles, and finally, the solver produces a merged and broad peak. To address this challenge, we propose Unrolling Soft Thresholding Network (USTNet) that uses a neural-enhanced prior in a data-driven manner for compressive model. Compared with the hand-picked l_1-norm prior, the neural-enhanced prior can learn a sparsifying transform and impose more vital sparsity constraint, producing finer peaks even when the two sound sources are close.

Our contributions are summarized as follows:

- We design MetaAng, a system capable of high-accuracy and high-resolution AoA estimation using 3D printed passive structures combined with sparse microphone arrays. To our knowledge, this is the first system that uses passive structures to enhance microphone arrays with spatial perception for inaudible sounds, while minimally affecting the processing of voice-frequency sounds.
- We propose a neural-enhanced compressive angle estimation algorithm to estimate the AoA leveraging the frequency diversity of the passive structure, which can support the concurrent and high-resolution angle estimation for multiple sources.
- We develop an optimization model aimed at identifying the optimal configuration of the passive structure. This model enables us to tailor unique encoding features for each incident angle, thereby enhancing the accuracy of angle estimation.
- We implement MetaAng on a COTS Bela platform and evaluate its performance in a variety of scenarios. Specifically, for inaudible sounds, the frequency diversity introduced by AMS enables a mean angle estimation error of 1.91 degrees even with only two microphones, and an error of 4.85 degrees when estimating 5 incident signals simultaneously. Compared with the baseline super-resolution angle estimation algorithm, the system can reduce the estimation error from 53.9% to 64.5%. Moreover, the system can maintain stable angle estimation performance for microphone arrays of different geometric types. Finally, we also evaluate the performance of our system in localization and tracking.

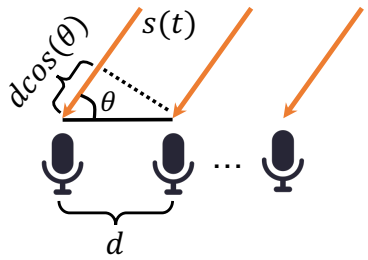
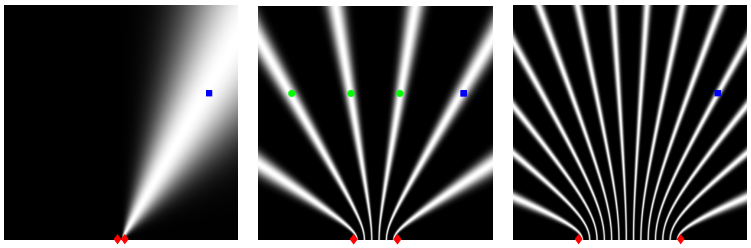


Fig. 2. Uniform linear array model and angle notations.



(a) $\lambda/2$ separation (b) 3λ separation (c) 7λ separation

Fig. 3. Increasing inter-element separation will cause ambiguous angles of arrival. As the separation of two microphones (red diamond marker) increases, the beams become finer, but causing ambiguity (some examples are marked by green dots) in localizing the true sound source (blue square marker).

2 PRELIMINARY STUDY

This section first explains the fundamentals of AoA estimation and the reasons behind ambiguous angles arising from sparse microphone arrays. Next, we explore the concept of acoustic metasurface and its impact on manipulating sound waves.

2.1 Basics of Angle Estimation

Angle Estimation Model. The estimation of AoA is typically achieved through a microphone array. Consider there is a tone wave x with wavelength λ arrives at the array at angle θ from the far field (*i.e.*, the sound wave arrives at each microphone in parallel), as shown in Fig. 2. Due to the difference in traveling distance at adjacent two microphones, the received signal has a phase difference $2\pi d\cos(\theta)/\lambda$, where d is the separation of two microphones. Assume there are N microphones and the received signal $y(\theta)$ at microphones can be modeled as follows:

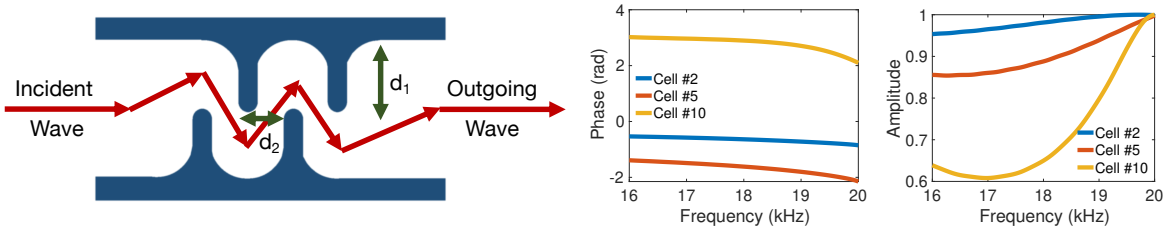
$$y(\theta) = S(\theta)x = [1, e^{-j2\pi d\cos(\theta)/\lambda}, e^{-j2\pi 2d\cos(\theta)/\lambda}, \ldots, e^{-j2\pi(N-1)d\cos(\theta)/\lambda}]^T x \tag{1}$$

where $S(\theta)$ is called the steering vector, which is a function of the angle of arrive θ and describes the phase shift of the signal x at microphones due to the spatial separation d.

The attenuation of power of received signals at each microphone can be ignored under the far-field assumption, thus phase shift plays the dominant role in the conventional angle estimation algorithms. Typically, we can search all possible angles and compensate for the corresponding phase shift, then find the angle that maximizes the power of the superposition of phase-aligned signals at each microphone. Specifically, we have the following model:

$$\max_{\hat{\theta}} S(\hat{\theta})^T x(\theta) \tag{2}$$

The peak value can be achieved at many angles if $2\pi d\cos(\theta)/\lambda = \phi_0 + 2k\pi$, where ϕ_0 can be any number within $(-\pi, \pi)$ and k can be any integer. To get the unambiguous solution, we have to ensure that the phase shift between adjacent microphones does not exceed π, *i.e.*, $|2\pi d\cos(\theta)/\lambda| \leq \pi$, such that $k = 0$ and yields a unique solution. To achieve this, we have to make the separation smaller than half the wavelength of the signal, *i.e.*, $d \leq \lambda/2$. However, for a larger separation, there will be multiple choices of k to achieve the peak value such that causes the ambiguous angles.



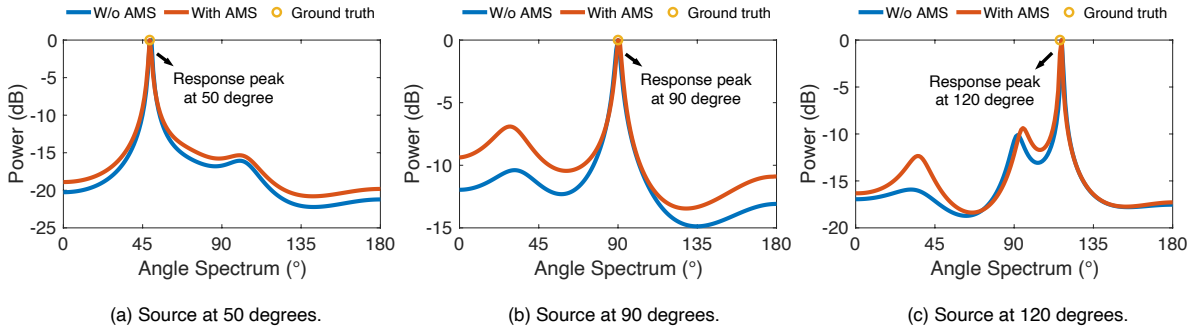
(a) Structure of AMS cell. (b) Amplitude response. (c) Phase response.

Fig. 4. The structure and response of AMS cell: (a) represents the structure of the AMS cell, while (b) and (c) are its corresponding amplitude and phase response, respectively.

Impact of Sparse Microphone Arrays. Signals from the ambiguous angles, *e.g.*, θ_1, θ_2, and θ_3, will yield the identical received signals at the microphone array (*i.e.*, the same phase shift corresponding to the reference mic), such that we have $x(\theta_1) = x(\theta_2) = x(\theta_3)$. Many popular algorithms are proposed in the literature for AoA estimation, such as MVDR [53] and MUSIC [20]. However, none of these approaches apply to distinguish identical signals from ambiguous angles. Fig. 3 illustrates this through some examples of possible angle estimations (*i.e.*, the different beams) with varying inter-element separations using a two-microphone setup, where the frequency of inaudible sound is set to be $20kHz$. In Fig. 3(a), when the separation equals half wavelength, *i.e.*, $\lambda/2$ (specifically $0.85cm$ for $20kHz$), a single beam is observed, enabling unambiguous AoA estimation, as indicated by the blue mark. However, increasing the separation to 3λ (*i.e.*, $5.15cm$) results in the emergence of 6 ambiguous beams, complicating the determination of the true AoA. This complexity escalates to 14 beams at a separation of 7λ (approximately $12cm$). The ambiguous angles pose significant challenges for finding the true AoAs of the sound sources.

2.2 Passive Acoustic Metasurface

Properties of Passive Acoustic Metasurface (AMS). Our passive structure is a 3D-printed planar acoustic metasurface that can manipulate the phase of acoustic waves. AMS comprises many unit cells, comprising a coiled-up internal structure to increase the path length of acoustic waves. Fig. 4(a) illustrates examples of unit structures [71] used in our system. Assume the incident wave passes through a unit cell from the left side, then the coiled-up path of the wave prolongs its traveling time through the cell, thereby controlling the phase delay and amplitude variations of the outgoing wave. The phase delay is determined by two essential parameters, d_1 and d_2, which should be tuned accordingly to the wavelength of the operating frequency. For any specific frequency (*e.g.*, $20kHz$ in our case), we can scale the size of the unit cell and vary the combination of these two parameters to design multiple different unit cells to manipulate the incident acoustic wave with complete phase control in the range from 0 to 2π. One practical approach to determine the values of $d1$ and $d2$ is using simulators like COMSOL [3], a widely recognized finite element-based multiphysics simulator. By precisely adjusting $d1$ and $d2$, we have identified 16 distinct cell types, each characterized by phase shifts ranging from 0 to 2π. Specifically, in our setup, the ranges of d_1 and d_2 are $0{\sim}6.4mm$ and $0{\sim}3.1mm$ for phase control, respectively. For each cell unit, our goal is to select a configuration that offers a phase shift closest to the desired offset. Using COMSOL, we have simulated each cell type's amplitude and phase response across a specific bandwidth, which assists in crafting unique response patterns for different incidence angles. Fig. 4(b) and Fig. 4(c) highlight the response of two cell types within the $16 - 20kHz$ frequency band, laying the groundwork for the AMS design approach.



(a) Source at 50 degrees. (b) Source at 90 degrees. (c) Source at 120 degrees.

Fig. 5. The impact of AMS on the AoA estimation of low-frequency sound, using a uniform linear array (internal unit spacing 4.25cm) when equipped without/with AMS. The results show that the existence of AMS has a negligible impact on the angle estimation of low-frequency sound.

Applications of Passive AMS. For AMS, each cell can act as a 'mini-antenna' to control the phase change at a specific unit cell, and the combination of different unit cells of AMS can beamform acoustic waves like a phased array. However, the significant difference between the passive AMS and an active phased array is that the function of AMS is fixed and cannot be changed after fabrication, while an active phased array can be altered flexibly. Though the passive AMS is not adjustable, it is still beneficial in many applications by carefully configuring its units. For example, we can focus the energy of acoustic waves from an omnidirectional speaker to a specific direction by placing a low-cost AMS in front of the speaker [35]. It is also possible to steer the direction of the focused beam by combining a passive AMS with a small phased array to equivalent the performance of a large phased array [71]. These applications are implemented at the speaker side to align the phase difference at each unit of the AMS and finally perform beamforming to increase signal power in a specific direction. In this paper, our basic idea is to adopt a passive AMS at the microphone side to bring more distinct spatial features.

Impact on Low-Frequency Sound. Since the unit cell design depends on the operating frequency (*i.e.*, for the inaudible frequency $20kHz$), it raises a question: Does AMS affect the array's spatial perception ability for low-frequency sounds? To explore this, we use a random AMS (the configuration of unit cells are randomly picked from the 16 designs) and a linear 4-microphone array with $4.25cm$ separation (corresponding to the half wavelength of $4kHz$ sound) and conduct the impact of AMS on the low-frequency sound AoA estimation. Specifically, we use a smartphone to loop-play pre-recorded audio from a participant as the sound source. Then, we collect signals at 50, 90, and 120 degrees, both with and without AMS, and apply the MVDR algorithm to estimate the incident angles. The angle estimation results, as shown in Fig. 5, reveal that the angle spectrum with AMS almost overlaps with that without AMS, demonstrating the minimal impact of AMS on low-frequency sounds. More detailed evaluations are presented in Sec. 5.1.4. We observe that the existence of AMS has a negligible impact on the angle estimation of low-frequency sound. The reason behind this is that the wavelength of human speech significantly exceeds the internal path length of the unit cell, making the phase delay introduced by AMS negligible. For inaudible sound, the wavelength is comparable to the internal path length of the unit cell, such that the phase delay introduced by AMS is significant and can be used to encode spatial features for inaudible sound. Next, we require leveraging AMS to encode the inaudible sound for each incident angle as its unique feature, and design algorithms to support accuracy and high-resolution AoA estimation, even with a large separation of microphone arrays.

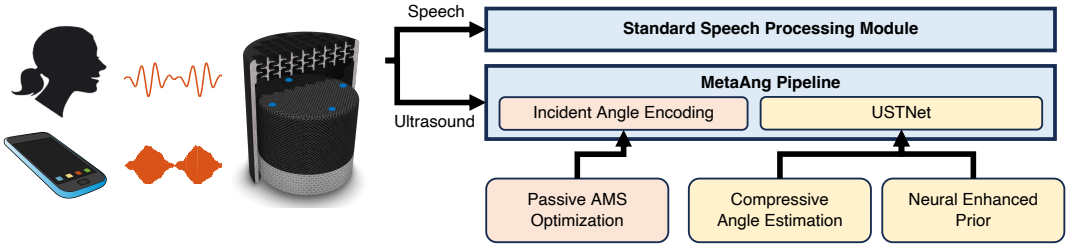


Fig. 6. Design overview of MetaAng.

3 METAANG DESIGN

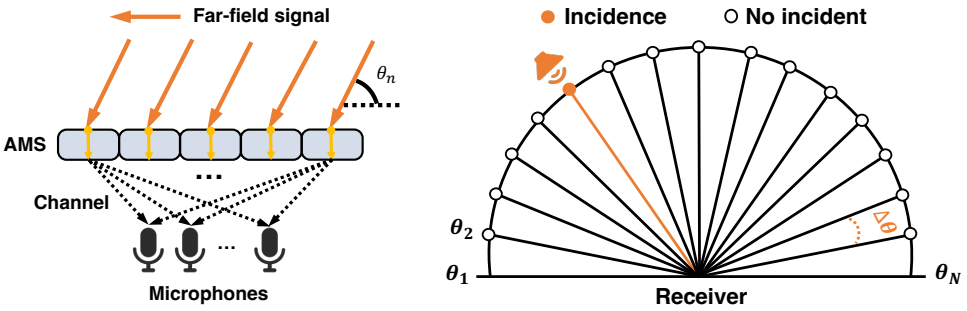
3.1 Design Overview

The primary objective of MetaAng is to address the challenges associated with AoA estimation for inaudible sounds, particularly when using a sparse microphone array such as a COTS smart speaker. The key innovation lies in harnessing the spatial encoding capability of a passive shell, also known as an acoustic metasurface, to mitigate ambiguities in the AoA estimation of inaudible sounds without compromising the spatial perception performance for low-frequency speech. The design overview of MetaAng is illustrated in Fig. 6. The smart speaker simply invokes the standard speech processing module (*e.g.*, speech direction estimation) for speech signals, because the passive shell minimally affects low-frequency speech. In contrast, for inaudible sounds, MetaAng incorporates two critical modules to achieve unambiguous AoA estimation, even with a significantly larger inter-element separation. These modules involve (i) encoding incident angle features using a passive shell and (ii) estimating AoAs with an Unrolling Soft Thresholding Network (USTNet). Specifically, MetaAng implements three key design elements to optimize performance. Firstly, we optimize the passive shell's configuration by considering the impact of frequency diversities, ensuring the collection of the richest spatial features (Sec. 3.3). Secondly, we introduce a compressive sensing model capable of estimating AoAs in the presence of the passive structure, making it possible to support concurrent estimation of multiple sources (Sec. 3.2). Thirdly, we enhance angular resolution by incorporating a neural-enhanced prior in the unrolling neural networks (Sec. 3.4). The last two design elements are seamlessly integrated into an explainable USTNet to infer AoAs from the spatially encoded features. This comprehensive approach ensures optimal performance in resolving AoA ambiguities while maintaining high-quality spatial processing capabilities of microphone arrays for low-frequency speech.

3.2 Angle Estimation Model with Acoustic Metasurface

In order to understand how to deduce the arrival angle using an AMS, we construct a basic angle response model. As shown in Fig. 7, we consider a receiver in free space, which consists of an AMS and M microphones with a known geometric configuration. We define the distribution of the incident waves' angles as $\theta = [\theta_1, \theta_2, \cdots, \theta_N]^T \in \mathbb{R}^{N\times 1}$. Each element θ_i within this space represents a potential arrival signal from the $i-th$ direction. The spacing between adjacent elements $\Delta\theta$ serves as the receiver's minimum angular resolution. Therefore, to estimate the arrival angle, it is essential to determine a vector, called direction profile $x = [x_1, x_2, \cdots, x_N]^T \in \mathbb{R}^{N\times 1}$, which represents the intensity of acoustic waves from all possible directions. If there is a signal from a particular direction, it will manifest as a corresponding peak in the direction profile.

For simplicity, we assume the presence of only one source in direction θ_n with normalized amplitude (the subsequent derivation is also applicable to multiple sources). The source under consideration emits at a frequency f (*e.g.*, tone signal[63]). Consequently, the value of the n^{th} element x_n in the direction profile should be a peak



(a) Angle response model at the receiver.

(b) Division of angle grid.

Fig. 7. Theoretical model of using AMS for angle estimation.

exceeding a predefined threshold, while the other elements are close to zero. Considering it as a far-field source model, the plane wave arriving at the front of the metasurface can then be modeled by

$$y_{front} = Sx \tag{3}$$

where $S = [s(\theta_1)\sin(\theta_1), s(\theta_2)\sin(\theta_2), \cdots, s(\theta_N)\sin(\theta_N)] \in \mathbb{C}^{P\times N}$, P is the number of the AMS's cells, and $\sin(\theta_n)$ represents the decomposed incident energy of acoustic waves impinging to the AMS. $s(\theta_n) = [1, e^{j2\pi\frac{d_m\cos\theta_n}{\lambda}}, \cdots, e^{j2\pi\frac{(P-1)d_m\cos\theta_n}{\lambda}}]^T$ is the steering vector of direction θ_n, where d_m is the distance between cells. After passing through the AMS, the sound field at the back of the metasurface can be denoted by

$$y_{back} = GSx \tag{4}$$

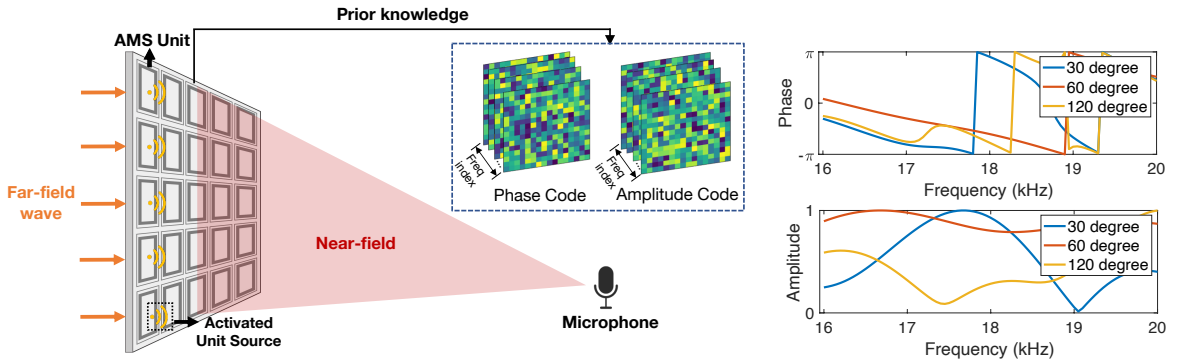
where $G \in \mathbb{C}^{P\times P}$ is the transfer function of AMS. Each element in the diagonal of G depicts the manipulation of sound in each cell, while all the non-diagonal elements are set to be zero. The diagonal structure of G is due to the fact that each cell only controls the sound waves passing through its chamber and there are no interactions across different cells. Finally, let $H \in \mathbb{C}^{M\times P}$ denotes the acoustic channel between the AMS and the microphone array, where M is the number of microphones. $H_{ij} = a(d_{ij})e^{-j2\pi d_{ij}/\lambda}$ defines the channel from the i^{th} cell to the j^{th} microphone. d_{ij} is the distance between these two points and $a(d_{ij})$ is the attenuation of signal after passing through this channel. Then the overall response of the microphone array can be modeled as follows:

$$y = HGSx + \epsilon = Ax + \epsilon \tag{5}$$

where ϵ is Gaussian white noise. $A = HGS \in \mathbb{C}^{M\times N}$ is called the measurement matrix and $M \ll N$ due to the limited number of microphones. The property of measurement matrix A is mainly determined by the design of metasurface G since the channel H and steering matrix S can not be tuned. Therefore, the received signals are determined by the AMS and the angles of arrival (*i.e.*, the direction profile x).

Our Goal. Once the configuration of the AMS is determined (*i.e.*, G is known), the measurement matrix A can serve as sampling prior knowledge for inferring the incident angle, since H and S can be derived from the relative locations of microphones and the AMS's cells. Our objective is to utilize the received signal y and the predefined measurement matrix A to determine the incident angle from the estimated *profile*, denoted as $\hat{x}$.

To solve such a linear inverse problem, a rudimentary response to this issue could be the direct implementation of the Least Squares (LS) method, formulated as $\hat{x} = (A^TA)^{-1}A^Ty$. However, the matrix A^TA typically lacks invertibility due to the inability of M microphones to gather sufficient constraints for determining N unknowns,



(a) AMS encoding at different frequencies.

(b) Receiver response at various incident angles.

Fig. 8. Incident angle encoding: (a) concept of incident angle encoding using AMS, while (b) is the amplitude response and phase response on the receiver.

thereby rendering the inverse problem ill-posed. Alternatives, such as the pseudo-inverse method, which entails augmenting with a minor full-rank matrix to achieve invertibility, might offer a feasible solution. Nevertheless, such methods tend to be vulnerable to noise and often struggle to yield accurate outcomes. From a general perspective, the estimation of the incident angle can be formulated as the following optimization problem:

$$\hat{x} = \arg\min_{x} \|Ax - y\|_2^2 \tag{6}$$

In order to achieve accurate angle estimation, we acquire additional information from two aspects to solve the underdetermined optimization problem in Eq. 6. Firstly, we utilize the diversity of frequency responses and configurations of AMS to better encode the incident angles, thereby accumulating more constraints for Eq. 6 (Sec. 3.3). Secondly, we explore the prior knowledge of the signal to further reduce the complexity of solving the optimization problem, such as utilizing the signal's sparse characteristics in the spatial domain (Sec. 3.4).

3.3 Efficient Spatial Encoding

The angle estimation accuracy of MetaAng is determined by the uniqueness of the incident angle features, where the efficiency of spatial encoding can be improved by (i) utilizing the frequency diversity and (ii) optimizing the configuration of AMS.

3.3.1 Utilizing Frequency Diversity. To enhance the effectiveness of the constraints provided by the measurements, we seek to utilize the frequency diversity of the AMS to increase the number of sampling points for each incident angle. For incident angle encoding, we arrange the cells linearly to form a 1D AMS, or in a rectangular grid for a 2D AMS configuration. Such an arrangement enables the modulation of incoming signals from different directions by introducing specific phase shifts and amplitude changes in each micro-area. In our approach, we prioritize the use of a regular 2D AMS pattern. Such consideration introduces two primary advantages. Firstly, the 2D AMS's ample surface area allows for effective interaction with the incident signals, which is crucial for modifying the signals received by the microphone. For instance, our signal propagation model simplifies by considering the signals initially interacting with the AMS and then propagating to the microphone, while marginalizing the negligible diffraction effects at the AMS's periphery. Secondly, regarding the far-field model, although the steering vectors are consistent across each row of the 2D AMS at a given angle, enhancing the number of rows can improve the encoding ability of AMS in the spatial domain. For the microphones, each unit

of the AMS effectively acts as an independent near-field source, and the signal received by the microphone is a superposition of signals from all units of the AMS. Thus, under the excitation of the incident signal, the 2D AMS structure allows us to encode the amplitude and phase of the signal over a larger spatial range, even though the steering vectors of each row are similar.

As shown in Fig. 8(a), the AMS can encode the amplitude and phase of incident signals at different frequencies, creating unique frequency response characteristics as identifiers for each incident angle, also known as sampling prior knowledge. The role of the AMS is to disrupt the uniformity of phase and amplitude in the signal propagation process, thereby facilitating effective encoding. Without the AMS, the signals on the AMS plane would maintain the same (or similar) amplitude and exhibit coherent phase variations upon reaching the AMS, making encoding ineffective, even with the use of multiple frequencies. Thus, the diversity in the AMS's frequency response is crucial for generating distinct frequency responses for different incident angles. Fig. 8(b) illustrates the amplitude-frequency and phase-frequency responses for various incident angles, highlighting the noticeable differences across these angles. Through this approach, we can extend the measurement matrix A into the frequency domain to $A(G)$, adding more constraints, where f signifies the encoded frequency index. In our work, we select the frequency range of $16 - 20kHz$ since acoustic signals in this band are generally inaudible to most people, and this range is where our designed AMS exhibits the most ideal response.

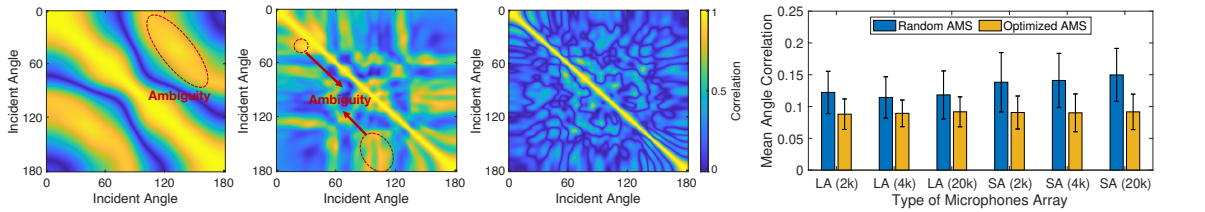
3.3.2 Configuration Optimization for Acoustic Metasurface. As previously mentioned, MetaAng has effectively harnessed the capabilities of the AMS to increase the frequency diversity of incident angle features. The similarity distribution map between various angle features, as depicted in Fig. 9(a) and Fig. 9(b), clearly demonstrates a substantial reduction in the degree of similarity among incident angle features configured with AMS compared to those without AMS. However, upon closer examination, we can intuitively identify that several regions still exhibit a relatively high level of feature similarity. Furthermore, the presence of broader diagonal areas on the map indicates the persistent similarity in features between neighboring angle features, which remains the principal limiting factor affecting estimation accuracy.

One straightforward approach to address this limitation is to consider the use of additional microphones or larger AMS setups, thereby enhancing our control over the signals. Nevertheless, the former option incurs additional costs and may deviate from our intended objectives, while the latter, although potentially effective, raises a significant question: "Have we fully exploited the capabilities of the AMS configuration currently in use?" Up to this point, our approach has primarily focused on employing randomly generated AMS, which may not be suitable for arrays with different geometric types shown in Fig. 9(d). To unlock the fully potential of AMS for incident angle encoding and maintain stable performance across different geometric types, we develop an optimization algorithm that aims to maximize the encoding capabilities of the current AMS.

We represent the cosine similarity between the i-th and k-th incident angle features (i.e., columns of $A(G)$) using $G(i,k)$. To comprehensively enhance the distinctiveness among angles while aiming to reduce overall correlation, the optimization problem can be formulated as $\min_{i \neq k} \sum G_{i,k}$. Furthermore, we introduce an additional loss function G_{max} into the optimization process to enhance the minimum differentiation between angles. This auxiliary loss function serves the dual purpose of constraining the optimization of the AMS configuration and ensures a baseline performance in angle estimation. Our ultimate optimization problem can thus be defined as:

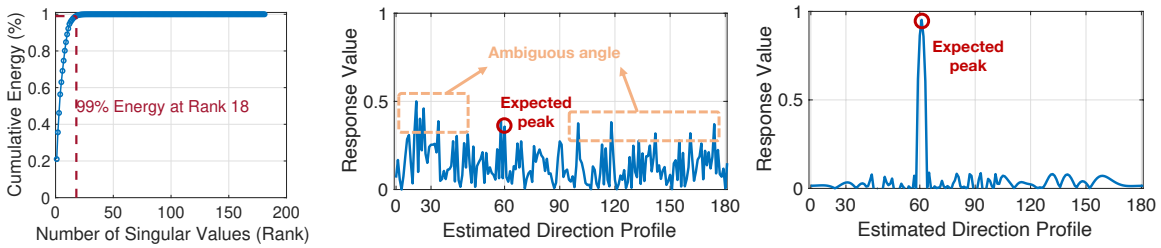
$$\min_{i \neq k} \sum G_{i,k} + \beta G_{max} \tag{7}$$

where β is the loss weights. The optimization of Equation 7 helps reduce the similarity of features between various incident angles, so that the algorithm can better identify each angle. The Fig. 9(c) displays the similarity distribution map of angle features after AMS optimization. It demonstrates the effectiveness of our optimization algorithm in significantly enhancing the distinctiveness of features between angles, thus unlocking the fully



(a) W/o AMS. (b) With random AMS. (c) With optimized AMS. (d) Comparision of angle correlation optimization across arrays with different geometric types.

Fig. 9. Configuration optimization: (a), (b) and (c) are the angle correlation map for w/o AMS, with random AMS and with optimized AMS, respectively. (d) is the comparison of angle correlation optimization across different geometric arrays. LA and SA represent linear arrays and square arrays, respectively. The inter-element spacing for each array is configured for three typical frequency bands (2k, 4k, 20k).



(a) Effective rank of measurement matrix. (b) Angle estimation using LS method. (c) Angle estimation using ISTA.

Fig. 10. Angle estimation using AMS-enhanced measurement matrix: (a) is the effective rank of measurement matrix, while (b) and (c) are the estimated direction profile using LS method and ISTA, respectively.

potential of AMS for incident angle encoding. Furthermore, as shown in Fig. 9(d), we can observe that while the encoding performance of AMS may vary with changes in array types, the optimized AMS significantly reduces angle correlation and maintains stability.

3.4 Neural-Enhanced Angle Estimation

In this section, we discuss how to solve the underdetermined optimization problem in Eq. 6 and achieve finer angular resolution.

3.4.1 Conventional Compressive Angle Estimation. The enhanced measurement matrix $A(G)$ using AMS encoding possesses sufficient constraints (being full rank) to solve Eq. 6, thereby facilitating angle estimation. Generally, the Least Squares method (LS) is a simple yet effective solver for such a problem. However, in practical angle estimation scenarios, the LS method encounters difficulties. As illustrated in Fig. 10(b), for a single incident signal at 60 degrees, LS fails to yield an accurate direction profile. The expected angular peak is almost indistinguishable, being obscured by multiple ambiguous peak values. This is frustrating as the current full-rank measurement matrix still fails to be effective straightforwardly. To explore further, we examine the effective rank of the measurement matrix, defined as the number of singular values occupying 99% of the energy. Fig. 10(a) shows that

despite A being full rank, its effective rank remains low, with a ratio of less than 10%, known as ill-conditioned. This analysis reveals the vulnerability of using the LS method and the measurement matrix $A(G(f))$ to solve for the direction profile x. Specifically, the low effective rank indicates that the distinctiveness of angular features is insufficient to offset the disruption caused by noise, leading to significant angular ambiguities when estimating angles with noisy received signals (e.g., device noise or background noise). This limitation motivates us to consider designing a more efficient algorithm for stable angle estimation.

Using Sparsity Prior. As mentioned in Sec. 3.2, we partition the available space into N angular regions. In an ideal scenario, we can reasonably assume that the incident sources are sparsely distributed in space. Let K represent the number of sources, where $K << N$. To mitigate the computational complexity of angle estimation, we can incorporate this sparsity assumption into our solution algorithm by introducing an L_1 regularization term, denoted as $\|x\|_1$. By promoting sparsity through L_1 regularization, our goal is to identify the few significant components in the signal that correspond to the actual incident angles, effectively filtering out noise and irrelevant data. Subsequently, we can estimate the angles by solving the following problem:

$$\hat{x} = \arg\min_{x} \|A(G)x - y\|_2^2 + \alpha\|x\|_1 \tag{8}$$

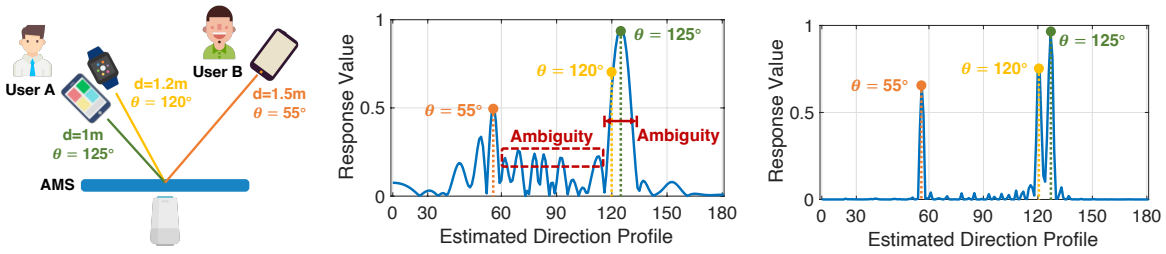
where $\|A(G)x - y\|_2^2$ computes the error of the reconstructed direction profile and the measurement, controlling the fidelity of $\hat{x}$. $\|x\|_1 = \sum_i |x_i|$ is the L_1 norm regularization term, under which we expect the object has a large number of coefficients that are near zero. Therefore, the L_1-norm imposes sparse constraint to the solutions and meanwhile, it also ensures the convexity of our optimization model simultaneously. Other regularization terms with L_p-norm ($p < 1$) may lead to better sparse constraints L_1-norm, but they are not convex [40]. As a result, L_1-norm is one of the most commonly used regularization terms in the various applications of compressive sensing [57, 61, 62]. α is a hand-picked hyper-parameter which controls the importance of the fidelity term and the sparsity term.

Basic ISTA Solver. The Iterative Soft Thresholding Algorithm (ISTA) is a widely employed optimization technique for solving the optimization problem from Eq. 8. Its effectiveness arises from its ability to promote sparsity within a signal while maintaining a relatively straightforward iterative structure. At its core, ISTA seeks to find a sparse representation of a signal by solving an optimization problem that includes an L_1 regularization term. This regularization term encourages most of the coefficients in the representation to become zero, effectively identifying the essential components of the signal while suppressing noise and irrelevant data. ISTA operates through a series of iterations where, in each step, it performs a soft thresholding operation on the coefficients. This operation shrinks the coefficients toward zero, leading to sparser representations and, consequently, improved recovery of the underlying signal. Specifically, ISTA solves the compressive angle estimation problem by iterating between the following update steps:

$$\begin{aligned} z^k &= \hat{x}^{k-1} - \rho A(G)^T (A(G)\hat{x}^{k-1} - y) \\ x^k &= Soft(z^k, \frac{\alpha}{2}) \end{aligned} \tag{9}$$

where k denotes the iteration index and ρ is the step size. z^k conducts the gradient descent, while $Soft(\cdot)$ is the soft-thresholding function with the definition as follows:

$$Soft(z, \lambda) = \begin{cases} z - \lambda & if \quad z > \lambda \\ 0 & if \quad -\lambda \le z \le \lambda \\ z + \lambda & if \quad z < -\lambda \end{cases} \tag{10}$$



(a) Multi-source setup. (b) Multi-source angle estimation using ISTA solver. (c) Multi-source angle estimation using USTNet.

Fig. 11. Angle estimation for multiple devices : (a) is the multi-device setup, while (b) and (c) are the estimated direction profile using basic ISTA and USTNet, respectively.

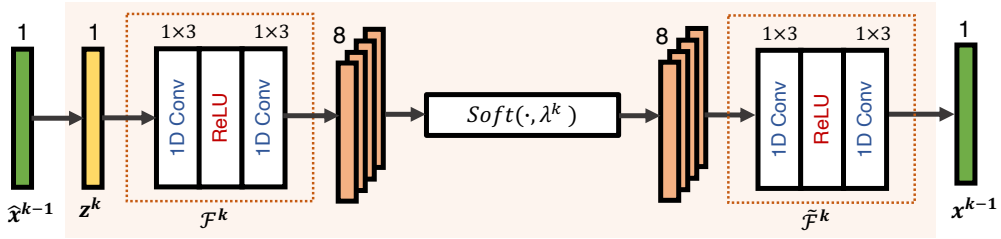
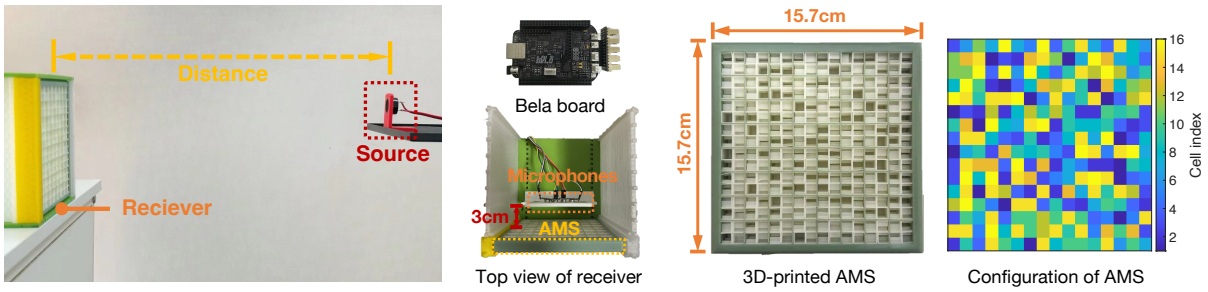


Fig. 12. The detailed structure of the kth unrolling layer of USTNet.

where λ represents the threshold. By using the ISTA solver, we can accurately estimate the precise incident angle. As shown in Fig. 10(c), a pronounced narrow peak appear at the expected angles, demonstrating the effectiveness of the sparsity prior the in single-source angle estimation.

3.4.2 Improve Angular Resolution with Neural Prior Enhancement. Our approach relies on the inherent frequency responses at different incident angles, which serve as critical angle features. The linear superposition property of MetaAng allows for its seamless extension to angle estimation scenarios involving multiple devices. Consider the scenario illustrated in Fig. 11(a), where User A possesses two devices requiring precise localization, including smartphones and smartwatches, while User B is similarly in need of tracking his smartphone. However, empirical testing results have not aligned with our initial expectations, as depicted in Fig. 11(b). This observation implies the existence of virtual responses exceeding the actual number of devices(*i.e.*, ambiguities). This phenomenon can be attributed to limitations in angle resolution, as some angles exhibit similar characteristics. The primary challenge in multi-device angle estimation here is the naive sparsity prior is insufficient to obtain precise solutions when dealing with multiple sources. This highlights the need for more sophisticated approaches in handling the complexity of estimating angles from multiple devices simultaneously.

To enhance MetaAng performance in the context of multiple devices and address the angle resolution challenge, we propose a neural-enhanced prior approach that learns a data-driven transformation function, ensuring sparsity in a more general manner. Therefore, we can achieve high performance angle estimation for multi-sourses and substantial improvements in the system's angle resolution, as visually represented in Fig. 11(c).



(a) Testbed. (b) Receiver. (c) Optimized AMS.

Fig. 13. Experimental setup of MetaAng.

To be specific, we propose an Unrolling Soft Thresholding Network (USTNet) inspired by [64] that leverage neural-enhanced prior to hold a general sparsity assumption. Specifically, we incorporate a learnable module into the optimization problem for compressive angle estimation. Consequently, we need to solve the following problem:

$$\hat{x} = \arg\min_{x} \|A(G)x - y\|_2^2 + \alpha\|\mathcal{F}(x)\|_1 \tag{11}$$

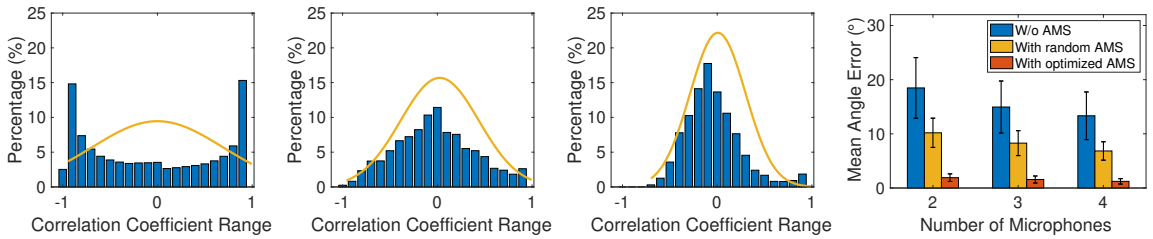
where $\mathcal{F}(\cdot)$ represents a learnable transformation function, we have incorporated the addition of two 1D convolution layers and a ReLU layer to enhance the sparsity of non-linear transformations. In order to further boost the algorithm's inference capabilities, we propose the expansion of the iterative algorithm, treating each iteration as a layer composed of learnable computational modules. Consequently, the inference process for each iteration is illustrated in Fig. 12, and it can be solved in closed form, as outlined below:

$$x^k = \tilde{\mathcal{F}}_k(\mathcal{F}_k(Soft(z^k, \lambda^k))) \tag{12}$$

$\tilde{\mathcal{F}}_k(\cdot)$ is the left-inverse version of $\mathcal{F}_k(\cdot)$, sharing a similar structural configuration. The threshold λ^k and the step size ρ^k can be learned to adapt to varying requirements at different stages of the inference process. The unrolled algorithm is achieved by setting a truncation length K (or the number of stages). To summarize, for any given inference layer (or iteration), its set of learnable parameters is denoted as $\Theta^k = \{\lambda^k, \rho^k, \mathcal{F}_k(\cdot), \tilde{\mathcal{F}}_k(\cdot)\}$, which can be acquired from real-world scenarios to ensure a stable neural-enhanced sparsity prior. Our final angle esimation algorithm consists of K concatenated unrolling layers.

4 IMPLEMENTATION

As depicted in Fig. 13(a), we develop a testbed using the commercial Bela platform [6]. Our system comprises a microphone array, an 3D-printed AMS, sound sources, and a Bela development board as the controller, as shown in Fig. 13(b). The geometric configuration of the microphone array is restructured through a 3D-printed panel. Unless otherwise stated, we maintain a square array of four microphones with a spacing of 4.25 cm optimized for 4kHz. We strategically place the AMS 3 cm in front of the microphones, aligning its center with the microphone array's center to ensure most signals pass through the AMS before reaching the microphones. The AMS fabricated following the optimized guidelines from Sec. 3.3.2, is shown in Fig. 13(c). We optimize AMS using the Adam optimizer in PyTorch, which does not require any dataset for this optimization step and can achieve quick convergence (less than 1 minute). Our AMS design is a 16×16 grid, consisting of 256 units, with each unit having 16 selectable states, as detailed in [71]. The AMS is manufactured using a commercial 3D printer. The printing cost is about 5 dollars, and the cost will be significantly reduced with mass production.



(a) W/o AMS. (b) With random AMS. (c) With optimized AMS. (d) Angle estimation.

Fig. 14. Overall performance of incident angle encoding. (a)-(c) is the correlation distribution of measurement matrix collected by testbed, and (d) is the angle estimation across various number of microphones using one source.

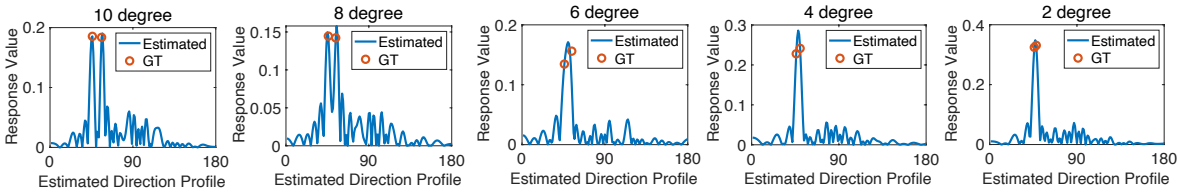
For algorithm configuration, we set a sound source to emit chirp signals ranging from 16 to 20 kHz at 1 degree intervals within a 0 to 180 degree range relative to the AMS. The period of the signal is 100ms, with a single chirp being transmitted for 40ms. Once the receiver receives the signal, it first passes through a bandpass filter to capture signals within the frequency band of interest, and then performs an FFT on the chirp within one period to obtain the spectral information of the signal. We collect these signals five times at varying distances, from 50 to 200 cm in 50 cm intervals. Then, we evaluate the performance of USTNet by employing five-fold cross-validation on the five sets of collected data. Moreover, the collected data at different angles and distances are linearly superimposed to simulate multi-source incidence scenarios, forming our training dataset. Specifically, we consider scenarios where the number of sources ranges from 1 to 5, ensuring that the amount of data for each scenario is consistent, meaning that our final dataset consists of $4 \times 181 \times 5 \times 5 = 18100$ entries. We collect data an additional time at a distance of 100cm to calibrate the measurement matrix A, which contains a total of 181 columns, each representing the spectral features of its corresponding angle. This step is necessary because the manufacturing errors of the AMS may lead to significant discrepancies between simulations and actual results. We set the number of stages in the USTNet to 10 and train our algorithm using the Adam optimizer on a server equipped with an NVIDIA GTX 3090 GPU. The total optimization time takes 1.3 hours. We test the inference latency of USTNet on a MacBook Pro laptop with Intel Core i5 and recorded it as 12.3 ms. Notably, our algorithm requires training only once unless the AMS or microphone array geometry changes.

Baseline. We choose four classic angle estimation algorithm as baselines, including DAS [7], MVDR [53], ESPRIT [20], and MUSIC [39]. DAS is a basic beamforming method that aligns and sums sensor signals to enhance specific direction signals and reduce noise. MVDR adapts to minimize noise while preserving signal quality in the desired direction. ESPRIT estimates multiple signal directions using rotational invariance of signal subspaces. MUSIC, a high-resolution technique, identifies wavefront directions by analyzing signal frequency and exploiting signal and noise subspace orthogonality.

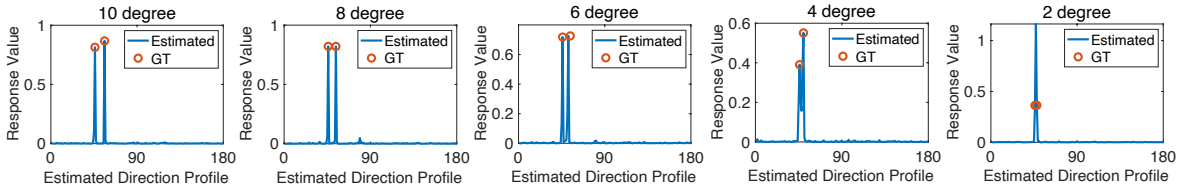
5 EVALUATION

5.1 Overall Performance

To thoroughly evaluate the performance of incident angle encoding, we explore three distinct configurations: (i) without any AMS, (ii) with a randomly generated AMS, and (iii) using our specially optimized AMS. Initially, we capture the characteristic feature of each angle at 1-degree intervals using our testbed. We then analyze the statistical distribution of the correlation between these encoded features at various angles, as illustrated in Fig. 14(a) to 14(c). The median absolute correlation of features across angles in configurations (i), (ii), and (iii) are found to be 0.71, 0.28, and 0.16, respectively. The distribution of angle feature correlation using the optimized AMS centers around zero and displays a slight standard deviation, a trait known to be advantageous



(a) Estimated directional spectrum using ISTA.



(b) Estimated directional spectrum using USTNet.

Fig. 15. AoA estimation performance under different source angle spacing between traditional ISTA and USTNet.

for solving compressive sensing problems, as mentioned in [5]. Generally, smaller average correlations indicate stronger specificity between angular features. To visually demonstrate the effectiveness of the AMS, we assess its performance in angle estimation with a single incident source. Fig. 14(d) shows that without the AMS, even after employing multiple sub-frequencies, there is a significant error in angle estimation, heavily dependent on the number of microphones. Introducing a random AMS considerably reduces this angular error by 44.8%, 44.6%, and 48.72%, respectively, when using 2, 3, and 4 microphones compared to the scenario without AMS. We observe that increasing the number of microphones enhances the effectiveness of the AMS. This improvement is attributed to the distinct roles played by the quantity of microphones and frequency encoding. Moreover, compared to configurations without an AMS, the optimized AMS notably reduces the angular error by 86.97%, 87.51%, and 89.29%, respectively. Experimental results demonstrate that employing the optimized AMS for incident angle estimation substantially decreases the dependency on the number of microphones. It maintains an average error of just 1.91 degrees, even with only two microphones.

5.1.1 Performance of Multi-Sources. Subsequently, we examine MetaAng's ability to handle multiple incident sources using a receiver with four microphones and an optimized AMS. We first test the distinguishability of two sources at different angular separations. As shown in Fig. 15, as the angle spacing decreases, the two expected angle response peaks of the traditional ISTA will stick together and produce many ambiguous responses. While using USTNet, the ambiguity problem is significantly alleviated, and high-quality angular responses are produced. The results show that our method can achieve an angular resolution of 4 degrees. Then, we verify the impact of the number of incident sources on the AoA estimation accuracy of MetaAng.We place incident sources at a fixed 100 cm distance from the receiver, increasing the number of sources from 1 to 5. The angular separation between these sources varies from 5 to 30 degrees in 5-degrees increments. As shown in Fig. 16, the mean angle estimation errors for 1 to 5 sources are 1.25, 1.52, 2.24, 3.17, and 4.85 degrees, respectively. Interestingly, despite using only four microphones, MetaAng successfully estimates up to five incident sources. Traditional angle estimation methods like MUSIC struggle with underdetermined linear problems, but our approach overcomes this by using angle encoding and sparsity knowledge. This capability allows for accurate multi-source angle estimation with limited microphones, which is ideal for smart home environments. It enhances spatial awareness in smart devices, improving their functionality in intelligent settings.

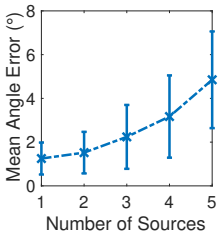


Fig. 16. Performance on varying number of incident sources.

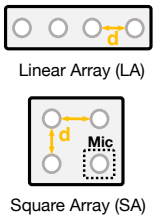


Fig. 17. Geometry types of microphone array.

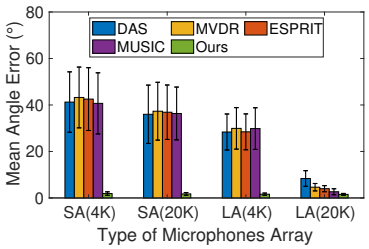


Fig. 18. Comparison with baseline angle estimation algorithms using different geometric types.

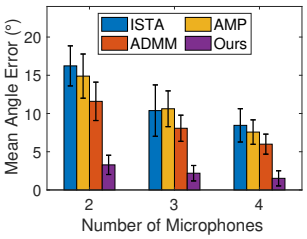
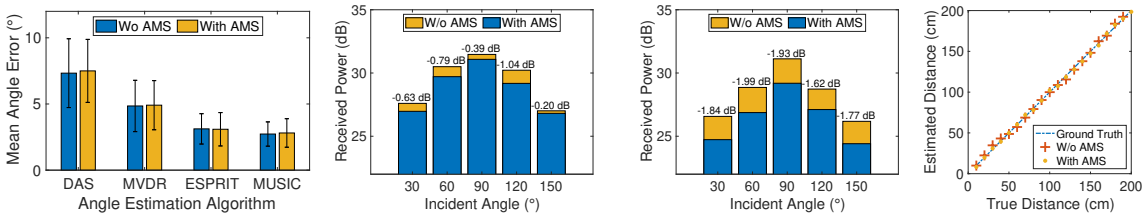


Fig. 19. Comparison with various compressive sensing algorithms using the same measurement matrix.

5.1.2 Performance on Various Array Geometry. Moreover, we demonstrate that MetaAng also reduces the constraints of array geometry and compare it with baseline angle estimation algorithms. These algorithms require a half-wavelength spaced array for accurate angle estimation, failing which they encounter destructive ambiguities. To validate this, we evaluate the angle estimation performance for two incident sources on four different geometric arrays, each comprising four microphones, as depicted in Fig. 17. For each geometric array, we recollect datasets to calibrate the measurement matrix and train USTNet. For testing, we only consider the scenario of two sources to simplify the influence of other factors. The array geometries include linear and square arrays for 4kHz and 20kHz frequency bands, respectively. As shown in the experimental results in Fig. 18, MetaAng demonstrates similarly low estimation errors across the four geometric arrays, with errors respectively at 1.89, 1.74, 1.61, and 1.53. In contrast, the performance of traditional algorithms shows sensitivity to the array geometry. With square arrays, where only two microphones are available in the azimuth direction, traditional super-resolution angle estimation algorithms struggle, showing nearly unusable performance regardless of whether they meet the half-wavelength element spacing. With linear arrays, where four microphones are available in the azimuth direction, the non-compliance with half-wavelength array geometry leads to angle estimation errors close to 30 degrees, and the performance among the algorithms does not show significant differences due to the presence of ambiguities. When the linear arrays comply with half-wavelength spacing, traditional algorithms function correctly, with angle estimation errors for the four algorithms at 8.37, 4.63, 3.99, and 2.97. Despite considerable improvements, our method still reduces the error by 53.9% compared to the best-performing MUSIC algorithm. Overall, our method is insensitive to the geometric construction of the array. A primary reason is that our AMS can be optimized according to the array's geometry, and it does not rely on strict phase delays between elements to estimate angles but extends the frequency dimension as an additional feature. This characteristic allows our system to seamlessly adapt to various commercial geometrically heterogeneous microphone arrays without affecting the microphone array's optimization for low-frequency voice applications.

5.1.3 Comparison with Various Compressive Sensing Algorithms. Our approach is compared with three classic compressive sensing algorithms: the original ISTA, Approximate Message Passing (AMP), and the Alternating Direction Method of Multipliers (ADMM), in estimating the incident angles of two sources using the same measurement matrix. As shown in Fig. 19, the results indicate improved performance across all algorithms with an increased number of microphones. Our method consistently yields the lowest error, achieving an average error reduction compared to ISTA, AMP, and ADMM by 81.2%, 79.0%, and 73.5%, respectively. This improvement is credited to additional learnable modules integrated into our solution process, which enhance priors through neural networks and utilize learned hyperparameters for superior denoising. These results verify that our design leads to more precise angle estimation using the same incident angle encoding.



(a) Low-freq angle estimation. (b) Low-freq transmission attenuation. (c) High-freq transmission attenuation. (d) Distance estimation

Fig. 20. Impact of AMS on acoustic sensing: (a) and (b) are the low-frequency sound angle estimation and transmission attenuation, respectively. (c) and (d) are the high-frequency acoustic transmission attenuation and distance estimation, respectively.

5.1.4 Effects on the Functionality of Low-Frequency Sound. A fundamental premise of our system design is not to adversely affect low-frequency usage. We assess the effect of the AMS on low-frequency applications, mainly focusing on voice incident angle estimation and AMS's influence on low-frequency energy transmission. Employing a four-microphone linear array designed for 4kHz, we estimate the angle of an incident source emitting a pre-recorded human voice. According to the results shown in Fig. 20(a), AMS has negligible impact on low-frequency voice angle estimation due to its minimal response to these frequencies. Additionally, we evaluate how AMS affects the transmission rate of low-frequency voice, playing the same voice from various directions. The energy attenuation, depicted in Fig. 20(b), shows an average of only 0.61 dB with AMS, a marginal level unlikely to affect human voice frequency applications significantly. Overall, our experiments confirm that AMS's presence has little effect on low-frequency voice, highlighting its potential for seamless integration with existing voice systems and opening up possibilities for exciting new applications.

5.1.5 Effects on the Functionality of High-Frequency Acoustic Sensing. Another aspect worthy of attention is the impact of the AMS on high-frequency sound waves. First, we examine high-frequency sound wave transmission (close to 20KHz) through AMS, as depicted in Fig. 20(c). Our AMS is designed to respond significantly to high-frequency sounds, resulting in an average energy loss of 1.83 dB. This loss could slightly reduce the effective range of acoustic perception in the air, but it is minor compared to the gains in incident angle estimation performance. Further, we explore AMS's impact on high-frequency sound wave distance estimation, a crucial perceptual metric. We measure the distance estimation of a sound source emitting chirp signals from 16-20kHz across distances of 5 to 200cm in 5cm intervals. As shown in Fig. 20(d), AMS's presence scarcely affects distance perception. This negligible impact arises because chirp signal-based distance estimation relies on the different frequencies of mid-frequency signals unaffected by AMS. Consequently, we can maintain accurate distance estimation while using AMS to improve the spatial perception of high-frequency sound waves.

5.2 Micro Benchmark

In this section, we focus on factors critical to our system's performance. This includes elements influencing the measurement matrix, such as frequency bandwidth, metasurface properties, and the distance between the metasurface and microphones. We also examine the impact of the number of layers in the unfolding algorithm stages, which is important for convergence.

5.2.1 Impact of Frequency Bandwidth. The frequency bandwidth directly impacts the number of rows in the measurement matrix. We investigate the effects of a 4kHz bandwidth, ranging from 20kHz to 16kHz. As shown in

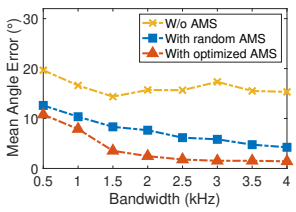


Fig. 21. Impact of frequency bandwidth.

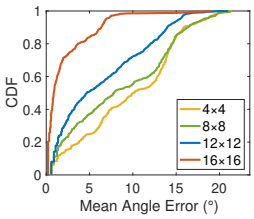


Fig. 22. Impact of AMS size.

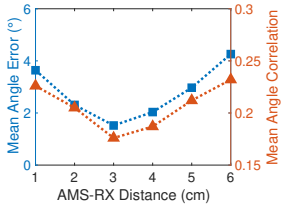


Fig. 23. Impact of AMS-Rx distance.

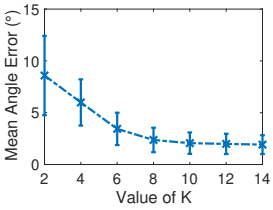


Fig. 24. Impact of layer number K.

Fig. 21, without AMS, increasing frequency bandwidth does little to improve angle estimation due to frequency coherence, which prevents them from acting as independent features. Therefore, a wider bandwidth does not significantly enhance angle-specific feature information. Conversely, using any metasurface, even a random one, increasing bandwidth improves the differentiation of incident angle features, thanks to the metasurface's inherent properties. With an optimized AMS, the expanded frequency bandwidth results in substantially improved angle estimation, which is evident in significantly lower errors than a random AMS. These findings highlight the importance of AMS, especially optimized ones, in leveraging increased frequency bandwidth to enhance angle estimation accuracy and resolution, contrasting with AMS-less systems where more bandwidth is needed to lead to better angle discrimination.

5.2.2 Impact of AMS Size. We investigate the impact of AMS size on the measurement matrix. The size of the AMS, determining its number of reconstructable units, directly influences the generation of the measurement matrix. Our exploration of angle estimation performance encompasses four different AMS sizes: 16×16, 12×12, 8×8, and 4×4. For each AMS size, we recollect data according to the method described in the implementation section to recalibrate the measurement matrix A and to train and evaluate the performance of USTNet. The results presented in Fig. 22 show that AMS size directly affects angle estimation performance. Generally, larger AMS units yield better performance. However, excessively large AMS sizes can incur higher costs and may be impractical for compact devices like smart speakers. Conversely, smaller AMS sizes have limited signal customization capabilities and pronounced diffraction effects at the edges, which can affect signal modulation before reaching the microphones. Therefore, selecting an AMS of appropriate size is crucial. It should be large enough to ensure accurate angle encoding and estimation yet compact enough for practical integration into various devices. This balance is essential for maximizing the AMS's angle estimation effectiveness while maintaining its practicality for broad application.

5.2.3 Impact of Distance between AMS and Microphone. In practical deployments, the distance between the AMS and the microphones is critical but often overlooked. The AMS should be close enough to the microphones to satisfy near-field conditions yet far enough to ensure that the microphones receive adequate energy from all cells. We optimize the AMS at distances from 1 to 6 cm and assess its angle estimation error. Results shown in Fig. 23 indicate that a 3 cm distance provides the best performance. Closer and further distances reduce AMS performance in encoding incident angles. Notably, decreasing the distance between the AMS and the receiver leads to a quicker drop in performance, especially because signals from cells at the AMS edges struggle to reach the microphones at closer distances. For effective integration of AMS in applications, maintaining an ideal distance of about 3 cm in front of the microphones is essential. This distance optimally balances angle encoding performance and effective signal capture by the microphones.

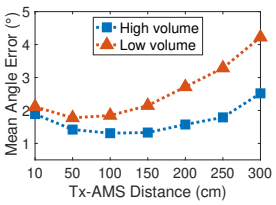


Fig. 25. Impact of varying source distance.

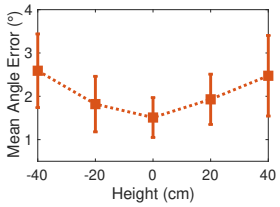


Fig. 26. Impact of varying source height.

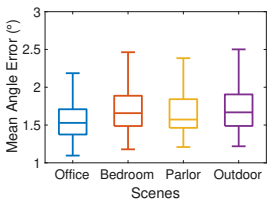


Fig. 27. Impact of environment migration.

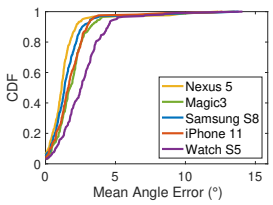


Fig. 28. Impact of varying commodity devices.

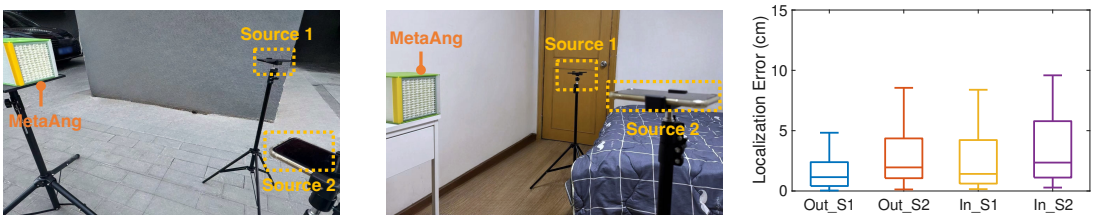
5.2.4 Impact of Number Layers for USTNet. We investigate the impact of the number of layers (iterations) in the unfolding algorithm, as depicted in Fig. 24. When the number of layers K is relatively tiny (less than 6), the algorithm remains unconverted, resulting in higher average angle estimation errors and standard deviations. Setting K to 10 typically leads the USTNet to reach convergence, and further increasing K does not significantly improve angle estimation accuracy. For balancing computational load with accuracy, we find that a setting of K to 10 is sufficient for most angle estimation tasks. This setting strikes a practical balance, ensuring effective and efficient algorithm performance without excessive computational complexity while delivering reliable angle estimation. This approach is ideal for real-world applications where accuracy and computational efficiency are crucial.

5.3 Robustness Analysis

5.3.1 Impact of Device Mobility. Device mobility is essential for practical angle estimation. We vary the distance between the source device and the AMS to test the stability of angle estimation performance at different distances. A device emitting signals at two different volumes is moved from 10 to 300 cm, with tests conducted at 10-degrees intervals within a 0 to 180-degrees range. The results, shown in Fig. 25, indicate a slight drop in angle estimation performance when the sound source is at 10 cm, likely due to far-field modeling distortion from the sound source being too close to the receiver. As the distance increases, the angle error first decreases and then rises, with the error increase at longer distances attributed to a lower signal-to-noise ratio (SNR) of the received signals, a situation exacerbated by lower volumes. However, at high volumes, our system can achieve an angle estimation error of less than 2.67 degrees even at 300 cm, aligning with the expected working distance for most active acoustic perception applications, demonstrating the system's robustness and adaptability to varying distances in practical scenarios.

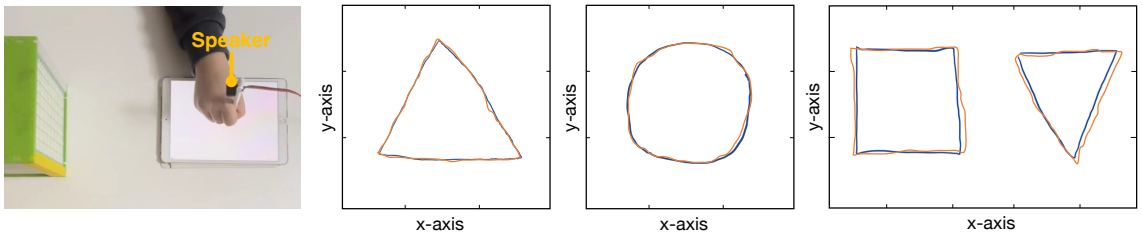
Given the mobility of sound source devices, their relative height to the receiver often varies. We place the sound source device at various heights to assess our system's performance with sound sources at different heights. The experimental results shown in Fig. 26 indicate that varying heights of the sound source can lead to a slight loss in angle estimation performance, resulting in a fluctuation in average angle estimation error ranging from 1.52 to 2.59 degrees across a height range of −40cm to 40cm. Currently, our system primarily focuses on azimuth angles in its estimations. However, the elevation angle (pitch) also has a corresponding response, a crucial direction for our future exploration.

5.3.2 Environment Migration. We test the robustness of our well-assembled and trained system across different environments. To do this, we evaluate our testing platform, initially set up in an office, directly in three other distinct environments without any extra adjustments. These environments are a bedroom, a parlor, and outdoors. The experimental results, shown in Fig. 27, indicate that the system maintains stable performance in these varied settings, with only a minor deviation in average error of 0.051 degrees. This robust performance is due to



(a) Outdoor setup. (b) Indoor setup. (c) Localization error.

Fig. 29. Performance of device localization. (a) and (b) are the outdoor and indoor setups, respectively. (c) shows the mean localization error, where S1 and S2 represent using one source and two sources, respectively.



(a) Tracking setup (b) User 1 (c) User 2 (d) User 1 and User 2

Fig. 30. Performance of MetaAng in acoustic tracking. (a) represents the tracking setup. (b) and (c) show the tracking results using a single source from User 1 and User 2, respectively. (d) illustrates the tracking result using two sources simultaneously from User 1 and User 2.

the angle encoding being solely reliant on the response of the AMS and microphones to the incident signals, independent of the environment. This feature allows our system to seamlessly deploy in various environments without additional costs for environment-specific adaptations. Such environmental independence is a significant advantage, ensuring the system's effectiveness in a wide range of real-world applications without requiring frequent recalibration or reconfiguration to suit different settings.

5.3.3 Types of Commodity Devices. We compare the impact of different sound source devices, selecting five commercial devices for evaluation: Google Nexus 5, Honor Magic3, Samsung S8, iPhone 11, and Watch S5. The experimental results, presented in Fig. 28, reveal that each device achieves average angle estimation errors of 1.17, 1.26, 1.22, 1.26, and 1.78 degrees, respectively. We note that smartphones maintain stable angle estimation performance. Smartwatches, while slightly less accurate than smartphones, still perform within a low error range, likely due to their limited sound volume output. These results demonstrate that our system is compatible with various commercial devices and consistently maintains stable performance in angle estimation. This versatility and reliability are vital for the practical application of our system across diverse environments, accommodating different types of sound source devices.

5.4 Case Study

In this section, we present the performance of MetaAng in practical applications, including localization and tracking.

5.4.1 Localization. We investigate the localization capabilities of MetaAng for both single-device and multi-device scenarios in diverse environments, specifically indoor and outdoor settings as illustrated in Fig. 29. In the single-device scenario, we use one smartphone, and in the multi-device scenario, we use two smartphones. For each scenario, we predefine the devices' initial positions as the origin and estimate their positions at 50 different points to calculate the error. The results detailed in Fig. 29(c) indicate that despite the higher indoor positioning error due to complex multipath environments, MetaAng achieves acceptable indoor positioning errors for both single and multiple devices, with average errors of 2.62 cm and 3.51, respectively. This performance underlines its suitability for a broad range of localization applications, demonstrating its robustness and reliability in varying environmental conditions.

5.4.2 Tracking. We develop a simple 2D air drawing interface using MetaAng. Fig. 30 (a) depicts the tracking setup, a small speaker is attached to the end of an Apple Pencil, with an iPad serving as the canvas for ground truth. We invite two participants to draw images on the iPad to assess the tracking capability of MetaAng. Furthermore, we ask them to draw simultaneously on two iPads to evaluate the multi-source tracking performance. As shown in Fig. 30 (b)-(d), the images drawn by MetaAng almost completely overlap with the ground truth, demonstrating its high precision in tracking. Moreover, MetaAng's ability to perform multi-source tracking with a single microphone array holds promise for a wide array of applications, such as human pose estimation and VR/AR.

6 DISCUSSION AND FUTURE WORK

6.1 2D Angle Estimation

We currently focus on azimuthal angle estimation, but elevation angle estimation is equally essential for scenarios requiring 3D positioning or imaging. Elevation angle estimation can be approached using a method similar to that for incident angle encoding, effectively increasing the number of unknowns in the measurement matrix. One potential solution is using more finely quantized AMS units, such as expanding from 16 to 32 options and enhancing the AMS's control over incident signals. Additionally, we are considering more efficient compressive sensing algorithms to handle a more significant number of unknowns under conditions of rank deficiency in the measurement matrix, which is one of our future work directions.

6.2 Energy Attenuation

The presence of AMS results in a slight energy attenuation in acoustic signals, which is nearly unavoidable. Our experiments have shown that this loss increases from low to high frequencies, leading to concerns about a reduction in acoustic sensing distance. The effect of AMS is minimal on low-frequency voices, but it is evident that the degree of attenuation varies in different directions, with the most severe attenuation occurring in the direction perpendicular to AMS. This necessitates the consideration of AMS's impact when designing spatial audio services for smart speakers, such as for AR/VR applications. On the other hand, although the energy attenuation of high-frequency inaudible sound signals is close to 2dB, because humans are less sensitive to these frequencies, it is possible to increase the perceived distance by raising the volume. Another potential approach is to use advanced signal processing techniques to improve the signal-to-noise ratio of microphones, which will be a part of our future work.

6.3 Extending to Passive Sensing

Many applications based on inaudible acoustic signals have been proposed, primarily categorized into active and passive sensing. Active sensing methodologies are centered around using the receiver end to discern the transmitting device's state, such as its position. In contrast, passive sensing revolves around extracting target information from signals backscattered by objects independent of the signal source. Although our current angle

estimation method primarily concentrates on active acoustic perception, its principles apply equally to passive sensing applications. Moving forward, we intend to expand the scope of our proposed methodology to encompass passive acoustic sensing applications, thereby contributing more holistically to the evolution of the acoustic perception research community.

6.4 User Friendliness of the Selected Frequency Band

In this work, we choose the frequency range of 16kHz to 20kHz, as many studies have shown that the majority of people might be insensitive to this frequency range. However, this might not apply to children or pets, especially since the modulation of chirp can lead to severe discomfort. There are a few potential solutions, one of which is using higher frequency acoustic sensing. Some advanced smart speakers are equipped with higher sampling rates, such as 192kHz [43], allowing for acoustic perception that is transparent and friendly to humans on higher frequency bands. Another solution is to adopt certain user-friendly special modulation methods, such as white noise modulation [49], which allows people to hear the sound without discomfort.

7 RELATED WORKS

Our work intersects with the following areas: This section presents some advances related to our works, including spatial perception of speech, acoustic sensing and angle estimation, and acoustic metasurface.

7.1 Human Sound Spatial Perception

The human auditory system (made up of two ears) allows a person to naturally discern the general direction of a sound source even in a noisy environment, which is also known as the cocktail party problem [17]. Inspired by this, researchers have developed various voice-based applications using microphone arrays, such as sound source localization [12, 47], speech enhancement [8, 69], and speaker extraction [72]. Though there also exist some works [28, 30] on single-channel speech applications based on deep learning, this is beyond the scope this paper. Compared with the single-channel system, the human auditory system (two channels) can collect spatial diversities, while more channels (such as an array with more than 2 microphones) can further improve the performance of these applications [45]. However, the performance of aforementioned applications mostly relies on the number of microphones and the inter-element spacing, while the commercial products tend to reduce the number of microphones to save cost and space (e.g., 4 microphones for Apple HomePod 2nd [1] and 3 microphones for Google Nest Mini [2]). Therefore, it is important to ensure the inter-element spacing of COTS smart speakers is large enough to collect spatial diverse features.

7.2 Acoustic Sensing and Angle Estimation

Wireless sensing has recently attracted significant attention in research community, and lots of innovative sensing algorithms and systems have been developed. Wireless signals with various modalities are explored for sensing/tracking, including acoustic [13, 55, 70, 71], WiFi [38, 48], RFID [16, 51], UWB [67], and mmWave signals [59]. Among these modalities, acoustic tracking using ultrasound signals only requires widely available speakers and microphones, while the other modalities require specialized and expensive hardware[32, 65]. Generally speaking, there are two types of ultrasonic acoustic tracking: (i) displacement tracking and (ii) AoA estimation. The former tracks the fine-grained displacement change of the target [56, 65], thus enables applications such as finger tracking [37] and gesture recognition [36]. The latter estimates the AoAs of the acoustic signal and enables applications such as localization [41] and spatial filtering [60]. Combining the techniques of displacement tracking and AoA estimation can further lead to room-scale tracking [34], multi-target tracking [22], and acoustic imaging [33]. As a consequence, the performance of AoA estimation is crucial for ultrasound acoustic sensing.

There are several AoA estimation algorithms using microphone arrays, including the intuitive conventional beamforming (CBF) [7], the minimum variance distortionless response (MVDR) [53], and the subspace-based supersolution methods [20, 39], etc. These approaches are widely used to improve the accuracy and resolution of AoA estimation, but their performance will be significantly decreased with sparse microphone arrays with ambiguity angles. Researchers attempt to cope this issue by using nonlinear array geometries [53], distributed microphone arrays [27], and utilizing the frequency diversities [21]. However, for ultrasonic acoustic tracking, the above solutions are not applicable to COTS smart speakers, due to the limited product size, number of microphones, and available bandwidth for inaudible sounds.

7.3 Acoustic Metasurface

Acoustic metasurfaces have been extensively studied for their versatile manipulation of acoustic signals, driving progress in communication [71], imaging [29], and noise reduction [14] applications. Broadly categorized as active [46] and passive [35], active AMS often involves intricate structures, leading to higher costs and larger sizes. In contrast, passive AMS, celebrated for their simplicity and ease of fabrication, have gained considerable attention, notably through techniques like 3D printing. Common designs for passive AMS units include coiled structures [9], Helmholtz resonators [73], and membrane structures [19]. Opting for coiled passive structures in our approach stems from their cost-effectiveness and practicality, in contrast to the more complex Helmholtz resonator structures and the manufacturing challenges associated with membranes. The utilization of coiled structures allows us to exert control over the phase of incoming acoustic signals by manipulating the lengths of coiling paths within each unit. This method proves crucial in applications like beamforming, where adjusting the coil paths across all units compensates for phase variations in the input signal. Our focus diverges from previous research primarily concentrated on cell design and signal-to-noise ratio enhancement, as evident in [35] and [71]. Instead, our primary interest lies in harnessing AMS for incident angle encoding, with the goal of significantly enhancing angle estimation performance.

8 CONCLUSION

This paper introduces MetaAng, a system that achieves high-accuracy angle estimation of inaudible ultrasonic signals using a passive structure and a small number of microphones, while retaining microphone arrays' low-frequency sound spatial perception ability. MetaAng integrates an acoustic metasurface to enhance the uniqueness of incident angles through frequency diversity and configuration optimization. Moreover, MetaAng introduces USTNet, incorporating two key technologies, i.e., compressive sensing and neural-enhanced priors, to improve the resolution and accuracy of angle estimation in multiple acoustic sources. The experimental results under various conditions and scenarios indicate that MetaAng demonstrates high precision and robust angle estimation performance, offering unique insights for applications based on inaudible acoustic spatial perception.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful feedback. This research was supported in part by the National Natural Science Foundation of China under Grant No. 62122095, 62341201, 62072472, U2336204 and 62202256, and by a grant from the Guoqiang Institute, Tsinghua University. It was also partially supported by the Sichuan Natural Science Foundation Project No. 24ZNSFSC0038.

REFERENCES

[1] HomePod (2nd generation). https://www.apple.com/homepod-2nd-generation/specs/.
[2] Nest Mini (2nd Gen). https://store.google.com/us/product/google_nest_mini?hl=en-US.
[3] COMSOL: simulate real-world designs, devices, and processes with multiphysics software from comsol. https://www.ti.com/product/LM386, 2023.2.

[4] M. Arar, C. Jung, J. Awad, and A. H. Chohan. Analysis of smart home technology acceptance and preference for elderly in dubai, uae. *Designs*, 5(4):70, 2021.

[5] Y. Arjoune, N. Kaabouch, H. El Ghazi, and A. Tamtaoui. A performance comparison of measurement matrices in compressive sensing. *International Journal of Communication Systems*, 31(10):e3576, 2018.

[6] Bela platform, 2017. https://bela.io.

[7] J. Benesty, J. Chen, and Y. Huang. Conventional beamforming techniques. *Microphone array signal processing*, pages 39–65, 2008.

[8] Y. Cao, S. SRIDHARAN, and M. MOODY. Speech enhancement using microphone array with multi-stage processing. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 79(3):386–394, 1996.

[9] T. Chen, J. Jiao, and D. Yu. Enhanced broadband acoustic sensing in gradient coiled metamaterials. *Journal of Physics D: Applied Physics*, 54(8):085501, 2020.

[10] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong. Multi-channel overlapped speech recognition with location guided speech extraction network. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 558–565. IEEE, 2018.

[11] R. V. Cox, S. F. D. C. Neto, C. Lamblin, and M. H. Sherif. Itu-t coders for wideband, superwideband, and fullband speech communication [series editorial]. *IEEE Communications Magazine*, 47(10):106–109, 2009.

[12] D. Desai and N. Mehendale. A review on sound source localization systems. *Archives of Computational Methods in Engineering*, 29(7):4631–4642, 2022.

[13] Y. Fu, S. Wang, L. Zhong, L. Chen, J. Ren, and Y. Zhang. Svoice: Enabling voice communication in silence via acoustic sensing on commodity devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 622–636, 2022.

[14] N. Gao, Z. Zhang, J. Deng, X. Guo, B. Cheng, and H. Hou. Acoustic metamaterials for noise reduction: a review. *Advanced Materials Technologies*, 7(6):2100698, 2022.

[15] N. Garg, Y. Bai, and N. Roy. Owlet: enabling spatial information in ubiquitous acoustic devices. In *Proc. of ACM MobiSys*, 2021.

[16] U. Ha, J. Leng, A. Khaddaj, and F. Adib. Food and liquid sensing in practical environments using rfids. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, pages 1083–1100, 2020.

[17] S. Haykin and Z. Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

[18] Q. Huang, G. Zhang, and K. Liu. Near-field source localization using spherical microphone arrays. *Chinese Journal of Electronics*, 25(1):159–166, 2016.

[19] J. Lan, X. Zhang, X. Liu, and Y. Li. Wavefront manipulation based on transmissive acoustic metasurface with membrane-type hybrid structure. *Scientific reports*, 8(1):14171, 2018.

[20] T. B. Lavate, V. Kokate, and A. Sapkal. Performance analysis of music and esprit doa estimation algorithms for adaptive array smart antenna in mobile communication. In *Computer and Network Technology (ICCNT), 2010 Second International Conference on*, pages 308–311. IEEE, 2010.

[21] A. Lazaro, D. Girbau, P. Moravek, and R. Villarino. A study on localization in wireless sensor networks using frequency diversity for mitigating multipath effects. *Elektronika ir Elektrotechnika*, 19(3):82–87, 2013.

[22] D. Li, J. Liu, S. I. Lee, and J. Xiong. Fm-track: pushing the limits of contactless multi-target tracking using acoustic signals. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 150–163, 2020.

[23] D. Li, J. Liu, S. I. Lee, and J. Xiong. Room-scale hand gesture recognition using smart speakers. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 462–475, 2022.

[24] X. Li, Y. Yang, Z. Ye, Y. Wang, and Y. Chen. 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*, pages 240–249, 2023.

[25] J. Lian, J. Lou, L. Chen, and X. Yuan. Echospot: Spotting your locations via acoustic sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–21, 2021.

[26] J. Lian, X. Yuan, M. Li, and N.-F. Tzeng. Fall detection via inaudible acoustic sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–21, 2021.

[27] M. U. Liaquat, H. S. Munawar, A. Rahman, Z. Qadir, A. Z. Kouzani, and M. P. Mahmud. Sound localization for ad-hoc microphone arrays. *Energies*, 14(12):3446, 2021.

[28] Y. Luo and N. Mesgarani. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE, 2018.

[29] F. Ma, Z. Huang, C. Liu, and J. H. Wu. Acoustic focusing and imaging via phononic crystal and acoustic metamaterials. *Journal of Applied Physics*, 131(1), 2022.

[30] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux. Whamr!: Noisy and reverberant single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE, 2020.

[31] R. J. Mailloux. *Phased array antenna handbook*. Artech house, 2017.

[32] W. Mao, J. He, and L. Qiu. CAT: high-precision acoustic motion tracking. In *Proc. of ACM MobiCom*, 2016.

[33] W. Mao, M. Wang, and L. Qiu. Aim: Acoustic imaging on a mobile. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, pages 468–481. ACM, 2018.

[34] W. Mao, M. Wang, W. Sun, L. Qiu, S. Pradhan, and Y.-C. Chen. Rnn-based room scale hand motion tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–16, 2019.

[35] G. Memoli, M. Caleap, M. Asakawa, D. R. Sahoo, B. W. Drinkwater, and S. Subramanian. Metamaterial bricks and quantization of meta-surfaces. *Nature Communication*, 2017.

[36] B. S. Moreira, A. Perkusich, and S. O. Luiz. An acoustic sensing gesture recognition system design based on a hidden markov model. *Sensors*, 20(17):4803, 2020.

[37] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota. FingerIO: Using active sonar for fine-grained finger tracking. In *Proc. of ACM CHI*, pages 1515–1525, 2016.

[38] Q. Pu, S. Gupta, S. Gollakota, and S. Patel. Whole-home gesture recognition using wireless signals. In *Proc. of ACM MobiCom*, 2013.

[39] R. O. Schmidt. A signal subspace approach to multiple emitter location spectral estimation. *Ph. D. Thesis, Stanford University*, 1981.

[40] I. Selesnick. Sparse regularization via convex analysis. *IEEE Transactions on Signal Processing*, 65(17):4481–4494, 2017.

[41] S. Shen, D. Chen, Y.-L. Wei, Z. Yang, and R. R. Choudhury. Voice localization using nearby wall reflections. In *Proc. of ACM MobiCom*, 2020.

[42] W. Shi, F. Jiang, S. Liu, and D. Zhao. Scalable convolutional neural network for image compressed sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12290–12299, 2019.

[43] Y. Su, F. Zhang, K. Niu, T. Wang, B. Jin, Z. Wang, Y. Jiang, D. Zhang, L. Qiu, and J. Xiong. Embracing distributed acoustic sensing in car cabin for children presence detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–28, 2024.

[44] K. Sun, T. Zhao, W. Wang, and L. Xie. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, pages 591–605, 2018.

[45] W. Sun. *From active to passive spatial acoustic sensing and applications.* PhD thesis, 2022.

[46] Z. Tian, C. Shen, J. Li, E. Reit, Y. Gu, H. Fu, S. A. Cummer, and T. J. Huang. Programmable acoustic metasurfaces. *Advanced functional materials*, 29(13):1808489, 2019.

[47] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau. Robust sound source localization using a microphone array on a mobile robot. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, volume 2, pages 1228–1233. IEEE, 2003.

[48] D. Vasisht, S. Kumar, and D. Katabi. Decimeter-level localization with a single wifi access point. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*, pages 165–178, 2016.

[49] A. Wang, J. E. Sunshine, and S. Gollakota. Contactless infant monitoring using white noise. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–16, 2019.

[50] J. Wang, D. Vasisht, and D. Katabi. Rf-idraw: virtual touch screen in the air using rf signals. *ACM SIGCOMM Computer Communication Review*, 44(4):235–246, 2014.

[51] J. Wang, J. Xiong, X. Chen, H. Jiang, R. K. Balan, and D. Fang. Tagscan: Simultaneous target imaging and material identification with commodity rfid devices. In *Proc. of ACM MobiCom*, pages 288–300. ACM, 2017.

[52] L. Wang, T. Gu, W. Li, H. Dai, Y. Zhang, D. Yu, C. Xu, and D. Zhang. Df-sense: Multi-user acoustic sensing for heartbeat monitoring with dualforming. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, pages 1–13, 2023.

[53] M. Wang, W. Sun, and L. Qiu. {MAVL}: Multiresolution analysis of voice localization. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*, pages 845–858, 2021.

[54] M. Wang, W. Sun, and L. Qiu. Mavl: Multiresolution analysis of voice localization. In *Proc. of NSDI*, 2021.

[55] S. Wang, L. Zhong, Y. Fu, L. Chen, J. Ren, and Y. Zhang. Uface: Your smartphone can" hear" your facial expression! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–27, 2024.

[56] W. Wang, A. X. Liu, and K. Sun. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pages 82–94. ACM, 2016.

[57] Y. Wang, J. Cao, and C. Yang. Recovery of seismic wavefields based on compressive sensing by an l1-norm constrained trust region method and the piecewise random subsampling. *Geophysical Journal International*, 187(1):199–213, 2011.

[58] R. Watanabe, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo. Dnn-based frequency component prediction for frequency-domain audio source separation. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 805–809. IEEE, 2021.

[59] T. Wei and X. Zhang. mTrack: high precision passive tracking using millimeter wave radios. In *Proc. of ACM MobiCom*, 2015.

[60] Y.-L. Wei and R. R. Choudhury. Estimating angle of arrival (aoa) of multiple echoes in a steering vector space. *arXiv preprint arXiv:2109.13072*, 2021.

[61] P. Xiao and B. Liao. Robust one-bit compressive sensing with weighted l1-norm minimization. *Signal Processing*, 164:380–385, 2019.

[62] Y. Xiao, Q. Wang, and Q. Hu. Non-smooth equations based method for l1-norm problems with applications to compressed sensing. *Nonlinear Analysis: Theory, Methods & Applications*, 74(11):3570–3577, 2011.

[63] J. Xiong, K. Sundaresan, and K. Jamieson. Tonetrack: Leveraging frequency-agile radios for time-based indoor wireless localization. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 537–549. ACM, 2015.
[64] D. You, J. Xie, and J. Zhang. Ista-net++: Flexible deep unfolding network for compressive sensing. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.
[65] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 15–28. ACM, 2017.
[66] F. Zhang, Z. Wang, B. Jin, J. Xiong, and D. Zhang. Your smart speaker can" hear" your heartbeat! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–24, 2020.
[67] F. Zhang, J. Xiong, Z. Chang, J. Ma, and D. Zhang. Mobi2sense: empowering wireless sensing with mobility. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, pages 268–281, 2022.
[68] H. Zhang, Q. Fu, and Y. Yan. Speech enhancement using compact microphone array and applications in distant speech acquisition. *Chinese Journal of Electronics*, 18(3):481–486, 2009.
[69] S. Zhang and X. Li. Microphone array generalization for multichannel narrowband deep speech enhancement. *arXiv preprint arXiv:2107.12601*, 2021.
[70] Y. Zhang, H. Pan, Y.-C. Chen, L. Qiu, Y. Lu, G. Xue, J. Yu, F. Lyu, and H. Wang. Addressing practical challenges in acoustic sensing to enable fast motion tracking. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*, pages 82–95, 2023.
[71] Y. Zhang, Y. Wang, L. Yang, M. Wang, Y.-C. Chen, L. Qiu, Y. Liu, G. Xue, and J. Yu. Acoustic sensing and communication using metasurface. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 1359–1374, 2023.
[72] S. Zheng, S. Zhang, W. Huang, Q. Chen, H. Suo, M. Lei, J. Feng, and Z. Yan. Beamtransformer: Microphone array-based overlapping speech detection. *arXiv preprint arXiv:2109.04049*, 2021.
[73] Y. Zhu and B. Assouar. Multifunctional acoustic metasurface based on an array of helmholtz resonators. *Physical review B*, 99(17):174109, 2019.