# Meta-Speaker: Acoustic Source Projection by Exploiting Air Nonlinearity

Weiguo Wang
Tsinghua University
wwg18@mails.tsinghua.edu.cn

Yuan He*
Tsinghua University
heyuan@tsinghua.edu.cn

Meng Jin
Shanghai Jiao Tong University
jinm@sjtu.edu.cn

Yimiao Sun
Tsinghua University
sym21@mails.tsinghua.edu.cn

Xiuzhen Guo
Zhejiang University
guoxiuzhen94@gmail.com

## ABSTRACT

This paper proposes Meta-Speaker, an innovative speaker capable of projecting audible sources into the air with a high level of manipulability. Unlike traditional speakers that emit sound waves in all directions, Meta-Speaker can manipulate the granularity of the audible region, down to a single point, and can manipulate the location of the source. Additionally, the source projected by Meta-Speaker is a physical presence in space, allowing both humans and machines to perceive it with spatial awareness. Meta-Speaker achieves this by leveraging the fact that air is a nonlinear medium, which enables the reproduction of audible sources from ultrasounds. Meta-Speaker comprises two distributed ultrasonic arrays, each transmitting a narrow ultrasonic beam. The audible source can be reproduced at the intersection of the beams. We present a comprehensive profiling of Meta-Speaker to validate the high manipulability it offers. We prototype Meta-Speaker and demonstrate its potential through three applications: anchor-free localization with a median error of 0.13 m, location-aware communication with a throughput of 1.28 Kbps, and acoustic augmented reality where users can perceive source direction with a mean error of 9.8 degrees.

## CCS CONCEPTS

• **Hardware** → **Sound-based input / output**; • **Human-centered computing** → *Interaction techniques*;

## KEYWORDS

Acoustic Field Manipulation, Air Nonlinearity, Speaker, Acoustic Sensing, Acoustic Communication, Localization

**ACM Reference Format:**
Weiguo Wang, Yuan He, Meng Jin, Yimiao Sun, and Xiuzhen Guo. 2023. Meta-Speaker: Acoustic Source Projection by Exploiting Air Nonlinearity. In *The 29th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '23), October 2–6, 2023, Madrid, Spain.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3570361.3613279

## 1 INTRODUCTION

Acoustic Field Manipulation (AFM) has immense potential for shaping and controlling the spatial distribution of mechanical energy within the medium. With AFM, it is possible to create personal sound zones where multiple listeners can hear their individual sounds without being disturbed by the sound in the other zones [7, 36]. Additionally, AFM enables precise manipulation of sound waves and their propagation, with great significance in designing spaces with great musical immersion [1, 13] and reducing noise pollution [15, 27].

There are two perspectives to achieving AFM: (1) *wave propagation* and (2) *source projection*. The former focuses on controlling the propagation of sound waves. The most common method is to leverage the obstacles (e.g. wall) to reflect waves, thereby changing their propagation direction. Recent development in acoustic metamaterials enables programmable control of wave propagation. For example, acoustic metamaterials can redirect reflection [12, 43], guide incident waves for "acoustic black holes" [8, 21, 51], or suppress scattering for "acoustic cloaking" [5, 24, 34].

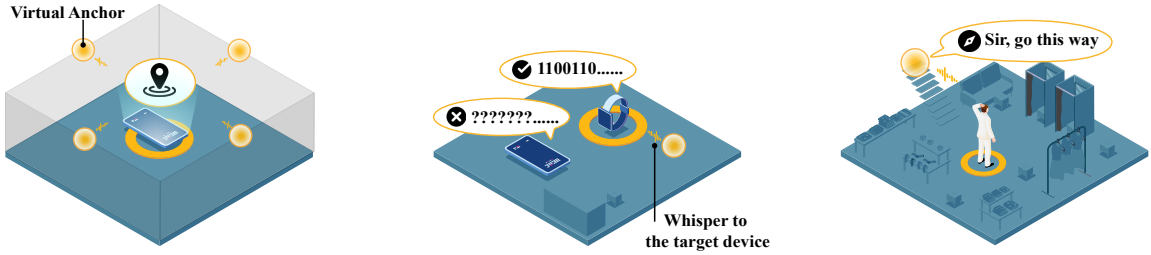
Source projection typically requires multiple distributed loudspeakers deployed in the space. By carefully designing the constructive and deconstructive combinations of sound waves, one or more independent sound zones can be created [2, 7, 28]. Generally, the more loudspeakers we deploy, the higher manipulability we gain for AFM. Besides distributed loudspeakers, parametric arrays (i.e., directional

*Yuan He is the corresponding author.



*ACM MobiCom '23, October 2–6, 2023, Madrid, Spain*

ACM ISBN 978-1-4503-9990-6/23/10.
https://doi.org/10.1145/3570361.3613279




(a) Anchor-free localization. (b) Location-aware communication. (c) Acoustic augmented reality.

**Figure 1: Illustrative applications of Meta-Speaker.**

speakers) [18, 44] are also used in AFM. They have the potential to improve freedom and flexibility in AFM, since they can project narrow audible beams, similar to lasers, and selectively manipulate a part of the acoustic field.

We present Meta-Speaker, a novel speaker with capabilities of projecting audible sources with high manipulability. Meta-Speaker offers distinct advantages over traditional loudspeakers and directional speakers in the following aspects: First, it enables precise control of the size of the audible region, down to a single point, which allows for finely-grained manipulation of the acoustic field. Second, the location of the audible source can be accurately manipulated, providing higher flexibility for AFM. Third, the source projected by Meta-Speaker is a physical presence in space, enabling humans and machines to perceive it as spatial audio.

Drawing inspiration from prior work on parametric arrays [18, 44], Meta-Speaker leverages the inherent nonlinearity of air to achieve its unique capabilities. Specifically, due to air fluidity, the propagation of a longitudinal wave (e.g., acoustic wave) can disturb the distribution of air molecules, resulting in uneven air density. In turn, the acoustic wave distorts itself as it travels through this heterogeneous medium—the air. This distortion offers an opportunity to reproduce an audible source from ultrasounds in the air: By accounting for the distortion and carefully modulating ultrasounds, we can harness the nonlinear distortion of ultrasounds to generate an audible source.

The design of Meta-Speaker employs two ultrasonic phased arrays. Each array transmits a narrow beam of ultrasound. When the two beams intersect, their nonlinear interaction, caused by the air can create an audible source at the intersection region. By varying the beamwidth of each array, Meta-Speaker can manipulate the size of the audible region. Meanwhile, the orientation of arrays can be steered to manipulate the intersection location and therefore manipulate the location of the audible region.

The manipulability of the projected acoustic source determines its usability in practice. Hence, we present a comprehensive profiling of Meta-Speaker's manipulability in three dimensions, namely spatial resolution, energy distribution, and frequency response. Based on this analysis, we validate that Meta-Speaker offers a high degree of manipulability with respect to the size and location of the audible source it reproduces, and can project multiple sources flexibly.

We present three proof-of-concept applications to demonstrate the great potential of Meta-Speaker: (1) **Anchor-free Localization.** Meta-Speaker can create multiple virtual anchors that broadcast acoustic beacons, by projecting audible sources at different locations, as shown in Fig. 1(a). This makes acoustic localization feasible without the need for physical anchors. (2) **Location-aware Communication.** Meta-Speaker enables communication in a spatially-selective manner, as illustrated in Fig. 1(b). Acoustic messages can be transmitted solely to a targeted device, while devices located elsewhere cannot perceive such messages. (3) **Acoustic Augmented Reality.** The physical presence of the reproduced audio in space allows humans to hear it spatially, as depicted in Fig. 1(c). This feature enables Meta-Speaker to interact directly with humans, e.g. by guiding people to destinations via spatial audios.

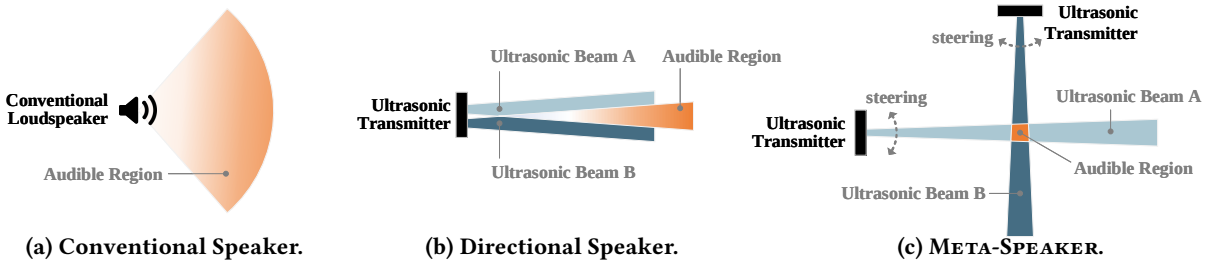
Our contributions are summarized as follows:

- We demonstrate the feasibility of projecting audible sources with separated ultrasonic beams, which enables unique capability of projecting sources with high manipulability.
- We present the design and implementation of Meta-Speaker. We conduct thorough analysis on its fundamental properties both theoretically and experimentally.
- Meta-Speaker will enable diverse applications. We showcase three examples: anchor-free localization, location-aware communication, and acoustic augmented reality.

**Roadmap.** Sec. 2 introduces and validates the idea of reproducing audible sources from ultrasounds. Sec. 3 introduces the design of Meta-Speaker, followed by its comprehensive profiling in Sec. 4. Three illustrative applications are presented in Sec. 5, Sec. 6, and Sec. 7. Sec. 8 gives the related work. Sec. 9 and Sec. 10 discuss and conclude this work.

## 2 SOUND FROM SILENCE

This section explains the air nonlinearity, from which we can reproduce an audible source from ultrasounds. It also introduces and validates the key idea of Meta-Speaker.




(a) Conventional Speaker. (b) Directional Speaker. (c) Meta-Speaker.

**Figure 2: Compared to the (a) conventional speaker and (b) directional speaker, (c) Meta-Speaker allows us to divide and multiplex the acoustic channel spatially at a much finer granularity.**

## 2.1 Key Idea

**Acoustic Nonlinearity in Air.** The airborne acoustic source is caused by the vibration of air molecules, which generates mechanical waves propagating in the air medium. Under the assumption that the air is a homogeneous medium, the air can be viewed as a linear system, whose output is the linear combination of the input. Suppose the input signal is $x(t)$, the output $y(t)$ can be expressed as

$$y(t) = A \cdot x(t), \tag{1}$$

where $A$ is the channel gain. This linear system model is widely accepted in the literature for its simplicity.

However, airborne acoustic signal exhibits nonlinearity inherently, as theoretically described by KZK equations [19, 52]. A fact is that *acoustic signal distorts itself as it propagates*: Given that the wave will cause the vibration of the air molecules, the air medium along the wave will experience different vibration velocities. This generally distributes the air medium unevenly and makes it heterogeneous. When propagating in such a heterogeneous medium, the acoustic signal experiences nonlinear distortion inevitably.

To account for this nonlinearity, we can use the Taylor series approximation, which yields an expression of the form:

$$y(t) = A_1 \cdot x(t) + A_2 \cdot x^2(t) + \underbrace{A_3 \cdot x^3(t) + \ldots}_{\text{can be ignored}} \tag{2}$$

where $A_1$, $A_2$, and $A_3$ are the channel gains of the first-, second-, and third-order terms, respectively. However, in practice, we can safely ignore third-order, and higher-order terms since their channel gains are rather weak compared to the first- and second-order terms [9].

By exploiting air nonlinearity, it is possible to reproduce an audible source from ultrasounds. Suppose a speaker plays two tones with frequencies $f_A$ and $f_B$, expressed as:

$$x(t) = \sin(2\pi f_A t) + \sin(2\pi f_B t). \tag{3}$$

It can be shown that the square term (second-order term) in Eq. (2) generates a waveform containing frequencies at the sums and differences between $f_A$ and $f_B$:

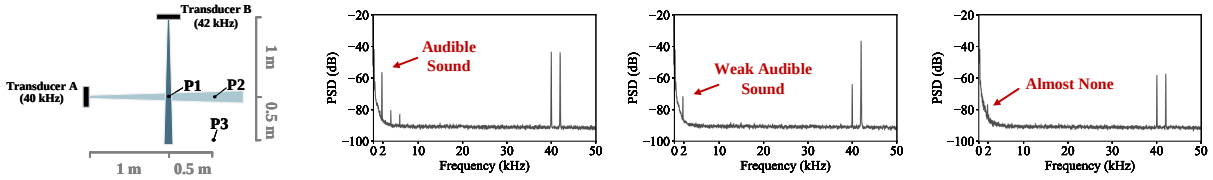
$$\begin{aligned}
&[\sin(2\pi f_A t) + \sin(2\pi f_B t)]^2 \\
&= \sin^2(2\pi f_A t) + \sin^2(2\pi f_A t) + 2\sin(2\pi f_A t)\sin(2\pi f_B t) \\
&= 1 - \frac{1}{2}\cos(2\pi 2 f_A t) - \frac{1}{2}\cos(2\pi 2 f_B t) + \cos(2\pi(f_A - f_B)t) \\
&- \cos(2\pi(f_A + f_B)t).
\end{aligned} \tag{4}$$

If both $f_A$ and $f_B$ are beyond audible frequency, say 20 kHz, we can safely ignore the terms with sum frequencies since they produce additional ultrasonic signals. The important thing is that, if we carefully choose $f_A$ and $f_B$ to ensure their difference $f_A - f_B$ is within the audible frequency range, the audible source can thus be reproduced.

The nonlinear behavior of air gives rise to *directional speakers* (a.k.a, parametric array [44], or audio spotlight [18]). Unlike conventional loudspeakers, which emit sound in all directions, as depicted in Fig. 2(a), the directional speaker can project a narrow audible beam, as shown in Fig. 2(b). To achieve this, the direction speaker uses ultrasonic transducers to play two narrow ultrasonic beams that overlap and modulate each other to produce the desired audible sound.

**Meta-Speaker.** In this paper, we propose a novel speaker, Meta-Speaker, which has the capability to project audible sources with a high level of manipulability, as illustrated in Fig. 2(c). Meta-Speaker also leverages air nonlinearity to reproduce audible sounds from ultrasounds. Our design involves two ultrasonic transmitters deployed distributedly. Each array transmits an ultrasonic beam. An audible sound is expected to be reproduced at the intersection of the beams.

It is worth noting that the beam separation in Meta-Speaker allows for a high degree of granularity in projecting audible sources, providing a high level of manipulability not available in traditional directional speakers. To illustrate this, let us compare the overlap region of ultrasonic beams transmitted by a directional speaker and Meta-Speaker (see Fig. 2(b) and (c)). In the case of the directional speaker, its ultrasonic beams are completely overlapped as they are emitted from the same ultrasonic transmitter, resulting in a fixed beam along a specific direction. Differently, Meta-Speaker allows for the ultrasonic beams to intersect, and the audible




(a) Recording places. (b) P1. (c) P2. (d) P3.

**Figure 3: As shown in (a), we record acoustic signals at three different places: P1, P2, and P3. A distinct audible frequency can be seen in (b), while it gets much weaker or even disappears in (c) and (d).**

region becomes the intersection of these beams. Due to the high directivity of the ultrasonic beams, the audible region can be as small as a point.

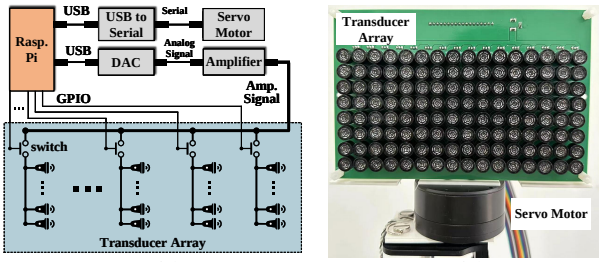
Moreover, by steering the orientations of the two ultrasonic beams, we can manipulate the location of the audible region, allowing us to project an audible source to a specific location. This capability enables us to spatially divide and multiplex the acoustic channel at a much finer granularity, surpassing both conventional and directional speakers.

## 2.2 Quick Validation

We conduct a proof-of-concept experiment to validate our proposed idea. In Fig. 3(a), we have two ultrasonic transmitters playing two sinusoidal tones with frequencies of 40 kHz and 42 kHz, respectively. As predicted by Eq. (4), we expect to hear a difference frequency of 2 kHz at the intersection of the beams (i.e., P1), and minimal to no sound at other locations (such as P2 and P3). To confirm this, we record signals at three different points (P1, P2, and P3) for analysis.

Prior to conducting our measurements, we must address the following two issues to record the signals truthfully. (1) Microphone nonlinearity, as reported in the literature [31, 32, 53], can result in the nonlinear behavior of some microphone hardware components, causing them to reproduce audible sounds from ultrasounds additionally. To mitigate this issue, we use a low-noise analog-to-digital converter (ADC) to directly sample the output signal of an electret sound sensor. To help measure and compare the power of the signals, we also disable the sound sensor's automatic gain control (AGC). (2) Sampling aliasing can occur if the sampling rate does not abide by the Nyquist sampling theorem, which can result in aliased frequencies in the sampled digital signal [26]. To avoid this issue, we set the ADC's sampling rate to 500 kHz.

The results of the proof-of-concept experiments depicted in Fig. 3(b)-(d) provide strong evidence in support of our idea. The PSD analysis clearly shows the presence of a distinct audible frequency of 2 kHz, which is the difference between the two ultrasonic tones at 40 kHz and 42 kHz. The presence of this frequency at the intersection of the ultrasonic beams (i.e., P1) indicates that the audible source is indeed successfully reproduced by leveraging air nonlinearity, as we hypothesized. On the other hand, the significant reduction in sound power at other locations away from the intersection (e.g., P2 and P3) further support our claims of fine-grained spatial manipulation of the acoustic field.




(a) (b)

**Figure 4: (a) Hardware design of the transmitter. (b) The transducer array sits on top of a servo motor.**

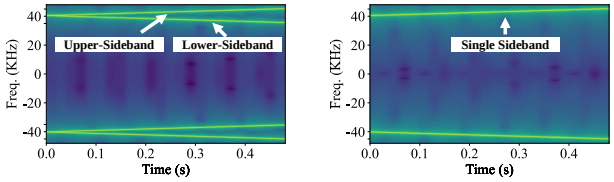
## 3 DESIGN

This section first presents the hardware design of the ultrasonic transmitter for Meta-Speaker. Then, the signal modulation for each transmitter is introduced.

### 3.1 Hardware

To implement Meta-Speaker, two ultrasonic transmitters that cooperate with each other are required. The hardware design of a transmitter is shown in Fig. 4(a). A Raspberry Pi 4B is used as the central controller for each transmitter and is connected to a 24-bit digital-to-analog converter (DAC), ESS9023, via a USB 2.0 interface. With a sample rate of 96 kHz, the DAC can output ultrasonic signals with frequencies up to 48 kHz. The output from the DAC is amplified using a 50 W Class-D power amplifier, TI TPA3116D2, which drives an ultrasonic array consisting of multiple transducers, Murata MA40S4S, connected in parallel with a 10 mm spacing.

The ultrasonic beam width can be adjusted by selectively activating a subset of the transducers. The transducer array has a rectangular layout with 16 columns and 8 rows, with each column connected to the amplified signal through a switch controlled by a GPIO pin on the Raspberry Pi. The transmitter can enable or disable a column of transducers by toggling the corresponding GPIO pin. To steer the ultrasonic beam, the array is mounted on top of a servo motor, K-Tech MF7015, as shown in Fig. 4(b). The Pi communicates with the motor using a UART protocol.




(a) Double Sidebands. (b) Single Sideband.

**Figure 5: Spectrum of up-converted audio.**

## 3.2 Signal Modulation

In Sec. 2.2, we have demonstrated the feasibility of reproducing simple sinusoidal audio in the air. Here, we focus on a general problem, that is, how to modulate the ultrasounds each transmitter emits to reproduce arbitrary audio $v(t)$.

**Amplitude Modulation.** One straightforward solution is to adopt the concept of amplitude modulation (AM), which is widely used in directional speakers [50].

In directional speakers, AM basically changes the envelope of an ultrasonic carrier according to the desired audible signal $v(t)$. Specifically, the modulated signal is given by

$$\begin{aligned} x(t) &= [1 + m \cdot v(t)] \cdot \sin(2\pi f_c t) \\ &= \underbrace{\sin(2\pi f_c t)}_{\text{carrier signal}} + m \cdot \underbrace{v(t) \cdot \sin(2\pi f_c t)}_{\text{up-converted signal}}, \end{aligned} \tag{5}$$

where $f_c$ is the carrier frequency, typically at 40 kHz, and $m$ denotes the modulation index. Intuitively, AM first up-converts the audio $v(t)$ to the ultrasonic frequency range by multiplying it with a sinusoidal tone at $f_c$. In order to make it audible, we need to additionally play a carrier signal $\sin(2\pi f_c t)$, which is used to down-convert the up-converted signal in the air by exploiting air nonlinearity.

Inspired by this, for META-SPEAKER, we can actually let two ultrasonic transmitters transmit the carrier and the up-converted audio separately. Specifically, the modulated signals for two transmitters A and B are as follows:

$$\begin{cases} x_A(t) = \sin(2\pi f_c t) & \text{(carrier)}, \\ x_B(t) = m \cdot v(t) \cdot \sin(2\pi f_c t) & \text{(up-converted audio)}. \end{cases} \tag{6}$$

It can be expected that when these two ultrasonic beams meet in the air at one point, the carrier will nonlinearly interact with the up-converted audio, thereby reproducing the desired audible signal $v(t)$.

**Self-Demodulation Problem and its Solution.** Unfortunately, the modulation method described above may lead to a self-demodulation problem: the undesired audible signals will be reproduced due to the presence of double sidebands in the up-converted audio.

To better understand this problem, let's consider the case where the desired reproduced audio, $v(t)$, is a chirp sweeping from 0 kHz to 5 kHz. As illustrated in Fig. 5(a), when $v(t)$ is multiplied by the 40 kHz tone, both its positive and negative basebands are shifted into the ultrasonic range centered at 40 kHz, creating two sidebands: an upper-sideband and a lower-sideband. Transmitting the modulated audio with double sidebands can be seen as transmitting two ultrasonic signals simultaneously: an up-chirp sweeping from 40 kHz to 45 kHz and a down-chirp sweeping from 40 kHz to 35 kHz.

Due to air nonlinearity, it can be expected that these two chirps will produce an additional undesired audible signal. What's more, the undesired audible signal will be projected in a direction rather than a spatial point, significantly undermining the spatial multiplexing granularity of META-SPEAKER.

To mitigate the above problem, we should reduce one sideband of the modulated signal. Our solution is straightforward: discarding the negative baseband of $v(t)$ before up-converting the signal. Specifically, let $\hat{v}(t)$ denote the Hilbert transform of $v(t)$. The analytic version of $v(t)$ with the negative baseband discarded, $v_a(t)$, can be obtained as

$$v_a(t) = v(t) + j\hat{v}(t). \tag{7}$$

Then, the up-converted signal transmitted by the transducer B in Eq. (6) can be updated to

$$\begin{aligned} x_B^{'}(t) &= m \cdot \mathrm{Re}\left\{ v_a(t) \cdot e^{j2\pi f_c t} \right\} \\ &= m \cdot \left[ v(t) \cdot \cos(2\pi f_c t) - \hat{v}(t) \cdot \sin(2\pi f_c t) \right], \end{aligned} \tag{8}$$

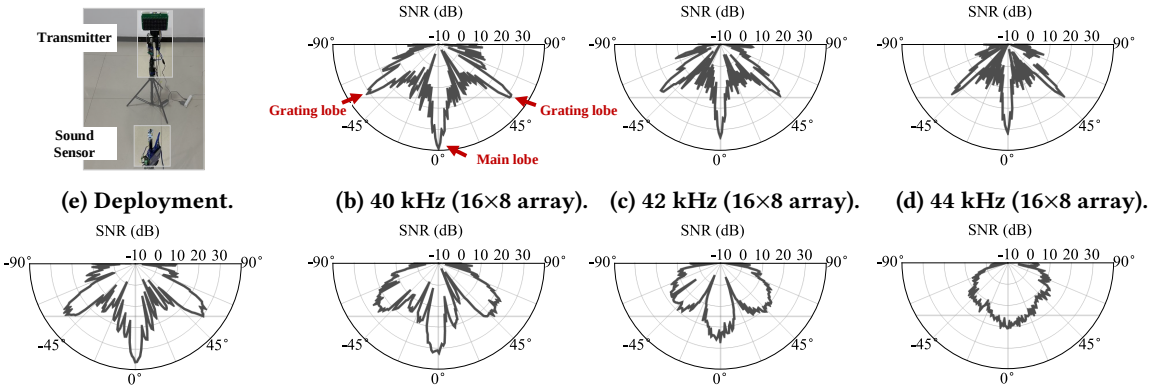
where $\mathrm{Re}\{\bullet\}$ denotes the operation that extracts the real part from a complex value. Fig. 5(b) illustrates the spectrum of $x_B^{'}(t)$. Clearly, the lower-sideband is significantly reduced, and there is only a single sideband left, thus suppressing the self-demodulation problem.

# 4 PROFILING

META-SPEAKER is expected to offer high manipulability over projecting audible sources, relying on the spatial resolution of its ultrasonic transmitter. Sharper ultrasonic beams result in a finer and more manipulable reproduced source. We validate this manipulability by analyzing spatial resolution in Sec. 4.1, and by studying energy distribution with varying spatial resolution in Sec. 4.2. Furthermore, Sec. 4.3 explores META-SPEAKER's frequency response, impacting the complexity of audible source reproduction.

## 4.1 Spatial Resolution

We study the spatial resolution of the ultrasonic array by analyzing its beampattern. Several factors determine the beampattern of an array, including wave frequency and array geometry. Shorter wavelengths typically result in narrower beampatterns. Ultrasonic waves inherently exhibit directionality due to their high frequency. Additionally, an array of transducers can further improve the directionality. As a rule of thumb, the relation between the 3-dB beamwidth and the number of parallel transducers $M$ is given by [23] 3-dB beamwidth $\propto 2 \cdot \sin^{-1}\left(\frac{1}{M}\right)$. This generally indicates




(e) Deployment. (b) 40 kHz (16×8 array). (c) 42 kHz (16×8 array). (d) 44 kHz (16×8 array).

(e) 40 kHz (8×8 array). (f) 40 kHz (4×8 array). (g) 40 kHz (2×8 array). (h) 40 kHz (1×8 array).

**Figure 6: Impact of frequency and array geometry on the beam pattern.**

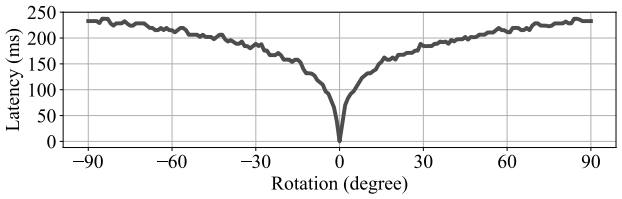



**Figure 7: Rotation angle v.s. Latency.**

that by varying the number of transducers, we can control the beamwidth and thus manipulate the spatial resolution.

**Observation 1.** *The transmitter can form a sharp beam, allowing it to pinpoint a direction precisely.*

To obtain the beampattern of the ultrasonic transmitter, we deploy a sound sensor at a distance of 1 m from the transmitter, as shown in Fig. 6(a). The transmitter continuously emits ultrasound while a servo motor rotates the ultrasonic array from -90 to 90 degree (The 0 degree is where the array right faces the sound sensor). At each angle, the motor temporarily stops, and we capture a 5-second sound recording. We calculated the SNRs for all angles to obtain the beampattern of the transmitter.

Fig. 6(b), (c), and (d) depict the beam patterns of ultrasound at different frequencies with the 16×8 array geometry (column×row). The 3-dB beamwidths for the 40 kHz, 42 kHz, and 44 kHz ultrasound frequencies are found to be only 3.05, 2.68, and 2.43 degrees, respectively. This ability supports Meta-Speaker reproducing the source at a fine granularity.

Additionally, We observe that the 40 kHz beam exhibits the highest gain. This is due to the fact that the resonant frequency of the transducer [25] is 40 kHz, which is its most efficient operating frequency. It is important to note that transducers with varying sensitivity to different frequencies can significantly impact the bandwidth of the generated audible source. This phenomenon will be explored in Sec. 4.3.

Besides the main lobe, the beam pattern also shows a few grating lobes, as marked in Fig. 6(b). These grating lobes are caused by spatial aliasing: Due to the packaging problem of our transducers; the 10 mm spacing is the minimum spacing we can set, which exceeds the half wavelength of the ultrasound (around 4.2 mm). We will discuss the impact of grating and side lobes further in Sec. 4.2.

**Observation 2.** *The transmitter offers flexible manipulability over its spatial resolution by enabling transducers selectively.*

In our design, Meta-Speaker can adjust its array geometry by selectively activating individual columns of transducers. Fig. 6(b), (e), (f), (g), and (h) compare the beam patterns for array geometries of 16×8, 8×8, 4×8, 2×8, and 1×8, respectively, at a frequency of 40 kHz. As the number of active columns of the array increases from 1 to 16, the resulting 3-dB beamwidth decreases from 24.97 to 3.05 degrees.

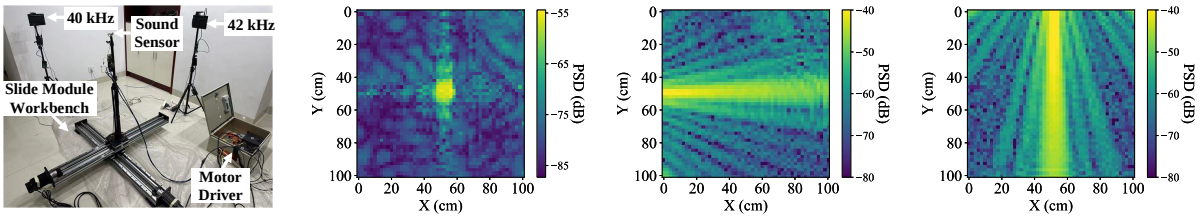
Apparently, the location of the reproduced source can be manipulated by steering the orientation(s) of the array(s).

**Observation 3.** *Meta-Speaker can quickly steer the beam orientation, allowing it to project multiple sources at different locations in a time-division manner.*

We assess the latency of the ultrasonic transmitter in steering the beam toward a new direction. Fig. 7 shows the steering latency (i.e., rotation duration) as a function of the rotation angle. Notably, the latency function is symmetric about the y-axis, indicating that the rotation duration depends only on the absolute rotation angle, not on the rotation direction. The maximum steering latency is less than 250 ms, meaning that Meta-Speaker can project a new source at a new location within 250 ms, given the fact that transmitters A and B can rotate simultaneously.

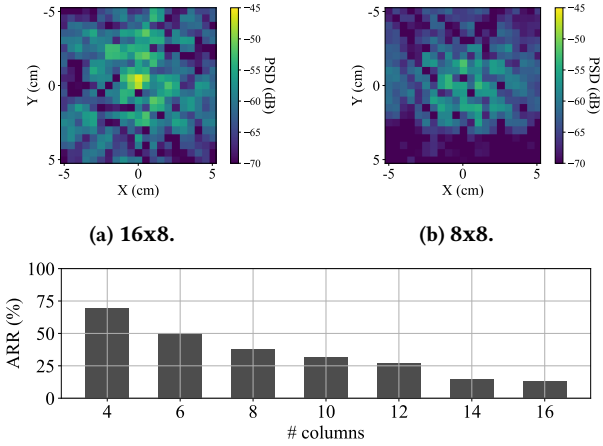
## 4.2 Energy Distribution

Next, we visualize the energy distribution of the reproduced source in space, from which we try to understand what the reproduced source looks like and how it propagates. We will also examine how the beampattern impacts the energy distribution of the reproduced source.




(a) Deployment. (b) 2 kHz. (c) 40 kHz. (d) 42 kHz.

**Figure 8: (a) A two-dimensional slide module workbench is used to carry the microphone to measure the signal energy across a 1m x 1m area with a grid size of 20 mm. Then, we generate the heatmap of signal energy of (b) the 2 kHz reproduced audio, (c) the 40 kHz ultrasonic beam, and (d) the 42 kHz ultrasonic beam.**




(a) 16x8. (b) 8x8.

(c) ARR as a function of array column

**Figure 9: The impact of array column on audible region size.**

The audible source in Meta-Speaker is reproduced at the area where ultrasonic beams intersect. We can expect that sharper ultrasonic beams result in a smaller intersection area, leading to a finer reproduced source. As demonstrated in Sec. 4.1, the transmitters implemented in Meta-Speaker deliver a high spatial resolution whereby the size of the reproduced source can be as fine as a single point. Moreover, the transmitter can selectively enable or disable the columns of the array and adjust the spatial resolution, providing additional manipulability over the size of the reproduced source.

**Observation 4.** *Meta-Speaker can reproduce a point-wise audible source at the intersection of two ultrasonic beams.*

To validate this, transmitters A and B are positioned according to the setup shown in Fig. 8(a). Transmitter A emits a 40 kHz tone, while transmitter B emits a 42 kHz tone. The array geometry adopted is 8×8. To precisely move the microphone, we employ a two-dimensional sliding module workbench with a slide stroke of 100 cm × 100 cm and a slide precision of 0.05 mm. The microphone is mounted on the workbench and moves across the entire operating range with a grid size of 2 cm × 2 cm. At each grid, a 3-second recording of sound is made, and the signal powers at frequencies of 2 kHz, 40 kHz, and 42 kHz are measured via PSD. The signal energy distribution for the reproduced audio at 2 kHz, and ultrasounds at 40 kHz and 42 kHz, are presented in Fig. 8(b), (c), and (d), respectively. The observations are as follows.

As seen in Fig. 8(b), a small, high-energy region is present at the center. We can further check that this region is exactly the intersection of two ultrasonic beams visualized by Fig. 8(c) and (d). The 3-dB reduction in energy of the reproduced audio occurs at a distance of approximately 2.19 cm from the highest energy point (The grating lobes in Fig. 6 are not shown here because the grating lobes are out of the moving area of the workbench).

Although a small amount of ultrasonic energy can leak along the ultrasonic beams, as shown in Fig. 8(c) and (d), the side lobes attenuate quickly as their index increases. When a grating or side lobe of one beam intersects with the main lobe of the other, this can result in a faint audible signal being produced (see Fig. 8(b)). However, the leakage is approximately 15-20 dB weaker than the audio produced by the two main lobes. Our evaluations show that the further the leakage is from the main lobe intersection, the weaker it becomes. Thus, it does not significantly impact the capability of projecting audible sources with fine-grained resolution.

**Observation 5.** *The reproduced audible signal follows the Huygens–Fresnel principle, and behaves as a new source of wavelet that spreads in all directions.*

Fig. 8(b) shows that, apart from the point where the ultrasonic beams intersect, we can detect a certain level of energy at 2 kHz signal in other areas. For instance, in areas that are 30 cm away from the intersection, the energy of the 2 kHz signal can remain 5-11 dB higher than the background noise.

To better understand the propagation characteristics of the reproduced sources, we should compare the constructive or destructive combination of reproduced sources generated from different regions of the air in the cases of directional speakers and Meta-Speaker:

In directional speakers, where the ultrasonic beams are parallel, the air molecules vibrate in phase along the direction of the beams. Specifically, two ultrasonic beams will create a series of Virtual Array Elements (VAEs). Each VAE

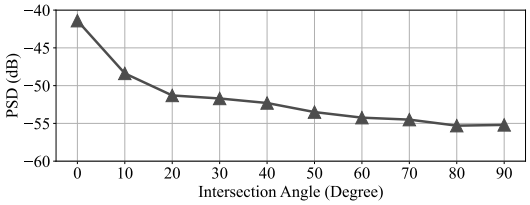



**Figure 10: Intersection angle vs. sound volume.**

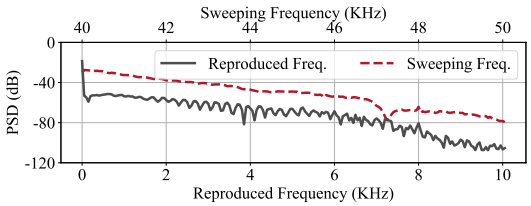



**Figure 11: Reproduced frequency vs. sweeping frequency.**

functions as a virtual source playing the reproduced source signal. Along the direction of the ultrasonic beams, the intrinsic time difference among VAEs precisely accounts for the required time shifts to ensure the constructive combination of the reproduced sources. In essence, the directional speaker behaves similarly to an *end-fire array*, which can focus the reproduced signal along the beam direction.

In Meta-Speaker, where the ultrasonic beams are separated and intersected, the reproduced sources played by different VAEs cannot be constructively combined. This is because the time differences among VAEs are out of order, making it impossible to consistently compensate for the time shifts in any particular direction. As a result, the reproduced source is not projected along a specific direction, but instead spreads in all directions.

We must acknowledge that, due to the absence of constructive combination, the power of the reproduced source in Meta-Speaker is comparatively weaker than that of a directional speaker. To quantify this, we conducted an experiment where we rotated the intersection angle of two ultrasonic beams, ranging from 0 degree (with the two beams parallel) to 90 degree (with the two beams perpendicular to each other). Fig. 10 shows that the volume of the reproduced sound (at 2 kHz) decreases as the intersection angle increases. For instance, when the angle reaches 90 degree, the volume is 16.8 dB lower than when the angle is 0 degree.

**Observation 6.** *The size of the audible region can be manipulated by adjusting the beamwidth of the transmitter.*

In order to investigate the relationship between the size of the audible region and the number of enabled columns of the transducer array, we conduct experiments and measure the energy distribution across a 10 cm × 10 cm area centered around the reproduced source, with a grid size of 0.5 cm × 0.5 cm. As depicted in Fig. 9(a) and (b), which show the energy distribution of the reproduced signal (2 kHz) when the array geometries are 16×8 and 8×8, respectively, the audible region size does change as the number of columns increases.

To quantitatively measure the audible region size, we define a new metric called audible region ratio (ARR). ARR is the ratio of the number of grids where the reproduced signal power is greater than $E_{\max}$ - 9 dB to the total number of measured grids, where $E_{\max}$ is the maximum signal power of all measured grids. Our results, presented in Fig. 9(c), indicate that as the number of columns increases, the beam becomes sharper and the audible region becomes finer, as evidenced by the decreasing ARR value. Therefore, we conclude that by tuning the beamwidth of the transmitter, the size of the audible region can be effectively manipulated.

## 4.3 Frequency Response

The bandwidth of the audible source reproduced by Meta-Speaker is mainly determined by the bandwidth of the ultrasonic transducers used. When the frequency of the transmitted ultrasound does not match the resonant frequency of the transducer (i.e., 40 kHz), the transmitter is less effectively excited and emits less ultrasonic energy. As a result, the power of the reproduced audible source is dampened.

**Observation 7.** *The signal power of the reproduced source decreases as the ultrasonic frequency increases.*

The frequency response curves in Fig. 11 clearly validate the observation. To conduct the validation, we keep transmitter A emitting a continuous tone of $f_A$ = 40 kHz, while we sweep the frequency of transmitter B, $f_B$, from 40 kHz to 50 kHz with a step size of 50 Hz. According to Eq. (4), the frequency of the reproduced audio is given by $f_B - f_A$ and is expected to range from 0 Hz to 10 kHz. In Fig. 11, we can see that the signal powers of both the ultrasound and the corresponding reproduced audios decrease as $f_B$ increases.

In addition, from the frequency response curve in Fig. 11, we can estimate the bandwidth of Meta-Speaker to be approximately 3.8 kHz, as the frequency with a flat response lies between the range of 200 Hz to 4 kHz. If the sweep frequency goes beyond 44 kHz, the signal power of both ultrasound and the reproduced audio decreases rapidly. This indicates that the ultrasound frequency should be less than 44 kHz. On the other hand, the frequency of the reproduced audio should be greater than 200 Hz to avoid the DC component.

In summary, the previously conducted profiling confirms that Meta-Speaker offers a high level of manipulability in both the size and location of the reproduced source.

The subsequent sections will present three distinct applications that exploit these unique capabilities: (1) Sec. 5 demonstrates how the ability to time-divisionally project multiple sources can be utilized to create multiple anchors for localization. (2) Sec. 6 shows how to transmit acoustic messages discreetly to a target device while remaining unheard to other devices by manipulating the source location.

(3) Sec. 7 showcases how Meta-Speaker can play real spatial audio to interact with humans since humans can hear the reproduced source spatially.

## 5 ANCHOR-FREE LOCALIZATION

Traditional acoustic localization systems typically rely on multiple distributed anchors broadcasting beacons to localize devices [16, 22, 29, 38]. However, with the unique capability of projecting sources in different locations, Meta-Speaker introduces the concept of *virtual anchors* for localization.

The advantage of virtual anchors lies in their flexibility to be projected to any desired location. In contrast, using physical anchors can lead to fluctuating localization performance and accuracy, depending on the target location relative to the anchor positions. With the ability to project virtual anchors anywhere, a target can be localized with the help of nearby virtual anchors, thus enhancing localization accuracy.

### 5.1 Design Issues

We first explain the estimation of distance difference of arrival (DDoA), from which we localize devices by trilateration.

**DDoA Estimation.** Suppose two ultrasonic transmitters, A and B, are deployed as shown in Fig. 12(a). By steering the orientation of transmitter A, we can project two distributed audible sources, thus acting as two virtual anchors time-divisionally (denoted as VA1 and VA2, respectively). Each virtual anchor will broadcast beacons.

The beacons transmitted by VA1 and VA2 may propagate through different distances, $d_1$ and $d_2$, before arriving at the microphone. Specifically, Fig. 12(b) demonstrates the timelines of signal transmission and reception, as well as motor steering. The DDoA between VA1 and VA2 is calculated as

$$\Delta d_{<1,2>} = d_2 - d_1 = \frac{t_4 - t_3}{c} - \frac{t_2 - t_1}{c} = \frac{t_4 - t_2}{c} - \underbrace{\frac{t_3 - t_1}{c}}_{\text{constant}}, \quad (9)$$

where $c$ denotes the sound speed. The term $\frac{t_3 - t_1}{c}$ in Eq. (9) is a constant as long as we fix the time interval between two transmitted beacons (i.e., $t_3 - t_1$). This means that the microphone device can directly estimate the DDoA $\Delta d$ from the time-difference-of-arrival (TDoA) of beacons (i.e., $t_4 - t_2$) without knowing the transmission timing of each beacon. Furthermore, with the far-field assumption [40, 41], the direction-of-arrival (DoA) can also be estimated as $\theta = \arccos(\frac{\Delta d}{d})$, where $d$ is the inter-distance between the virtual anchors.

Note that achieving time synchronization among virtual anchors with Meta-Speaker is an easy task. To strictly fix the time interval between beacon 1 and 2 (i.e., $t_3 - t_1$), we insert $(t_3 - t_1 - T_{beacon}) \cdot f_s$ zero samples between beacon samples that are streamed to transmitter A's DAC, where $f_s$ is the sampling rate, $T_{beacon}$ is the beacon duration. Since

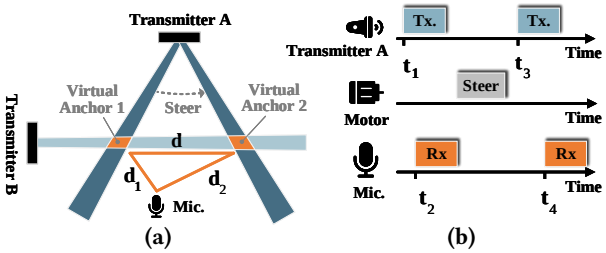



**Figure 12: Illustration of (a) virtual anchors and (b) the timelines of beacon transmission and reception.**

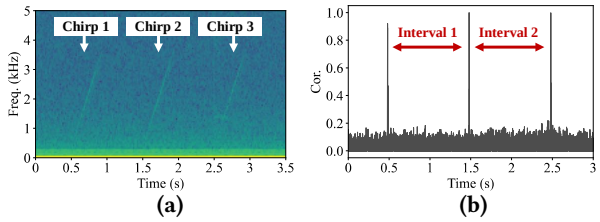



**Figure 13: (a) Spectrum of the recorded audio. (b) Using cross-correlation to detect beacons.**

the ADC generates samples following clock ticks strictly, we can maintain a precise time interval between beacons, preventing fluctuation of the constant term $\frac{t_3 - t_1}{c}$ in Eq. (9).

Meanwhile, to realize the synchronization between the beacon transmission and motor steering, we should ensure that the steering of the transducer array starts after VA1 finishes transmitting the beacon, and ends before VA2 begins transmitting the beacon. A gap of 500 ms is set between beacons 1 and 2 to offer sufficient time for motor steering, as the steering latency is no more than 250 ms (see Sec. 4.1).

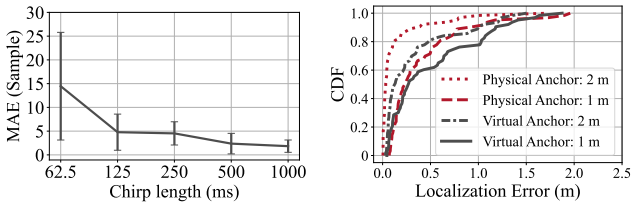
**Trilateration.** Similarly, transmitter B can also steer the beams, creating another virtual anchor (denoted as VA3). Therefore, another DDoA $\Delta d_{<2,3>}$ between VA2 and VA3 can thus be estimated. With the trilateration method, the device's coordinates can thus be solved based on two estimated DDoAs. Formally, we define the coordinates of VA 1, 2, and 3 as $V_1 = (-d/2, d/2)$, $V_2 = (d/2, d/2)$, and $V_3 = (d/2, -d/2)$, respectively. The coordinate of the device $P$ can be calculated by solving the following equations [17, 42]

$$\begin{cases} |PV_1 - PV_2|^2 = \text{abs}(\Delta d_{<1,2>}) \\ |PV_2 - PV_3|^2 = \text{abs}(\Delta d_{<2,3>}) \end{cases}, \quad (10)$$

where $|\bullet|^2$ is the Euclidean distance, $P = (P_x, P_y)$ is the device coordinates, and the signs of $P_x$ and $P_y$ depends on the signs of $\Delta d_{<1,2>}$ and $\Delta d_{<2,3>}$.

### 5.2 Validations

**Setups.** For beacon transmission, we adopt a chirp signal sweeping from 500 Hz to 4000 Hz as the beacon The default chirp length is 500 ms. We generate ultrasonic sounds for




**(a) TDoA error.** **(b) Localization error.**

**Figure 14: Localization results.**

transmitters by following Eq. (12) and Eq. (13). For beacon reception, a commercial microphone, Seeed Stuido Respeaker [35], is adopted to record the signals at a sampling rate of 48 kHz. Fig. 13(a) demonstrates the spectrum of the recorded audio, from which we can observe the chirps transmitted by VA 1, 2, and 3. To detect these chirps, we first use a band-pass filter to suppress frequencies below 200 Hz or above 4000 Hz. Subsequently, with the idea of the matched filter, we take the audible chirp as the template and correlate it with the received samples to compute the cross-correlation function (CCF). As demonstrated by Fig. 13(b), we can observe distinctive correlation peaks at the locations of beacons in CCF, from which we can estimate DDoAs according to Eq. (9) and then localize the microphone using Eq. (10).

**Methodologies.** We evaluate the localization performance in a 4 m × 4 m room. Transmitters A and B are separately placed next to two adjacent walls. Three virtual anchors are projected in the center of this room. We place the commercial microphone at different locations. We vary the beacon length and the inter-distance of VAs ($d$) to assess their impacts on localization performance. For comparison, we place three loudspeakers at the same locations as the virtual anchors to act as physical anchors.

**Results.** Fig. 14(a) shows the impact of chirp length on time difference of arrival (TDoA) estimation (We calculate DDoA from TDoA). With the increase in chirp length, the correlation peaks tend to be stronger and sharper, allowing the receiver to more accurately estimate the arrival time of each beacon. We vary the chirp length from 62.5 ms to 1000 ms to evaluate the mean absolute error (MAE) of TDoA. During experiments, the inter-distance of VAs is set to 1 m. As expected, we observe that the TDoA error decreases as the chirp length increases. When the symbol length increases to 500 ms, the TDoA error decreases to 2.37 samples (about 0.049 ms at a 48 kHz sampling rate), which is reasonable.

Fig. 14(b) compare the localization performances of virtual anchors reproduced by Meta-Speaker and physical anchors projected by loudspeakers. The chirp length is fixed to 500 ms. The effective aperture sizes of both the virtual and physical anchors increase as the inter-distance is increased, resulting in more accurate localization results [10]. As expected, increasing the inter-distance leads to more accurate localization results, with the median errors decreasing from 0.27 m to 0.13 m for the virtual anchors and from 0.23 m to 0.03 m for the physical anchors. While the physical anchors outperformed Meta-Speaker slightly, the latter has the advantage of flexibility in adjusting the inter-distance of virtual anchors at runtime. This means that Meta-Speaker can further improve the performance by manipulating the locations of virtual anchors.

## 6 LOCATION-AWARE COMMUNICATION

Meta-Speaker enables location-aware communication in a spatial-selective manner. That is, we can send acoustic messages only to devices in a specific area, while devices outside of that area can hardly receive these messages. This capability can physically establish secure communication links for low-end devices which are vulnerable to eavesdropping.

### 6.1 Design Issues

In the following, we demonstrate a toy example of transmitting information using the Frequency Shift Keying (FSK) method and then explain how to enhance transmission security through carrier frequency hopping.

**FSK.** The idea is to encode data bits by shifting the frequency of the reproduced audio $v(t)$. Suppose there are $N$ data bits to be encoded, The frequency of the reproduced audio $v(t)$ for FSK symbol $s$ can be represented as

$$f_v^{(s)} = K \cdot \Delta f + f_{\min}, \tag{11}$$

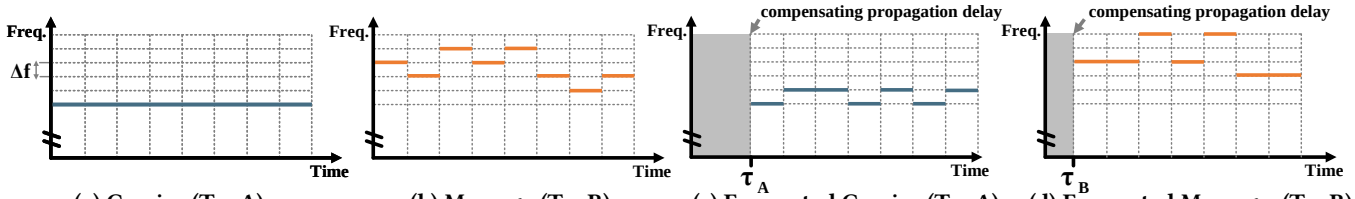
where K is an integral value the range of $[0, 2^N - 1]$, $\Delta f$ is the frequency shift resolution, and $f_{\min}$ is the minimum frequency for suppressing the DC component (refer to Sec. 4.3).

As illustrated by Fig. 15(a) and (b), we can transmit the audible source $v(t)$ by allowing transmitter A to play a carrier tone with a frequency of $f_c$ and allowing transmitter B to play the single-sideband modulation of $v(t)$ based on Eq. (8). Furthermore, by steering these two beams and having them meet at the target device's location, we can transmit the encoded audio to that device spatial-selectively.

**Encryption.** However, this design is still vulnerable to eavesdropping attacks. For instance, an attacker could record the modulated ultrasounds (Fig. 15(B)) by directly deploying an ultrasonic microphone right in front of transmitter B.

To address this vulnerability, we propose encrypting the frequency shift using the location of the target device. This method ensures that the modulated frequency shifts can only be received by the device located in the target area, while other devices receive meaningless frequency shifts.

Inspired by the basic idea of frequency-hopping spread spectrum (FHSS), we rapidly change the carrier frequency among $Z$ distinct frequencies for each symbol $s$. Specifically,




(a) Carrier (Tx. A). (b) Message (Tx. B). (c) Encrypted Carrier (Tx. A). (d) Encrypted Message. (Tx. B)

**Figure 15: Illustrative spectrums of (a) message and (b) carrier for FSK, and (c) message and (d) carrier for encrypted FSK.**

the carrier frequency for symbol $s$ can be denoted as

$$f_A^{(s)} = f_c + z^{(s)} \cdot \Delta f, \tag{12}$$

where $z$ denotes a discrete random variable that takes values from 0 to $Z-1$ with equal probability. Fig. 15(c) shows the carrier frequencies when $Z = 2$. The frequency of the modulated ultrasounds can be described as

$$f_B^{(s)} = f_A + f_v^{(s)} = f_c + z^{(s)} \cdot \Delta f + K \cdot \Delta f + f_{\min} \tag{13}$$

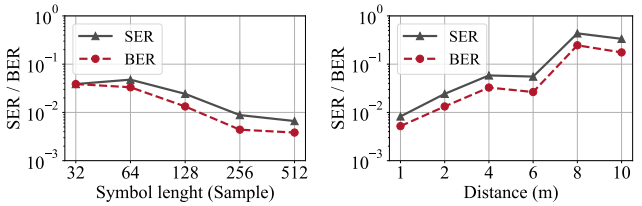
Because transmitters A and B are located in different places, the ultrasounds transmitted by them will experience different propagation delays before arriving at the target device. This generally requires that transmitters A and B should compensate for these propagation delays, as shown in Fig. 15(c) and (d). This design ensures that each symbol,i.e., frequency shift, can be extracted correctly by down-conversion with the appropriate carrier frequency.

Our design encrypts the symbols implicitly. This is because each symbol, or frequency shift, is jointly determined by the ultrasounds transmitted by transmitters A and B. Therefore, an eavesdropper would need to record both ultrasounds with two ultrasonic microphones to be able to intercept the transmission, which increases the risk of being detected. Additionally, the attacker may not be able to extract the correct frequency shifts without knowing the target device's location, since they cannot deduce the compensation delays.

## 6.2 Validations

**Setups.** We conducted experiments with a communication bandwidth of 2.56 kHz, which spans from 1.44 kHz to 4 kHz. The sampling rate $f_s^{\text{mod}}$ for modulation is 40.96 kHz. We provide a range of FSK symbol length (denoted as $L_{\text{sym}}$) options, 32, 64, 128, 256, and 512 samples, corresponding to data rates of 1.28, 1.28, 0.96, 0.64, and 0.4 kbps.[1] The gray coding is used for encoding FSK symbols. These modulated signals will be further resampled to 96 kHz to match the sample rate of the DAC in the ultrasonic transmitter. The reference carrier frequency $f_c$ is 40 kHz. Following Eq. (12) and Eq. (13), we generate ultrasonic samples played by transmitters A and B. To ensure that two transmitters are tightly synchronized, their controllers (i.e., Raspberry Pi) are connected via GPIO.

[1]Let us take the symbol length of 512 samples as an example to calculate the data rate. Since we use FFT size with the same size as the symbol length, the frequency resolution $\Delta f = f_s^{\text{mod}}/L_{\text{sym}} = 80$ Hz. Each FSK symbol contains $\log_2 B/\Delta f = 5$ bits. The data rate is given by $f_s^{\text{mod}}/L_{\text{sym}} \times 5 = 400$ bps.



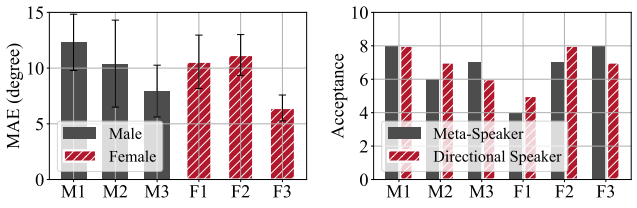

(a) Symbol length (b) Projection distance

**Figure 16: Impact of (a) symbol length and (b) projection distance on SER and BER.**

For the receiver, we use a commercial microphone, Seeed Stuido Respeaker [35], to record the signals at a sampling rate of 48 kHz. The signals are processed offline using Python. To decode FSK symbols, the receiver first shifts the signal's frequency by -1.44 kHz. It then employs a low-pass filter to extract the signal below 2.56 kHz. Subsequently, it performs FFT and peak detection to identify FSK symbols, from which it decodes the data bits.

**Results.** Fig. 16(a) shows the impact of symbol length on the symbol error rate (SER) and bit error rate (BER). The distances from the microphone to transmitters A and B are fixed at 2 m each. The microphone is deployed at the intersection of two ultrasonic beams. The symbol length varies from 32 to 512 samples. For symbol lengths of 32 or 64 samples, the SER and BER are approximately 3.50%, which is a reasonable performance. The robustness can be further improved by increasing the symbol length. As shown in the figure, when the symbol length is increased to 256 samples, the SER and BER drop to only 0.67 % and 0.38 %, respectively.

Fig. 16(b) shows the impact of audio projection distance. We fix the distance between the microphone and transmitter A to 2 m, while varying the distance between the microphone and transmitter B from 1 m to 10 m. The microphone is kept at the intersection of the beams, and the symbol length is fixed at 128 samples. As expected, the performance deteriorates with the increase of the projection distance. When the projection distance is no more than 6 m, we can achieve reasonable performance. However, when the distance is larger




**(a) Direction Est. Accuracy.** **(b) Acceptance Score.**

**Figure 17: Evaluations of Acoustic AR.**

than 8 m, the performance becomes undesirable. For example, as the projection distance reaches 10 m, the SER and BER respectively increase to 33.50 % and 17.53 %, respectively.

What's more, when the microphone is deployed 1 m away from the intersection of the ultrasonic beams, We can only detect almost meaningless symbols. According to our measurements, the SERs are between 84.3 % and 93.6 % as the symbol length is 128 samples, which validates that Meta-Speaker can physically achieve a spatial-selective transmission.

## 7 ACOUSTIC AUGMENTED REALITY

The reproduced sound is a physical presence in space, allowing not only microphones to pick it up, but also humans to hear it and perceive its location. This offers a novel way to interact with humans.

### 7.1 Design Issues

Here, we focus on the task of playing spatial audio (i.e., auditory cues) with Meta-Speaker to provide guidance and navigation services for people, especially for the elderly or the visually impaired. In contrast to spatial audio delivered through earphones, which involves complex computations of head-related transfer functions (HRTF) to deceive human hearing [48, 49], Meta-Speaker offers a more straightforward approach to creating real spatial audio through the projection of point-wise sources. The design is simple: to direct the user's attention towards a particular direction, Meta-Speaker plays auditory cues from that direction.

### 7.2 Validations

The IRB approval is obtained for the following experiments.

**Setups.** We invite 6 volunteers (3 males and 3 females) to join our experiments. During experiments, each volunteer is required to close their eyes, and sits 3 m away from transmitters A and B. We play the auditory cues from the volunteer's one directions[2], and ask he/she to pin the source direction with fingers. To address the issue of front-back ambiguity in spatial hearing [46], each volunteer is given the opportunity to reduce ambiguity by rotating their head and body to reorient themselves and hear the cue again (up to 3 repetitions). We then measure the direction error by comparing the direction estimated by the volunteer and the ground truth in the horizontal dimension. We conduct 5 trials with random directions ranging from -90 to 90 degrees for each volunteer. Subjective evaluations are also conducted to compare Meta-Speaker and a commercial directional speaker (W-SPEAKER WS-V2.0) in terms of sound quality. The volunteers rate the acceptance of sound on a scale of 0-10 (10 being perfect).

**Results.** Fig. 17(a) displays the mean absolute error of the direction estimated by each volunteer. All of the errors are below 15 degrees, with an average error of 9.8 degrees. This demonstrates that Meta-Speaker has the ability to produce sound sources that can be spatially perceived by humans with reasonable accuracy, given the psychophysical fact that humans can perceive the direction of sounds with an accuracy of about 5 to 11.8 degrees [3]. Additionally, all volunteers feedback that they hear and distinguish the cues crystally. Also, Fig. 17(b) shows that the average acceptance scores of Meta-Speaker and the the directional speaker are 6.7 and 7.0 respectively, indicating that Meta-Speaker achieves a comparable acceptance as that of the directional speaker.

## 8 RELATED WORK

**Acoustic Field Manipulation.** The idea of AFM is to manipulate the spatial distribution of mechanical energy propagating in various media. AFM can be accomplished through multiple approaches, including source projection [7, 13, 18, 36, 44] and wave propagation [5, 8, 12, 21, 24, 34, 43]. Meta-Speaker offers a revolutionary approach for AFM. The unique advantage of Meta-Speaker lies in its capability to physically project fine-grained audible sources with precise location manipulation. This sets it apart from traditional methods and has the potential to fundamentally change the way of manipulating the acoustic field.

**Acoustic Nonlinearity.** The field of acoustics has a vast body of research focused on exploring nonlinearity [47], in terms of the reception sensors (e.g., microphones), or the medium through which the sound wave travels (e.g., air). (1) Recent works [4, 30, 33, 39, 53] demonstrate the feasibility of sending inaudible voice commands to voice-enabled devices by exploiting the hardware nonlinearity in microphones. The microphone nonlinearity can also be used for improving sensing granularity [6]. (2) Meanwhile, the air is also nonlinear [33]. Based on the KZK equations [19, 52], which describe the self-distortion of acoustic signals as they propagate through a medium [18, 37, 44], we have the opportunity to reproduce audible sources from ultrasounds. This gives birth to directional speakers [18, 44] that can project audible sources along a narrow line (or beam). Meta-Speaker

[2]The height of the projected auditory cue is set 0.3 m above the volunteer's head to avoid any obstruction from their body, and the distance between the projected sound and the volunteer is kept at 0.5 m.

takes it a step further in the sense that it can project audible sources with high manipulability, down to a point.

**Acoustic Location-aware Communication.** Recent work SpotSound [14] also achieves spatial-selective transmission, in which a message can be encoded to ensure that only the target receiver can decode the valid message. The difference between Meta-Speaker and SpotSound is that SpotSound accomplishes spatial-selective transmission logically (i.e., signal precoding), while Meta-Speaker does so physically (i.e., manipulable source projection).

## 9 DISCUSSION AND FUTURE WORK

**Safety Concerns.** A common concern is that ultrasounds may create heat as they propagate through human tissues. Fortunately, thanks to the high impedance mismatch between the air and the human skin, about 99.9% of the airborne energy will be reflected by the skin, rather than absorbed by the tissues [45]. Furthermore, there is no evidence to suggest that ultrasound around 40 kHz and below 120 dB SPL (or 135 dB SPL) has adverse effects on human hearing, such as temporary threshold shift [11, 20].

**Grating Lobes.** The beampattern of our system contains undesirable grating lobes, as discussed in Sec. 4.1. The cause of this issue is the minimum inter-sensor distance of 10 mm imposed by our ultrasonic transducers (Murata MA40S4S), which is larger than the half-wavelength. We have to clarify that the source reproduced by the side lobes will be extremely weak, even if exists. The reason is two-folded: First, the signal strength of the side lobes is significantly lower than that of the main lobe. The difference is typically from 10 dB to 20 dB. Second, due to the low volume, the side lobes contribute much less to the air non-linearity, as explained above. Nevertheless, it is possible to use smaller transducers, such as Murata MA40H1S-R (measuring 5.2 mm $\times$ 5.2 mm), to mitigate the issue of spatial aliasing.

**Background Sounds.** Background sounds have negligible impact on air non-linearity. The propagation of sound induces vibrations in the air molecules, leading to an uneven distribution of the air. A higher sound volume results in larger vibration amplitudes of the air molecules, consequently amplifying the non-linearity. For notable impact on non-linearity, the sound volume needs to exceed 90 SPL, which is far beyond the typical volume of background sounds.

**Limited Bandwidth.** The bandwidth of Meta-Speaker, approximately 3.8 kHz, is restricted by the bandwidth of the ultrasonic transducers utilized. It is feasible to combine transducers with their resonant frequencies spanning a wide frequency range to expand the bandwidth.

**Orthogonal Frequency Division Multiplexing (OFDM).** In the context of location-aware communication, we opt for FSK as the modulation method due to its simplicity of use and robust performance. However, it is worth noting that OFDM modulation presents a viable alternative that might further enhance communication throughput.

**Concurrent Multi-sources Projection.** Currently, Meta-Speaker is limited to projecting only one audible source at a time. However, it is possible to explore multi-beam techniques or take advantage of multipath effects to concurrently project multiple sources. We leave it for future work.

**Digital Phased Array.** In order to reduce steering latency, it is possible to steer the ultrasonic beam digitally using digital phase shifts instead of a mechanical motor. However, this approach comes with a trade-off. As the steering angle increases, the digital phased array experiences a decrease in array gain and an increase in beam width.

**Non-Line-of-Sight (NLOS).** Meta-Speaker is vulnerable to signal blockage. The non-linear behavior of air, which is crucial for Meta-Speaker's functionality, becomes significant only when the ultrasonic wave reaches a certain volume level. If the ultrasonic wave is blocked, the reproduction of the source will not occur as a substantial portion of the ultrasonic signal strength is lost due to blockage reflection.

**Multipath.** The impact of multipath propagation (of ultrasonic beams) varies depending on the number of reflections experienced by different paths. Paths that undergo multiple reflections may have a negligible impact due to the substantial loss of signal energy after each reflection. On the other hand, paths experiencing only a single reflection may exhibit noticeable reflections as the majority of the signal strength is preserved [40, 41]. Therefore, careful consideration of the room's geometry during deployment is crucial to mitigate the potential impact of single-reflected paths.

**More Applications.** With Meta-Speaker, a multitude of potential applications in various fields become possible. For instance, it can be utilized for active noise cancellation by projecting an inverted noise wave directly to the noise source. Additionally, it can be employed for acoustic sensing by projecting sensing signals from various angles, providing a more comprehensive and precise sensing solution.

## 10 CONCLUSION

This paper demonstrates the feasibility of reproducing audible sources from ultrasounds in the air by exploiting air nonlinearity. We propose a novel device, Meta-Speaker, that can project audible with high manipulability in spatial granularity. We prototype Meta-Speaker and demonstrate its potential by presenting three illustrative applications.

## ACKNOWLEDGMENTS

We thank our anonymous shepherd and reviewers for their insightful comments. This work is partially supported by the National Science Fund of China under grant No. U21B2007.

## REFERENCES


[1] Leo Leroy Beranek. 2004. *Concert halls and opera houses: music, acoustics, and architecture*. Vol. 2. Springer.

[2] Terence Betlehem and Paul D Teal. 2011. A constrained optimization approach for multi-zone surround sound. In *Proceedings of the 36th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'11)*. IEEE, Prague, Czech Republic.

[3] Blauert and Jens. 1983. *Spatial hearing : The psychophysics of human sound localization*. The MIT Press.

[4] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou. 2016. Hidden voice commands. In *Proceedings of the 25th USENIX Security Symposium*. USENIX, Berkeley, CA, USA.

[5] Huanyang Chen and Che Ting Chan. 2007. Acoustic cloaking in three dimensions using acoustic metamaterials. *Applied Physics Letters* 91, 18 (2007), 183518–183518–3.

[6] Xiangru Chen, Dong Li, Yiran Chen, and Jie Xiong. 2022. Boosting the sensing granularity of acoustic signals by exploiting hardware non-linearity. In *Proceedings of the 21st ACM Workshop on Hot Topics in Networks (HotHets'22)*. ACM, Austin, Texas, USA.

[7] Joung-Woo Choi and Yang-Hann Kim. 2002. Generation of an acoustically bright zone with an illuminated region using multiple sources. *The Journal of the Acoustical Society of America* 111, 4 (2002), 1695–1700.

[8] A. Climente. 2012. Omnidirectional broadband acoustic absorber based on metamaterials. *Applied Physics Letters* 100, 14 (2012), 041106.

[9] David G Crighton. 1979. Model equations of nonlinear acoustics. *Annual Review of Fluid Mechanics* 11, 1 (1979), 11–33.

[10] John C Curlander and Robert N McDonough. 1991. *Synthetic aperture radar*. Vol. 11. Wiley, New York.

[11] Andrew Di Battista. 2019. *The effect of 40 kHz ultrasonic noise exposure on human hearing*. Universitätsbibliothek der RWTH Aachen.

[12] Changlin Ding, Xiaopeng Zhao, Huaijun Chen, Shilong Zhai, and Fangliang Shen. 2015. Reflected wavefronts modulation with acoustic metasurface based on double-split hollow sphere. *Applied Physics A* 120, 2 (2015), 487–493.

[13] Dolby. 2023. 7.1.4 Overhead speaker setup. https://www.dolby.com/about/support/guide/speaker-setup-guides/7.1.4-overhead-speaker-setup-guide/. (2023). Accessed: 2023-01-01.

[14] Tingchao Fan, Huangwei Wu, Meng Jin, Tao Chen, Shangguan Longfei, Xinbing Wang, and Chenghu Zhou. 2023. Towards Spatial Selection Transmission for Low-end IoT devices with SpotSound. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking (MobiCom'23)*. ACM, Madrid, Spain.

[15] Colin N Hansen. 1999. *Understanding active noise cancellation*. CRC Press.

[16] Andy Harter, Andy Hopper, Pete Steggles, Andy Ward, and Paul Webster. 1999. The anatomy of a context-aware application. In *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking (MobiCom'99)*. ACM, Seattle, Washington, USA.

[17] Yuan He, Weiguo Wang, Luca Mottola, Shuai Li, Yimiao Sun, Jinming Li, Hua Jing, Ting Wang, and Yulei Wang. 2023. Acoustic Localization System for Precise Drone Landing. *IEEE Transactions on Mobile Computing* (2023). Early Access.

[18] Holosonics. 2023. Audio Spot Light. https://www.holosonics.com/. (2023). Accessed: 2023-01-01.

[19] VP Kuznetsov. 1971. Equations of nonlinear acoustics. *Soviet Physical Acoustics* 16 (1971), 467–470.

[20] Ben W Lawton. 2001. *Damage to human hearing by airborne sound of very high frequency or ultrasonic frequency*. Health & Safety Executive.

[21] Li, Rui-Qi, Zhu, Xue-Feng, Liang, Bin, Yong, Zou, Xin-Ye, and Cheng. 2011. A broadband acoustic omnidirectional absorber comprising positive-index materials. *Applied Physics Letters* (2011).

[22] Qiongzheng Lin, Zhenlin An, and Lei Yang. 2019. Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices. In *Proceedings of the 25th Annual International Conference on Mobile Computing and Networking (MobiCom'19)*. ACM, Los Cabos, Mexico.

[23] Robert J. Mailloux. 2017. *Phased array antenna handbook*. Artech house.

[24] G. W. Milton, M. Briane, and J. R. Willis. 2010. On cloaking for elasticity and physical equations with a transformation invariant form. *New Journal of Physics* 8, 10 (2010), 248.

[25] Murata. 2023. MA40S4S. https://www.murata.com/en-global/products/productdetail?partno=MA40S4S. (2023). Accessed: 2023-01-01.

[26] Alan V Oppenheim, John R Buck, and Ronald W Schafer. 2001. *Discrete-time signal processing. Vol. 2*. Upper Saddle River, NJ: Prentice Hall.

[27] Alan V Oppenheim, Ehud Weinstein, Kambiz C Zangi, Meir Feder, and Dan Gauger. 1994. Single-sensor active noise cancellation. *IEEE Transactions on Speech and Audio Processing* 2, 2 (1994), 285–290.

[28] Mark Poletti. 2008. An investigation of 2-d multizone surround sound systems. In *Proceedings of the 125th Audio Engineering Society Convention*. Audio Engineering Society, San Francisco,CA, USA.

[29] Nissanka B Priyantha, Anit Chakraborty, and Hari Balakrishnan. 2000. The cricket location-support system. In *Proceedings of the 6th annual international conference on Mobile computing and networking (MobiCom'00)*. ACM, Boston, Massachusetts, USA.

[30] N. Roy, H. Hassanieh, and R. R. Choudhury. 2017. BackDoor: Making Microphones Hear Inaudible Sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'17)*. ACM, Niagara Falls, NY, USA.

[31] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'17)*. ACM, Niagara Falls, NY, USA.

[32] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Inaudible Voice Commands: The {Long-Range} Attack and Defense. In *Proceedings of the 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI'18)*. USENIX, Renton, WA, USA.

[33] N. Roy, S. Shen, H. Hassanieh, and R. R. Choudhury. 2018. Inaudible Voice Commands: The Long-Range Attack and Defense. In *Proceedings of the 14th USENIX Symposium on Networked Systems Design and Implementation (NSDI'18)*. USENIX, Renton, WA, USA.

[34] J. Sanchez-Dehesa. 2008. Approach Proposed for Acoustic Cloaking in Two Dimensions. *Noise Regulation Report: The Nation's Only Independent Noise Control Publication* 9 (2008), 35.

[35] Seeed. 2023. ReSpeaker 6-Mic Circular Array kit. https://wiki.seeedstudio.com/ReSpeaker_6-Mic_Circular_Array_kit_for_Raspberry_Pi/. (2023).

[36] Mincheol Shin, Sung Q Lee, Filippo Maria Fazi, Philip Arthur Nelson, Daesung Kim, Semyun Wang, Kang-Ho Park, and Jeongil Seo. 2010. Maximization of acoustic energy difference between two spaces. *Acoustical Society of America* 128(1) (2010), 121–131.

[37] Soundlazer. 2023. The Open Source, Hackable Parametric Speaker. https://www.kickstarter.com/projects/richardhaberkern/soundlazer. (2023).

[38] Yimiao Sun, Weiguo Wang, Luca Mottola, Ruijin Wang, and Yuan He. 2022. AIM: Acoustic Inertial Measurement for Indoor Drone Localization and Tracking. In *Proceedings of ACM SenSys, Boston, USA, November 6-9, 2022*.

[39] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields. 2015. Cocaine noodles: exploiting the gap between human and machine speech recognition. In *Proceedings of the 9th USENIX Conference on Offensive Technologies*. USENIX, Washington, DC, USA.

[40] Weiguo Wang, Jinming Li, Yuan He, and Yunhao Liu. 2020. Symphony: localizing multiple acoustic sources with a single microphone array. In *Proceedings of ACM SenSys, Virtual Event, Japan, November 16-19, 2020*. 82–94.

[41] Weiguo Wang, Jinming Li, Yuan He, and Yunhao Liu. 2022. Localizing Multiple Acoustic Sources With a Single Microphone Array. *IEEE Transactions on Mobile Computing* (2022). Early Access.

[42] Weiguo Wang, Luca Mottola, Yuan He, Jinming Li, Yimiao Sun, Shuai Li, Hua Jing, and Yulei Wang. 2022. MicNest: Long-Range Instant Acoustic Localization of Drones in Precise Landing. In *Proceedings of ACM SenSys, Boston, USA, November 6-9, 2022*.

[43] Wenqi Wang, Yangbo Xie, Bogdan-Ioan Popa, and Steven A Cummer. 2016. Subwavelength diffractive acoustics and wavefront manipulation with a reflective acoustic metasurface. *Journal of Applied Physics* 120, 19 (2016), 195103.

[44] Peter J Westervelt. 1963. Parametric acoustic array. *The Journal of the acoustical society of America* 35, 4 (1963), 535–537.

[45] CHRISTOPHER WIERNICKI and WILLIAM J KAROLY. 1985. Ultrasound: biological effects and industrial hygiene concerns. *American Industrial Hygiene Association Journal* 46, 9 (1985), 488–496.

[46] Frederic L. Wightman and Doris J. Kistler. 1999. Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America* 105, 5 (1999), 2841–2853.

[47] Chen Yan, Xiaoyu Ji, Kai Wang, Qinhong Jiang, Zizhi Jin, and Wenyuan Xu. 2022. A Survey on Voice Assistant Security: Attacks and Countermeasures. *Computing Surveys* 55, 4 (2022), 1–36.

[48] Zhijian Yang and Romit Roy Choudhury. 2021. Personalizing head related transfer functions for earables. In *Proceedings of the 35th ACM Special Interest Group on Data Communication (SIGCOMM'21)*. ACM, Virtual Event.

[49] Zhijian Yang, Yu-Lin Wei, Sheng Shen, and Romit Roy Choudhury. 2020. Ear-ar: indoor acoustic augmented reality on earphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (MobiCom'20)*. ACM, London, UK.

[50] Masahide Yoneyama, Jun-ichiroh Fujimoto, Yu Kawamo, and Shoichi Sasabe. 1983. The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design. *The Journal of the acoustical society of America* 73, 5 (1983), 1532–1536.

[51] Zhang Yongzhao, Yang Lanqing, Wang Yezhou, Wang Mei, Chen Yichao, Qiu Lili, Liu Yihong, Xue Guangtao, and Yu Jiadi. 2023. Acoustic Lens for Sensing and Communication. In *Proceedings of the 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI'23)*. USENIX, BOSTON, MA, USA.

[52] EA Zabolotskaya. 1969. Quasi-plane waves, in the nonlinear acoustics of confined beams. *Soviet Physical Acoustics* 15 (1969), 35–40.

[53] Guoming Zhang, Chen Yan, Xiaoyu Ji, Taimin Zhang, Tianchen Zhang, and Wenyuan Xu. 2017. DolphinAtack: Inaudible Voice Commands. In *Proceedings of the 24th ACM Conference on Computer and Communications Security*. ACM, Dallas, Texas, USA.