VISAR: Projecting Virtual Sound Spots for Acoustic Augmented Reality Using Air Nonlinearity

JUNTAO ZHOU[†], Shanghai Jiao Tong University, China
YIJIE LI[*][†], Shanghai Jiao Tong University, China
YIDA WANG[†], Shanghai Jiao Tong University, China
DIAN DING[†], Shanghai Jiao Tong University, China
YU LU[†], Shanghai Jiao Tong University, China
YI-CHAO CHEN[*][†], Shanghai Jiao Tong University, China
GUANGTAO XUE[†], Shanghai Jiao Tong University, China

Augmented reality that integrates virtual content in real-world surroundings has attracted lots of concentration as the growing trend of the metaverse. Acoustic augmented reality (AAR) applications have proliferated due to readily available earphones and speakers. AAR can provide omnidirectional engagement through the all-around sense of spatial information. Most existing AAR approaches offer immersive experiences by playing binaural spatial audios according to head-related transfer functions (HRTF). These involve complex modeling and require the user to wear a headphone. Air nonlinearity that can reproduce audible sounds from ultrasound offers opportunities to achieve device-free and omnidirectional sound source projection in AAR. This paper proposes VISAR, a device-free virtual sound spots projection system leveraging air nonlinearity. VISAR achieves simultaneous tracking and sound spot generation while suppressing unintended audio leakages caused by grating lobes and nonlinear effects in mixing lobes through optimization. Considering multi-user scenarios, VISAR also proposed a multi-spot scheduling scheme to mitigate the mutual interference between the spots. Extensive experiments show the tracking error is 7.83cm and the orientation estimation error is 10.06°, respectively, envisioning the considerable potential of VISAR in AAR applications.

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

Additional Key Words and Phrases: Acoustic augmented reality, Sound source projection, Air nonlinearity

ACM Reference Format:
Juntao Zhou, Yijie Li, Yida Wang, Dian Ding, Yu Lu, Yi-Chao Chen, and Guangtao Xue. 2024. VISAR: Projecting Virtual Sound Spots for Acoustic Augmented Reality Using Air Nonlinearity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 147 (September 2024), 30 pages. https://doi.org/10.1145/3678546

[*] Both Yijie Li and Yi-Chao Chen are corresponding authors.
[†] Also with Shanghai Key Laboratory of Trusted Data Circulation and Governance in Web3.

Authors' addresses: Juntao Zhou, Shanghai Jiao Tong University, Shanghai, China, juntaozhou@sjtu.edu.cn; Yijie Li, Shanghai Jiao Tong University, Shanghai, China, yijieli@sjtu.edu.cn; Yida Wang, Shanghai Jiao Tong University, Shanghai, China, yidawang@sjtu.edu.cn; Dian Ding, Shanghai Jiao Tong University, Shanghai, China, dingdian94@sjtu.edu.cn; Yu Lu, Shanghai Jiao Tong University, Shanghai, China, yulu01@sjtu.edu.cn; Yi-Chao Chen, Shanghai Jiao Tong University, Shanghai, China, yichao0319@gmail.com; Guangtao Xue, Shanghai Jiao Tong University, Shanghai, China, gt_xue@sjtu.edu.cn.

ACM 2474-9567/2024/9-ART147
https://doi.org/10.1145/3678546

1 INTRODUCTION

While virtual reality (VR) in the metaverse is engineered to provide an engaging and immersive experience, it cannot interact with existing surroundings. Augmented reality (AR), as a significant supplement to virtual reality applications, in which virtual content is seamlessly integrated with real-world surroundings, has recently attracted lots of attention from academics and companies. One particular aspect is especially for acoustic augmented reality (AAR), which focuses on enhancing real-world environments with virtual soundscapes. These enhancements enrich the user's sensory experience without the need for visually obstructive equipment, allowing for a more natural interaction with the real world.

AAR has been increasingly deployed in recent decades, and developers are exploring practical applications, especially navigation. Let us imagine the following scenario: *Bob is visiting a museum and wants to go to a gallery. A voice emerges from where he should go: "Follow me." Bob does not need to look up on his smartphone or look for the indicator. All he needs to do is just follow the perceived direction of the sound source to reach the gallery.* Fig. 1 shows a virtual sound spot source to be utilized in navigation, offering omnidirectional engagement through 360° sensing of spatial information, which demonstrates the enormous potential for direction guideline and communication with surroundings, especially for visually impaired assistance.

Existing work has made great efforts to offer an enhanced navigational aid through spatial audio cues. An important representative research area is related to the Head Related Transfer Function (HRTF) [18, 26, 70], which is widely adopted to create spatial audio using earphones. By leveraging HRTF, users can obtain sounds carrying spatial information. Most works leverage generalized HRTF [18, 26] and few attempts [70] measure personalized HRTF for more realistic spatial perception. Companies like Microsoft [43] and Apple [2] similarly provide information with synthesized binaural audio, creating effects of 3D sounds. However, involving HRTF is a complicated process that should consider personalities, and almost all the work requires the aid of headphones. Ideally, AAR without device assistance will bring users more immersive experiences as well as reduce additional device connections.

Opportunities that offer AAR applications without carrying any equipment come from the sound source projection by leveraging **air nonlinearity** [49]. Air nonlinearity refers to the phenomenon that differential low-frequency component is automatically demodulated from multiple high-frequency components during the propagation of high-energy sound waves in the air. Thus, the audible sound will be reproduced from the ultrasound, which becomes the core of parametric arrays and directional speakers. Highly directional loudspeakers such as SoundLazer [33], HoloSonics [25], and Focusonics [12], composed of an ultrasound transducer array, all based on air nonlinearity to achieve sound projection along with specific direction. However, they all project **beam-shaped** sound source, meaning the sound appears through the entire transmission path, which prevents them from producing omnidirectional auditory enhancements. Meta-Speaker [63] and Audio Hotspot Attack [29]explore the feasibility of using air nonlinearity to generate a sound spot or an audible area, but they did not consider practical issues like additional interference caused by grating lobes. Moreover, the demand for providing AAR for multiple users simultaneously is also gradually increasing to reduce the cost of deployment.

In this work, we present Visar, a **device-free** virtual sound **spots** source projection system for acoustic augmented reality leveraging air nonlinearity. Unlike those beam-shaped projections from directional speakers, Visar has the capabilities for projecting fine-grained sound spots (down to $15cm \times 15cm$ area) in the space. The key idea of Visar is to split the high-frequency components and project them through two transducer arrays. The arrays are deployed at different locations, and the emitting beams are pointed in various directions to control the intersection points. Only the *intersections* of the beams suffer air nonlinearity, further emitting audible sounds. Specifically, Visar utilizes *phased arrays* to support digital beam steering, which achieves fast and accurate spot position adjustment. In this way, the spots can be projected at any angle around the users. Moreover, Visar also achieves multiple spots for multi-user scenarios through optimization and beam redistribution.

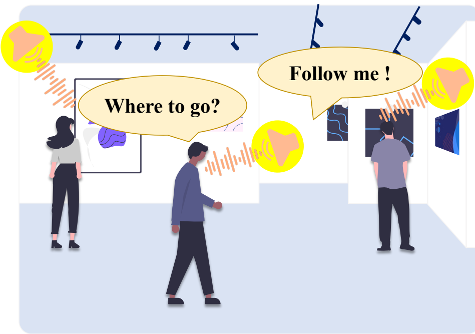


Fig. 1. Motivation of utilizing VISAR to project virtual sound spot sources to realize direction guidance in AAR scenarios such as navigation.

Taking into account the user experience and practicality, VISAR must meet several essential requirements: Firstly, the system should track the users and play sounds simultaneously. Secondly, audible sound spots should only appear around the users, and audible sounds in unintended areas involving the sound along the beam path or in other intersections should be avoided. Thirdly, the VISAR needs to support multiple spots that play different contents for multi-user scenarios.

Challenges: Nevertheless, implementing VISAR remains challenging due to several factors: **1) Limited bandwidth:** The limited bandwidth of ultrasound transducers prevents tracking and playback using the same band. Otherwise, there will be audible leakage due to the tracking process; **2) Self-demodulation:** The up-converted modulated signal from an array would suffer the self-demodulation problem, resulting in sound leakage along the path. **3) Grating lobes:** Due to element spacing greater than half a wavelength, the grating lobes may cause multiple spots, affecting direction discrimination. **4) Mutual interference:** When multiple sound spots need to be projected, interference will not only occur between spots but also between beams due to the nonlinearity effects. Moreover, the audible sound spots should be scheduled to guarantee each spot is heard by only one user.

Our approach: To address the above issues, we propose the following approaches. Firstly, we present a real-time acoustic tracking scheme (Sec. 5.1) for concurrent tracking and playback. We propose a frequency-division idea that leverages ultrasound transducers of different central frequencies. Secondly, we designed the pipeline for virtual sound spot projection (Sec. 5.2). Specifically, we explore the specific spot position by modeling the space and calculating the beam steering angle for each array. Single sideband modulation scheme is used to reduce sound leakage along the path. Thirdly, an optimization-based beamforming scheme (Sec. 5.3) involving grating lobe suppression for unintended sound spot elimination and wide-nulling to confront mutual interference between the beams. Finally, in order to mitigate the interference between the spots, we further consider multi-spot scheduling (Sec. 5.4), involving the transmission content redistribution and extraneous spots' locations adjustment. Extensive experiments verified the capability for use in AAR applications that VISAR can realize $7.83cm$ tracking error and 10.06° orientation estimation error.

Our contributions can be summarized as follows:

- We studied the feasibility of projecting virtual sound spots from ultrasound in the air by leveraging air nonlinearity. Furthermore, we proposed VISAR system, supporting multi-spot in multi-user scenarios for AAR applications.
- We present a frequency-division idea to achieve concurrent tracking and projection, supporting the requirements in AAR applications.

- We proposed an optimization-based beamforming scheme to overcome the unintended audio leakage caused by grating lobes as well as interference from beams in the space.
- We consider multi-spot scheduling involving the transmission content redistribution and extraneous spots' locations adjustment to mitigate the mutual interference from spots in multi-user scenarios.
- We implemented and evaluated VISAR. Extensive experiments show VISAR can realize 7.83cm tracking error and 10.06° orientation estimation error, demonstrating the efficiency of VISAR.

Roadmap. The rest of the paper is organized as follows. Sec. 2 discusses related works. Sec. 3 introduces the working principle and Sec. 4 shows the feasibility of VISAR. In Sec. 5, we present our approaches in detail. Sec. 6 presents our evaluation results. Finally, Sec. 7 and Sec. 8 discuss the future works and conclude the work.

2 RELATED WORK

2.1 Acoustic Augmented Reality Application

Acoustic augmented reality (AAR) applications have attracted much attention in recent years, including but not limited to audio visualization [44, 51, 59], acoustic haptics [31], acoustic navigation [16, 56, 71], and sound field modeling [70]. Companies like Microsoft [43] and Apple [2] provide spatial information by creating effects of 3D sound. In particular, Head Related Transfer Function (HRTF) [18, 26, 70] performs a significant role in creating realistic spatial audios through a binaural headphone. By combining spatial audio and location-aware context, Ear-AR [71] leverages earphones and IMU information to achieve indoor localization and guide users to specific directions. Different from earphone-based AAR systems such as EAR-AR, VISAR does not require additional equipment to track users; the system uses acoustic signals for device-free user tracking. Acoustic tracking can be achieved by Time of Flight (ToF) [48, 57], Doppler effect [27, 74], and phase shift [47, 75], in which Frequency Modulated Continuous Wave (FMCW) exhibits good auto-correlation, high spectral efficiency, simple demodulation, etc [42, 46, 60]. VISAR utilizes frequency division to achieve simultaneous acoustic tracking and sound source projecting. Unlike those spatial audio involving complex HRTF models, VISAR can directly project the real sound source in the air.

The target of VISAR is similar to Ear-AR, both are to implement AAR applications such as user guidance. However, Ear-AR requires users to wear an additional headset to achieve acoustic augmented reality. In contrast, VISAR does not require additional wearables (i.e., headphones) or involve complex HRTF models. VISAR straightforwardly projects real-true sound sources that can guide AR users or the visually impaired, which improves user comfort.

2.2 Acoustic Nonlinearity

Acoustic nonlinearity [65, 66] is ubiquitous in sound generation, propagation, and reception stages. It can occur in electronics like speakers and microphones [55, 77], as well as in propagation media like air [15] and water [78].

Hardware nonlinearity. Nonlinearity occurs in the diaphragm and amplifier of the microphone when transmitting ultrasound. Recent works [9, 54, 55, 77] show the feasibility of injecting inaudible voice commands by exploiting the hardware nonlinearity. The microphone nonlinearity can also improve sensing granularity [9] due to the wider spectral information that nonlinearity brings.

Media nonlinearity. Researches [67, 69] verified the distortion of sound as it propagates in non-linear media, including air, which suggests ideas for demodulation of ultrasound in air. Projects such as SoundLazer [33], HoloSonics [25], and Focusonics [12] proposed highly directional loudspeakers based on air non-linearity. In addition to air, nonlinearity can also occur in the water [34, 72]. MuDiS [39] developed a multi-directional speaker to project different audio content in different directions. Meta-Speaker [63] and Audio Hotspot Attack [29]explore the feasibility of using air nonlinearity to generate a sound spot or an audible area, this is a new idea of generating nonlinearity through the distributed arrangement of speakers.

Application. The nature of sound transmission underwater has broadened the application of acoustic nonlinearity in underwater acoustic engineering, using parametric arrays to enable applications such as sub-bottom profile measurement [17, 28], underwater acoustic communication [34, 72], and detection of buried targets [6, 37]. Furthermore, the human body can also be used as a nonlinear medium for bone conduction speech transmission based on ultrasonic hearing AIDS [11, 36] and earphones [35].

In this work, VISAR also makes use of air nonlinearity to project virtual sound spots rather than beams, which achieves fine-grained audible sound zones compared to existing directional speakers. VISAR also combines air nonlinearity with human tracking and localization to realize a real-time acoustic augmented reality application. Unlike Meta-Speaker, which does not consider multi-user scenarios, VISAR supports multi-user scenarios. Besides, Meta-Speaker requires mechanical rotation while VISAR uses digital steering to enhance mobility and is more affordable at $450, compared to Meta-Speaker's $600 due to the additional Servo Motor.

2.3 Acoustic Tracking and Localization

Acoustic tracking and localization have been studied extensively. Acoustic tracking can achieve high accuracy owing to the low propagating speed and long wavelength of sound [40]. Furthermore, acoustic localization becomes more attractive when it comes to indoor localization and subtle movement tracking. Acoustic tracking and localization can be easily applied with speakers and microphones already embedded in intelligent devices, adding to its attractiveness in VR, AR [42, 75], health care [46, 52], smart home applications [38], and human-computer interaction [13, 14, 47, 74]. Acoustic tracking and localization can be achieved by leveraging Time of Flight (ToF) [20, 48, 57], Doppler effect [27, 74], and phase shift [47, 75].

Frequency-modulated continuous wave (FMCW) is a chirp signal commonly used in acoustic tracking and localization, in which the signal frequency changes linearly with time. Compared to other waveforms, FMCW shows good autocorrelation properties, high spectral efficiency, and simple demodulation [60]. CAT [42] develops a novel distributed FMCW to overcome the synchronization problem between the transmitter and the receiver in traditional FMCW. ApneaApp [46] also utilizes FMCW to capture the subtle frequency shift of breathing. Millisonic [60] further improves the performance of 1D acoustic tracking and localization in the presence of multipath by extracting distance information from the instantaneous FMCW phase.

3 BACKGROUND

In this section, we introduce the core intuition of VISAR, which utilizes air nonlinearity and phased arrays to reproduce sound from ultrasound.

3.1 Sound from Ultrasound

The sound propagation can be seen as a linear system in most cases. If the input sound is $s(t)$, then the received signal S_r can be represented by $S_r = A_1 s(t)$. Here, A_1 represents a complex gain of the transmitting channel in the air. However, when the amplitude of the sound wave is large enough, the propagation process will exhibit nonlinear properties, as theoretically called KZK equation [10]. As shown in Fig. 2, the received signal S_r combined with nonlinearity expressed in terms of $s(t)$ is

$$S_r = A_1 s(t) + A_2 s^2(t) + A_3 s^3(t) + \cdots \tag{1}$$

where $A_1, A_2, A_3, \cdots$ represent the channel gain of the first-, second-, third-, and higher order nonlinear terms. The square (second-order) term offers the opportunity to reproduce the sound from ultrasound, while the higher-order terms are extremely weak and negligible.

With the help of nonlinear characteristics, a directional audible sound speaker can be achieved by modulating an ultrasound wave through an ultrasound transducer array [77]. Assuming that the transmitted signal has two single-tone components ω_a and ω_b, expressed as $s(t) = \cos(\omega_a t) + \cos(\omega_b t)$. According to Eq. 1, the received

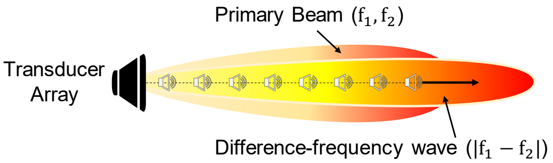

Primary Beam (f_1, f_2)
Transducer Array
Difference-frequency wave ($|f_1 - f_2|$)

Fig. 2. An illustration of air nonlinearity that the transducer array can generate audible differential frequency beam from a high energy primary beam composed of high frequencies.

signal (omit the second-order term) can be derived as

$$\begin{aligned} S_r =& A_1 s(t) + A_2 s^2(t) \\ =& A_1(\cos(\omega_a t) + \cos(\omega_b t)) + \\ & A_2(1 + \frac{1}{2}(\cos(2\omega_a t) + \cos(2\omega_b t)) \\ & + \cos((\omega_a + \omega_b)t) + \cos((\omega_a - \omega_b)t)) \end{aligned}$$

Thus, the received sound wave due to nonlinearity mathematically contains multiple frequencies ω_a, ω_b, $2\omega_a$, $2\omega_b$, $\omega_a+\omega_b$, $\omega_a - \omega_b$. Since human hearing is not sensitive to high-frequency sounds, it can be regarded as a low-pass filter. When ω_a and ω_b are two close high-frequency signals (i.e., $41kHz$ and $40kHz$), only the frequency $\omega_a - \omega_b$ (i.e., $1kHz$) lies in audible band, as shown in Fig. 2.

3.2 Phased Array

A phased array [21] combines signals constructively using beamforming techniques from N antennas or transducers. The resulting wavefront of a phased array can be steered and shaped by adjusting the phase and amplitude of the signals sent by each element. Specifically, for an phased array with each element i ($1 \le i \le N$) emitting a wave with amplitude A_i and phase ϕ_i, and the distance between adjacent elements is d, the emitted wave can be expressed as

$$W(\theta, \phi) = \sum_{i=1}^{N} A_i e^{j(\omega t + k\vec{r_i} \cdot \vec{n})}$$

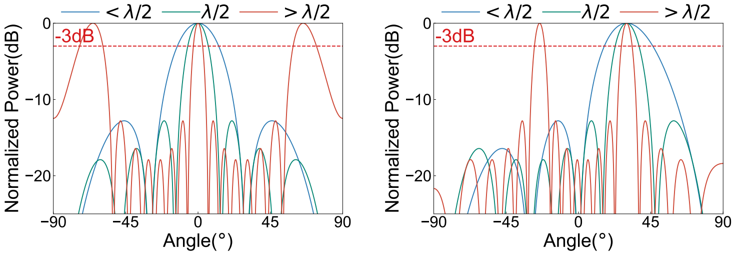
where θ and ϕ are the angles of the direction of interest relative to the array axis, j is the imaginary unit, ω is the angular frequency, k is the wave number, $\vec{r_i}$ is the position vector of element i relative to the array center, and $\vec{n}$ is the unit vector in the direction of interest. The phase difference between adjacent elements is given by:

$$\Delta\phi = kd\cos\theta$$

By adjusting the phase shift of each element relative to its neighbors, the phased array can control the direction and shape of the emitted wavefront, allowing it to focus on a specific target or steer the beam in a desired direction.

The performance of a phased array is strongly affected by the spacing between elements. Fig. 3 shows the beam pattern of an 8-channel phased array emitting $40kHz$ and targeting 0° and 30°; beam width becomes more expansive when the spacing is less than half-wavelength, and the grating lobe appears when the spacing is above half-wavelength. The grating lobe will significantly influence the performance of creating virtual sound spots, for which we will suppress the grating lobes by optimization in Sec. 5.3.

Key idea. In this paper, we propose VISAR, combining both air nonlinearity and phased arrays. VISAR consists of two-phased arrays to steer the beams. The two beams play different high-frequency sounds separately and meet each other at a spot. Due to air nonlinearity, audible sounds are produced at the intersection spots. Unlike



(a) Target 0°

(b) Target 30°

Fig. 3. The beam patterns of an 8-channel phased array emitting $40kHz$ and target 0° and 30° while varying the spacing between elements (above half-wavelength, half-wavelength, and below half-wavelength). Grating lobes appear when the spacing is above half-wavelength.

traditional directional speakers that emit sound in the whole propagation path, Visar produces audible sounds in fine-grained sound zones. In the next section, we explore the feasibility of Visar.

4 PRELIMINARY STUDY

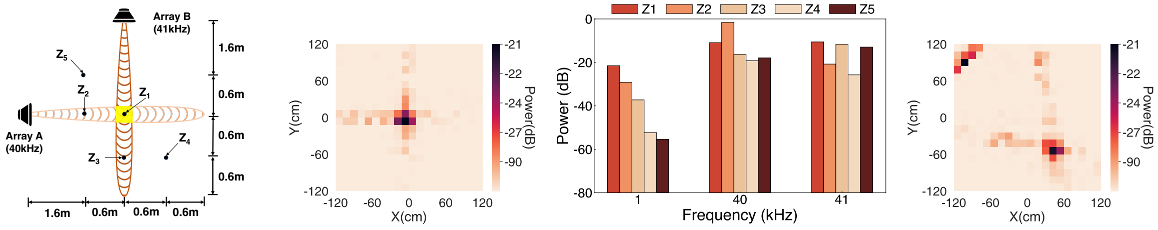
In this section, we conduct experiments to preliminarily study the feasibility of Visar. The target for Visar is to produce audible sounds that are only limited in manipulated small areas. We mainly focus on two questions: 1) Can Visar produce audible sounds only at the target points and not elsewhere? 2) What is the relationship between the transducer arrays' parameters and the audible zones' properties?

Setup. we built an initial prototype of Visar from theory to practice. For the *transmitter* side, two transducer arrays emit different single-tone sinusoidal waves with frequencies of $40kHz$ and $41kHz$, respectively. Each array contains $8 \times 8 = 64$ ultrasound transducers and the beam points straight ahead. Two arrays are placed vertically on two sides of the space. Therefore, the intersection point is at the center. (shown as Fig. 4(a)). For the *receiver* side, an electret condenser microphone is used to record the acoustic signals. To confront the aliasing issue of insufficient sampling rates, we set $192kHz$ as the analog-to-digital converter(ADC) sampling rate. Moreover, we also turn off the automatic gain control (AGC) to fairly compare the signals.

4.1 Virtual Sound Spot Projection

We first verified whether Visar could generate a virtual sound spot only at the target location (i.e., the central area). The test space is within $2.4m \times 2.4m$ with a resolution of $0.12m \times 0.12m$. Fig. 4(b) shows the heatmap of the intensity of $1kHz$. An audible zone appears at the center. Furthermore, we select 5 representative points (shown as Fig. 4(a)): a point at the intersection of the beam ($Z1$), two points on the path of one of the beams ($Z2$,$Z3$), and two points not on the path of either beam ($Z4$,$Z5$). We drew the intensity of five points and compared the signal strength. The result (shown in Fig. 4(c)) shows that the power of $1kHz$ at $Z1$ outperforms other points by $11dB$. In other points, especially $Z4$ and $Z5$, the sound is significantly weaker than $Z1$. This demonstrates the feasibility of Visar for producing small regions of audible sound through two transducer arrays.

We then discuss the impact of grating lobes on Visar. Since the spacing between array elements is wider than half the wavelength of the carrier frequency, grating lobes are generated. As both arrays emit beams toward 0° in Fig. 4(b), the desired spot will be generated in the center of the area. The grating lobes of the beam will appear near ±60°, and the extraneous spot generated by the intersection of the grating lobes is far away from the activity area we set. If the emission angle of the array deflects with the movement of the user, the intersection of the grating lobes will appear in the area and interfere with the user. When the emission directions of the two



(a) Preliminary setup (b) Heatmap of $1kHz$ (c) Power of sound at different points (d) Grating lobes crossover.

Fig. 4. (a) An illustration of a preliminary setup; (b) shows the heatmap that an obvious audible $1kHz$ component is produced at the intersection Z_1 of the beam. (c) shows the audible sound at Z_1 is much stronger than Z_2, Z_3, Z_4 and Z_5. This demonstrates the feasibility of producing a virtual sound spot using VISAR. (d) shows the heatmap that when the transmit beams are deflected by 10°, there is not only a target spot but also a spot generated by the intersection of the grating lobes.

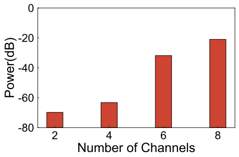
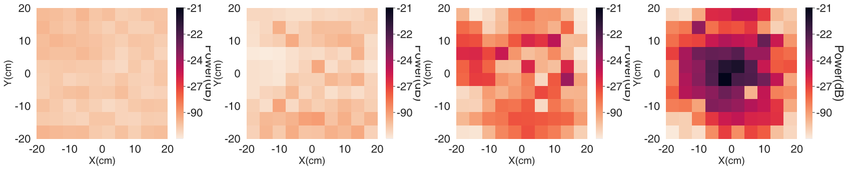


Fig. 5. The signal strength under the impact of channel numbers.



(a) 2 × 8 (b) 4 × 8 (c) 6 × 8 (d) 8 × 8

Fig. 6. The heatmap of the audible sound area under the impact of channel numbers.

arrays are deflected by 10°, as shown in Fig. 4(d), the sound spot is projected to the lower right of the active area. However, an extraneous spot due to the intersection of grating lobes will appear in the upper left. To solve the grating lobes, we introduce an optimization algorithm to suppress the grating lobes in Sec. 5.3.

4.2 Pre-analysis of Audible Zone

While VISAR can produce an audible area, the properties of the area will significantly affect the performance in acoustic augmented reality applications. Specifically, the number of channels will affect not only the amount of energy but also the beamwidth, thereby affecting the intersection area. To this end, we change the number of channels from 1 × 8 to 8 × 8 and study the impact of signal strength and the granularity of the audible zone.

4.2.1 Signal strength. Fig. 5 shows the strength of audible $1kHz$ at the target point. When the number of channels is less than 2, the energy is too small to be heard. As the number of channels gradually increases, the audible sound gets louder.

4.2.2 Granularity. As for the granularity, we chose an area in the center for a more refined measurement. The resolution is $4cm \times 4cm$ for a $40cm \times 40cm$ area. Fig. 6 shows the heatmaps of audible energy under the impact of channel numbers. The results indicate that when the channel number is 2 and 4 (Fig. 6(a) and Fig. 6(b)), the audible sound is almost non-existent. For the sound from 6 channels (Fig. 6(c)), its energy is not concentrated even though it is audible. However, when the channel number reaches 8, the energy concentrates down to $16cm \times 16cm$ area as shown in Fig. 6(d). It is because the transducer array generates a sharper beam pattern as the number of channels increases, and the intersection area becomes finer-grained.

Remarks. Our preliminary study shows VISAR can project a finer-grained virtual sound spot source in the air. Noticing that the replicated auditory signal adheres to the Huygens–Fresnel principle, it can be discerned that it

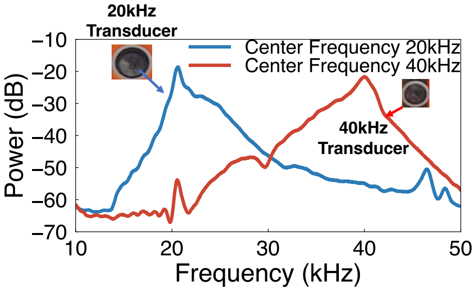


Fig. 7. Frequency response of two ultrasound transducers of different central frequencies ($20kHz$ and $40kHz$).

functions as a novel source of wavelet which disseminates omnidirectionally [63]. This preliminary provides the ability to finely control the positions of sound spots, which offers many opportunities in AAR applications. In practice, we chose 8×8 elements in this work for they can offer both enough volume and enough finer audible sound zones.

5 VISAR FRAMEWORK

In this paper, we present VISAR, a real-time virtual sound spot projection system utilizing air nonlinearity through two transducer arrays. In practice, VISAR should meet several key requirements: i) the system needs to simultaneously track the users' locations and project the virtual sound spots. ii) the system should avoid sound leakage at locations other than the target points. iii) the system should support multiple spots playing individual contents for multiple users.

The proposed VISAR system is composed of four parts: **1) Real-time acoustic tracking (Sec. 5.1)**: introduce how to achieve concurrent user tracking and audio projecting. **2) Virtual sound spot projection (Sec. 5.2)**: present the pipeline to model the space and calculate the beam steering angle for each array. **3) Beam optimization (Sec. 5.3)**: optimize the beam weights to eliminate unintended spots and mitigate the mutual interference from beams. **4) Spot scheduling (Sec. 5.4)**: eliminate the mutual interference from spots in multi-user scenarios.

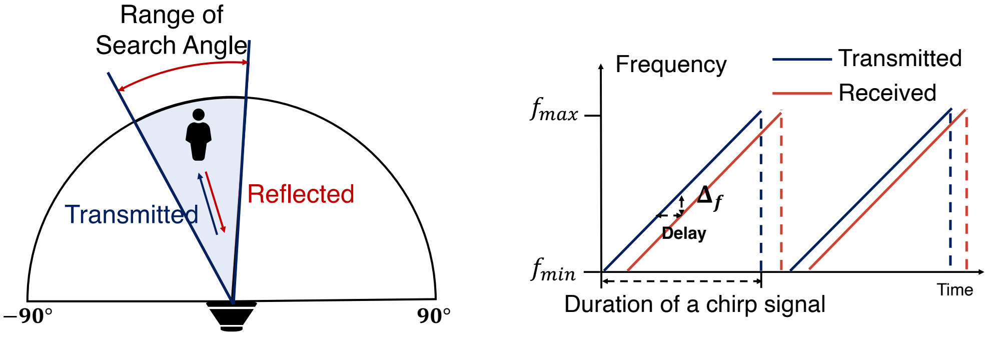
Real-time acoustic tracking determines the user's position as well as the location of the projected sound source. The geometric operation of each array beam intersection is clarified through the virtual sound spot projection. The actual emission depends on the beam optimization and spot scheduling.

5.1 Real-time Acoustic Tracking

In the context of Acoustic Augmented Reality (AAR) applications, precise real-time user tracking is of paramount importance. The tracking module must operate seamlessly alongside sound spot projection, ensuring that the tracking process remains acoustically inconspicuous without introducing additional audible sounds.

5.1.1 Frequency division. We present a frequency-division method to achieve human imperceptible tracking. Fig. 7 shows the frequency response of two transducers whose central frequencies are $20kHz$ and $40kHz$, respectively. The frequency responses are recorded $0.5m$ away from the transducers. Each transducer has only $4kHz$ bandwidth, and the response bands of the two transducers do not overlap.

The combination of frequency ranges has to satisfy two inaudible conditions. First, the transmitted signals should be inaudible, meaning the tracking frequency f_t and projecting frequency f_p range higher than audible bands ($f_t, f_p > 16kHz$). In addition, the difference between the two frequency ranges due to air nonlinearity should also lie in the inaudible band. ($f_p - f_t > 16kHz$). Specifically, we chose the frequency range of $18kHz$ to $22kHz$ for tracking and $40kHz$ to $44kHz$ for producing sound spots, respectively.



(a) Chirp steering range (b) Chirp Signals

Fig. 8. An illustration of the tracking process, including chirp steering range and an example of chirp signals. We first scan a large range and obtain the user's location. Then, we perform small range scans due to users' movement. After each scan, we update the scan range to ensure the continuous tracking of the system.

5.1.2 Tracking process. The tracking module contains a transducer array and a time-synchronized microphone. The tracking process can be divided into FMCW chirp steering and distance calculation.

Chirp steering. The FMCW chirps are sent through the phased array to find the angle, namely θ. At the beginning of the process, the beam sweeps all angles to determine the initial position. Once the user is located (shown as Fig. 8(a)), since the movement will not change drastically, VISAR only needs to search the nearby angle ranges (i.e., empirically set ±10°) to reduce the tracking time (i.e., for only $0.5s$ a round). The chirp length is set as $0.1s$, and the steering resolution is set as 5°.

Distance calculation. Next, we leverage an FMCW-based approach [42] to calculate the distance, shown as Fig. 8(b). The frequency of the FMCW chirp linearly changes as $f = f_{min} + \frac{Bt}{T}$, where f_{min} is the minimum frequency, B is the bandwidth, T is the sweep time. The phase can be derived by integration of frequency $\phi(t) = 2\pi(f_{min}t + \frac{Bt^2}{2T})$. The transmitted signal is $s_t(t) = \cos(2\pi(f_{min}t + \frac{Bt^2}{2T}))$. The chirp signal propagates through the air, encounters the user, and reflects to the microphone. Suppose the time delay from sending a signal to receiving the signal is t_d^i, and the received signal is:

$$s_r(t) = \sum_{i \in I} \alpha^i \cos(2\pi(f_{min}(t - t_d^i) + \frac{B(t - t_d^i)^2}{2T}))$$

where α is the attenuation factor of the channel, and I contain different reflection paths. Then we mix the received signal with the transmitted signal and go through a low pass filter to obtain differential frequency components:

$$s_{mix}(t) = \sum_{i \in I} \alpha^i \cos(2\pi f_{min} t_d^i + \frac{\pi B(2tt_d^i - (t_d^i)^2)}{T})$$

When the distance between the user and the transmitter is d, then $t_d = \frac{2d}{c}$, where c represents the sound speed, the frequency spectrum of $s_{mix}(t)$ contains both the direct path and the echoed signal. In this case, the frequency component f_d of the echoed signal is determined by the second largest peak, as the direct path always has the highest signal strength. Therefore, the distance d can be derived as $d = \frac{f_d cT}{2B}$.

Modeling space. We model the entire space in a two-dimensional coordinate system centered at $(0, 0)$. The transducer arrays are placed at $(-t, 0)$ and $(0, t)$, respectively. Since we have two sets of transducer arrays, we could obtain two pairs of angles and distances (θ_1, d_1) and (θ_2, d_2) from above, where $\theta_1, \theta_2 \in [-90°, 90°]$. The user

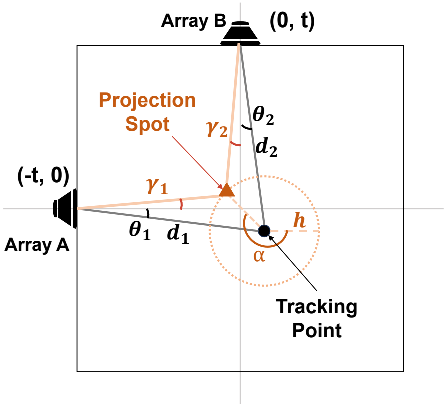


Fig. 9. Suppose the guide direction α is known, we can model the space for tracking (black lines) and determine the projection directions of two transducer arrays (orange lines).

tracking estimation located in the space will be $L_1 = (d_1 \cos\theta_1 - t, -d_1 \sin\theta_1)$ and $L_2 = (-d_2 \sin\theta_2, t - d_2 \cos\theta_2)$. We select the average of two points as the user's coordinates $L_u = (L_1 + L_2)/2$.

5.1.3 Multi-user splitting. The above delineates the process of tracking within a single user. Given the need for VISAR to project sound spots applicable to each user, extending this tracking to a multi-player environment is essential. We assume the presence of n_{user} users, and the dual transmitters can derive two distinct sets of coordinates, denoted as $\mathbb{L}_1 = \{L_1^k | 1 \le k \le n_{user}\}$ and $\mathbb{L}_2 = \{L_2^k | 1 \le k \le n_{user}\}$. We identify and pair the two proximally closest points from two coordinate sets to ascertain each user's location. Subsequently, calculating the mean of these paired points yields the positions of the respective users.

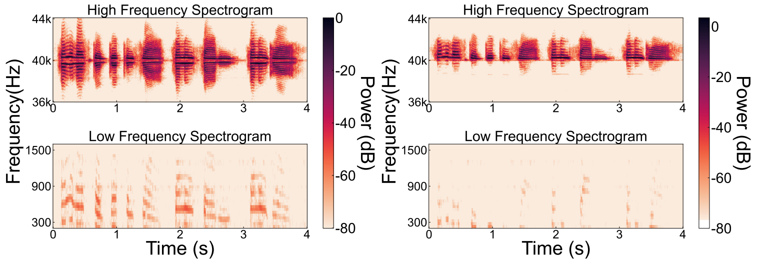
5.2 Virtual Sound Spot Projection

After obtaining the users' locations, the next step is to project virtual sound spots at the specific positions. Note that we project the sound spot to the front of the target's head, so there will be no issues that users block the spot. This process consists of two parts: i) determine the projection directions of two transducer arrays; ii) determine the signal modulation scheme.

5.2.1 Projection direction. The projection spots should be controlled based on the users' positions. We set the coordinates of the tracking position L_u as (a, b). Shown as Fig. 9, suppose the specific direction that VISAR wants to guide the user to is denoted as $\alpha \in [0°, 360°]$. The projection distance away from the user is set to h. In practice, the h is set to $20cm$, considering the audio quality and orientation estimation (refer to Sec. 6.2). In this case, the spot needs to be projected at $(a + h\cos\alpha, b - h\sin\alpha)$. Then we can calculate the projection direction. For the transducer array at $(-t, 0)$, the beam steering angle is $\gamma_1 = \arctan(\frac{b-h\sin\alpha}{t+a+h\cos\alpha})$, and for the transducer array at $(0, t)$, the beam steering angle is $\gamma_2 = \arctan(\frac{a+h\cos\alpha}{t-b+h\sin\alpha})$.

5.2.2 Signal modulation. In order to reproduce audible sound from ultrasound, one straightforward solution is amplitude modulation [73]. Suppose the audio we want to reproduce is $m(t)$ and a carrier frequency f_c, the modulated signal of modulation depth α is:

$$s(t) = \underbrace{\alpha m(t)\cos(2\pi f_c t)}_{Array\ A} + \underbrace{\cos(2\pi f_c t)}_{Array\ B} \tag{2}$$



(a) Amplitude modulation

(b) USB modulation

Fig. 10. The spectrogram of amplitude modulation and USB modulation. USB modulation performs weaker audio leakage in the transmission path.

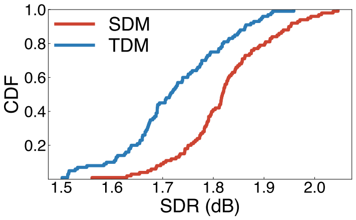


Fig. 11. Compare SDR of spatial-division multiplexing and time-division multiplexing.

We split the signal in Eq. 2 into two parts: an up-converted part $s_A(t) = \alpha m(t)\cos(2\pi f_c t)$ transmitted by transducer array A, and a carrier wave $s_B(t) = \cos(2\pi f_c t)$ transmitted by transducer array B. The two beams will intersect, and the audible sound will be demodulated at the spot.

However, the up-converted part $\alpha m(t)\cos(2\pi f_c t)$ contains double sidebands and suffers from self-demodulation problems, which results in possible sound leakages along the transmission path. To weaken this issue, we utilize single-sideband modulation (SSB modulation). Only one sideband is retained in SSB modulation. For an upper sideband (USB) modulation, the formula is given

$$s_A(t)_{usb} = m(t)\cos(2\pi f_c t) - \hat{m}(t)\sin(2\pi f_c t)$$

where $\hat{m}(t)$ is the Hilbert Transform of $m(t)$. Fig. 10 tests the sounds in $1m$ away from the array and shows the noise level of amplitude modulation and USB through the transmission, and we can see that USB modulation outperforms the AM.

5.2.3 Multi-beam beamforming. To achieve VISAR, we need multi-beam beamforming to support the projection of multiple spots. We use spatial-division multiplexing (SDM) [64] to transmit audio to different spatial areas, enabling beams pointing in different directions to reuse the same frequency band. Specifically, the beam weight vector for transmitting one beam to direction θ can be expressed as $\mathbf{w}_\theta = \frac{1}{\sqrt{n}}[\omega_\theta^1, \omega_\theta^2, ..., \omega_\theta^n]^T$. Assuming there are M directions needs to transmit concurrently, the corresponding phase shifts and audio contents can be expressed as $\mathbf{w}_{\theta_1}, \mathbf{w}_{\theta_2}, \cdots, \mathbf{w}_{\theta_\mathrm{M}}$ and $a_{\theta_1}, a_{\theta_2}, \cdots, a_{\theta_M}$, respectively. The SDM will transmit the sum of audio that has undergone the corresponding phase shift, which is $\frac{1}{M}\sum_i^M a_{\theta_i} e^{-j\mathbf{w}_{\theta_\mathrm{i}}}$. In this way, VISAR can transmit multiple beams to support multi-user scenarios. Note that the interference and leakage still occur between multiple beams and the suppression will be presented in Sec. 5.3.

Another intuitive multi-beamforming solution is time-division multiplexing (TDM) method [61], which transmits each beam alternately in time slices. we compare the difference between SDM and TDM. Specifically, each beam is emitted alternately with a time slice of $0.01ms$. We measured the SDR (defined in Sec. 6.1) and as shown in Fig. 11, we found that the SDR is smaller than $2dB$ in both SDM and TDM, which means that the leakage caused by beams interference is large without optimization. This is because the limitations of the ultrasonic vibrator diaphragm cannot change quickly between different signals [23], so the vibration residue of the ultrasonic vibrator in the previous time slice will lead to possible leakage. Compared to TDM, SDM has the advantage of directly superimposing audio, so the number of supported users will theoretically be greater than that of time division multiplexing. Especially, in the case where each user has a time slice, more users will bring frequent

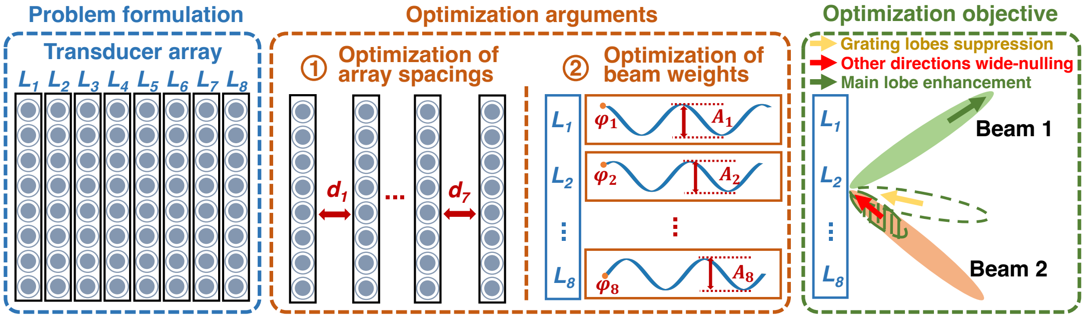


Fig. 12. Description of beam optimization. We optimize the ultrasonic transducer array on a column basis. The arguments of optimization are the distances between the columns of the ultrasonic transducer and the beam weight of the audio transmitted, including amplitude and phase. The objective of optimization is to suppress the grating lobe, perform wide-nulling in other possible audio playback directions, and ensure that the main lobe is strong enough.

switching and longer intervals. Therefore we chose SDM to transmit the audio concurrently and introduce a beam optimization method to suppress the leakage.

5.3 Beam Optimization

Visar is designed to project one specific sound spot to each user. While the above processes including tracking and projection allow the system to project the sound spots and track the users, the grating lobes are disturbing. Since the diameter of ultrasonic speakers is greater than the half wavelength of our operating frequency, as mentioned in Sec. 3.2, grating lobes with approximate amplitude as the main lobe will appear at other angles. A grating lobe may intersect with another main lobe or grating lobe, resulting in unintended sound spots.

5.3.1 Unintended leakage production. In scenarios involving multiple users, wherein multiple transmitters emit a plurality of beams, a notable issue arises from the interference caused by the mixing of lobes, in addition to the interference from grating lobes. Specifically, the interaction between a side lobe, which is distinct from and smaller than the grating lobes, and the main lobe in a different direction results in the generation of undesirable acoustic outputs.

For simplicity and without loss of generality, we consider the example of two beams. In a phased array system operating at a central carrier wave frequency of f_c, the audio signals intended for reproduction along two directions, θ_1 and θ_2, are denoted as $h_1(t)$ and $h_2(t)$ respectively. The output from the speaker that transmits up-converted audio in these two directions can be represented as $s_1(t) = h_1(t)\cos(2\pi f_c t)$ and $s_2(t) = h_2(t)\cos(2\pi f_c t)$. Assuming the presence of side lobes from the first beam in the direction of the second beam, and denoting α as the amplitude ratio of the side lobe to the main lobe, the side lobes can be represented as $\alpha s_1(t) = \alpha h_1(t)\cos(2\pi f_c t)$. Consequently, in the direction of the second beam, this results in the generation of an additional nonlinear component expressed as $\alpha s_1(t)s_2(t) = \alpha h_1(t)h_2(t)\cos^2(2\pi f_c t)$, which includes the term $\alpha h_1(t)h_2(t)$, representing an undesirable acoustic output.

In this section, we will explain our **beam optimization** methodology to surmount the limitations, specifically addressing the issues of grating lobes and the interferences caused by multiple beams.

5.3.2 Problem formulation. We consider a setup where an $n-$channel parametric array is anticipated to orient towards a predefined direction set, denoted as Θ. Within this set, a particular beam directed along θ is characterized by the beam weight ω_θ which is a complex number corresponding to the transmitted signal's scaling factor and phase shift. Let d_m denote the spacing between each channel, constituting an array D with $N-1$ spacing values. Therefore, the signal received R, which is directed towards a set of receiving directions Φ, can be deduced in accordance with

$$R = K\omega_\theta$$

where R is defined as an array congruent in length to Φ, where each constituent element corresponds to the signal received from a specific direction ϕ. Additionally, K is introduced as a $\Phi \times N$ matrix, delineating the steering vectors pertinent to the N channel transducer array.

Our goal is to optimize **the beam weight** ω_θ as well as **the spacings D** for each channel to suppress **grating lobes** and **the interferences caused by multiple beams**. This optimality is quantified in terms of the power P of the received signal, where the power is calculated using the expression $P = 10 \log 10 |W|^2$. Therefore, we define our objectives as follows (shown as Fig. 12):

Grating lobes suppression: The attenuation of grating lobes is extremely significant in reducing unintended spots. The set of grating lobes is represented by $\Theta_{grating}$, and the energy associated with the grating lobes can be expressed mathematically as follows:

$$L_{grating} = \sum \mathcal{P}(P\mathcal{M}(\Theta_{grating}))$$

where the function denoted as $\mathcal{M}(\cdot)$ is defined to yield a mask sequence that is characterized by an identical length to that of P. Within $\mathcal{M}(\cdot)$, values are assigned a binary state: a value of 1 is attributed to those within the specified input range, while all remaining elements are assigned a value of 0. The function $\mathcal{P}(\cdot)$ is used to calculate the peak value of the grating lobe.

Other directions wide-nulling: In addition to the grating lobes, we next solve the interference caused by multiple beams. Since some other directions Θ_{other} need to project audio as well, as elucidated in Section 5.3, these auxiliary directions emerge as leakage directions. The optimization approach presented herein is meticulously crafted to mitigate such leakage across a specified angular expanse, a technique called *wide-nulling*. Considering an angular interval centered around Θ_{other}, with a breadth of 2γ degrees, the following equation is proposed for the restraint of these angles:

$$L_{other} = \sum P\mathcal{M}([\Theta_{other} - \gamma, \Theta_{other} + \gamma])$$

The selection of the parameter γ and its implications will be further expounded in Section 6.2.

Mainlobe enhancement: To prevent the above optimization steps from sacrificing the main lobe energy, we maximize the beam power directed towards the designated angle θ. It is imperative to consider the beam's width ξ, ensuring it is not excessively narrow, as this could result in the sound being perceptible to only one ear. To this end, we assume an optimal beam width, characterized by an angular interval of 2ξ degrees, thus defining the main lobe's directional range as $[\theta - \xi, \theta + \xi]$. The following expression can represent the cumulative power within this main lobe direction:

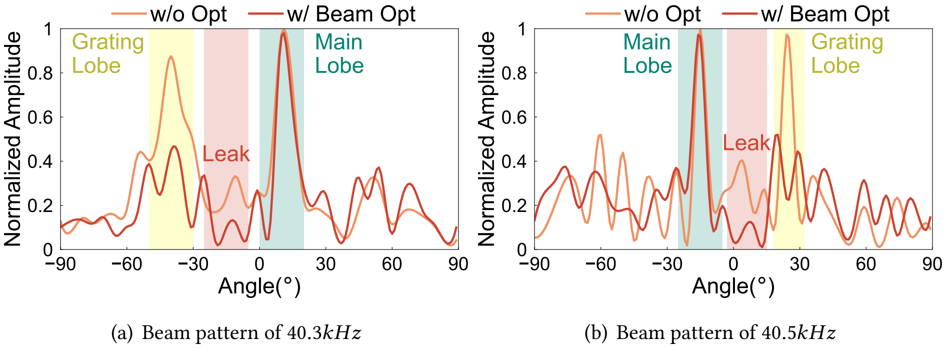
$$L_{main} = \sum P\mathcal{M}([\theta - \xi, \theta + \xi])$$

where ξ is empirically set to 10 degrees, based on the rationale that a total angular range of 20 degrees is conducive to ensuring auditory perception by the human ear.

By integrating the above items, we have the following optimization model:

$$\begin{aligned} \min_{\omega_\theta, D} \quad & \lambda L_{grating} + \nu L_{other} - L_{main} \\ s.t. \quad & \begin{cases} |\omega_\theta| \leq 1 \quad (\theta \in \Theta) \\ d_m \geq L \quad (1 \leq m \leq N-1). \end{cases} \end{aligned} \tag{3}$$

where λ and ν represent the importance of loss items. $|\omega_\theta| \leq 1$ are constraints on the normalized sound magnitude generated from ultrasonic transducers. d_m is the spacing between $transducer_m$ and $transducer_{m+1}$, larger than an ultrasonic transducer's diameter, L. The objective formulation elucidates that it is imperative to diminish the



(a) Beam pattern of $40.3kHz$

(b) Beam pattern of $40.5kHz$

Fig. 13. The beam pattern of $40.3kHz$ and $40.5kHz$ using optimization and not using optimization.

intensity of the grating lobe and concurrently curtail disturbances in alternate directions to the fullest extent feasible while preserving the energy of the principal lobe.

5.3.3 Wideband Beamforming. To extend the narrowband beamforming optimization to wideband signals, which ensure Visar to transmit audios, we use the codeword of each frequency to weigh the signal for each transducer and sum up the weighted signal of all frequencies to reconstruct the acoustic signal transmitted by each transducer.

To guarantee that the main lobe values of each frequency are consistent and avoid distortion, we add the following penalty terms to the optimization of wideband beamforming:

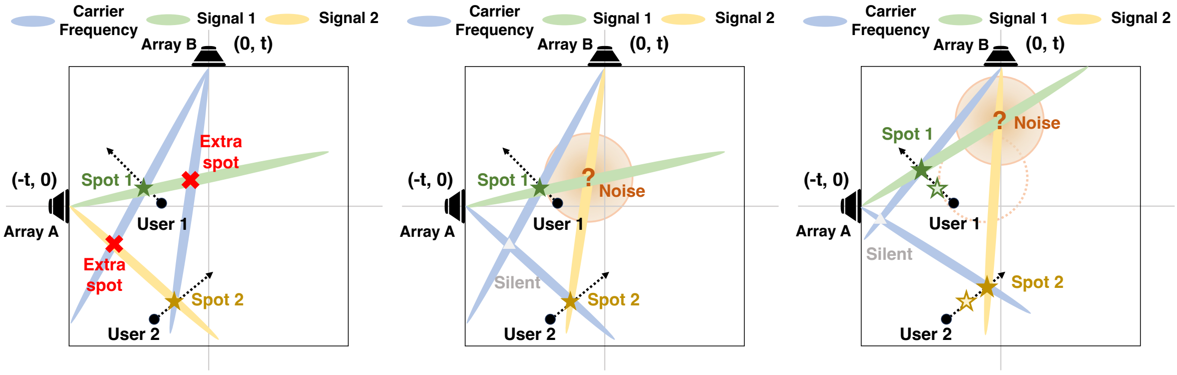
$$L_{wideband} = var\left\{F_i, F_{i+1}, ..., F_{n_f}\right\}$$

where F denotes the frequency bins and n_f is the number of frequencies.

5.3.4 Optimization Solution and Time Consumption. The goal of beam optimization is to optimize the beam weight as well as the spacings for each channel to suppress grating lobes and the interferences caused by multiple beams. We use the gradient descent method (the optimizer is Adam) to gradually find the optimal beamforming beam weights and array spacing that can minimize grating lobes and multi-beam interference while maintaining the main lobe. Note that the beam weight vector is unrelated to the audio content, our method can fix the array spacing and store the beam weights in advance.

We optimize the angles at intervals of 5° from −45° to 45° in turn, which results in 19 cases of array spacing optimization. We take the array spacing that occurs the most times in all 19 cases as the fixed array spacing for the system. After fixing the array spacing, we optimize the beam weights for combinations of directions in turn and store these beam weights (which occupied only 2.44MB of storage space and took 20 minutes to pre-calculate the beam weights). When we need to send to specific directions, we can configure the parameters of beamforming by searching the beam weight table. The computation time only requires calculating the lookup time and hardware delay for setting the beam weights, which is **within 0.004s**.

5.3.5 Verification. To evaluate the effectiveness of beam optimization, we applied the beam weight to the up-converted audio and adjusted the layout of the ultrasound array according to the optimized spacing. We measured the strength of the received signal using a microphone apart at $2m$ from the transducer arrays, with intervals of 5°, and obtained the beam pattern. Fig. 13 shows the case where two beams are sent at different angles, and the grating lobe is effectively suppressed, as well as the leak tends to be weakened, but at the cost of some side lobes being weakly enhanced.



(a) Without spot scheduling (b) With redistribution (c) With adjustment

Fig. 14. The description of spot scheduling, where the blue beam is the carrier wave, the green beam is the up-converted audio played to User 1, and the yellow beam is the up-converted audio played to User 2. In the case of two users, the original solution would generate two extra spots. If the contents are redistributed, the extra spots will be reduced, but there are still audible noises. By adjusting the position of the target spot, we keep the audible noises as far away from the user as possible.

5.4 Spot Scheduling

In multi-user scenarios, practically implementing VISAR also needs to consider the potential problems due to additional sound spots. Since beam optimization only focuses on whether the content played by the beam itself is clean, without considering the inevitable spatial beam crossing in multi-user scenarios, such beam crossing will generate additional sound source points. This cannot rely solely on beam optimization to minimize the interference of these extraneous sound spots to users. Specifically, in scenarios designed to accommodate n users, there will be less than or equal to n^2 intersection points (some beams may not intersect) of auditory rays within the spatial domain. The challenge lies in ensuring that each user perceives sound spots emanating exclusively from their designated direction because the presence of extraneous sound spots can adversely impact the user's auditory discernment and overall listening experience. To address this issue, we introduce a comprehensive **spot scheduling** strategy that encompasses two primary facets: firstly, redistribute content emitted by a transmitter, which reduces the number of irrelevant sound spots, and secondly, adjust the spatial distance between extraneous sound spots and the users.

5.4.1 Transmission content redistribution. VISAR is disturbed by extraneous sound spots, and in this part, we explore strategies to reduce these spots. Consider a scenario with two users, as depicted in Figure 14: An ultrasonic array, Array A, broadcasts up-converted audio signals targeting two distinct directions for the users; Another array, Array B, projects carrier signals along two carriers concurrently. Subsequently, this generates two extraneous sound spots in the vicinity, each bearing significance.

In Figure 14(a), the additional sound spot located at the lower left is instrumental in delivering the audio required by user 2. However, this emission is potentially perceptible to user 1, thus engendering auditory interference. By reconfiguring the distribution of the emissions from the ultrasonic arrays, as illustrated in Figure 14(b), we assign Array B the task of transmitting the carrier wave in one trajectory and the up-converted audio in another. Array A is assigned a similar function. This redistribution eliminates the additional spot proximal to User 1, as the intersecting 40k carriers fail to produce audible sound. Another extraneous spot is a confluence of two up-converted audio signals. This amalgamated sound is characterized as noise with reduced volume, owing to its lower intensity relative to the carrier, thereby minimally impacting the users. In this redistribution paradigm, the interference experienced by the users is markedly attenuated.

We set up an algorithmic framework for facilitating transmission content redistribution so that the system can quickly compute the redistribution outcomes across many users and their respective movements. Considering a scenario involving n users, two arrays will engender the emission of n rays, culminating in generating n^2 sound spots. Under such circumstances, aside from the spots requisite for user engagement, an additional $n_{extra} = n^2 - n$ spots emerge, denoted as $\mathbb{P}_{extra} = \{sp_i | 1 \leq i \leq n_{extra}\}$.

Concurrently, these spots are categorized with specific labels (silent, noise, or meaningful sound). This classification is predicated on the premise that the orientation of beams in both directions dictates the nature of the extraneous spots, with the intersection of these beams presenting the auditory possibilities. Our aspiration in the optimization process is to strategically position silent spots proximate to users while maintaining a maximal distance from spots characterized by noise or meaningful sound. Assuming the presence of user j at the location p_j, we propose the following objective function to refine and optimize the content redistribution process:

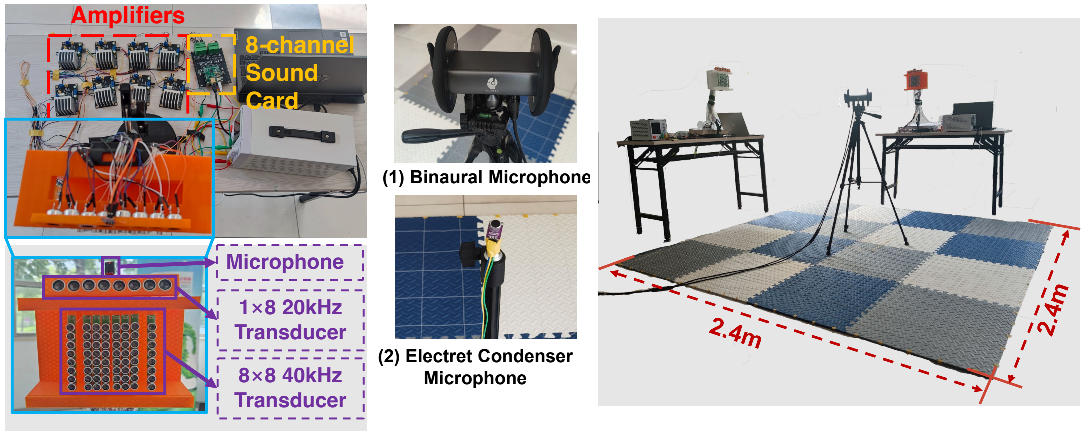
$$\min \ L_{redistribution} = \sum_i \sum_j \frac{\sigma}{\|sp_i - p_j\|_2} \quad , \quad \sigma = \begin{cases} \sigma_{silent} & \text{if } sp_i \text{ is silent,} \\ \sigma_{noise} & \text{if } sp_i \text{ is a noise,} \\ \sigma_{meaningful} & \text{if } sp_i \text{ is a meaningful spot.} \end{cases}$$

where σ is a penalty term related to extraneous spot content. Through the simulation results of different situations, we empirically determined that σ_{silent}, σ_{noise}, and $\sigma_{meaningful}$ are 0.001, 0.01, and 0.1 respectively. Such a penalty term can constrain us to achieve the requirement that the distances between the noise and meaningful spots are as far away from the user as possible. We choose the content redistribution result that minimizes $L_{redistribution}$. This result describes the direction in which each array should play the carrier wave or the up-converted audio.

We use an enumeration method to find the optimal redistribution result. Specifically, for N users, the decision of which array to send carriers to the users and which array to send up-converted audio to the users needs to be determined N times. Then, we need to calculate the loss function of redistribution optimization for 2^N cases one by one and take the allocation configuration with the lowest loss. For 4 users, only 16 operations are needed, which takes **less than 0.0002s**.

5.4.2 Extraneous spots' locations adjustment. In the preceding section, we delineated a methodology for diminishing the incidence of excessive sound spots by redistributing the playback content across various arrays in each direction. Nevertheless, this redistribution might not adequately account for the spatial separation between each spot and the user. Despite the redistribution, this distance between spots and users remains constant, potentially leading to auditory disturbances perceived by the users, as shown in Fig. 14(b).

To address this, we introduce a method for adjusting these extraneous spots. Recall from Sec. 4 that the diameter of each spot approximates 30 cm. This implies that a modest repositioning of the spot towards the user could only have a little impact on their auditory perception of themselves. We posit that if a spot sp_j is projected towards a user at p_j, it should satisfy the condition $d_{min} \leq \|sp_j - p_j\|_2 \leq d_{max}$, ensuring the user can discern the auditory content emanating from the spot with clarity. As shown in Fig. 14(c), Altering the position of the projected spots in this manner consequentially modifies the placement of extraneous spots within the entire area. We aim to maximize the distance between the unwanted sound-emitting spot and the user. Consequently, our optimization target involves adjusting the separation between each user's projected spot and the user. This can



(a) Transmitter (b) Receiver (c) Test Scenario

Fig. 15. The experiment setup of Visar.

be quantitatively expressed through the following objective function:

$$\max_{sp_j} \; L_{adjustment} = \sum_i \sum_j \|sp_i - p_j\|_2$$

$$s.t. \quad \begin{cases} sp_i \text{ is a noise or meaningful spot} \\ d_{min} \leq \|sp_j - p_j\|_2 \leq d_{max}, \forall j. \end{cases}$$

where sp_i is the location of the extraneous spot mentioned in Sec. 5.4.1. The constrained spatial adjustment, including d_{min} and d_{max}, is discussed in Sec. 6.2

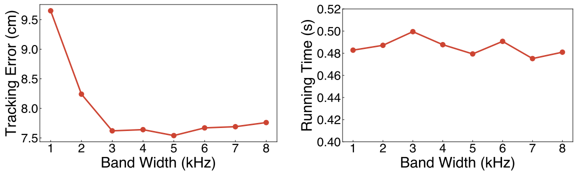
We use gradient descent to search for the optimal distance between the spot and the user, to maximize the distance between the extraneous spot and the user. Furthermore, we use the time it takes to calculate the adjustment once to evaluate the time complexity of the algorithm, and the time required for the algorithm to calculate the optimal distance between different users and their respective spots is **within 0.062s**.

Summary. The calculation time required by optimizations is within 0.07s (including beam optimization, transmission content redistribution, and extraneous spots' locations adjustment), which can meet the standards of user movement.

6 EVALUATION

6.1 Experiment Setup

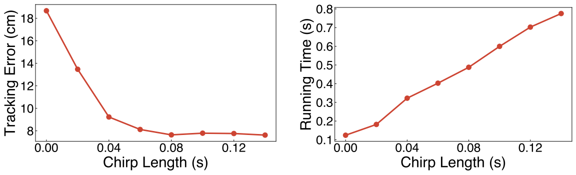
Prototype: Our proposed Visar consists of two groups of hardware setup. Each hardware setup (*transmitter side*) (shown as Fig. 15(a)) contains a microphone module ADMP401 [50], and an ultrasound transducer array which is composed of 8×8 transducers with central frequency of $40kHz$ (TR4010HC-1) and 1×8 transducers with central frequency of $20kHz$ (EU16AOF20H12T). Each column of ultrasound transducers is connected with a class D amplifier OPA541 [30] which supports a maximum power of $50W$. An 8-channel sound card is used to play the audio to achieve beamforming, and the microphone is also connected to realize synchronous tracking. For *receiver side* (shown as Fig. 15(b)), we chose two receiver settings for different aspects of measurement. One is to use an electret condenser microphone whose amplifier is disabled to avoid hardware non-linearity to record



(a) Tracking error

(b) Running time

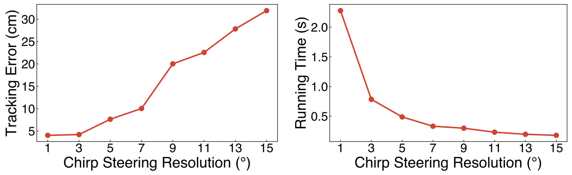
Fig. 16. Impact of chirp band width.



(a) Tracking error

(b) Running time

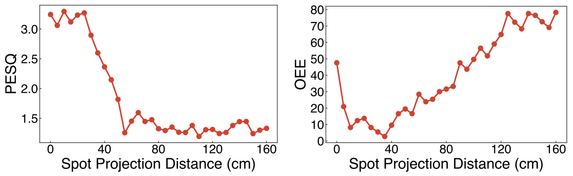
Fig. 17. Impact of chirp length.



(a) Tracking error

(b) Running time

Fig. 18. Impact of chirp steering resolution.



(a) PESQ

(b) OEE

Fig. 19. Impact of spot projection distance.

the acoustic signal in the whole space to evaluate the distribution of sound energy. Another setting is to simulate the Angle-of-Arrival (AoA) heard by a user using a dummy head (binaural) microphone (Binal 2 [22]).

Deployment: Our system is deployed on a $5m \times 5m$ floor in a closed and unobstructed room, where the active area is $2.4m \times 2.4m$ in the center, and the measurement resolution is $0.1m \times 0.1m$ (shown as Fig. 15(c)). The transducer array is placed $1m$ away from the side due to the focused beam of the ultrasound transducer.

Testing sounds: We played 50 English sentences generated from Text-to-Speech (TTS) generator (30 pieces) and vocal recording clips (20 pieces). Each sentence has between 10 to 50 words and includes both female and male voices. Each sentence was recorded for 10 times.

Performance metric: To evaluate the performance of Visar, we defined 5 metrics for evaluating different aspects.

1) Tracking error: The tracking error is used to evaluate the error between the location Visar track and the ground truth location. Assuming that the position tracked by Visar is at p_{track} and the user's real position is at p_{truth}, the tracking error can be calculated as $\|p_{track} - p_{truth}\|_2$

2) Signal-to-disturbance ratio (SDR): SDR is defined as the ratio of the total energy at the target zone ($20cm \times 20cm$) to the energy at the strongest area besides targets. The larger the SDR, the more pronounced the sound effect produced by the spot. The goal of Visar is to project a spot at the target point and none elsewhere.

3) Signal-to-noise ratio (SNR): SNR is employed to assess speech quality objectively, quantifying the noise with the desired signal. We calculate SNR between a reference clean signal, denoted as x, and the received audible signal, designated as $\hat{x}$, and SNR is defined as: $\mathrm{SNR}(x, \hat{x}) = 10 \log_{10} \left(\frac{\|x\|_2^2}{\|x - \hat{x}\|_2^2} \right)$. Notably, a scale-invariant SNR is utilized to minimize the effects of scaling on the assessment.

4) Perceptual evaluation of speech quality (PESQ): PESQ is a standardized objective method for assessing speech quality and generates a score ranging from −0.5 to 4.5 [1]. PESQ is based on a psychoacoustic model that simulates the human auditory system's response to audio signals, and the score reflects the sound quality.

5) Orientation estimation error (OEE): To evaluate the performance of using Visar to guide a direction, we used a dummy head (binaural) microphone to calculate the AoA of the sound spot produced by Visar. The orientation estimation error (OEE) is the difference between the target angle and the calculated AoA.

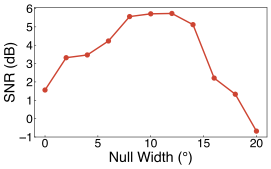


Fig. 20. SNR under impact of null width

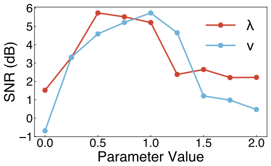
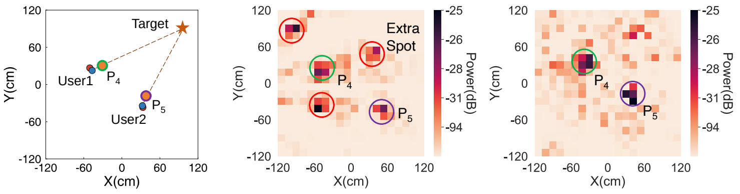


Fig. 21. SNR under impact of λ and v



(a) Guidance (b) w/o Optimization (c) w/ Optimization

Fig. 22. An example of an AR application that guides 2 users to find an item. Red dots represent the users' real positions, while blue dots represent the positions tracked by VISAR. With our optimization and scheduling scheme, VISAR can project two distinct sound (outlined by green and purple circles respectively) to tell the users where the object is.

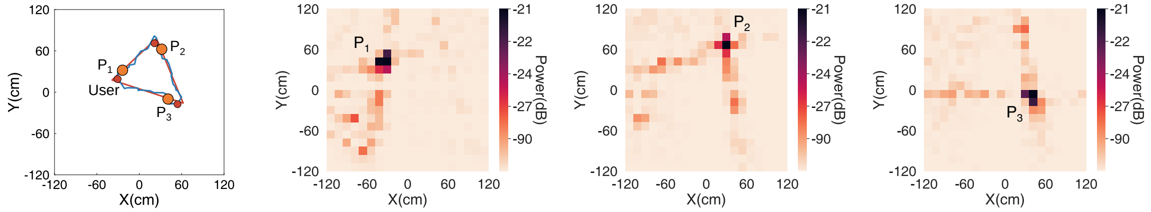
6.2 Micro Benchmark

6.2.1 Chirp Bandwidth. To evaluate the impact of the bandwidth of chirp on the tracking accuracy of VISAR, we set the FMCW signal's bandwidth to $1kHz$ to $8kHz$. Fig. 16(a) shows the tracking error under different bandwidths. We observe that when the bandwidth is less than $3kHz$, the tracking error rate is high because the narrow bandwidth makes the peak generated by the echo signal not prominent enough. When the bandwidth is greater than or equal to $3kHz$, the tracking error fluctuates in a small range of around $7.5cm$. We chose a bandwidth of $4kHz$ because excessive bandwidth may generate audible noise due to nonlinearity, such as $24kHz$ and $40kHz$ generating a high-frequency noise of $16kHz$. Different bandwidths have little impact on the running time of VISAR, as shown in Fig. 16(b), the relationship between bandwidth and the running time is approximately a straight line.

6.2.2 Chirp length. Chirp length is the duration of one cycle of an FMCW signal. Generally, the longer the chirp length, the stronger the tracking anti-interference ability. However, it also brings longer tracking time. We vary the chirp length from 0.02s to 0.16s to see the tracking error of VISAR. In Fig. 17(a), when the chirp length is longer than 0.08s, the accuracy of tracking can be guaranteed. However, at the same time, we need to ensure that the tracking time is as short as possible to match human movement. In Fig. 17(b), a running time of 0.5s is acceptable, so we chose a chirp length of 0.1s.

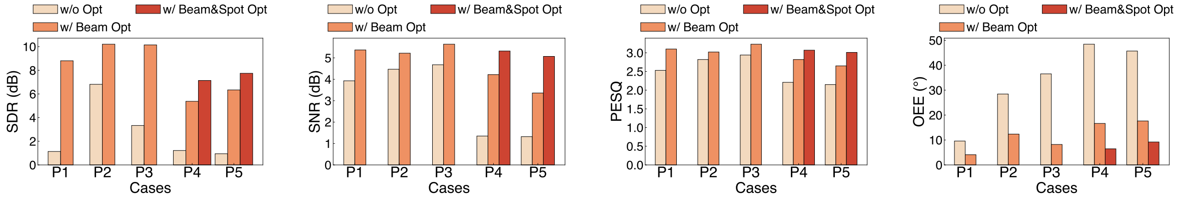
6.2.3 Chirp steering resolution. VISAR uses phased array scanning to determine the user's orientation. Fig. 18(a) shows that the finer the chirp steering resolution, the higher the tracking accuracy. If the scanning interval is 1 degree, the tracking error will decrease to about 5cm. However, As shown in Fig. 18(b), excessive resolution leads to extremely long running time, and VISAR will lose real-time performance. Therefore, setting the scanning accuracy at 5 degrees is a choice that balances tracking error and running time.

6.2.4 Spot projection distance. When using VISAR, users should be able to successfully identify the direction of the sound source and hear the sound content as much as possible. We use PESQ and OEE to measure speech intelligibility and source direction recognizability, respectively. We used a binaural microphone whose structure is close to the human ear for sound recording. We projected the sound source at different distances from the binaural microphone to test the metrics of the received sounds. From Fig. 19(a), we can see that the received signal is easiest to understand when the sound source is 0-40cm from the binaural microphone. The audio becomes no longer evident as the sound source gradually moves away. At the same time, Fig. 19(b) indicates that if the sound source is too close, the sense of direction expressed is not apparent, so an appropriate relative distance between the sound source and the person will make it easier for people to distinguish the direction of the sound and hear



(a) Trajectory (b) Heatmap at P_1 (c) Heatmap at P_2 (d) Heatmap at P_3

Fig. 23. An example of an AR application that guides along a specific trajectory. VISAR projects a sound to tell a user the direction. Red dots represent the users' real positions, while the blue connection represent the positions tracked by VISAR. Three points are selected to show the corresponding heat map.



(a) SDR (b) SNR (c) PESQ (d) OEE

Fig. 24. The overall performance of VISAR regarding SDR, SNR, PESQ and OEE.

it. We chose 20cm as the spot projection distance for VISAR. In addition, we determine the value of d_{min} to 10cm and the value of d_{max} to 30cm in Sec. 5.4 to ensure the user's listening experience.

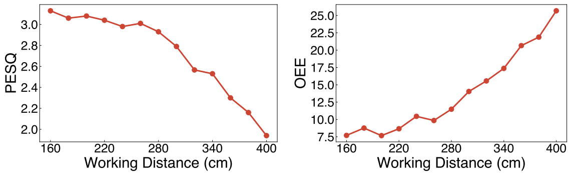
6.2.5 Null width. VISAR implements a wide-nulling beam optimization technique to mitigate interference from multiple directions, so we conduct experiments to evaluate the effects of varying null widths, ranging from 0° to ±20°. Figure 20 presents SNR concerning different null widths. It is observed that a narrow null width results in a suboptimal SNR, specifically 5.5dB. This diminished SNR can be attributed to the spatial separation between the ears. Also, an increase in null width leads to the same results, evidenced by the decline in performance within the main lobe. Based on these findings, a null width of ±10° is selected.

6.2.6 Beam optimization weights. We investigate the impact of the weights λ and ν, as outlined in Equation 3. A grid search is performed to ascertain the optimal parameters, setting λ and ν in the range of [0,2] with a step increment of 0.25. Figure 21 illustrates that both parameters have a maximum value in SNR within a given range, which leads us to choose the parameter value corresponding to this maximum SNR. Consequently, optimal weights of $\lambda = 0.5$ and $\nu = 1$ are selected for the optimization process.

6.3 Overall Performance

During our evaluation, we tested the tracking and sound projection performances of VISAR under two experiment scenarios, as shown in Fig. 23(a) and Fig. 22(a), respectively.

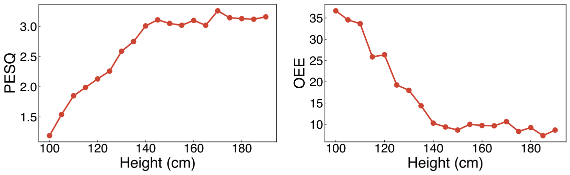
In the first experiment, a single user walks in a triangle trajectory guided by the projected sound source in the activity area. At each turning point of the trajectory, VISAR projects a sound source towards the next vertex of the triangle. In the second experiment, two users stand still and try to find an item in a specific location. The user trajectory is the straight line between the two users and the target point. VISAR projects sound in the directions toward the target location to lead the users to find the target. Fig. 23(a) and Fig. 22(a) show that VISAR can perform device-free tracking accurately, with an average tracking error of 7.5 cm. Fig. 23(b) - Fig. 23(d) shows 3 heat maps



(a) PESQ

(b) OEE

Fig. 25. Impact of working distance.



(a) PESQ

(b) OEE

Fig. 26. Impact of user height.

Table 1. Performance comparison with Meta-Speaker.

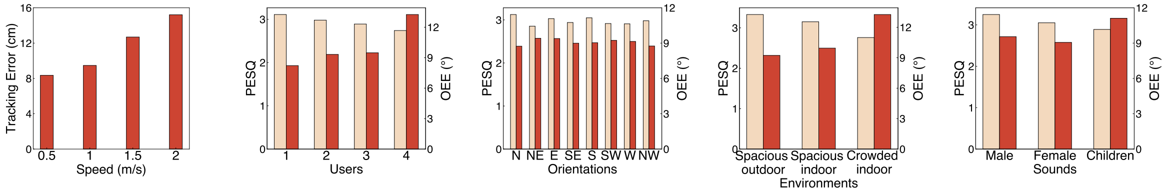
System	Number of Users	SDR	SNR	PESQ	OEE
Meta-Speaker	1	2.07	5.63	3.27	8.24
Visar	1	9.41	5.24	3.12	8.57
Visar	3	7.32	3.63	2.83	10.15

of the projected sound source power in Trajectory 1, demonstrating Visar's ability to accurately project a single sound source in the whole space. Fig. 22(b) and Fig. 22(c) shows the heat map of the power of different sound sources in Trajectory 2, and with our optimization and scheduling scheme, the extra spots generated by the grating lobe interaction (upper left) and the extra points generated without redistribution are well suppressed, which indicates that Visar strives to avoid additional leaks within the user's auditory range, while accurately projecting two distinct content sound sources.

We further measure and compare the metrics of the projected sound sources at the 5 selected locations in Fig. 23 and Fig. 22 without any optimization, with the addition of beam optimization, and with the addition of beam optimization and spot scheduling, as shown in Fig. 24. It can be found that beam optimization can significantly improve SDR because it eliminates abundant sound source points that grating lobes may generate. The increase in SNR and PESQ and the decrease in OEE indicate that beam optimization also reduces mutual interference. The optimization of points has also improved various indicators because redistributing audio content reduces additional sound source points, resulting in an improvement in SDR. In contrast, the other three indicators are due to noise being moved away, allowing users to hear clean audio.

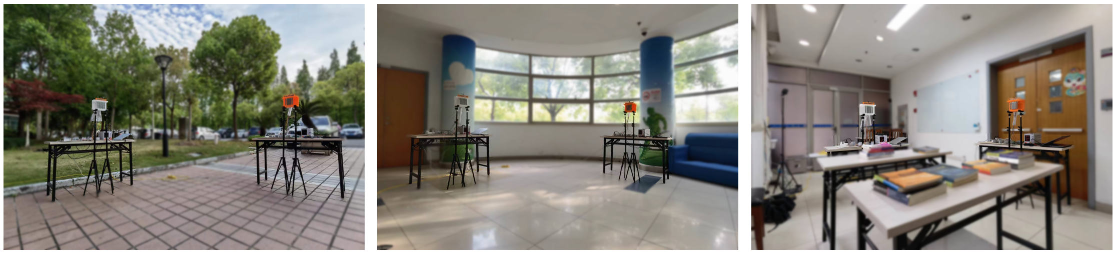
We also compare the performance with Meta-Speaker [63], which realizes movable virtual sound source projection by mechanically rotating two transducer arrays. We all use the prototype of 8 × 8 transducer arrays. To simulate the effect of Meta-Speaker, we physically rotate the two arrays. We calculate the mean value of each metric for different number of users, and as shown in Tab. 1, it can be found that the SDR of Meta-Speaker is lower than that of the Visar, which suggests that other individuals in the space are more likely to hear the leakage, due to the intersection of the grating lobes. In the case of 1 user, Meta-Speaker has a slightly higher SNR, PESQ, and OEE due to the physical rotation that allows the arrays to emit more focused beam, but Visar using digital beam steering is also quite available. However, in some special cases, Meta-Speaker does not optimize away additional audible sources, which can affect the user's listening quality. Since Meta-Speaker does not consider multiple users, we only test Visar in the case of multiple users. We take the 3-user case as an example, for a more detailed multi-user performance in Sec. 6.3.4. We can find that the metrics are only slightly reduced, which proves the feasibility of Visar support for multiple users.

6.3.1 Impact of working distance. To evaluate the maximum working distance, Visar projects 3 sound spots and we measure PESQ and OEE corresponding to the spot farthest from the two arrays. The distance on the horizontal axis refers to the distance between the spot and the farthest array. As shown in Fig. 25, we can find that after 2.8*m*, the PESQ decreases significantly and is less than 2.5, and the OEE also increases considerably and is greater



(a) Mobility of users (b) Number of users (c) Orientation of users (d) Environments (e) Sound types

Fig. 27. Impact of various practical influencing factors.



(a) Spacious outdoor (b) Spacious indoor (c) Crowded indoor

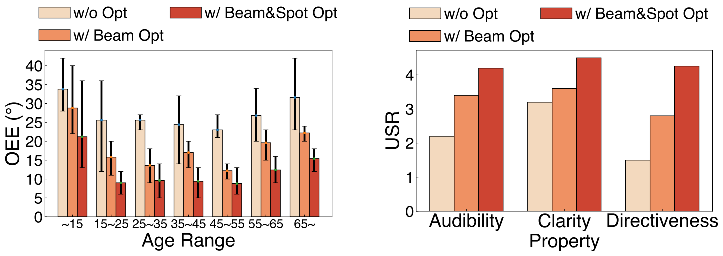
Fig. 28. Different environments scenes.

than 15°, which may mislead the user's movement at such a distance, suggesting that **the maximum working distance supported by the system is about 2.8m**, which is already suitable for some practical scenarios, such as museums, libraries, etc. Because the working distance is related to the power, a larger size of the transducer array with greater transmission power is required to support the transmission of ultrasonic waves to longer distances.

6.3.2 Impact of user height. To evaluate the impact of users' height, we conducted experiments. Specifically, we fixed the height of the sound spot to $2m$, placed the binaural microphone at different heights, and measured PESQ and OEE. As shown in Fig. 26, our system can support **heights from 1.35m to 1.9m** in which PESQ is greater than 2.5 and OEE is less than 15°. This can cover the height range of most users [53]. For some special cases, such as children or disabilities, one possible solution is to add additional arrays dedicated to them. Another potential solution is to use a 64-channel independent speaker array so that the array can perform beamforming at height. However, it will introduce additional costs and we leave it in our future work.

6.3.3 Impact of mobility of users. To verify that Visar can achieve dynamic tracking of users, we tested the tracking error of users at different moving speeds, as shown in Fig. 27(a). Since it is difficult for users to control a specific walking speed, we loaded a mechanical trolley with the binaural microphone. We controlled it to move from the lower left end of the field to the upper right end of the field at different speeds (0.5m/s to 2m/s), while the normal walking speed of a normal person is around $1.2m/s$ [5]. It was found that under normal moving speed, Visar can maintain good performance and accurately project sound sources. But if the movement speed becomes faster, the tracking error of Visar increases slightly.

6.3.4 Impact of number of users. We evaluated the impact of different user numbers on Visar and observed the results shown in Fig. 27(b). When the number of users is small, the users' sense of hearing and direction recognition can be well guaranteed. However, if the number of users increases, the distance between different spots will be shorter due to space limitations, affecting user recognition.



(a) OEE

(b) User Ratings

Fig. 29. The user study contains OEE measurement and USR from three aspects: audibility, clarity, directiveness

6.3.5 Impact of orientation of users. Visar supports various user orientations, as shown in Fig. 27(c). We evaluated eight directions, such as north, northeast, east, etc., and found that the user's orientation has no impact on the system. This is because the sound spot is projected to the front of the user's head, which is not affected by the user's orientation.

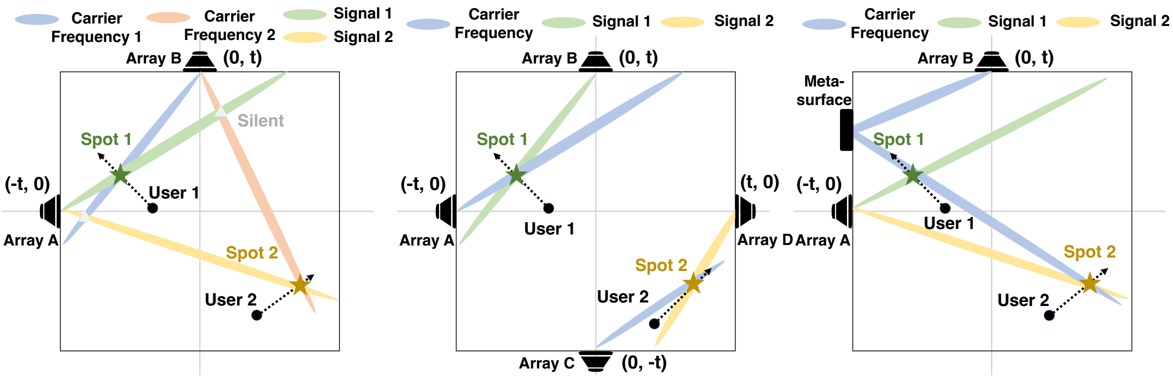
6.3.6 Impact of environments. We evaluated Visar in three environments as shown in Fig. 28: spacious outdoor, spacious indoor, and an indoor space with several decorations. In a crowded indoor environment, our experimental scenario is a 6m * 8m office with some occlusions including tables and chairs, with Visar placed at the center of two adjacent walls and projected onto users sitting on chairs. From Fig. 27(d), we can see that the effect is better in open environments, but if it is more crowded, the effect will slightly decrease due to multipath effects or signal attenuation caused by excess items.

6.3.7 Impact of sound types. We also evaluated the impact of different sounds and observed the results shown in Figure 27(e), which includes three types of male, female, and child voices. The results showed that male voices were better than other types because lower frequencies were more dominant and this feature fits the band-limited ultrasonic transducer. Despite this, for all types of speech, *PESQ* still reaches around 3, and OEE is around 10, which verifies Visar's ability to play speech.

6.4 User Study

We invited 35 volunteers, including a wide age range (8 to 73 years old) to evaluate Visar. We asked the volunteers to stand in the test area and reach preset target points unknown to them. Visar projects the sound sources to guide the multi-users to this location. When the user reaches the specified location, the game ends. We recorded the OEE of the system and invited users to rate audibility, clarity, and directiveness on a scale of 1 to 5, with higher scores signifying superior experience. We choose the mean value as user study ratings (USR).

The results of the user study integrating feedback from all 35 volunteers are shown in Fig. 29(a) and Fig. 29(b). It can be noticed that after both beam optimization and spot scheduling, OEE is reduced, indicating that the direction recognition is improved, but there are differences among diverse individuals. Specifically, the OEE of children under the age of 15 is around 20 since some of them are much shorter than the minimum height (1.35m) that the system can support. In addition, the OEE of middle-aged and elderly people over the age of 55 is also higher than that of young adults due to the decrease in their hearing. The audibility and clarity of the system improved with the proposed optimization scheme, and the directiveness was also augmented due to the reduction of interference, proving the effectiveness of Visar, which was acceptable for most people.



(a) Using multiple carrier frequencies (b) Using more transducer arrays (c) Using reflective metasurfaces

Fig. 30. The possible cases of using potential solutions to alleviate more extraneous spots.

7 DISCUSSION

Visar innovatively takes use of air nonlinearity to realize real-time virtual sound spots projection for augmented reality. Next, we discuss the performance and limitations of Visar in some specific scenarios:

1) Capability of more users. With more users, the number of extraneous spots will increase significantly. We believe that in some particular situations, the following solutions can alleviate this problem in more user scenarios.

- **Using multiple carrier frequencies.** The premise for this method to effectively reduce extraneous spots is that the differential frequency signal generated by two carrier signals based on air nonlinearity is inaudible. The human cannot hear the sound over $20kHz$ [2], the carrier frequencies fc_a and fc_b should satisfy $|fc_a - fc_b| > 20kHz$. In this case, as long as the number of carriers is the same as the number of users, there is a chance to realize interference-free spots. Fig. 30(a) shows the case of using two different carrier frequencies to project spots to two users. The reason why there is silence in the irrespective intersections of the beams is that the signal frequency difference between the two beams is greater than $20kHz$ and the users cannot hear it.
  However, it meets several practical problems: *a) Frequency response:* Due to the narrow bandwidth of ultrasonic transducers, it requires transducers of different central frequency bands to support multiple carrier frequencies. As a result, the structure of the entire array will significantly increase. *b) Half-wavelength spacing law:* The spacing of the transducer array is required to be less than half the wavelength of the emitted signal to not produce grating lobes. For a transducer array with a fixed spacing, an increase in the frequency of the emitted signal will cause the signal half-wavelength to fall below that spacing, resulting in the grating lobes [3]. Furthermore, the number of grating lobes increases in the signal frequency, leading to more extraneous spots.
- **Using more transducer arrays.** Arranging more transducer arrays in the field can to some extent reduce the extraneous spots caused by beam intersection, as more arrays can enhance the flexibility of beam distribution, including carrier and upconverted sound waves. As shown in Fig. 30(b), in the case of two additional arrays, each of the two arrays can project the sound spot for one user without interference.
  This method faces some practical challenges: *a) Extraneous spots:* Even with multiple arrays, beam crossings will still exist, and extraneous spots will inevitably be generated in the area. Therefore, it still needs further optimization to mitigate this issue. *b) Additional cost and deployment:* More arrays require more hardware devices such as sound cards and groups of amplifiers, which will bring additional cost and deployment difficulties.

- **Using reflective metasurfaces.** Another option that may work is the deployment of reflective acoustic metasurfaces [32, 76]. The acoustic metasurface has two advantages: first, it can support flexible outgoing wave directions after reflection. Second, it can enhance the output energy of the outgoing wave compared to general reflective surfaces (e.g., walls). Specifically, one metasurface is placed at the edge of the field and the beam can be sent to the metasurface to achieve the predetermined target through the reflection to avoid unintended beam intersection, as shown in Fig. 30(c). It has the advantages of reducing layout complexity and saving costs to cover the whole space.
  However, it faces several challenges and limitations: *a) Metasurface size:* The area of the metasurface should be large enough to cover most of the sound wave emitted by the transducer array. This limits the portability and easy deployment of the system. *b) Challenging design:* Passive metasurface design is challenging because it requires a combination of different cell designs to allow the metasurface to support a feasible frequency bandwidth. Active metasurface is expensive as it requires additional electronic hardware.

In comparison, we propose a spot scheduling method to rearrange the content of these spots. It redistributes the beams of the two arrays and fine-tunes the position of the sound spot to minimize the impact of extraneous spots on users in real time, which has the advantage of being easy to deploy, effective in cost, and suitable for different scenarios. In some special cases, we could combine the spot scheduling method with the above solutions to achieve better performance or coverage for specific applications in the future.

2) Varying signal power. The attenuation of acoustic signal propagation follows power-law frequency-dependent acoustic attenuation [24], and the formula is $P(d + \Delta d) = P(d)e^{-\alpha(\omega)\Delta d}$. Among them, d is the initial position, Δd is the propagation distance relative to the initial position, P is the pressure, ω is the angular frequency, and $\alpha(\cdot)$ is the attenuation coefficient. It can be found that the power at different distances does change, but as shown in Fig. 25, within our maximum working range of 2.8m, the PESQ and OEE indicators can show that users can clearly hear the sound emitted by the spot and correctly determine the direction. Therefore, the energy has very little impact on hearing. To keep spots at different distances maintaining the same power, we can adjust the power of multiple beams. Without loss of generality, we take two beams as an example. Assume that one beam is transmitted to users at a distance Δd_1, and another beam is transmitted to users at a distance Δd_2. To make $P_{beam1}(d + \Delta d_1) = P_{beam2}(d + \Delta d_2)$, we can calculate the initial power $P_{beami}(d)$ of different beams, and assign the proportion of the initial power to the power ratio of the beam, thereby minimizing the difference in sound volume heard by users at different distances.

3) Maximum working Distance. In some specific scenarios, a longer distance is required, and the maximum working distance depends on the size of the transducer array and the energy that each transducer can emit. We derive the theoretical maximum distance step by step:

- **Acoustic Beamforming:** Assume that the sound pressure level (SPL) of the sound wave emitted by a transducer is P_t, and the transducer array contains n_t transducers, which need to support n_u users. According to the amplitude superposition property of beamforming [7], the SPL of one of the beams in the multi-beam transmitting process is $P_{beam} = P_t + 20\log_{10}\frac{n_t}{n_u}$.
- **Acoustic transmission attenuation:** Assuming that we need to support a maximum distance of d_{max} in a real scenario, we calculate the SPL after the acoustic wave has been transmitted to d_{max} as P_{dmax}. According to the law of attenuation [24], we have $P_{dmax} = P_{beam} * e^{-\alpha\omega^{\eta}d_{max}}$, where ω is the angular frequency and α and η are real which can be obtained from the acoustic attenuation due to sound wave dispersion [3] and the absorption of sound wave by air [4].
- **Nonlinear effect:** We consider the SPL of an audible source generated by two sound waves with P_{dmax}. According to the theory of nonlinearity [19] and our experiments in Sec. 4.1, it can be empirically assumed that some of the energy will be lost, and the SPL of an audible source $P_{au} = P_{dmax} - P_{non}$, where $P_{non} \approx 15dB$.

- **Human ear audibility:** P_{au} needs to be perceived by the human ear. The SPL of a weak sound wave that the human ear can perceive is $20dB$ [58], in order to let the human ear hear comfortably, we require $P_{au} \geq 30dB$.

From above, We can get the theoretical maximum working distance

$$d_{max} = \frac{\ln \frac{P_{au}+P_{non}}{P_{beam}}}{-\alpha\omega^{\eta}}$$

To achieve a working distance of $3.6m$, the size of the transducer array is about 16×16, and the energy emitted by each transducer is 146dB, which is commercial off-the-shelf [41]. To support a larger area, we suggest distributed placement of the transducer array, and this arrangement method we leave for future work.

4) Limited bandwidth. Typically, the bandwidth of sound waves emitted by ultrasound transducers is below $4kHz$ (shown as Fig. 7), which limits the frequency range of audible sound that can be recovered. Although this is enough for the restoration of vocal sounds, we hope that VISAR can further support the music. It is possible to combine multiple transducers of various central frequencies to achieve wider bandwidth.

5) Focused beam. Due to the structure of the ultrasound transducer, it always emits a focused beam, and the element spacing is larger than half-wavelength of the sound wave, thus limiting the coverage. VISAR effectively suppresses the affection of grating lobes caused by wide spacing. For the focused beam, one possible solution is to use an acoustic metasurface to redirect the output beam.

6) Environmental noise. a) Impact on spot generation. Most of the environmental noise is in the low-frequency band, and the system's operating frequency band is around 40kHz, so the generation of sound spots will not be affected. **b) Impact on user hearing.** General environmental noise is random and comes from a distance. Since our sound spot is very close to people (about 30cm), the user can still clearly hear the sound from the spot and identify it.

7) Non-line-of-sight. If there is an obstruction blocking any beam, it will be difficult to form a sound source in the current system. However, we can exploit natural reflectors [45] or metasurfaces [32] in the environment to bypass obstacles. Specifically, we can calculate the reflection angle between the array, the reflector, and the user, so that the array emits the beam to the reflector at a predetermined angle, and then the beam is reflected by the reflector to the target sound source formation position, thereby forming the sound source normally.

8) Multipath. The multipath effect may cause new additional spots to be generated due to beams reflected by the environment. For complex environments, the way to deal with multipath is to perform channel estimation and flexibly control the volume and phase based on the channel estimation results.

9) Users' paths intersection. if in some complex scenarios, such as when users' paths intersect at nearby points, we need to distinguish the next direction of each user. To this end, we can introduce acoustic-based user authentication methods to distinguish between multiple users, such as gait authentication [68], respiration authentication [8], heartbeat authentication [62], etc. We will leave it to future work.

8 CONCLUSION

In this paper, we present VISAR, a novel virtual sound spot projection system to guide directions for acoustic augmented reality applications using air nonlinearity. We introduce the feasibility of reproducing fine-grained audible sound spots from ultrasound through the air. VISAR can achieve $7.83cm$ precision control of sounding area manipulation, and the orientation error estimation error reaches 10.06°, which indicates effectiveness on augmented reality applications, such as navigation and finding items.

ACKNOWLEDGMENTS

We would like to express our gratitude to all anonymous reviewers for their constructive comments. This work is supported in part by the NSFC (61936015, 62072306).

REFERENCES

[1] 2001. *BS.1387: Method for objective measurements of perceived audio quality.* ITU. https://www.itu.int/rec/R-REC-BS.1387-1-200111-I/en

[2] Apple. [n. d.]. Airpods Pro (2nd generation). https://www.apple.com/airpods-pro/.

[3] Keith Attenborough. 2014. Sound propagation in the atmosphere. *Springer handbook of acoustics* (2014), 117–155.

[4] HE Bass, H-J Bauer, and LB Evans. 1972. Atmospheric absorption of sound: Analytical expressions. *The Journal of the Acoustical Society of America* 52, 3B (1972), 821–825.

[5] Richard W Bohannon and A Williams Andrews. 2011. Normal walking speed: a descriptive meta-analysis. *Physiotherapy* 97, 3 (2011), 182–189.

[6] David Boulinguez and André Quinquis. 2002. 3-D underwater object recognition. *IEEE journal of oceanic engineering* 27, 4 (2002), 814–829.

[7] Paolo Castellini and Milena Martarelli. 2008. Acoustic beamforming: Analysis of uncertainty and metrological performances. *Mechanical systems and signal processing* 22, 3 (2008), 672–692.

[8] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2017. BreathPrint: Breathing acoustics-based user authentication. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 278–291.

[9] 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*. 53–59.

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

[11] Bruce H Deatherage, Lloyd A Jeffress, and Hugh C Blodgett. 1954. A note on the audibility of intense ultrasonic sound. *The Journal of the Acoustical Society of America* 26, 4 (1954), 582–582.

[12] Focusonics. [n. d.]. Focusonics Directional Speakers. https://www.focusonics.com/. Accessed on May 1, 2023..

[13] Yongjian Fu, Shuning Wang, Linghui Zhong, Lili Chen, Ju Ren, and Yaoxue Zhang. 2022. SVoice: Enabling Voice Communication in Silence via Acoustic Sensing on Commodity Devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 622–636.

[14] Yongjian Fu, Yongzhao Zhang, Yu Lu, Lili Qiu, Yi-Chao Chen, Yezhou Wang, Mei Wang, Yijie Li, Ju Ren, and Yaoxue Zhang. 2024. Adaptive Metasurface-Based Acoustic Imaging using Joint Optimization. In *The 22nd ACM International Conference on Mobile Systems, Applications, and Services*.

[15] Woon-Seng Gan, Jun Yang, and Tomoo Kamakura. 2012. A review of parametric acoustic array in air. *Applied Acoustics* 73, 12 (2012), 1211–1219.

[16] Zhihui Gao, Ang Li, Dong Li, Jialin Liu, Jie Xiong, Yu Wang, Bing Li, and Yiran Chen. 2022. Mom: Microphone based 3d orientation measurement. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 132–144.

[17] Olav Rune Godø, Kenneth G Foote, Johnny Dybedal, and Eirik Tenningen. 2009. Observing Atlantic herring by parametric sonar. *The Journal of the Acoustical Society of America* 125, 4 (2009), 2718–2718.

[18] Corentin Guezenoc and Renaud Seguier. 2020. HRTF individualization: A survey. *arXiv preprint arXiv:2003.06183* (2020).

[19] Mark F Hamilton. 1986. Fundamentals and applications of nonlinear acoustics. *VOLUME HI: BACKGROUND MATERIALS* (1986), 82.

[20] Hao Han, Shanhe Yi, Qun Li, Guobin Shen, Yunxin Liu, and Ed Novak. 2016. AMIL: Localizing neighboring mobile devices through a simple gesture. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 1–9.

[21] S. Haykin. 1985. *Array signal processing*.

[22] HEADREC. [n. d.]. Binal 2 datasheet. https://headrec.com/products/binal-two.

[23] Jarmo Hietanen, Pentti Mattila, Jyrki Stor-Pellinen, Fabio Tsuzuki, H Vaataja, Ken Sasaki, and Mauri Luukkala. 1993. Factors affecting the sensitivity of electrostatic ultrasonic transducers. *Measurement Science and Technology* 4, 10 (1993), 1138.

[24] Sverre Holm et al. 2019. *Waves with power-law attenuation*. Vol. 714. Springer.

[25] Holosonics. [n. d.]. Holosonics Audio Spotlight Series. https://www.holosonics.com/. Accessed on May 1, 2023..

[26] Hongmei Hu, Lin Zhou, Hao Ma, and Zhenyang Wu. 2008. HRTF personalization based on artificial neural network in individual virtual auditory space. *Applied Acoustics* 69, 2 (2008), 163–172.

[27] Wenchao Huang, Yan Xiong, Xiang-Yang Li, Hao Lin, Xufei Mao, Panlong Yang, and Yunhao Liu. 2013. Accurate indoor localization using acoustic direction finding via smart phones. *arXiv preprint arXiv:1306.1651* (2013).

[28] Victor F Humphrey, Stephen P Robinson, John D Smith, Michael J Martin, Graham A Beamiss, Gary Hayman, and Nicholas L Carroll. 2008. Acoustic characterization of panel materials under simulated ocean conditions using a parametric array source. *The journal of the*

acoustical society of America 124, 2 (2008), 803–814.

[29] Ryo Iijima, Shota Minami, Yunao Zhou, Tatsuya Takehisa, Takeshi Takahashi, Yasuhiro Oikawa, and Tatsuya Mori. 2021. Audio Hotspot Attack: An Attack on Voice Assistance Systems Using Directional Sound Beams and its Feasibility. *IEEE Transactions on Emerging Topics in Computing* 9, 4 (2021), 2004–2018. https://doi.org/10.1109/TETC.2019.2953041

[30] Texas Instruments. [n. d.]. Data Sheet OPA541. https://www.ti.com/lit/ds/symlink/opa541.pdf.

[31] Louis Jackowski-Ashley, Gianluca Memoli, Mihai Caleap, Nicolas Slack, Bruce W Drinkwater, and Sriram Subramanian. 2017. Haptics and directional audio using acoustic metasurfaces. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*. 429–433.

[32] Xue Jiang, Yong Li, Dean Ta, and Weiqi Wang. 2020. Ultrasonic sharp autofocusing with acoustic metasurface. *Physical Review B* 102, 6 (2020), 064308.

[33] Soundlazer kickstarter. 2016. https://www.kickstarter.com/projects/richardhaberkern/soundlazer.

[34] Byung-Chul Kim and I-Tai Lu. 2000. Parameter study of OFDM underwater communications system. In *OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference Proceedings (Cat. No. 00CH37158)*, Vol. 2. IEEE, 1251–1255.

[35] SE Kim, JH Hwang, TW Kang, SW Kang, and SW Sohn. 2012. Generation of audible sound with ultrasonic signals through the human body. In *2012 IEEE 16th International Symposium on Consumer Electronics*. IEEE, 1–3.

[36] Martin L Lenhardt, Ruth Skellett, Peter Wang, and Alex M Clarke. 1991. Human ultrasonic speech perception. *Science* 253, 5015 (1991), 82–85.

[37] Kevin D LePage and Henrik Schmidt. 2002. Bistatic synthetic aperture imaging of proud and buried targets from an AUV. *IEEE Journal of Oceanic Engineering* 27, 3 (2002), 471–483.

[38] Yijie Li, Xiatong Tong, Qianfei Ren, Qingyang Li, Lanqing Yang, Yi-Chao Chen, Guangtao Xue, Xiaoyu Ji, and Jiadi Yu. 2023. AUDIOSENSE: Leveraging Current to Acoustic Channel to Detect Appliances at Single-Point. In *2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 240–248.

[39] Yijie Li, Juntao Zhou, Dian Ding, Yi-Chao Chen, Lili Qiu, Jiadi Yu, and Guangtao Xue. 2024. MuDiS: An Audio-independent, Wide-angle, and Leak-free Multi-directional Speaker. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 263–278.

[40] Manni Liu, Linsong Cheng, Kun Qian, Jiliang Wang, Jin Wang, and Yunhao Liu. 2020. Indoor acoustic localization: A survey. *Human-centric Computing and Information Sciences* 10 (2020), 1–24.

[41] Robert Malkin, Brian Kappus, Benjamin Long, and Adam Price. 2023. On the non-linear behaviour of ultrasonic air-borne phased arrays. *Journal of Sound and Vibration* 552 (2023), 117644.

[42] Wenguang Mao, Jian He, and Lili Qiu. 2016. Cat: high-precision acoustic motion tracking. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 69–81.

[43] Microsoft. [n. d.]. SoundScape. https://www.microsoft.com/en-us/research/product/soundscape/.

[44] Christopher Morse, Adam Chernick, Zeyu Ren, Sabrina Naumovski, and Luke Gehron. 2019. Sound space: Communicating acoustics through interactive visualization. In *2019 IEEE Games, Entertainment, Media Conference (GEM)*. IEEE, 1–4.

[45] Philip M Morse and Richard H Bolt. 1944. Sound waves in rooms. *Reviews of modern physics* 16, 2 (1944), 69.

[46] Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel Watson. 2015. Contactless sleep apnea detection on smartphones. In *Proceedings of the 13th annual international conference on mobile systems, applications, and services*. 45–57.

[47] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1515–1525.

[48] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. 2007. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *Proceedings of the 5th international conference on Embedded networked sensor systems*. 1–14.

[49] F Joseph Pompei. 2002. *Sound from ultrasound: The parametric array as an audible sound source*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[50] ADMP401 POWER. [n. d.]. Data Sheet ADMP401. *POWER* 8616 ([n. d.]), 00.

[51] Virginia Puyana-Romero, Lilian Solange Lopez-Segura, Luigi Maffei, Ricardo Hernández-Molina, and Massimiliano Masullo. 2017. Interactive soundscapes: 360-video based immersive virtual reality in a tool for the participatory acoustic environment evaluation of urban areas. *Acta acustica united with acustica* 103, 4 (2017), 574–588.

[52] Kun Qian, Chenshu Wu, Fu Xiao, Yue Zheng, Yi Zhang, Zheng Yang, and Yunhao Liu. 2018. Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices. In *IEEE INFOCOM 2018-IEEE conference on computer communications*. IEEE, 1574–1582.

[53] M Roser, C Appel, and H Ritchie. 2019. Human height. Our world in data. 2019. *Availble online at: https://ourworldindata. org/human-height* (2019).

[54] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. BackDoor: Making Microphones Hear Inaudible Sounds *(MobiSys '17)*. Association for Computing Machinery, New York, NY, USA, 2–14. https://doi.org/10.1145/3081333.3081366

[55] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Inaudible voice commands: The long-range attack and defense. In *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)*. 547–560.

[56] Swapnil Sayan Saha, Sandeep Singh Sandha, Siyou Pei, Vivek Jain, Ziqi Wang, Yuchen Li, Ankur Sarker, and Mani Srivastava. 2022. Auritus: An open-source optimization toolkit for training and development of human movement models and filters using earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–34.

[57] Adam Smith, Hari Balakrishnan, Michel Goraczko, and Nissanka Priyantha. 2004. Tracking moving devices with the cricket location system. In *Proceedings of the 2nd international conference on Mobile systems, applications, and services*. 190–202.

[58] Greta C Stamper and Tiffany A Johnson. 2015. Auditory function in normal-hearing, noise-exposed human ears. *Ear and hearing* 36, 2 (2015), 172–184.

[59] Michael Vorländer, Dirk Schröder, Sönke Pelzer, and Frank Wefers. 2015. Virtual reality for architectural acoustics. *Journal of Building Performance Simulation* 8, 1 (2015), 15–25.

[60] Anran Wang and Shyamnath Gollakota. 2019. Millisonic: Pushing the limits of acoustic motion tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[61] Han Wang, Jiming Tang, Zhipeng Wu, and Yu Liu. 2022. A Multibeam Steerable Parametric Array Loudspeaker for Distinct Audio Content Directing. *IEEE Sensors Journal* 22, 13 (2022), 13640–13647.

[62] Lei Wang, Kang Huang, Ke Sun, Wei Wang, Chen Tian, Lei Xie, and Qing Gu. 2018. Unlock with your heart: Heartbeat-based authentication on commercial mobile phones. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 3 (2018), 1–22.

[63] Weiguo Wang, Yuan He, Meng Jin, Yimiao Sun, and Xiuzhen Guo. 2023. Meta-Speaker: Acoustic Source Projection by Exploiting Air Nonlinearity. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[64] Yi Weng, Ezra Ip, Zhongqi Pan, and Ting Wang. 2016. Advanced spatial-division multiplexed measurement systems propositions—from telecommunication to sensing applications: a review. *Sensors* 16, 9 (2016), 1387.

[65] Peter J Westervelt. 1951. The theory of steady forces caused by sound waves. *The Journal of the Acoustical Society of America* 23, 3 (1951), 312–315.

[66] Peter J Westervelt. 1957. Scattering of sound by sound. *The Journal of the Acoustical Society of America* 29, 2 (1957), 199–203.

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

[68] Wei Xu, ZhiWen Yu, Zhu Wang, Bin Guo, and Qi Han. 2019. Acousticid: gait-based human identification using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–25.

[69] Jun Yang, Khim-Sia Tan, Woon-Seng Gan, Meng-Hwa Er, and Yong-Hong Yan. 2005. Beamwidth control in parametric acoustic array. *Japanese journal of applied physics* 44, 9R (2005), 6817.

[70] Zhijian Yang and Romit Roy Choudhury. 2021. Personalizing head related transfer functions for earables. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*. 137–150.

[71] 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*. 1–14.

[72] Jingwei Yin, Xiao Zhang, and Yiming Zhou. 2015. Differential pattern time delay shift coding underwater acoustic communication using parametric array. *The Journal of the Acoustical Society of America* 137, 4 (2015), 2214–2214.

[73] 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.

[74] Sangki Yun, Yi-Chao Chen, and Lili Qiu. 2015. Turning a mobile device into a mouse in the air. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. 15–29.

[75] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th annual international conference on mobile systems, applications, and services*. 15–28.

[76] Kexin Zeng, Zhendong Li, Zichao Guo, and Zhonggang Wang. 2023. Reconfigurable and Phase-Engineered Acoustic Metasurfaces for Broadband Wavefront Manipulation. *Advanced Physics Research* (2023), 2300128.

[77] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. DolphinAttack: Inaudible Voice Commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Dallas, Texas, USA) *(CCS '17)*. Association for Computing Machinery, New York, NY, USA, 103–117. https://doi.org/10.1145/3133956.3134052

[78] Hanyun Zhou, SH Huang, and Wei Li. 2020. Parametric acoustic array and its application in underwater acoustic engineering. *Sensors* 20, 7 (2020), 2148.