Model-based Head Orientation Estimation for Smart Devices

QIANG YANG, The Hong Kong Polytechnic University, China
YUANQING ZHENG*, The Hong Kong Polytechnic University, China

Voice interaction is friendly and convenient for users. Smart devices such as Amazon Echo allow users to interact with them by voice commands and become increasingly popular in our daily life. In recent years, research works focus on using the microphone array built in smart devices to localize the user's position, which adds additional context information to voice commands. In contrast, few works explore the user's head orientation, which also contains useful context information. For example, when a user says, "turn on the light", the head orientation could infer which light the user is referring to. Existing model-based works require a large number of microphone arrays to form an array network, while machine learning-based approaches need laborious data collection and training workload. High deployment/usage cost of these methods is unfriendly to users. In this paper, we propose HOE, a model-based system that enables Head Orientation Estimation for smart devices with only two microphone arrays, which requires a lower training overhead than previous approaches. HOE first estimates the user's head orientation candidates by measuring the voice *energy radiation pattern*. Then, the voice *frequency radiation pattern* is leveraged to obtain the final result. Real-world experiments are conducted, and the results show that HOE can achieve a median estimation error of 23 degrees. To the best of our knowledge, HOE is the first model-based attempt to estimate the head orientation by only two microphone arrays without the arduous data training overhead.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile computing design and evaluation methods.*

Additional Key Words and Phrases: acoustic sensing, head orientation, smart devices

ACM Reference Format:
Qiang Yang and Yuanqing Zheng. 2021. Model-based Head Orientation Estimation for Smart Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 136 (September 2021), 24 pages. https://doi.org/10.1145/3478089

1 INTRODUCTION

Recently, we have witnessed the prosperity of smart devices and their applications in homes. Most of them equip with microphone arrays that enable interaction with users by voice commands. As a friendly interface to access smart devices, it is intuitive to use for users, especially for the elderly, handicapped, and disabled people. To provide better services and attract more customers, smart device companies have developed a lot of new technologies to infer users' context based on the captured voice commands. For example, some companies leverage acoustic sensing to infer the user's location [11, 39]. The research community also pays close attention to this trend and proposes many innovative voice localization technologies [10, 12, 14, 30, 38, 47]. Knowing a user's location helps to narrow down the possible set of voice commands and provide customized services to users. Same as the location, head orientation also provides important contextual information:

*This is the corresponding author.

Authors' addresses: Qiang Yang, qiang.yang@connect.polyu.hk, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong, China; Yuanqing Zheng, yuanqing.zheng@polyu.edu.hk, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong, China.

2474-9567/2021/9-ART136 $15.00
https://doi.org/10.1145/3478089

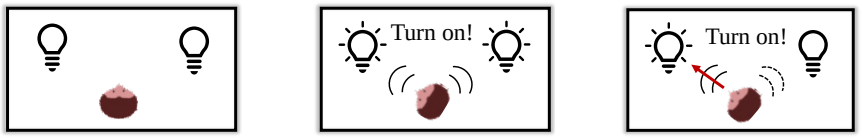


Fig. 1. An example application scenario for head orientation estimation. (a) Two voice-controlled lights in a home. (b) The user would like to turn on the left light, but all lights receive this voice command and become bright. (c) With head orientation estimation, the left light could be turned on as the user intended to.

(1) **Multi-device Wakeup Arbitration.** Nowadays, most families own more than one smart voice-controlled device, such as smart speakers, smart lamps, smart TVs, *etc.* Without head orientation information, these devices may suffer from the multi-device confusion problem in practice. As illustrated in Fig. 1(a) and Fig. 1(b), imagine that we have two smart lights in a room. When they receive "turn on the light" as a voice command, they may wonder which light to turn on. If the smart lights can infer the user's head orientation, they can turn on the exact light as the user intended to (Fig. 1(c)).

(2) **Meeting diarization.** By inferring the location and head orientation of a user, the smart microphones in a meeting room can figure out Alice is actually talking to Bob but not Charlie sitting in different orientations. Thus, the meeting diarization will be more clear on the task assignment and conversation log.

(3) **Additional application scenarios.** For example, disabled people could control their wheelchairs with the head orientation when they are equipped with voice devices [31]; Verbally indoor navigation is also possible when there are several smart devices deployed in a building that could give directional instruction like "the office is on your left side" when users ask the destination. Moreover, we generally speak towards smart devices when we intentionally interact with them, so smart devices can filter out the commands not facing them, such as the sound from TV or computer in case of the ghost waking-up by mistake. We believe that head orientation as a voice contextual knowledge would inspire and benefit more applications in the future.

However, at present, most works on head orientation estimation adopt vision-based approaches utilizing cameras to monitor the human head orientation [4, 35]. Such approaches however raise privacy concerns in home environments. Existing model-based acoustic methods typically require hundreds/dozens of microphone arrays densely deployed in monitoring areas [1, 5, 18, 22, 37]. The deployment cost of such large array networks consisting of so many microphones is prohibitive for practical usage scenarios [45]. Moreover, these methods perform exhaustively search hence cannot work in real-time due to high computational overhead [2]. Machine learning-based acoustic approaches require fewer arrays but laborious data collection and labeling efforts to train a learning model [3, 41, 42, 49]. For example, Soundr [49] leverages a head-mounted VR device to collect 700+ *min* ground truth data to train a neuron network, which is also not friendly to users. Therefore, we may ask a question: *could we estimate head orientation with fewer arrays, as well as lower training overhead?*

In this paper, we propose HOE, a Head Orientation Estimation system with only two microphone arrays. HOE is model-based, which means it does not require arduous data collecting or training overhead. Besides, compared with existing model-based methods, it significantly reduces the number requirement of arrays. Intuitively, the human voice energy is mainly radiated to the head front direction, while the energy radiated to the side and opposite direction is generally weaker. HOE models the voice radiation pattern based on this fact, and estimates a user's head orientation with the voice signals received by two microphone arrays. Although intuitive and simple in concept, it entails tremendous challenges in practice:

Noise and reflection interference. The key enabler underlying head orientation estimation is the correct energy measurement for matching with the theoretical voice radiation pattern. However, interfering with ambient noise (*e.g.*, air-conditioners or fans) and reflections, the energy measured by microphone arrays may substantially differ from the expected radiation pattern if not handled properly.

Energy attenuation. The energy of voice signals varies at different positions and directions due to the propagation attenuation. Therefore, we must compensate for the energy to the further array before performing head orientation estimation. However, voice signal attenuation is very complicated in practice, since it is affected by many factors such as distance, signal frequency, directions, and so on [8].

Orientation ambiguity. To reduce the deployment cost, HOE only utilizes two microphone arrays. However, using fewer microphone arrays (*e.g.*, 2) will result in ambiguity in the estimation result. For example, when two arrays measured the same energy levels, we cannot distinguish if a user is speaking towards the middle of arrays or in the opposite direction. This ambiguity problem may significantly affect the final estimation result if not properly resolved.

HOE addresses the above challenges by proposing the following techniques:

(1) To mitigate the impact of reflection interference, microphone arrays are beamformed to the direction of the user's position, since beamforming can enhance the voice signal from the user and suppress the signal from other directions (*e.g.*, reflections). Background noise is mostly within low-frequency bands, and the signal in high-frequency band has a better directivity as well as a less reflection effect [43]. Therefore, HOE leverages the high-frequency component of the beamformed voice signal to perform head orientation estimation. Thus, background noise can be effectively mitigated.

(2) Although indoor voice attenuation is hard to model in theory, we could perform a one-off parameter training to approximate the attenuation pattern for each room, since the location of smart devices would not change frequently. We investigate the attenuation effect caused by both distance and orientation, and propose an adaptive compensation model considering both factors into account. By properly compensating for the received voice signals, HOE could mitigate the attenuation impact of different distances and orientations.

(3) To resolve the ambiguity, we study the distribution of two ambiguous orientations and find that all ambiguities are always symmetrical: facing or backing arrays. Furthermore, arrays would receive more high-frequency energy when the user is facing them [33]. Based on this key observation, we could check the proportion of high-frequency component energy received by the arrays to perform disambiguation.

The main contributions of this paper are summarized as follows:

- We propose HOE, the first model-based effort on head orientation estimation with two microphone arrays to the best of our knowledge.
- We present an adaptive compensation model for voice signals considering the effect not only from distances but also orientations. We also propose an approach that utilizes the voice frequency radiation pattern to tackle the orientation ambiguity problem.
- HOE is implemented and evaluated in real-world experiments. The results show that HOE can achieve an overall median angular error of 23°, which is promising to provide new context information (*i.e.*, head orientation).

The rest of this paper is organized as follows. In Sec. 2, we briefly introduce the capability of commodity smart devices and voice assistants, and give a definition of the target problem. Followed by Sec. 3, we summarize the design space of related works and highlight our novelty. In Sec. 4, we describe the detailed system design of HOE. Our system is implemented and evaluated in Sec.5. We discuss some limitations and future work in Sec. 6. Finally, Sec. 7 concludes this paper.

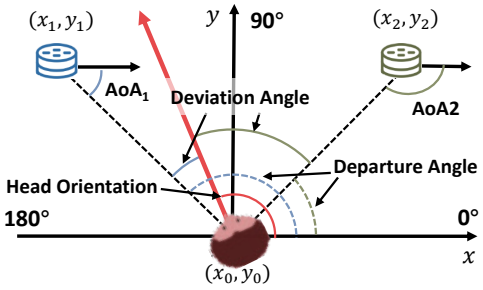


Fig. 2. Problem illustration of the head orientation estimation. The aim of HOE is to estimate the orientation, *i.e.*, the angle between the speaking direction (red arrow) and x direction.

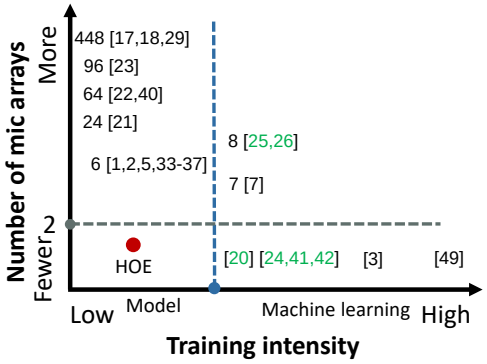


Fig. 3. Design Space: comparing with related work. The digits before citations are the number of microphones/arrays used in the corresponding work. The literature marked in green means they conduct researches on loudspeakers instead of humans.

2 BACKGROUND AND PROBLEM DEFINITION

Smart devices are generally equipped with a microphone array to enable voice interaction. They are usually triggered by a name or phrase, such as “Alexa, ...” or "Hello, ..." which are termed as *Keywords*. Nowadays, commercial voice assistants in smart devices provide many built-in functions including Keyword Spotting (KWS), Angle of Arrival (AoA), and so on [32]. Keyword Spotting detects the keyword to wake up the device and start a conversation, and AoA could estimate the Angle of Arrival of voice, indicating the direction of the speaking user.

With two microphone arrays (referring to smart devices hereinafter), the user’s position could be localized by finding the intersection point of two AoAs. As shown in Fig. 2, AoA_1 and AoA_2 are crossed at (x_0, y_0), which indicates the user’s location. Voice localization with microphone arrays has been extensively studied [1, 7, 13, 16, 21, 28, 36, 38, 47], so we can build HOE on them directly. Head orientation, however, has not been thoroughly studied yet. Previously, voice localization only focuses on the distance or angle between the user and the microphone array, but ignores the angle between the array and the user’s *speaking direction*. We give a formal definition of our target problem as follows:

Problem definition. Fig. 2 illustrates a head orientation estimation scenario. The user’s position (x_0, y_0) could be localized by existing methods. In a local coordinate, the **head orientation** is defined as the angle between the speaking direction (red arrow) and x direction. Furthermore, the directions of the microphone array with respect to human are termed as the **departure angle**, indicating the Line of Sight (LOS) departure directions of voice. the **deviation angle** defines the deviation from the departure angle to the user’s head orientation. HOE aims to *estimate the head orientation of a voice command*, so the orientation estimation error should be as small as possible to meet the practical requirement of applications.

3 RELATED WORK

Ideally, a head orientation estimation method should deliver a high estimation accuracy with a few microphones and low training overhead. Many researchers have made great efforts to achieve this goal. Fig. 3 summarizes the existing works and our proposed solution in a design space. We categorize existing works according to the number of needed microphones and their training overhead as follows.

3.1 More arrays & low training intensity

J. M. Sachar *et al.* [29] used a Huge Microphone Array (HMA) consisting of 448 microphones distributed in a laboratory to estimate a user's head orientation. Such HMA-based methods [17, 18, 22, 23] could detect differences in the energy from microphones and accurately estimate head orientation. However, they need a large number of microphones and incur high deployment costs. Several works [1, 5, 21, 36, 37] utilized the GCC-PHAT [15] based method to estimate a user's head orientation by searching all possible locations and orientations, finding a maximum in the 3D space. This exhaustive search leads to high computation costs and is not suitable for real-time applications. Another multi-microphone approach [2, 33–35] is based on HLBR [34], a frequency-domain metric related to the head orientation. These model-based approaches do not involve much training overhead. However, they typically need to deploy a large number of microphone arrays *at each wall around a room* in order to cover all possible directions. For example, [22, 40] deploy a 64-microphone array, and [34, 36] utilized six T-shape arrays (4 microphones in each array). Therefore, these methods cannot be deployed in ordinary homes.

3.2 More arrays & high training intensity

With the development of machine learning (ML) techniques, many researchers applied them to improve the performance of the head orientation estimation. Brutti *et al.* [7] utilized the Nearest Neighbors to classify loudspeaker orientations by seven 4-microphone arrays. [25, 26] deployed eight T-shape arrays in a room, and trained a neural network to estimate the orientation of a loudspeaker. These methods need to collect training data which incur high overhead for users, and still require a large number of microphone arrays.

3.3 Fewer arrays & high training intensity

Following that, various machine learning based methods have been proposed to reduce the number of needed microphones in head orientation estimation. Many works trained a classification model to predict the orientation of a loudspeaker rather than a real user [20, 24, 41, 42]. [20] distinguishes whether the user is talking to the array, which is less usable in our applications. Soundr [49] and [3] estimate real human head orientation with one microphone array. Soundr needs a massive amount of training data (*e.g.*, 700+ *min*) collected with a VR device to train a workable neural network. However, it requires a dedicated VR headset and does not perform well if it has not been trained for a given environment or a user [49]. [3] is a state-of-the-art for head orientation estimation, which extracted acoustic features to train a tree, but it can only predict the relative orientations (*i.e.*, deviation angle in Fig. 2) instead of absolute orientations as HOE. Although these ML-based methods can reduce the number of microphone arrays from dozens to two or even one, the training overhead including data collection, manual labeling, and training workload increases substantially. In contrast, HOE proposes a model-based method to estimate absolute head orientation with two arrays, and does not need a laborious data collection and training overhead.

4 HOE SYSTEM DESIGN

In this section, we introduce the detailed system design. We start with an overview of HOE, followed by the description of functional components. In each subsection, we discuss some practical challenges and present our solutions. Finally, we summarize the whole pipeline of head orientation estimation.

4.1 System overview

Fig. 4 illustrates the overview of HOE. HOE consists of three components: *Energy Compensation*, *Orientation Estimation*, and *Orientation Disambiguation*. When a user would like to deliver a voice command, he/she speaks a keyword to wake up voice assistants such as "Hello, HOE". The microphone arrays of smart devices capture the voice command by Keyword Spotting and locate the user leveraging the Built-in Preprocessing function.

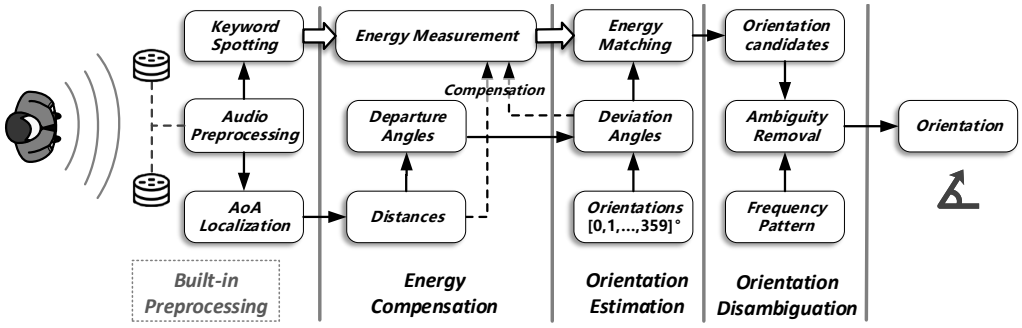


Fig. 4. Overview of HOE.

Next, the distances and departure angles of the user could be calculated with the known locations of smart devices. Following that, the *Energy compensation* component compensates for the energy measured by two arrays due to the attenuation loss from the distance and deviation angle differences. Then, *Orientation Estimation* utilizes an energy matching method to figure out head orientation candidates with ambiguity. In *Orientation Disambiguation*, the ambiguity is resolved by the frequency radiation pattern of voice, and eventually, HOE outputs a final orientation result.

HOE only utilizes the audio segment of the wake-up word for orientation estimation. Voice commands may be different from each other but have the same preceded wake-up word for the smart devices from the same vendor. Thus, we can conveniently adapt HOE to different smart devices. Moreover, a wake-up word lasts about 500 *ms*, thus we can assume that the user's head keeps static in such a short period. In the rest of this section, instead of introducing the Energy Compensation module first, we start with the Orientation Estimation and then raise the reason why HOE needs energy compensation.

4.2 Orientation Estimation

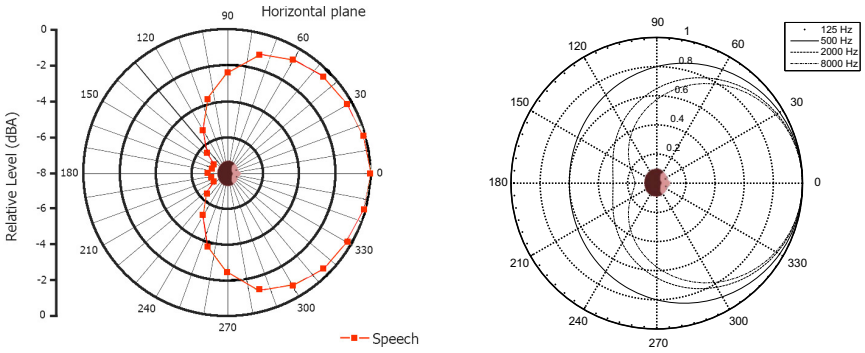
Our head orientation estimation method is based on the fact that the user's voice propagation is anisotropic, which means that we have different measurements when voice is radiated in different directions. To this end, we build voice radiation patterns to model this anisotropic property of the human voice, including an energy radiation pattern and a frequency radiation pattern.

Energy Radiation Pattern. The average energy of human voice is not uniform in all directions. More energy is radiated in the user's forward direction than towards the side or rear directions [9]. As shown in Fig. 5(a) borrowed from [2], blocked by the face and head, the voice energy suffers about -2 dB attenuation on the side of the user, as well as more than -8 dB attenuation behind the body. This kind of voice energy radiation presents basically a cardioid-like attenuation pattern, which can be mathematically parameterized as follows [6]:

$$w(\theta) = 8\left[\left(\frac{1+cos(\theta)}{2}\right)^{\rho} - 1\right] \tag{1}$$

where θ is the deviation angle of the microphone array, and $w(\theta)$ is the energy attenuation (dB) in θ degree compared to the front direction. The exponent ρ determines the directivity level of voice radiation. When $\rho = 0$, the voice radiation pattern is omnidirectional. Fig. 6 shows a common attenuation pattern modeled by Eq. 1 where $\rho = 1$.

Suppose a case that two arrays have the same distance to a user as a bird-eye view shown in Fig. 7. Two microphone arrays are settled at the two sides in front of the user. With the known positions of microphone



(a) Energy pattern (b) Frequency pattern

Fig. 5. Voice radiation pattern with different directions (bird-eye view). (a) Energy pattern [2]: more energy is radiated in the user's forward direction than other directions. (b) Frequency pattern [43]: high-frequency signals ($f \geq 2KHz$) have more notable directivity, but low-frequency signals are almost omnidirectional.

arrays and the user localized before, the distances and departure angles of the two microphone arrays can be computed by geometry. Let E_1 and E_2 denote the energy received by microphone array #1 and #2, respectively. Intuitively, when a user speaks a voice command, the energy in different directions would attenuate following the radiation pattern $w(\theta)$. Therefore, the energy of the signals received by two microphone arrays would be different, and presents an attenuation pattern as the dashed line. Thus, we can formulate a loss function and minimize the residue to estimate the head orientation Θ by searching all possible angles:

$$\Theta = \underset{\theta_1,\theta_2}{argmin} \left\| w(\theta_1) - w(\theta_2) - 10log10(\frac{E_1}{E_2}) \right\|^2 \tag{2}$$

where θ_1, θ_2 are the deviation angles of two arrays associated with the head orientation Θ. Here we omit the A-weighting [43].

4.3 Energy Compensation

Note that the energy matching method above only works where the distances between two arrays and the voice source are the same. It also assumes the signal propagates freely in a 3D space without noise or interference. However, the propagation becomes more complicated in practice, especially in indoor scenarios. In the following, we propose corresponding solutions to tackle these challenges.

4.3.1 Mitigate the impact of noise and interference. We present the results of an empirical study to describe how we mitigate the impact of noise and interference. As shown in Fig. 8(a), two microphone arrays are placed on two sides with equal distances to the user ($d_1 = d_2$), and the departure angles of them are 150°and 30°, respectively. A user first speaks a command towards 90° (black arrow), and then repeats this command towards 135° (grey arrow). The energy values measured by the two arrays are presented in the top subfigure of Fig. 8(b). The first peak corresponds to the first command, so does the second one.

When the head orientation is 90°, the absolute deviation angles of the two arrays are equal. Therefore, the measured energy level of the two arrays should be similar in theory, but we observe that the energy of array #1 is slightly higher than that of array #2. When the orientation changes to 135° (the second peak), the user deviates to array #1 and thus we expect a higher power obtained by that array. However, the measured energy levels remain

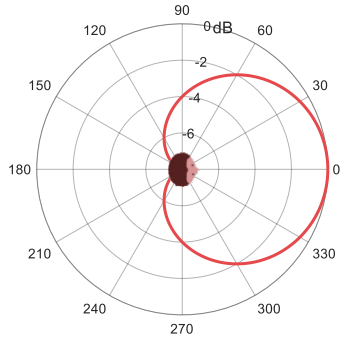


Fig. 6. The voice energy radiation pattern modeled by Eq. 1. The energy radiated to 0° has 0 dB attenuation, and it drops to -8 dB at most as the deviation angle increases to 180° (rear direction).

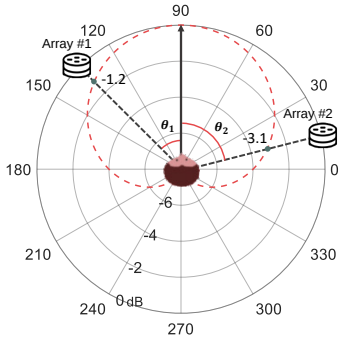


Fig. 7. Two microphone arrays are placed at the same distance from a user. When the user speaks a voice command, two arrays will receive different voice energy levels due to the anisotropic radiation pattern.

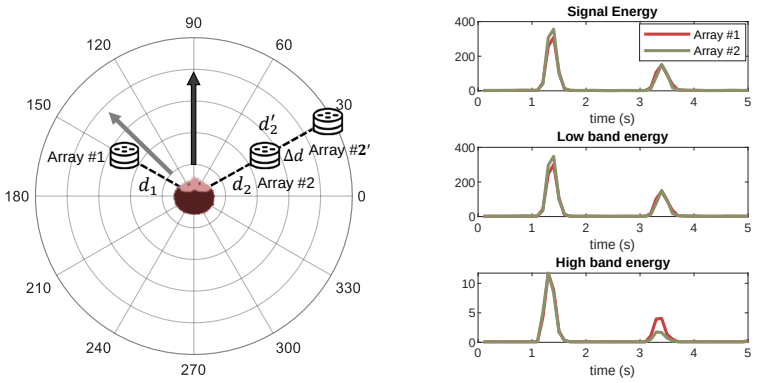
almost the same as shown in the second peak of Fig. 8(b). The main reason is that microphone arrays measure both user's voice and background interference in the environment. Besides the background noise, large objects such as walls and furniture can reflect the voice signal back to microphone arrays. Therefore, it is challenging to infer the head orientation from the measured power levels with interference. To deal with this problem, we first perform beamforming to the user's direction with microphone arrays, since it could enhance the signal from a specific direction as well as suppress the interference from other directions (*e.g.*, reflections) spatially. Then, we use the voice frequency radiation pattern of the human voice to further mitigate noise.

Frequency Radiation Pattern. The human voice is produced by the vocal cords in the throat and radiated out through the mouth. The low-frequency component has a longer wavelength and is low directional due to the *diffraction effect*. In contrast, the wavelength of the high-frequency component is short. As a result, the high-frequency component is highly directional compared with the low-frequency one.

As illustrated in Fig. 5(b) [43], the low-frequency signal like 125 Hz practically has no directivity, *i.e.*, the signal is emitted almost uniformly to all directions, while the signal with higher frequencies (*e.g.*, 2 KHz) exhibits a notable directional radiation pattern. As such, the signal with high frequency has fewer reflections than low frequencies due to the higher directivity. Besides, the high-frequency signal also suffers less from noise, since the ambient noise generally lies in low-frequency bands (lower than 2 KHz). Therefore, we choose 2 KHz as a threshold to separate the beamformed signal into two components: the low-frequency and high-frequency bands. The energy values of these two components are illustrated in the middle and bottom subfigures of Fig. 8(b). We can see that the low-frequency part contributes the vast majority of energy and present a similar pattern with the raw signal. In contrast, for the high-frequency component, the energy level is quite low but matches the energy pattern as we expected. This result hints that the beamforming and high-frequency characteristic can effectively mitigate the impact of noise and reflection interference.

4.3.2 Distance Attenuation compensation. In the above estimation model (Eq. 2), we assume the distances from a user to two microphone arrays are the same. However, generally, the distances are different in practice, so we need to carefully compensate for the energy attenuation before applying the estimation model.

According to the *inverse square law*, the energy level is inversely proportional to the square of the distance between the voice source (*i.e.*, mouth) and the microphone array. However, signal attenuation is more complicated in practice and challenging to accurately model, since it is affected by many factors (*e.g.*, frequency, orientation, room interior, distance, reflection, *etc.* [8]), especially for the wide-band voice signal. These factors are associated



(a) Illustration of two different head orientations when speaking the same commands

(b) The energy of different frequency bands measured by two arrays

Fig. 8. Energy measurement with two different orientations when a user speaks the same command "Hello". (a) The user speaks commands to 90° (black arrow) and 135° (gray arrow) (b) the energy measurement in different frequency bands of two arrays when the orientation equals to 90° (first peak) and 135° (second peak).

with both user (*e.g.*, voice frequency) and room (*e.g.*, reverberation and interior). Considering that the room of smart devices usually keeps fixed, for each user, we can conduct a one-off parameter training to approximately estimate the attenuation pattern in each room.

We performed an empirical experiment to investigate the distance energy attenuation effect. As shown in Fig. 8(a), suppose two arrays (array #2 and array #2′) and the user are in a straight line. d_2 and $d_2'(= d_2 + \Delta d)$ are the distances from the voice source to two arrays. Generally, the voice interaction distance between human and smart devices is about $1 \sim 3m$. In our experiment, microphone array #2 is set to the reference array, 100 cm far away in front of the user. d_2' is set varied from 120 cm to 300 cm (with a 20 cm step). Users were asked to repeat five commands *towards the direction* from the reference array #2 to the target array #2′ at each distance. Considering that the energy of spoken voice command is unstable and unknown for each time, the attenuation could only be measured as a relative quantity. Therefore, We aim to explore the relationship between the energy ratio ($\frac{E_2'}{E_2}$) and distance ratio ($\frac{d_2'}{d_2}$).

Fig. 9 shows the experiment result of two users in two different rooms: an office and a meeting room. Each colored point represents one command measurement. We can see that the energy ratios of near positions (*e.g.*, distance ratio equals 1.2) are very close for different users/rooms. However, for the same user 1, the energy attenuates faster in the meeting room than in the office. The reason is that the meeting room is almost three times larger than the latter one, so there are fewer blocks and reflections. Moreover, the energy attenuation for different users in the same room presents a similar pattern (user 1, 2 in the office) with a slight difference. We believe the similarity is because of the same room acoustics, but the difference is attributed to the varied human physiological voice (*i.e.*, user diversity). We also find that the energy ratio measurements at each distance fluctuate more largely in the office than that in the meeting room, since the energy stability also suffers from blocks and reflections in smaller rooms. Although the energy ratio has fluctuations among five repetitions, we can see a clear trend, that is, the energy ratio has a quadratic relationship with the distance ratio. As such, a quadratic curve can be fitted to mathematically formulate this attenuation pattern:

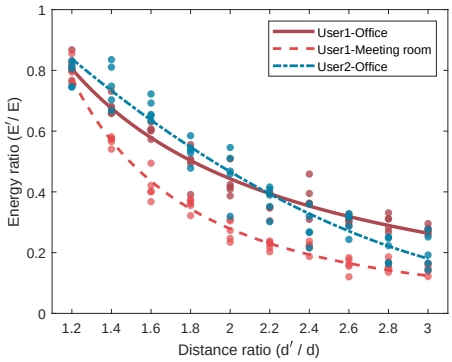


Fig. 9. Distance attenuation of different users and rooms. Each dot represents one measurement.

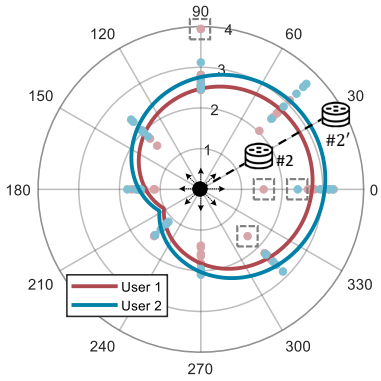


Fig. 10. Orientation attenuation of different users. The dots in gray box are outliers.

$$\frac{E'}{E} = h\left(\frac{d'}{d}\right)^{-2} + i\left(\frac{d'}{d}\right)^{-1} + j \tag{3}$$

Here we drop subscripts for the sake of simplicity. h, i, and j are constant factors associated with the room and user. Therefore, users could conduct a one-off parameter training mentioned above to approximate the distance attenuation pattern in a room for themselves before using HOE. In this circumstance, if the distances from the user to two microphone arrays are not equal after localization (*e.g.*, array #1 and #2′), HOE can compensate the energy for one array to a comparable level with another one. It is equivalent to logically *"move"* one array to the position with the same distance as another array to the user (array #2′ →#2, which is termed as the *equal-distance array*) to mitigate the distance attenuation effect.

4.3.3 Orientation Attenuation compensation. It is noted that the experiment above was conducted where the head orientation (*i.e.*, speaking direction) was always towards the line linking two microphone arrays. In other words, both deviation angles of array #2 and # 2′ equal to 0°. However, when the head orientation is not aligned with two arrays (array #2→#2′), the deviation angles would also contribute to the attenuation accordingly.

We conducted another experiment to investigate the energy attenuation with different head orientations in an office. In this experiment, the target array #2′ was fixed at 2.4 m far away from the user. We labeled eight orientations anticlockwise from 0° to 315° with a 45° spacing step (shown as the black arrows in Fig. 10). Two users are asked to speak five voice commands in each direction. The degree ticks in Fig. 10 denote different head orientations, and the radius ruler indicates the measured high-frequency band energy ratio (E/E') between two arrays #2 and #2′, each dot represents one measurement.

Intuitively, these points are expected to distribute uniformly across different orientations, *i.e.*, the energy ratio should be almost the same since the positions of the user and arrays are not changed, meaning that the distance ratio (d'/d) of two arrays keeps constant. However, the experiment result presents a different pattern from what we expect, and the only changed factor is the user's head orientation, further said, the deviation angles of the arrays accordingly. This result indicates that the orientation also has an impact on the energy attenuation, especially for high-frequency signals. In Fig. 9, we can see the energy ratio of user 2 is higher (lower for E'/E) than user 1 at this distance ($d'/d = 2.4$). The result in Fig. 10 is consistent with this observation, where user 2 has a maximal energy attenuation (3.1), which is higher than user 1 (2.8) towards the arrays' direction (30°). The energy ratio of both users decreases gradually along with the head orientation turns left until to the opposite direction to the arrays (210°). The decrease speed is proportional to the user's directivity factor. The reason

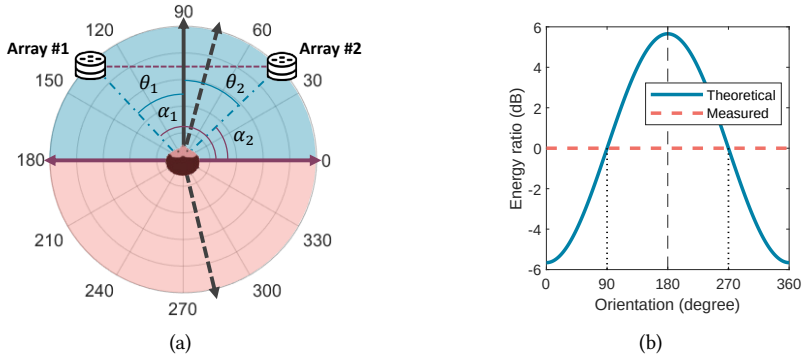


Fig. 11. Illustration of the orientation ambiguity. (a) Two ambiguous orientation are always symmetrical with the boundary (purple solid line with arrows) (b) There are two intersection points for the theoretical energy ratio (blue solid line) and the measured one (red dashed line), which causes the ambiguity.

is that high-frequency signals are more directional, so their radiation range is narrow. As a result, the signal energy attenuation approximately confirms with the Eq. 3 at the exact front of the user's head orientation, but high-frequency signals attenuate more at the side or behind of the user. Thus, when the user faces back to the two arrays, reflections and a part of relative low-frequency component are the dominant part in the recordings, so the energy ratio nearly reaches 1. The relationship between the energy ratio and deviation angle can be approximately fitted as a Gaussian-like pattern:

$$ER = r \cdot \exp\left(-\frac{\theta^2}{2k^2}\right), ER \geq 1 \tag{4}$$

where ER is the angular energy ratio, θ is the deviation angle of the array, and r is the maximum energy ratio when the deviation angle is zero, which can be obtained by Eq. 3 with the distance ratio. k is the orientation attenuation factor associated with the user's physiological feature. Considering that the energy level of the closer array is hardly lower than the far array, ER should be greater than or equal to 1. The empirical studies above show that we can conduct a one-off parameter training to approximate the distance and orientation attenuation patterns for different users in different rooms, which are used to jointly compensate for the energy loss in orientation estimation.

4.4 Disambiguation

In the following, we will first give the reason why ambiguity happens. And then, a frequency pattern-based approach will be proposed to resolve the ambiguity problem and find the real head orientation.

4.4.1 Why ambiguity. Because of the symmetry of the voice radiation pattern (Fig. 5), with only two arrays, the result estimated by Eq. 2 is not unique but has ambiguity. Referring to Fig. 11(a), we illustrate a typical ambiguity case. Two microphone arrays are placed in the 45° and 135° directions with respect to a user. As per Eq. 1, the theoretical energy ratio (*i.e.*, $w(\theta_1) - w(\theta_2)$) with different orientations presents a symmetric shape as the blue solid line exhibited in Fig. 11(b). Suppose that the user speaks towards 90° (black arrow), the measured energy ratio should be 0 dB (red dashed line in Fig. 11(b)). Therefore, there are two intersection points corresponding to 90° and 270°. That is to say, Eq. 2 would have two solutions that lead to ambiguity.

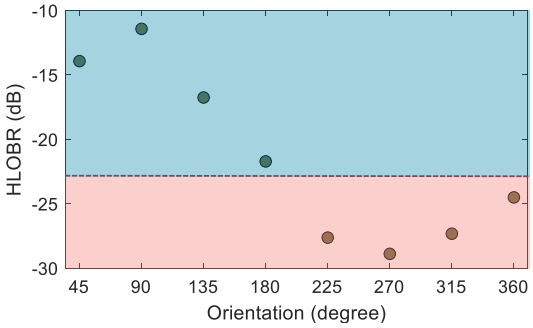


Fig. 12. HLOBR values of different orientations. A threshold could be used to detect if the user faces or backs arrays.

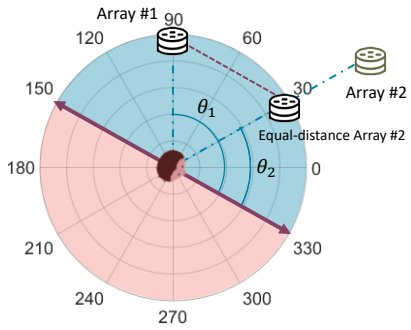


Fig. 13. A general example of the ambiguity when two arrays are placed with different distances and departure angles.

In particular, the ambiguous orientations are always symmetrical with the angle of $\frac{\pi+\alpha_1+\alpha_2}{2}$ or $(\frac{\pi+\alpha_1+\alpha_2}{2}+\pi)$, where α_1, α_2 are the departure angles of two arrays. Thus, two half-circles are divided by the symmetric axis. For example, the ($180^\circ \leftrightarrow 360^\circ$) axis in Fig. 11(a). We term these axis directions as the *boundary direction*, and we define the *front half-circle* as the one that contains arrays. Consequently, the ambiguity problem is equivalent to distinguish whether the real orientation is towards the *front half-circle* area or not. For instance, when the user speaks towards 90°, HOE will report two preliminary estimation results as indicated by the black dashed lines due to a tiny error, which are symmetrical with the boundary. The next step is to discriminate the user is speaking to the blue front half-circle area or red rear half-area to recognize the real orientation.

4.4.2 Disambiguation with the frequency pattern. We address this problem based on the following key observation. The blue *front half-circle* area is always the orientation range *towards* two arrays, while another red half-circle area is always *back* to the arrays. As we mentioned human frequency radiation pattern in Sec. 4.3.1, the low-frequency signal is almost omnidirectional, while the high-frequency signal has much higher directivity. As a result, when a user speaks towards arrays, the arrays would "hear" more high-frequency components than back to the arrays. However, we cannot utilize the high-frequency energy value alone to discern whether the user is speaking towards the arrays or not, since the human speaking volume may vary each time. [34] proposed High and Low Band Ratio (HLBR) as a metric associated with head orientations, which is calculated by dividing the energy of the low-frequency band by the one of the high-frequency band. HLBR utilizes the relative energy value which is less sensitive to the absolute voice volume as well as the distance. However, the range parameters separating high and low bands require to be tuned case by case carefully. To deal with this problem, we measure the energy Ratio between the High octave band and Low Octave Band (HLOBR). The Octave filterbank is commonly used to model how the human ear weights the spectrum and mimic how humans perceive loudness by psychoacoustic perceptual criteria [27, 48]. HOE measures the HLOBR as the energy ratio between the $8th$ and $3rd$ octave band whose center frequencies are 4 KHz and 125 Hz, respectively. As a result, we do not need to laboriously tune the band separation parameters person by person. Fig. 12 shows the summed HLOBR of two arrays when a user speaks towards eight different orientations. We can see that a boundary (*i.e.*, threshold) could be set to separate the orientations into a blue area and into a red area. Therefore, by comparing the HLOBR value of a voice command with this threshold, HOE can distinguish the head orientation towards the arrays or not, and further remove ambiguity.

The HLOBR threshold however may vary among different users due to the human pitch difference. Thus, users could measure their own threshold by speaking wake-up words towards boundary directions position for initialization before running HOE. As illustrated in Fig. 11(a)), when users speak towards the boundary directions

(*i.e.*, 180° or 360°), two deviation angles $\theta_1 + \theta_2 = \pi$, which means θ_1, θ_2 are always supplementary angles. HLOBR values along with different orientation are almost centrosymmetric. Therefore, we can regard that the HLOBR threshold is independent of array position. Fig. 13 shows a general example where two arrays (#1 and #2) are placed with different distances and different departure angles. As we mentioned before, the energy compensation procedure is equivalent to logically "move" the far array to the same distance as an *equal-distance array*. In this circumstance, we can see that the deviation angles of two arrays θ_1, θ_2 are still *supplementary* angles when the user speaks to the boundary direction (*e.g.*, 330°). Note that the HLOBR is a coarse-grained metric related to the orientation. It can not measure the head orientation directly, although it can work as a two-category classification problem for disambiguation. Moreover, if the rough spatial information of head orientation (*e.g.*, position of intended devices, room space constrain) is known in advance, we can leverage this prior knowledge to exclude the ambiguity as well.

4.5 Summary

4.5.1 Parameter Configuration and Personalization. Indoor signal attenuation is associated with both user (*e.g.*, voice frequency) and room (*e.g.*, reverberation and interior). Users can conduct a one-off training to approximately estimate the attenuation pattern in different rooms, including distance attenuation and orientation attenuation. The whole procedure requires collecting about 100 samples.

Different rooms. We first mark a series of points (generally with a 20cm step) from 1 m to 3 m in front of the user. One mic array is fixed at 1 m location, and another array is placed at every point marked before. For each point, a user repeats wake-up words five times towards the array direction. Accordingly, HOE computes the energy ratio between two arrays and fits a quadratic curve to approximate the attenuation pattern. More repetitions and more fine-grained distance intervals are better for more accurate pattern approximation.

Different users. The orientation attenuation is related to the user's physiological factors. So users can speak wake-up words five times towards eight 45°-spacing directions while keeping two arrays static, then HOE measures the energy ratios to fit a Gaussian orientation attenuation pattern. Likewise, more repetitions and smaller direction intervals are better for estimation. Users are also required to speak extra five voice commands towards the "boundary direction" to measure their own HLOBR thresholds for disambiguation.

4.5.2 HOE pipeline. We refer to Fig. 8(a) to describe the whole procedure of HOE. Suppose there are two microphone arrays #1 and #2′, then HOE would like to estimate the user's head orientation. The algorithm in a glance is summarised as Alg. 1.

0) **Initialization**. Users perform the personalization step to initialize the attenuation parameters and HLOBR threshold. The departure angles of two arrays α_1, α_2, and the user's position are calculated by the built-in pre-processing module.

1) **Configuration**. HOE calculates the distances d_1, d_2' from the user to two arrays. Suppose the head orientation is Θ, then deviation angles θ_1, θ_2 and corresponding radiation function $w(\theta_1)$, $w(\theta_2)$ can be calculated.

2) **Distance attenuation compensation**. If $d_1 = d_2'$, HOE goes to step 4. If not (suppose $d_1 < d_2'$), HOE should compensate for the energy attenuation for array #2′. So, if $\theta_2 = 0$, HOE compensates for the distance energy loss for array #2′ by $ER = r$ according to Eq. 3, and then go to step 4. If not, go to the next step.

3) **Orientation attenuation compensation**. If $\theta_2 \neq 0$, the orientation will also cause attenuation. In this way, a new energy ratio ER could be computed by Eq. 4 with r and deviation angle θ_2.

4) **Orientation Estimation**. HOE calculates the residue by Eq. 2, goes back to step 2 with the next orientation until searching all possible directions. The orientation candidates could be estimated by minimizing the residue.

5) **Disambiguation.** Finally, HOE computes the summed HLOBR of the two arrays and compares it with the pre-measured HLOBR threshold. The ambiguity could be removed and then HOE outputs the final estimated orientation Θ.

Algorithm 1: Head Orientation Estimation

Input: Recorded signals, positions of two microphone arrays, user's position reported by pre-processing, attenuation parameters, and HLOBR threshold.
Output: Head orientation Θ of the user;

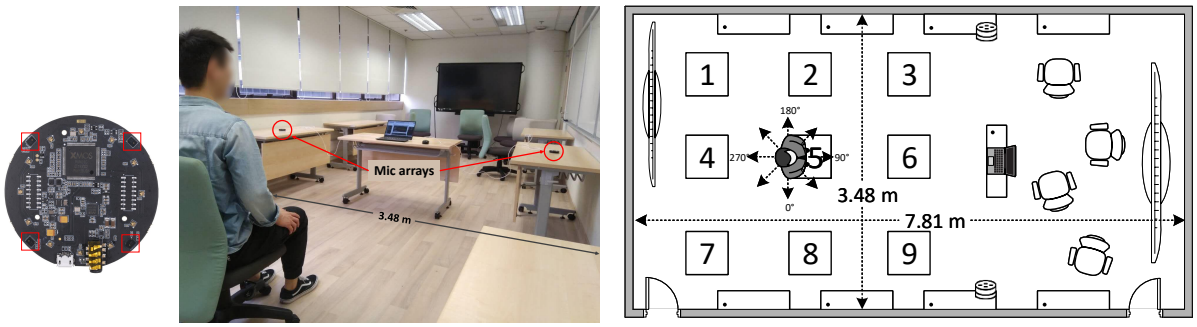
```
Calculate the distances d_1, d_2' and departure angles α_1, α_2 from the voice source to two arrays #1, #2';
Initialize orientation Θ = 0;
for Θ = 0 to 359 do
    Compute deviation angles θ_1 = α_1 − Θ, θ_2 = α_2 − Θ of two arrays;
    Calculate theoretical energy radiation patterns w(θ_1), w(θ_2) of two arrays;
    if d_1 ≠ d_2' then // we suppose d_1 < d_2'
        if θ_2 = 0 then
            Compensate the distance attenuation ER = r according to Eq. 3;
        else
            Compensate the distance and orientation attenuation ER with Eq. 3 and Eq. 4
        end
    else
        ER = 1;
    end
    Calculate the compensated energy ratio E_1/(ER·E_2') and corresponding residue;
end
Estimate head orientation candidates minimizing the residue using Eq. 2;
Remove ambiguity with the HLOBR threshold and obtain the final head orientation Θ.
```

5 IMPLEMENTATION AND EVALUATION

5.1 Implementation and experiment setting

As commercial smart devices like Alexa Echo do not output recorded raw audio data, we implemented HOE with the Seeed Respeaker USB microphone array v2.0 [32]. As shown in Fig. 14(a), it consists of four omnidirectional microphones placed in a circular shape and supports USB Audio Class 1.0 (UAC 1.0). The sampling rate was set to 16 KHz which covers most of the voice frequency bands. Two arrays were connected to a ThinkPad X1 laptop with an Intel i7-10510 CPU (4.9 GHz at boost clock) by cables for data collection and processing. We run HOE and process signals in MATLAB.

We recruited ten participants (5 male, 5 female, mean age 27) and conducted experiments in an office ($7.81m \times 3.48m$) and a meeting room ($10.61m \times 7.62m$). Two arrays were settled on the desks near the wall. The experiment setting in the office is shown in Fig. 14(b). There is some furniture around the room, such as e-boards, desks, and several chairs. Before the experiment, we have already measured and marked the locations and corresponding orientations on the ground in advance as the ground truth. Users were asked to sit at nine labeled positions (1~9) separated by 1 m and speak wake-up commands in eight different directions from 0° to 315° with a 45° step as illustrated in Fig. 14(c). These commands start with two keywords: "Hello" or "Alexa". Each command was repeated three times in each direction per position. Users completed parameters training of attenuation and HLOBR threshold before conducting the experiment, but these samples collected for the parameter configuration are not used in the evaluation.



(a) (b) (c)

Fig. 14. Experiment setting. (a) A Seeed Respeaker microphone array v2.0 with four mics. (b) Experiment illustration in an office. (c) Experiment setting.

5.2 Performance metrics

In the following, we evaluate the performance of HOE in various experiment settings. Before that, we first introduce some evaluation metrics for head orientation estimation following the common agreement of the CHIL consortium [33, 46].

- **Mean Average Error (MAE)**: the mean average angle error of the head orientation estimation.
- **Correct Classification (CC)**: CC measures the percentage of the estimations of head orientation within the nearest sector of ground truth, where the 2D plane is divided into eight 45° sectors.
- **Correct Classification within a Range (CCR)**: CCR measures the percentage of estimated head orientations within the nearest sector of ground truth and adjacent two sectors.

Generally, MAE reveals the fine-grained estimation performance, while CC and CCR evaluate the coarse-grained orientation estimation ability.

5.3 Overall estimation performance

Fig. 15 illustrates the Cumulative Distribution Function (CDF) of HOE's orientation estimation errors in all experiments. The result is obtained with the general directivity factor $\rho = 1$ for all participants. Overall, the median error is 23°, and 90% errors are less than 64°. In addition, we present HOE without ambiguity, which gives the theoretical upper bound of HOE if the ambiguity can be completely resolved. As the green dashed line shows, its median error is 22°. Compared with HOE, they are almost the same, while the 90%-percentage point of later is 54°, and no error is larger than 150°, which has a large enhancement. As we discussed in Sec. 4.4, the ambiguous orientations always distribute in two half-circle parts, which would cause some big errors. The experiment results show that HOE has an amenable estimation accuracy for orientation-aware applications and the disambiguation could effectively decrease the probability of large errors.

5.4 Impact of participants

We also tested the head orientation estimation performance across different participants. As shown in Fig. 16, the overall MAE across ten users is 28.9°. Due to different physiologic directivity factors, the performance across them varies slightly, where the highest and lowest MAE among ten participants are 35.8° and 23.6°, respectively. Considering the minute 3.6° standard deviation, we can claim that users have a similar performance. Fig. 17 displays the CC/CCR of different users. The overall class classification rate among ten participants is 50.2%,

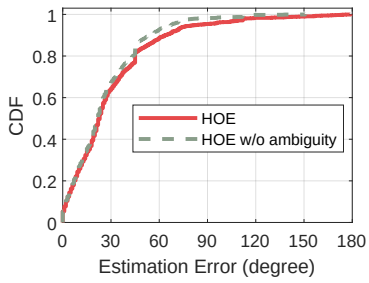


Fig. 15. CDF of HOE orientation estimation errors.

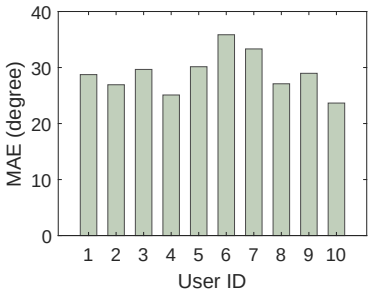


Fig. 16. Overall MAE across different users.

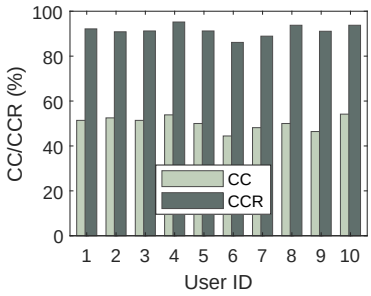


Fig. 17. Overall CC/CCR across different users.

corresponding to the percentage where MAE is lower than 22.5°. According to the definition, CCR is always larger than CC, which ups to 91.4% across all experiments. This demonstrates that HOE could accurately report a coarse-grained direction of a voice command.

5.5 Impact of directivity factor

Different people have different voice radiation patterns because of physiological factors such as mouth size, pitch, head, *etc.*. To investigate the impact of the directivity pattern of different users, we tuned the directivity factor ρ in Eq. 1 from 0.5 to 2. In Fig. 18, we show the results of three users whose optimal ρ are larger than 1. Evidently, the estimation error changes with the parameter variation, and superior results are seen when ρ equals 1.25, 1.6, and 1.7 for users 3, 7, and 6, respectively. Overall, the mean average error reduces by 1.3° across different users after adopting the optimal directivity factors, about 4% compared to the case before. This result demonstrated that the directivity factor does have an influence on the head orientation estimation accuracy. Users are suggested to use the default parameter $\rho = 1$ for initialization first, and HOE could search the optimal directivity factors for them after a period of use-and-feedback.

5.6 Impact of orientations

Fig. 19 illustrates the MAE of different orientations. Overall, the front-back directions (90° ↔ 270°) have lower estimation errors than the left-right aspects (*i.e.*, 180° ↔ 360°). Besides, 225° is the corner direction, the MAE of which is larger than the direction to the door (315°). This result indicates that a complex environment could make a negative effect on the estimation result. It is because that large objects like walls and furniture would block and corrupt signal propagation, which leads to energy fluctuation that does not accord with the expected energy attenuation pattern (Fig. 5(a)). CC (red sectors) and CCR (grey sectors) of different orientations are exhibited in Fig. 20. We can see that front-back directions have a higher CC value (70%) than left-right aspects (40%). As for CCR, the average value is 91%, and there is no obvious difference among all orientations, which confirms that HOE has a viable orientation estimation ability.

5.7 Impact of ambiguity

To evaluate the ability of disambiguation, we define a metric ADR (Ambiguity Detection Rate), which means the rate of correctly detecting the real head orientation from ambiguous candidates. We explore the relationship between ADR and estimation errors (*i.e.*, MAE), and position ten participants in an ADR-MAE coordinate. As shown in Fig. 21, each point represents a user with the corresponding index. ADR varies from 84.3% to 93.1% across different participants with an average of 89.1%. Moreover, as we discussed in Sec. 5.3, higher ADR could

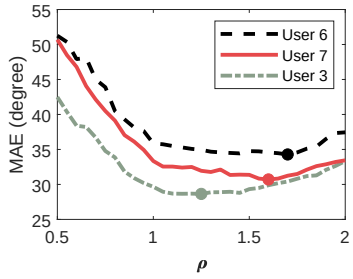


Fig. 18. Overall MAE across different directivity factors (ρ) for different users.

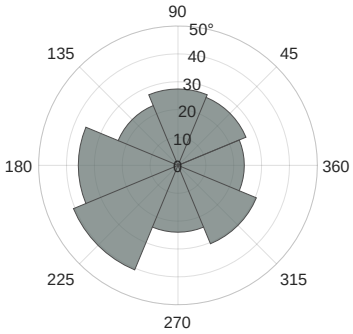


Fig. 19. Overall MAE across different head orientations.

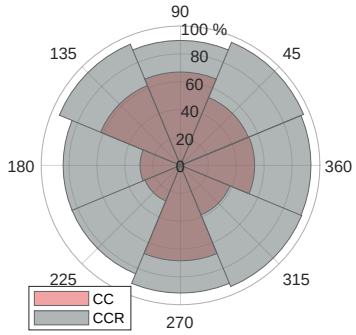


Fig. 20. Overall CC (red) and CCR (gray) across different head orientations.

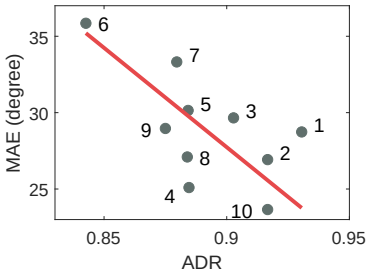


Fig. 21. Strong positive correlation between MAE and ADR.

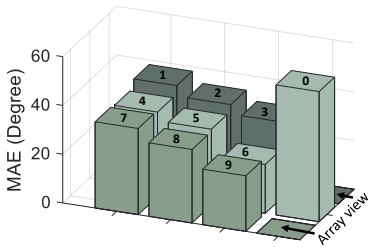


Fig. 22. Overall MAE across different locations (the view via arrays).

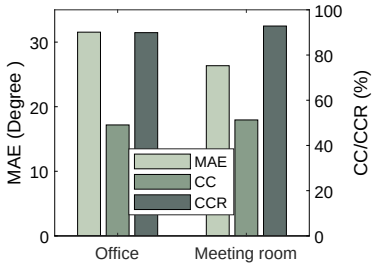


Fig. 23. HOE Performance in different rooms.

mitigate big errors and improve the overall performance. Evidently, we could infer a strong positive correlation between ADR and MAE, that is to say, the higher ADR, the lower estimation error. This result guides us to further advance the HOE performance by improving the ambiguity detection rate.

5.8 Impact of locations/rooms

The estimation performances of HOE at different locations in the office are shown in Fig. 22. To illustrate the disambiguation of HOE, we also conducted an additional test at location 0 where it is directly between two microphone arrays. The perspective view is from the array side, and the index of each bar represents the corresponding location in Fig. 14(c). The height of bars refers to the value of MAE. We see that the performance depends on the location. Specifically, nearer locations (*e.g.*, 3, 6, and 9) have a lower MAE than farther ones (*e.g.*, 2, 5, 8 or 1, 4, 7). It is because that farther positions suffer more from energy attenuation. As such, it is challenging for HOE to compensate for the energy precisely. Another finding is that the locations in the middle (*i.e.*, 4, 5, and 6) have lower estimation errors than the ones on the two sides (locations 1, 2, 3 and 7, 8, 9). This result indicates that the locations near the walls experience more complex voice propagation, which also causes inaccurate voice energy compensation. As a result, we see the largest errors at the two corners (*i.e.*, locations 1 and 7). There is a blind region of head orientation estimation where a user is between two arrays (*e.g.*, location 0), we can see that the estimation error increases to 53.3°. In this case, there is no *front half-circle* containing two arrays for disambiguation. So HOE can only randomly guess between two ambiguous orientation estimations, which leads

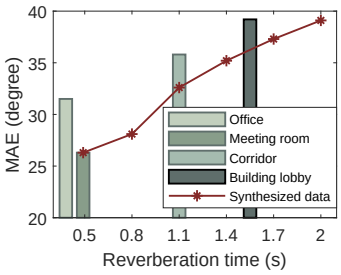


Fig. 24. Performance of different reverberation times.

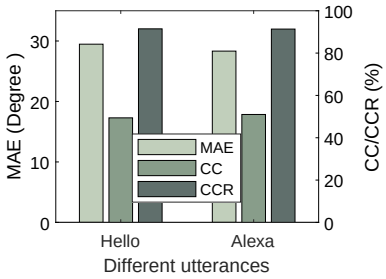


Fig. 25. Performance of different utterances.

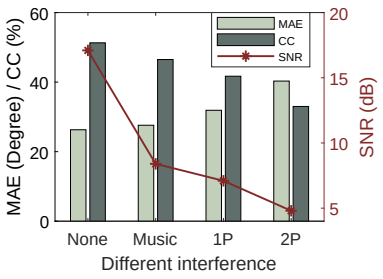


Fig. 26. Performance with different interference.

to a poor estimation performance. In practice, this problem can be mitigated by placing microphones in locations where such ambiguity could be avoided, or cooperating with other microphone arrays in the room if available.

We also compare the performance between two rooms: an office and a meeting room. As shown in Fig. 23, MAE corresponds to the left y-axis, and the right y-axis refers to CC/CCR. We can see that MAE of the meeting room is 26.3°, lower than office (31.5°). CC/CCR of the meeting room are 51.2% and 92.8% respectively, which are slightly higher than ones of the office. The reason is that the size of the meeting room is almost three times larger than the office, and there are fewer blocks and reflections. Therefore, the energy difference measured by arrays can match with the expected voice diffusion model more accurately.

5.9 Impact of reverberation time

We measured the reverberation time (RT60) of the two rooms above with ARTA software, and they are 0.38 *s* and 0.49 *s* respectively. By measuring the RT60 of many labs, offices, classrooms, and lecture halls, we found that all of them are within 0.6 *s*. To explore the HOE performance with different reverberation times, we synthesized the recordings with different RT60s based on the data collected in the meeting room. We also conducted the experiment in a building lobby and a closed corridor with RT60s of 1.1 *s* and 1.55 *s* respectively to investigate the HOE performance in real acoustically-wet environments.

The evaluation result is shown in Fig. 24. According to the Sabine equation [50], with the same room material, a higher reverberation time indicates a larger room volume. Therefore, when the reverberation time is within a low range (*e.g.*, $< 0.5s$), we can see a performance improvement for large rooms with fewer blocks and reflections (*e.g.*, meeting room vs. office). However, the estimation error increases gradually with an incremental RT60, since the wet component in recordings mainly makes a negative effect on the energy measurement and compensation. When the RT60 is 2 *s*, the MAE achieves 39.1°. Furthermore, the field experiments perform worse than synthesized recordings. The result is that besides reverberations, the energy measurement also suffered from noise like elevators and reflections from different interiors.

5.10 Impact of utterance

The performance of different utterances is illustrated in Fig. 25. Overall, the performance CC/CCR of these two commands are almost the same, while the MAE of "Alexa" (28.3°) is a little lower than the command "Hello". The reason may be that "Alexa" has more syllables than another one, making it easier to be captured by microphone arrays. Nowadays, most companies design a 4-syllable wake-up command for their voice assistant in smart devices, which is more effective to trigger.

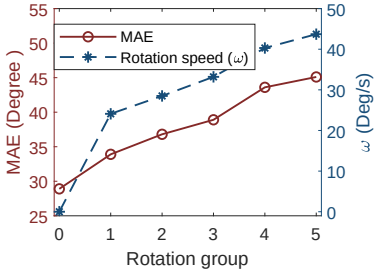


Fig. 27. Performance with different head rotation speeds.

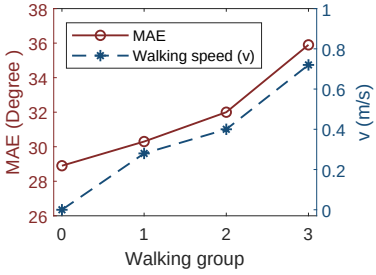


Fig. 28. Performance with different walking speeds.

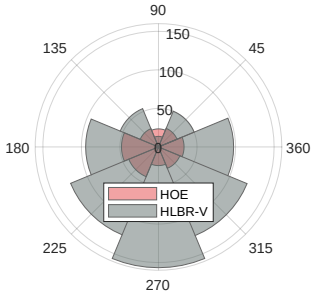


Fig. 29. MAE of HOE and HLBR-V [33] across different orientations.

5.11 Impact of interference

To evaluate the robustness to noise, we used an EARISE AL-202 loudspeaker placed near the wall, at the position where 3.6 *m* away from the user, and 3.1 *m* away from the right mic array. The volume of playing music is set to 60 *dB*. The voice SNR measured at the microphone is about 9 *dB*, which means the voice still dominates in recordings. The MAE of HOE slightly increases by 1.3°. Accordingly, CC decreases by 2.7%. This result indicates HOE is robust to the daily background music, since HOE employs beamforming to mitigate the noise from other directions and enhance the voice effectively. Considering that the distance is quite long and music noise may diffuse, we conducted another experiment interfering with one and two persons to further evaluate the HOE under directional near interference source. Specifically, we asked human interferers to read books with a normal volume, and walked around the target user while keeping a 1.5 *m* social distance when the user speaks commands.

Fig. 26 illustrates the performance of HOE with different interference conditions. When the voice SNR drops to 7.1 dB with the interference of one person (1P, female), the MAE decreases by 4.3° accordingly. The reason is two-fold. On the one hand, the energy of near human interference is comparable and even higher than the target user, leading to strong energy turbulence. On the other hand, the sound fields of multiple directional voice sources would overlap with each other and corrupt the original attenuation pattern of the target user. With two human interference (2P, female and male), the performance further deteriorates with a lower SNR of 4.8 dB. As a result, HOE can hardly measure the energy level correctly, leading to a decrease to 40.3° and 33% for MAE and CC, respectively. Besides former reasons, we also found that HOE misdetected the wake-up word occasionally in this case, since it may be overwhelmed in voice from interferers. Therefore, we do not suggest users use HOE in a very noisy (especially multi-person) scenario. Fortunately, when the user is listening to music or without interference, HOE provides a satisfying estimation result.

5.12 Impact of head rotation and movement

HOE estimates the user's head orientation only with the recording clip of the wake-up word. Generally, a wake-up word is very short and lasts about 500 *ms*. Therefore, we assume that the user's head keeps almost static for such a small duration. However, head orientation may change due to the subconscious motion. To further investigate its effect, we conducted two experiments with head rotation and walking. The first one is that users sat on a chair and were asked to speak voice commands towards eight directions with five different levels of rotation speeds. The second experiment is conducted where users walked from a 3 *m*-far wall towards microphone arrays with three different levels of speed, speaking three voice commands while keeping a fixed orientation. A camera was used to record the experiment and calculate the time spent as well as ground truth orientations.

Fig. 27 shows the HOE performance with different head rotation speeds. The rotation speed was controlled by users themselves and hard to follow a certain value, so we grouped all data into five different speed groups, and calculated the average rotation speed ω corresponding to each group. Group 0 means static. We can see that the MAE of HOE increases with the increasing rotation speed. This result is not surprising, since a tiny head rotation will lead to a large orientation shift. For example, when the average rotation speed is 33.2 $degree/s$ in group 3, the orientation shifted in a wake-up word duration (about 0.5 s) is 16.6 degree, and MAE raises to 38.9° accordingly.

The estimation results with different walking speeds are shown in Fig. 28. We also divided experiments into three groups with different levels of speed. Similarly, the higher walking speed is, the higher MAE is. The estimation error increases to 35.9° when the user walks with a speed of 0.72 m/s in group 3. Even though the user's orientation kept fixed during walking, the speed resulted in a location shift, which further caused errors in departure angle measurement and energy compensation. Moreover, we can see that the MAE caused by walking does not increase so fast as the rotation, since generally, the walking speed of users is not high when users interact with smart devices. Therefore, the distance shift is relatively shorter than the range between the user and smart device, and the effect of walking is not so notable as the head rotation which directly changes the orientation estimation.

5.13 Comparison with the model-based method

We implemented HLBR-V [33] as a model-based benchmark for comparison with HOE. This method regards the direction from the user to the array as a directional vector, whose norm equals the HLBR measurement of this array. Summing up all direction vectors of microphone arrays, it can estimate the head orientation as the direction of the summed vector. The mean estimation error of HOE (red sectors) and HLBR-V (gray sectors) are shown in Fig. 29. The overall MAE of latter is 89.3°, which is 50.4° higher than HOE. We can see that MAE of the front direction of this benchmark method is 13°, lower than HOE instead, since this range is the positive intersection area of two array direction vectors. However, the estimation error increases dramatically when the head orientation deviates from the front direction and ups to 156°when facing back to arrays. The reason is two-fold. On the one hand, HLBR-V cannot accurately estimate the head orientation when the number of microphone arrays is small (*i.e.*, a few direction vectors). Therefore, this method generally requires many arrays around the room covering all directions. On the other hand, the HLBR value is unstable since it highly relies on the thresholds separating the high and low-frequency bands.

5.14 Comparison with the ML-based approach

To make a comprehensive comparison with ML-based approaches, we implemented the state-of-the-art ML-based work [3] and evaluated it on our collected data. [3] utilizes the same microphone array as HOE and extracts hundreds of features to predict head orientation. As specified in [3], we implemented an EXTRA-trees classifier with 1000 estimators. We note that [3] is an 8-category classification problem. In contrast, HOE is a regression problem, which reports a degree-level estimation result. As such, we cannot compare the two methods directly on the basis of estimation error. Instead, we choose CC as the performance metrics. Since [3] is tested on one array, we first implemented [3] with the data collected from the left array (denoted as 'Left array'). Furthermore, we also implemented other two versions of the baseline: testing this left-array-trained model on the data collected by the right array, denoted as the 'Right array (L)'; and extending the one-array version and train it with the data of both arrays (noted as 'two arrays'). We compared HOE with [3] in four cases: per user, cross user, cross room, and training overhead. Specifically, for the baseline method, we left other users'/room's data to train the ML model and test it on the target user's/room's data. For HOE, we utilized the average value of other users'/room's profiles (*i.e.*, attenuation parameters and HLOBR threshold) to run the estimation algorithm on the target user's/room's data. The results are shown in Fig. 30.

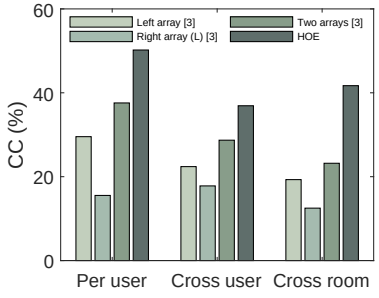


Fig. 30. Performance comparison between HOE and UIST'20 [3].

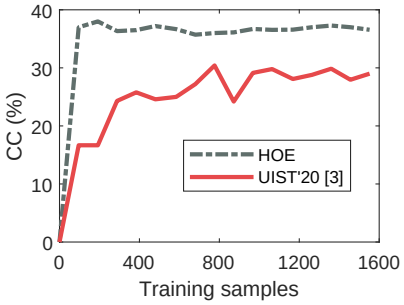


Fig. 31. Performance comparison along with different training sample sizes.

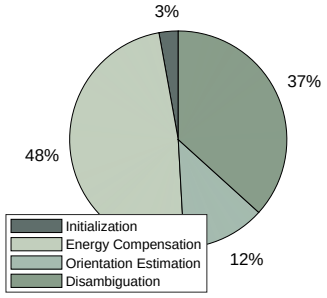


Fig. 32. Processing time of each component of HOE.

Per-user case. The CC of the baseline method on the left array is 29.5%, which is far lower than HOE (50.2%). This is because [3] is designed for relative orientation estimation (*i.e.*, deviation angles in our Problem Definition of Sec. 2), while the problem is to estimate the absolute orientation. We also test the 'Right array (L)' model, which CC drops to 15.5%. The reason is that [3] only learns the knowledge of relative orientation. However, some relative orientations became contrary when the reference coordinate is changed from the left array to the right array, although the absolute orientation (*i.e.*, true class label) remains fixed. The CC of 'Two arrays' model increases to 37.6%. This result indicates that combining the features of two arrays can improve the classification performance, since two arrays avoid the relative orientation confusion problem. However, this performance is still lower than HOE, because most features used in [3] are highly correlated with locations and reverberations (*e.g.*, auto/cross-correlation) but not with head orientations.

Cross-user case. As expected, the performance of most methods decreases in the cross-user case due to the generalization problem. The CC of 'Left array' and 'Two arrays' models decrease by 7.1% and 8.9%, respectively. The HOE performance also presents a considerable fall to 36.9%, by 13.3% compared to the CC before. In Fig. 9 we can see that the attenuation of different users in the same room is close although with variations. Therefore, the HLOBR threshold of users plays a more significant role in the performance than attenuation parameters, since HOE relies on the HLOBR threshold to remove the ambiguous orientations. However, the HLOBR threshold is quite user-dependent, so HOE performs worse in this case. An interesting finding is that the CC of 'Right array (L)' model increases slightly by 1.3%. We infer that cross-user data break the symmetrical orientation confusion problem of [3] to some extent, which increases the generalization of the ML model instead. But overall, the performance of HOE is still higher than the baseline approach. One reason is that the disambiguation of HOE is essentially a binary classification problem. Thus, it guarantees a half ambiguity detection ratio even though with a wrong HLOBR threshold.

Cross-room case. The performance change in cross-room cases is similar with cross-user ones compared with the per-user conditions: all models experience a decrease in CC as we expected. It is worth pointing out that HOE has a better cross-room performance than one in the cross-user task, since although the room attenuation parameters are different in cross-room cases, the HLOBR thresholds of users are close. Thus, estimation results have fewer large errors.

Training overhead. We tested [3] and HOE with a varied amount of training data. The result is shown in Fig. 31. The CC of [3] increases with the number of training data, and keeps nearly constant at about 28% when the size increases up to 1000. The performance of HOE reaches up to 37% at the beginning with the training size of 100 and remains stable. The reason is that HOE is a model-based method, and 100 samples are enough for the

one-time parameter training configuration. By contrast, ML-based methods need lots of data to train a ML model. This result indicates that HOE can achieve a higher estimation accuracy with the minimum training overhead.

5.15 Processing time

Fig. 32 shows the processing time of each component of HOE. Overall, HOE takes around 57.3 *ms* for one voice command. Specifically, the initialization takes around 3%, including the localization and distance/departure angle calculation. HOE takes 27.5 *ms* for energy compensation, since the filtering operation is computation-intensive. Orientation estimation only takes about 7.1 *ms*, while 37% of total time is used for disambiguation. This part requires filtering signal into high-frequency and low-frequency bands. Considering the powerful computation of commercial smart devices, we believe that HOE is capable to estimate the head orientation in real-time.

6 LIMITATION AND DISCUSSION

Different environments. We conducted most experiments in lab settings, *i.e.*, offices and meeting rooms in a university. They are constrained environments compared to our family scenarios such as a noisy living room or kitchen. We conducted more experiments in a corridor and a building lobby, and the result in Sec. 5.9 shows that the performance of HOE presents a degradation since it becomes challenging to measure an accurate energy level with high reverberations. Moreover, the energy measurement also suffers in a noisy environment. Therefore, we suggest users use HOE in an acoustically-dry environment to achieve higher accuracy. We also hope that we can work with the community to further improve its applicability and robustness in the future.

Head motion and Interference. The current version of HOE is not resilient to head rotation or movement. HOE performs head orientation using the recording clip of the wake-up word only. Generally, a wake-up word is very short, so we assume that the user's head keeps almost static for such a small duration. However, head orientation may change due to the subconscious motion, which unavoidably leads to performance degradation. Moreover, HOE performance also decreases with loud interference like a human voice, since the wake-up word may be overwhelmed and misdetected. Current HOE cannot handle multiple users speaking simultaneously. Fortunately, when the user is listening to background music or without interference, HOE could provide a satisfying estimation result.

Number of smart devices. In this paper, we design and implement HOE with two microphone arrays. They are utilized to localize the user's position, perform beamforming to eliminate interference, and measure the relative energy difference. According to the report [19], current U.S. households with smart speakers own an average of 2.6. Along with smart speakers, many smart devices equipped with microphone arrays for voice interaction. For example, TCL P717 Android TV integrates a 4-mic array [44]. We believe the design principle and proposed models are not limited to specific types of smart devices. We plan to enhance our head orientation method by leveraging more smart devices in the future.

7 CONCLUSION

The head orientation enables smart devices to sense additional context information of voice commands. It is no doubt that more novel interactions will emerge with the directional voice information, especially for smart home appliances, smart meeting rooms, or smart care for handicapped people. In this paper, we propose HOE, a model-based approach that estimates head orientation by two microphone arrays with a minimum training overhead. The energy radiation pattern of voice is used to compensate for the energy attenuation and estimate head orientation. We also propose a frequency radiation pattern-based method to resolve the estimation ambiguity problem. To the best of our knowledge, HOE is the first model-based method to estimate head orientation with two microphone arrays. We believe HOE is promising to bring the head orientation to various ubiquitous context-aware applications for smart devices.

8 ACKNOWLEDGMENTS

This work is supported by the Hong Kong GRF under grant PolyU 152165/19E and by the NSFC General Program under grant No. 61872081. Yuanqing Zheng is the corresponding author.

REFERENCES

[1] Alberto Abad, Carlos Segura, Duàn Macho, Javier Hernando, and Climent Nadeu. 2006. Audio person tracking in a smart-room environment. In *Ninth International Conference on Spoken Language Processing*.

[2] Alberto Abad, Carlos Segura, Climent Nadeu, and Javier Hernando. 2007. Audio-based approaches to head orientation estimation in a smart-room. In *Eighth Annual Conference of the International Speech Communication Association*.

[3] Karan Ahuja, Andy Kong, Mayank Goel, and Chris Harrison. 2020. Direction-of-Voice (DoV) Estimation for Intuitive Speech Interaction with Smart Device Ecosystems. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1121–1131.

[4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[5] Alessio Brutti, Maurizio Omologo, and Piergiorgio Svaizer. 2005. Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays. In *Ninth European Conference on Speech Communication and Technology*.

[6] Alessio Brutti, Maurizio Omologo, and Piergiorgio Svaizer. 2011. Inference of acoustic source directivity using environment awareness. In *2011 19th European Signal Processing Conference*. IEEE, 151–155.

[7] Alessio Brutti, Maurizio Omologo, Piergiorgio Svaizer, and Christian Zieger. 2007. Classification of acoustic maps to determine speaker position and orientation from a distributed microphone network. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 4. IEEE, IV–-493.

[8] Linsong Cheng, Zhao Wang, Yunting Zhang, Weiyi Wang, Weimin Xu, and Jiliang Wang. 2020. AcouRadar: Towards Single Source based Acoustic Localization. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 1848–1856.

[9] Wing Tin Chu and A C C Warnock. 2002. Detailed directivity of sound fields around human talkers. (2002).

[10] Travis C Collier, Alexander N G Kirschel, and Charles E Taylor. 2010. Acoustic localization of antbirds in a Mexican rainforest using a wireless sensor network. *The Journal of the Acoustical Society of America* 128, 1 (2010), 182–189.

[11] Zoey Collier. 2016. Beco Focuses on Developing a Spatially-Aware Alexa Skill. https://developer.amazon.com/blogs/alexa/post/Tx1BPHXBLZV5ZVN/beco-focuses-on-developing-a-spatially-aware-alexa-skill

[12] Ionut Constandache, Sharad Agarwal, Ivan Tashev, and Romit Roy Choudhury. 2014. Daredevil: indoor location using sound. *ACM SIGMOBILE Mobile Computing and Communications Review* 18, 2 (2014), 9–19.

[13] Eleftheria Georganti, Tobias May, Steven van de Par, Aki Harma, and John Mourjopoulos. 2011. Speaker distance detection using a single microphone. *IEEE transactions on audio, speech, and language processing* 19, 7 (2011), 1949–1961.

[14] Carlos T Ishi, Jani Even, and Norihiro Hagita. 2013. Using multiple microphone arrays and reflections for 3d localization of sound sources. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Ieee, 3937–3942.

[15] Charles Knapp and Glifford Carter. 1976. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing* 24, 4 (1976), 320–327.

[16] Teemu Korhonen. 2008. Acoustic localization using reverberation with virtual microphones. In *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC)*. Citeseer, 211–223.

[17] Avram Levi and Harvey Silverman. 2009. A robust method to extract talker azimuth orientation using a large-aperture microphone array. *IEEE transactions on audio, speech, and language processing* 18, 2 (2009), 277–285.

[18] Avram Levi and Harvey F Silverman. 2008. A new algorithm for the estimation of talker azimuthal orientation using a large aperture microphone array. In *2008 IEEE International Conference on Multimedia and Expo*. IEEE, 565–568.

[19] National Public Media. 2019. The Smart Audio Report 2019. https://www.nationalpublicmedia.com/uploads/2020/01/The-Smart-Audio-Report-Winter-2019.pdf Accessed October 19, 2020.

[20] Menno Müller, Steven van de Par, and Joerg Bitzer. 2016. Head-Orientation-Based Device Selection: Are You Talking to Me?. In *Speech Communication; 12. ITG Symposium*. VDE, 1–5.

[21] Bob Mungamuru and Parham Aarabi. 2004. Enhanced sound localization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 34, 3 (2004), 1526–1540.

[22] Kazuhiro Nakadai, Hirofumi Nakajima, Kentaro Yamada, Yuji Hasegawa, Takahiro Nakamura, and Hiroshi Tsujino. 2005. Sound source tracking with directivity pattern estimation using a 64 ch microphone array. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1690–1696.

[23] Hirofumi Nakajima, Keiko Kikuchi, Toru Daigo, Yutaka Kaneda, Kazuhiro Nakadai, and Yuji Hasegawa. 2009. Real-time sound source orientation estimation using a 96 channel microphone array. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 676–683.

[24] Alberto Yoshihiro Nakano and Phillip Mark Seymour Burt. 2013. Estimation of user orientation using GMMs for multiple voice-command devices environments. In *International workshop on telecommunications (IWT2013)*.
[25] Alberto Yoshihiro Nakano, Seiichi Nakagawa, and Kazumasa Yamamoto. 2009. Automatic estimation of position and orientation of an acoustic source by a microphone array network. *The Journal of the Acoustical Society of America* 126, 6 (2009), 3084–3094.
[26] Alberto Yoshihiro Nakano, Kazumasa Yamamoto, and Seiichi Nakagawa. 2009. Directional acoustic source's position and orientation estimation approach by a microphone array network. In *2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop*. IEEE, 606–611.
[27] Acoustical Society of America. 2004. *American National Standard Specification for Octave-band and Fractional-octave-band Analog and Digital Filters*. Standards Secretariat, Acoustical Society of America.
[28] Flávio Ribeiro, Demba Ba, Cha Zhang, and Dinei Florêncio. 2010. Turning enemies into friends: Using reflections to improve sound source localization. In *2010 IEEE International Conference on Multimedia and Expo*. IEEE, 731–736.
[29] Joshua M Sachar and Harvey F Silverman. 2004. A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4. IEEE, iv—-iv.
[30] Janos Sallai, Will Hedgecock, Peter Volgyesi, Andras Nadas, Gyorgy Balogh, and Akos Ledeczi. 2011. Weapon classification and shooter localization using distributed multichannel acoustic sensors. *Journal of systems architecture* 57, 10 (2011), 869–885.
[31] Akira Sasou. 2009. Acoustic head orientation estimation applied to powered wheelchair control. In *2009 Second International Conference on Robot Communication and Coordination*. IEEE, 1–6.
[32] Seeed. 2020. ReSpeaker Mic Array v2.0. https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0// Accessed October 29, 2020.
[33] C Segura. 2011. Speaker Localization and Orientation in Multimodal Smart Environments. *UPC, Barcelona, PhD Thesis* (2011).
[34] Carlos Segura, Alberto Abad, Javier Hernando, and Climent Nadeu. 2008. Speaker orientation estimation based on hybridation of GCC-PHAT and HLBR. In *Ninth Annual Conference of the International Speech Communication Association*.
[35] Carlos Segura, Cristian Canton-Ferrer, Alberto Abad, Josep R Casas, and Javier Hernando. 2007. Multimodal head orientation towards attention tracking in smartrooms. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 2. IEEE, II—-681.
[36] Carlos Segura and Javier Hernando. 2014. 3D joint speaker position and orientation tracking with particle filters. *Sensors* 14, 2 (2014), 2259–2279.
[37] Carlos Segura and Francisco Javier Hernando Pericás. 2012. GCC-PHAT based head orientation estimation. In *13th Annual Conference of International Speech Communication Association*. 1–4.
[38] Sheng Shen, Daguan Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury. 2020. Voice localization using nearby wall reflections. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.
[39] Harshavardhan Sundar, Weiran Wang, Ming Sun, and Chao Wang. 2020. Raw Waveform Based End-to-end Deep Convolutional Network for Spatial Localization of Multiple Acoustic Sources. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4642–4646.
[40] Piergiorgio Svaizer, Alessio Brutti, and Maurizio Omologo. 2012. Environment aware estimation of the orientation of acoustic sources using a line array. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 1024–1028.
[41] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. 2011. Single-channel head orientation estimation based on discrimination of acoustic transfer function. In *Twelfth Annual Conference of the International Speech Communication Association*.
[42] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. 2012. Estimation of talker's head orientation based on discrimination of the shape of cross-power spectrum phase coefficients. In *Thirteenth Annual Conference of the International Speech Communication Association*.
[43] Ivan Jelev Tashev. 2009. *Sound capture and processing: practical approaches*. John Wiley & Sons.
[44] TCL. 2021. P717 Series. 4K UHD ANDROID TV. https://www.tcl.com/hk/en/products/p717/p717-50.html.
[45] Masahito Togami and Yohei Kawaguchi. 2010. Head orientation estimation of a speaker by utilizing kurtosis of a DOA histogram with restoration of distance effect. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 133–136.
[46] Alex Waibe11, Hartwig Steusloff, Rainer Stiefelhagen, et al. 2005. CHIL: Computers in the human interaction loop. (2005).
[47] Weiguo Wang, Jinming Li, Yuan He, and Yunhao Liu. 2020. Symphony: localizing multiple acoustic sources with a single microphone array. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 82–94.
[48] Cheng-Yen Yang, Chih-Wei Liu, and Shyh-Jye Jou. 2016. A systematic ANSI S1. 11 filter bank specification relaxation and its efficient multirate architecture for hearing-aid systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 8 (2016), 1380–1392.
[49] Jackie Yang, Gaurab Banerjee, Vishesh Gupta, Monica S Lam, and James A Landay. 2020. Soundr: Head Position and Orientation Prediction Using a Microphone Array. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[50] Robert W Young. 1959. Sabine reverberation equation and sound power calculations. *The Journal of the Acoustical Society of America* 31, 7 (1959), 912–921.