MAVL: Multiresolution Analysis of Voice Localization

Mei Wang*, Wei Sun*, Lili Qiu
The University of Texas at Austin

Abstract

The ability for a smart speaker to localize a user based on his/her voice opens the door to many new applications. In this paper, we present a novel system, MAVL, to localize human voice. It consists of three major components: (i) We first develop a novel multi-resolution analysis to estimate the AoA of time-varying low-frequency coherent voice signals coming from multiple propagation paths; (ii) We then automatically estimate the room structure by emitting acoustic signals and developing an improved 3D MUSIC algorithm; (iii) We finally re-trace the paths using the estimated AoA and room structure to localize the voice. We implement a prototype system using a single speaker and a uniform circular microphone array. Our results show that it achieves median errors of 1.49^o and 3.33^o for the top two AoAs estimation and achieves median localization errors of $0.31m$ in line-of-sight (LoS) cases and $0.47m$ in non-line-of-sight (NLoS) cases.

1 Introduction

Motivation: The popularity of smart speakers has grown exponentially over the past few years due to the increasing penetration of IoT devices, voice commerce, and improved Internet connectivity. The global smart speaker market is estimated to grow at a rate of 21.12% annually and reach 19.91 billion US dollars in 2024.

The ability to localize human voice benefits smart speakers in many ways. First, knowing the user's location allows the smart speaker to beamform its transmission to the user so that it can both hear from and transmit to a faraway user. Second, the user location gives context information, which can help us better interpret the user's intent. For example, as shown in Figure 1, when the user issues the command to turn on the light, the smart speaker can resolve the ambiguity and tell which light to turn on depending on the user's location. In addition, knowing the location also enables location based services. For instance, a smart speaker can automatically adjust the temperature and lighting condition near the user. Moreover, location information can also help with speech recognition and natural language processing by providing important context information. For example, when a user says "orange" in the kitchen, it knows that refers to a fruit; when the same user says "orange" elsewhere, it may interpret that as a color.

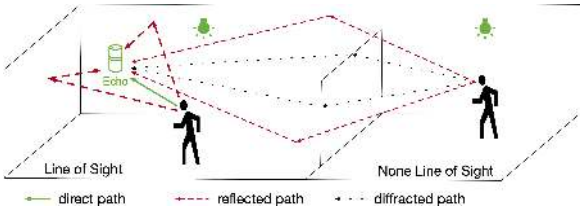


Figure 1: Illustration of application for MAVL under multiple coherent incoming paths in LoS and NLoS scenarios.

There have been a number of interesting works on motion tracking and localization using audio [23, 25, 29, 32, 41, 44, 50, 52], RF [43, 45, 48] and vision-based schemes [8, 53], etc. Cameras cannot be deployed everywhere at home for privacy concerns. Device-based tracking requires carrying a device, which is not convenient for people at home. Device-free RF is interesting, but requires large bandwidth, many antennas, or mmWave chips to achieve high accuracy, which is not easy to deploy at home. Meanwhile, acoustic-based tracking has also been shown to achieve high accuracy. In the past few years, acoustic tracking accuracy has improved from centimeter level [50] to millimeter level [23, 29, 41, 44]. These works focus on tracking users by emitting specially designed acoustic signals. These signals are mostly in inaudible frequency range $16kHz$-$22kHz$.

Challenges: Despite significant acoustic based tracking works, localizing human voice poses new challenges:

- Many of the existing systems require transmission of known signals (*e.g.*, chirps, OFDM symbols, sine waves).

*Both authors contributed equally to this work

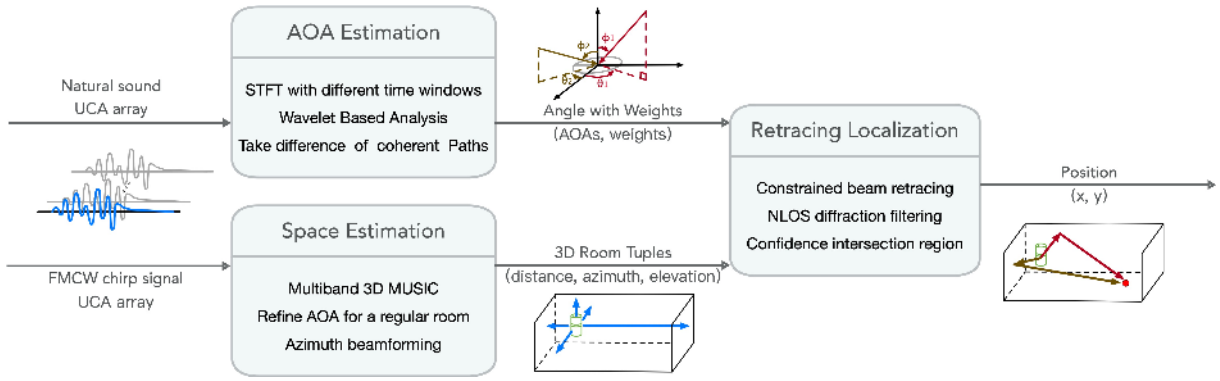


Figure 2: MAVL system involves a three-step process. (1) estimate AoA from multiple paths, (2) recover room structure by actively emitting wideband chirps, (3) localize the voice by retracing back the estimated AoA based on room structure.

In comparison, we can neither control nor predict users' voice signals, including their timing, frequency, and content. This makes it challenging to apply traditional channel estimation and distance estimation based methods.

- In order to localize a user, we need to estimate angle of arrival (AoA) of multiple propagation paths so that we can trace back these paths to localize the user. The signals traversing via multipath are coherent, which significantly degrades the accuracy of AoA estimation methods (*e.g.*, MUSIC requires all signals be independent).
- To enable retracing the location using multiple AoAs, we also need to estimate the indoor environment first. However, depth sensors are not widely deployed at home and vision-based approaches raise privacy concerns.
- The user may not be in line of sight (LoS) from the smart speaker (*e.g.*, the user is behind a wall or in a different room). Localizing the user in NLoS using acoustic signals remains an open problem due to low SNR and detoured propagation paths.

Our approach: In this paper, we build a novel indoor voice localization system, MAVL, by retracing multiple propagation paths that the user's sound traverses, as shown in Figure 2. First, we estimate AoAs of the multiple paths traversed by the voice signals from the user to the microphone array on the smart speaker. The multipath may include the direct path (if available) and the reflected paths. Second, we estimate the indoor space structure (*e.g.*, walls, ceilings) by emitting wideband chirps to estimate the AoA and distance to the reflectors in the room (*e.g.*, walls). Third, we re-trace the propagation paths based on the estimated AoA of the voice signals and the room structure to localize the voice.

We choose AoA since it eliminates the need of distance estimation, which is challenging when we do not know the transmission signals. We use a microphone array widely available on a smart speaker to collect the received signals. While there have been many AoA algorithms proposed, the low frequency of voice signals and the presence of coherent paths pose significant new challenges. To reduce coherence and separate paths, we capture the voice signal that finishes fast so that the signal traversing via the shortest path has small or no overlap with those traversing via the longer paths. We cannot control how many words a user speaks. Instead, we could select the voice signals that occupy some frequencies for a short time period. This requires good time and good frequency resolution. Since there is no single method that can simultaneously achieve good time and good frequency resolutions, we perform wavelet and STFT analyses over different time windows to benefit from both transient signals with low coherence and long signals with high cumulative energy. We further use differencing to cancel the signals in the time-frequency domain to reduce coherence, thereby improving the AoA accuracy.

Next we need to estimate the room contour, *i.e.*, the distances and direction of the walls. Researchers have used depth sensors [2, 15, 31], cameras [9, 18, 21] or multiple sensors [7, 12, 49] to estimate the indoor room contour. However, these systems require extra sensors and some need significant computation cost. It also raises significant privacy concerns. Acoustic has been applied to image the shape of objects [20, 24, 47]. It is promising to use acoustic signals to capture the room contour. Our system emits wide-band Frequency Modulated Continuous Waves (FMCW) chirps and propose the wide-band 3D MUSIC to estimate multiple propagation paths simultaneously. The wide bandwidth not only improves distance resolution, but also allows us to leverage the frequency diversity to estimate AoAs of coherent signals. We improve the AoA estimation by leveraging the assumption of a rectangle room (which is common in real world scenarios), and improve the distance estimation to the wall by using beamforming.

Finally, we develop a constrained beam retracing algorithm based on the estimated AoA candidates and room structure.

We localize the user at the intersection between the propagation paths with only one-time reflection. Our retracing can effectively identify the plausible user location.

We implement and evaluate our AoA and localization approaches in an anechoic chamber, conference room, bedroom and living room. Our results show that our AoA estimation yields median errors of 1.49^o and 3.33^o for the top two paths in LoS, and 2.75^o and 6.49^o in NLoS. Moreover, our retracing algorithm can localize the user with a median error of $0.31m$ in LoS and $0.47m$ in NLoS.

The contributions can be described as follows:

1. We develop a multi-resolution analysis to estimate the AoA of multipath. It combines STFS over different window sizes and wavelet to reduce coherence between signals.
2. We develop an effective method to estimate room structure and retrace the user based on the estimated AoA and room structure.
3. We implement a system to actively map indoor rooms and localize voice sources using only a smartspeaker without additional hardware. Our prototype system can localize voice in both LoS and NLoS. To our knowledge, this is the first indoor sound source localization system that works for None-Line-of-Sight (NLoS) scenarios.

2 Primer on AoA Estimation

In this section, we introduce AoA estimation problem, existing approaches, and challenges.

2.1 Antenna Array Model

We can estimate the AoA using an antenna array. The antenna array can take different forms, such as uniform circular array (UCA), uniform linear array (ULA), or even non-uniform array. This paper uses a uniform circular array consisting of N microphones as shown in Figure 3. The circle has a radius of r. The azimuth and elevation angles of signal arrival are θ and ϕ, respectively.

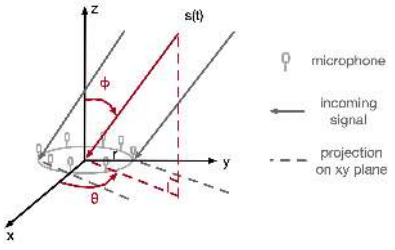


Figure 3: UCA Array model and angle notations.

A general model for the received signal of a single source is

$$x(t) = a(\theta,\phi)s(t) + n(t), \tag{1}$$

where a is the array steering vector and $n(t)$ is the noise vector. The steering vector for UCA is as follows:

$$a(\theta,\phi) = [1, e^{j2\pi\frac{f}{c}rcos(\theta)sin(\phi)}, \ldots, e^{j2\pi\frac{f}{c}r(N-1)cos(\theta)sin(\phi)}]^T. \tag{2}$$

where f is center frequency and c is sound propagation speed. For M independent source signals $S(t) = [s_1(t), \ldots, s_M(t)]^T$, we can extend the steering vector to a steering matrix, $A(\theta,\phi) = [a(\theta_1,\phi_1), \ldots, a(\theta_M,\phi_M)]$, where the i_{th} column is the steering vector associated with the i_{th} signal.

2.2 AoA Estimation Algorithms

There are several AoA estimation algorithms, including phase [43], MUSIC [35], ESPIRIT [17], and beamforming. The subspace based MUSIC algorithm is the most accurate. To apply MUSIC, we calculate the auto-correlation matrix R for the received signals x as $x^H x$, where x^H is conjugate transpose of x and R has the size $N \times N$. Following that, we apply eigenvalue decomposition to R, and sort the eigenvectors in a descending order in terms of the magnitude of corresponding eigenvalues. The space spanned by the first M eigenvectors is called *signal space*, and the space spanned by the other eigenvectors is called *noise space*. Let R_N denote the noise space matrix, whose the i_{th} column is the i_{th} eigenvector in the noise space. It can be shown that

$$R_N^H \cdot a(\theta_0,\phi_0) = 0, \tag{3}$$

when θ_0 and ϕ_0 are the incoming azimuth and elevation angles [35]. Based on this property, we can define a pseudo-spectrum of the mixed signals as

$$p(\theta,\phi) = \frac{1}{a(\theta,\phi)^H R_N R_N^H a(\theta,\phi)}. \tag{4}$$

Then we can estimate the AoA by locating peaks in the pseudo-spectrum.

2.3 Modeling Multipath Propagation

Now we consider signals under multipath propagation. Most traditional AoA estimation algorithms have the assumption that the signal sources should be independent. In contrast, our system requires estimating the AoA of multipath and have to handle coherent signals. To capture multipath effects, we introduce a channel matrix $H(\alpha,d) = [h(\alpha_1,d_1), \ldots, h(\alpha_M,d_M)]^T$, where α_i, d_i, and $h(\alpha_i,d_i) = \alpha_i\frac{d_0}{d_i}e^{j2\pi\frac{f}{c}d_i}$ are the attenuation, propagation delay, and channel of the $i-th$ path, respectively. The received signal $x(t)$ under multipath is as follows:

$$x(t) = A(\theta,\phi)H(\alpha,d)s(t) + n(t), \tag{5}$$

For the array model under multipath in Equation 5, we define a transformation matrix $T = A * H$ to capture the array manifold matrix A and propagation paths H. The transformation matrix T is

$$T_{i,j,k} = \alpha_j \frac{d_0}{d_j} e^{j2\pi \frac{d_j}{\lambda_k}} e^{j2\pi \frac{r}{\lambda_k}(i-1)cos(\theta_j)sin(\phi_j)} \quad (6)$$

where $1 \leq i \leq N$ denotes the microphone index, $1 \leq j \leq M$ denotes the j_{th} arrival path, and k denotes the frequency bin index. The transformation matrix T takes three dimensions: spatial dimension i, path delay in time dimension j, and frequency dimension k, which allows us to perform cancellation in the time-frequency domain.

The received signal from all incoming paths to microphone m_i on frequency f_k is

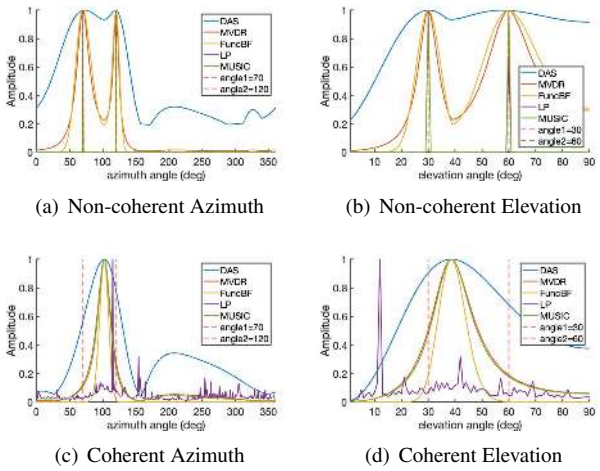
$$x(t)_{i,k} = \hat{T}_{i,k} * s(t) + n(t). \quad (7)$$

where $\hat{T}_{i,k} = \sum_{1 \leq j \leq M} T_{i,j,k}$. In order to estimate the AoA of multipath, we need to deconvolve $\hat{T}_{i,k}$ to each propagation path $T_{i,j,k}$.

2.4 Challenges

Coherent signals: A major source of AoA error comes from the coherence in the incoming signals. In our context, the received signals come from the same voice source and only differ in their propagation paths. Such strong correlation can significantly degrade the AoA estimation accuracy. We quantify the impact of coherent signals on several well-known AoA estimation schemes in the frequency range of human voice. We use a UCA with radius of $9.6cm$, which is approximately the half wavelength of $2kHz$. The two signals are $(70, 120)$ and $(30, 60)$ in the azimuth and elevation angles. Figure 4(a) and (b) are the azimuth and elevation power profiles of five AoA algorithms for two non-coherent signals and Figure 4(c) and (d) are profiles of two coherent multipath signals coming from the same source. MUSIC performs the best in all scenarios. However, when coherence occurs, the estimation errors increase in all algorithms. For example, the two peaks in MUSIC merge into one peak in this case and LP even gives incorrect results.

Impact of frequency: The low frequency of the voice also accounts for part of the error. Existing acoustic tracking schemes (*e.g.*, [23,25,44]) use frequency at $16kHz$ or higher. In comparison, human voice is typically below $6kHz$ [27, 33] and most energy is concentrated in $100Hz$-$3kHz$. The corresponding wavelength ranges between $11cm$ and $3.4m$. The resolution of angle of arrival is determined by the antenna separation distance normalized by the wavelength. Therefore, with centimeter level separation between the microphones and dm-level wavelength, the AoA resolution is very coarse.



(a) Non-coherent Azimuth (b) Non-coherent Elevation

(c) Coherent Azimuth (d) Coherent Elevation

Figure 4: Comparison of power profiles for different AoA algorithms in non-coherent (a,b) and coherent (c,d) scenarios. Coherence makes peaks merged and introduces error.

Summary: The above evaluation shows that MUSIC is competitive for AoA estimation accuracy. However, the accuracy is still insufficient to support coherent low-frequency voice signals. Motivated by these observations, next we will design approaches to explicitly address these major challenges.

3 Multipath Voice Localization

We decompose our approach into the following three steps: (i) estimate the AoA of coherent low-frequency voice signals, (ii) estimate the room structure, (iii) retrace the paths to localize the user. Below we describe each step in turn.

3.1 AoA Estimation of Voice Signals

As shown in Section 2, we should address two major challenges in AoA estimation of human voices: (i) received signals are strongly correlated and (ii) limited resolution due to the low frequency of human voice. Below we describe our sections in turn.

Limitation of existing work: Recently, Voloc [37] proposed an iterative-delay-and-cancellation algorithm to align and cancel the correlated paths to separate multipath signals in the time domain. Their first step, called ICA, is to estimate the AoA of the first reflection by using the initial recording samples before mixing with the second reflection. However, this method introduces two major problems. First, in order to cancel in the time domain, we need to use a small enough time window during which only samples from the direct path are included, usually only tens of samples. A small number of samples limits the AoA estimation accuracy. Moreover, hu-

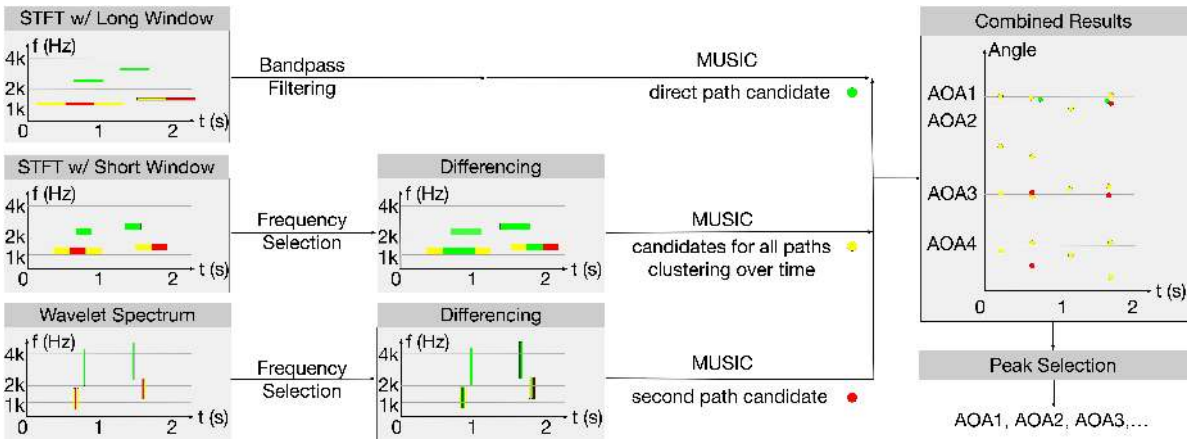


Figure 5: Illustration for multi-resolution analysis algorithm. We perform wavelet and STFT analyses over different time windows followed by the differencing component for small windows. We synthesize the combined results to select the final AoA results.

man voice ramps up slowly. This means the beginning cleaner audio samples for AoA estimation have low SNR, which also limits the accuracy. In addition, the cyclic autocorrelation property of human voice is large, which indicates small alignment error introduces large cancellation error. Therefore, Voloc reports over 10 degrees error for the first path AoA and relies on their second step, which uses joint optimization based on wall geometry to refine the estimation result. This has several limitations: (i) its standalone AoA estimation has limited accuracy, and (ii) the second step requires exploring a large search space, which is very time consuming (*e.g.*, hours to estimate wall parameters and 5 seconds to localize voice).

Overview: Different from [37], we use time-frequency analysis to reduce coherence in voice signals since signals that differ in either time or frequency will be separated out. As the transformation matrix $T_{i,j,k}$ shown in Equation 6, the IAC algorithm in Voloc aligns phases for each microphone i to cancel path delays d_j and get the second reflected path. We first separate coherence in across different frequency bins, and then cancel the paths in each frequency bin by taking the difference between the two consecutive time windows. This is especially useful for voice signals since different pitches may occur at different time. An important decision in time-frequency analysis is to select the sizes of time window and frequency bin to perform the analysis.

On one hand, aggregating the signals over a larger time window and larger frequency bin improves SNR and in turn improves the AoA estimation accuracy according to the Cramer-Rao bound [38]. On the other hand, a larger time window and larger frequency bin also mean more coherent signals. Moreover, the frequency of voice signals varies unpredictably over time, which makes it challenging to determine a fixed time window and frequency bin.

To separate paths with different delay, we desire good time resolution. Small time windows have good time resolution, but poor frequency resolution. To separate paths with different frequencies, we desire good frequency resolution. Small frequency bins have good frequency resolution, but poor time resolution. Therefore, there is no single time window or frequency bin that works well in all cases.

To address this challenge, we use multi-resolution analysis as illustrated in Figure 5. Specifically, we use Short Term Fourier Transform (STFT) with different window sizes and wavelet as they are complementary to each other. Our first method performs STFT using a large time window and feeds the spectrogram to MUSIC. While STFT results with large window have more coherent signals, which results in more outliers, their peaks also include points that are close to the ground truth, likely due to the stronger cumulative energy. Our second method is to perform frequency analysis using smaller windows and take difference between adjacent windows to reduce the coherent signals and improve AoA estimation under coherent multipath. Our third method uses wavelet. It has higher time resolution for relatively high frequency signals. This allows us to capture the transient voice signals that has low or no coherence, thereby reducing outliers in MUSIC AoA estimation. However, since transient signals have low cumulative energy and cause non-negligible AoA estimation errors, we combine Wavelet with STFT with different window sizes. Below we elaborate these three methods.

STFT using a large window size: We perform STFT using a larger time window. A larger window yields higher SNR and hence higher accuracy according to the Cramer-Rao bound [38]. On the other hand, a larger window tends to have more coherent multipath, which may degrade the accuracy. This is shown in Figure 4(c), where we see a merged peak near the ground truth. So this approach can provide information about the AoA of the direct path, but not sufficient on its

own.

STFT using a short window: Using a smaller time window gives good time resolution and helps separate paths with different delays. We choose to use a smaller time window and select the evanescent pitches in the time-frequency domain to reduce error from coherence. The next step is to further reduce coherent signals by taking difference between two consecutive time windows for each antenna. This cancels the paths with different delay in the time-frequency domain, and is more effective than cancelling in the time-domain alone. If the difference between two adjacent windows is greater than the delay difference of any two paths, this process can remove the old paths. This cancellation is not perfect since the amplitude may vary over time and each window may contain different sets of paths. Nevertheless it reduces coherence in a short time window.

Wavelet based analysis: Wavelet is a multi-resolution analysis. We can use both short basis functions to isolate signal discontinuities and long basis functions to perform detailed frequency analysis. It has super resolution for relatively high frequency signals. Transient signals in small time window have less energy and may yield large errors. To improve the accuracy, we also take difference of wavelet spectrum in the two consecutive time windows to further reduce the coherence.

Comparison: We compare the AoA derived from applying MUSIC to STFT and wavelet. Figure 6 shows the result for the case where a woman speaks at 2.4m away from the microphone array. The dashed red lines are ground truth AoAs of different paths. The STFT results without taking difference, shown in the blue circles, deviate from the right angles due to coherence even after using different window sizes. The wavelet results without taking difference are plot as yellow circles, which also deviates a lot from red dashed lines because of low energy. The stared orange and purple points are the AoA estimates derived from MUSIC when we apply differencing to STFT and wavelet, called STFT Diff and Wavelet Diff methods. Compared with the original results (shown in blue and yellow circles), differencing brings the estimation closer to the ground truth angles (shown as dashed lines). It is interesting to observe that there are false peaks in STFT Diff but the peaks in the Wavelet Diff are all close to the ground truth though STFT Diff may have peaks closer to the ground truth than the wavelet. This suggests that it is beneficial to combine STFT Diff and wavelet Diff results.

Final algorithm: Figure 5 shows our final algorithm. For each algorithm, we derive the results using different time windows. Then we compute weighted cluster of these points where the weight is set according to the magnitude of the MUSIC peak. We select the top K clusters from each algorithm. Our evaluation uses $K = 6$. To combine the results across

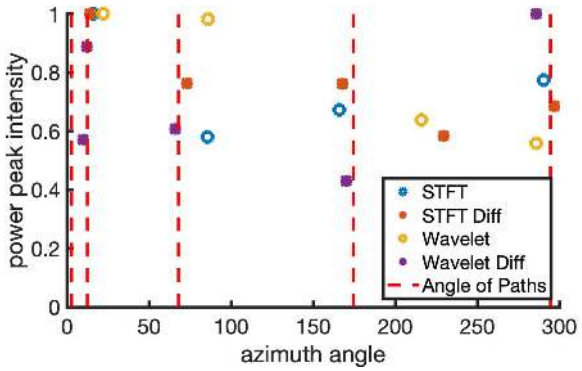


Figure 6: Comparison of AoA derived from STFT, Wavelet with and without differencing.

different algorithms, we use nearest neighbors. Since STFT with a large window provides more stable results without significant outliers, we use them to form the base. For each point in the base, we search for the nearest neighbor in the results of the other two methods as they contain both more accurate real peaks and outlier peaks. Finally, we pick the top P peaks from the selected nearest neighbors as the final AoA estimates. Our evaluation uses $P = 5$.

Algorithm 1 Multi-resolution analysis algorithm.

```
function [AoAs, w] = MultiResolutionAoA(signal)
Bandpass filter in voice frequency range
spectLong = STFT(signal,LongWindow);
spectShortDiff = diff(STFT(signal,ShortWindow));
spectWaveletDiff = diff(Wavelet(signal));
Select frequency and time ranges based on spectrograms
for method in STFTLong,STFTDiff,WaveletDiff do
    for time in SelectedTimeSlots do
        for frequency in SelectedFrequencies do
            forward backward smoothing;
            compute MUSIC profile;
        end for
        accumProfile = SUM(profile)
        [results,weights] = findPeaks(accumProfile);
        estimate candidateAoAs_m and weights_m;
    end for
end for
AoAs = select top P peaks from candidateAoAs_m for m=1..3
```

3.2 Room Structure Estimation

In order to localize the user, we need not only the AoAs of the propagation paths of the voice signals, but also the room structure information so as to retrace back the paths. In this section, we estimate the room contour using wideband 3D MUSIC algorithms. We improve the accuracy by leveraging constraints of the azimuth AoA and applying beamforming.

3.2.1 3D MUSIC

The smart speaker estimates the room structure once unless it is moved to a new position. The smart speaker estimates room structure by sending FMCW chirps. Let f_c, B and T denote the center frequency, bandwith, duration of the chirp. Upon receiving the reflected signals, it applies the 3D MUSIC algorithm.

We generalize 2D Range-Azimuth MUSIC algorithm [5,6,22] to 3D joint estimation of distance, azimuth AoA and elevation AoA. 3D MUSIC has better resolution than 2D MUSIC since the peaks that differ in any of the three dimensions are separated out. Our basic idea is to transform the received signals into a 3D sinusoid whose frequencies are proportional to the distance and a function of the two angles. We extend the steering vector to have three input parameters: distance R, azimuth angle θ, and elevation angle ϕ.

$$\hat{a}(R,\theta,\phi) = e^{j2\pi\frac{r}{c}f_c \sin\phi\cos(\theta-\frac{2\pi i}{N})+j4\pi\frac{RB}{cT}N_sM_sT_s}, \qquad (8)$$

where i is the array index, N is the number of microphones, r is the radius of the microphone array, c is sound speed, N_s is the subsampling rate, M_s is the temporal smoothing window and T_s is the time interval.

3.2.2 Our Enhancements

However, there are several challenges in applying the 3D MUSIC algorithm to indoor environments. First, the number of microphones and their sizes are both limited, which limits the resolution of 3D MUSIC. Second, there is significant reverberation in indoor scenarios. Third, large bandwidth is required to get accurate distance estimation, but MUSIC requires narrowband signals for AoA estimation. Therefore, we develop three techniques to improve the 3D MUSIC algorithm: (i) leveraging frequency diversity, (ii) incorporating the fact that rooms are typically rectangular shaped, and (iii) using beamforming to improve distance estimation.

Multiband 3D MUSIC: We use FMCW signals from $1kHz$ to $3kHz$ for AoA estimation. To satisfy the narrowband requirement in the MUSIC algorithm [35], we divide the 2 KHz bandwidth into 20 subbands each with $100Hz$. Since the frequency of FMCW signal increases linearly over time, we can divide the FMCW signal into multiple subbands in the time domain, run 3D MUSIC in each subband, and then sum up the MUSIC profiles from all subbands.

In order to use the $100Hz$ subband for 3D MUSIC, we should properly align the transmission signal with the received signal so that they span the same subband. The alignment is determined by the distance. Therefore, we search over the azimuth and distance for a peak in the 3D MUSIC profile obtained by mixing the received signal with the transmitted signal that is sent δT ago, where δT is the propagation delay and determined based on the distance.

We use the azimuth AoA and distance output from the 3D MUSIC. Figure 7 shows an example of azimuth-distance profile. Note that we adjust the elevation angle to the horizontal AoA since the elevation AoA estimation from the UCA (which has all antennas on the same horizontal plane) is not very accurate. However, despite a larger error in elevation AoA, the 3D MUSIC is more effective in separating the paths than the 2D MUSIC.

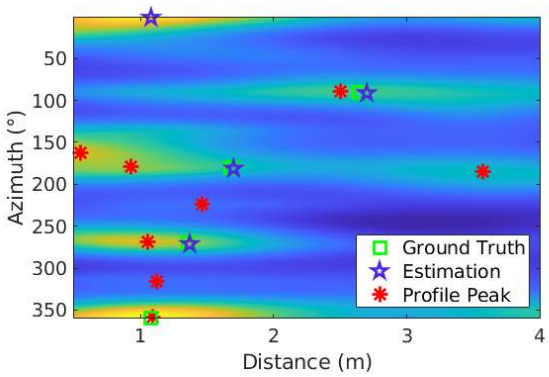


Figure 7: An example of azimuth-distance profile from real trace. Azimuths are accurate and distances requires further the fine granularity search.

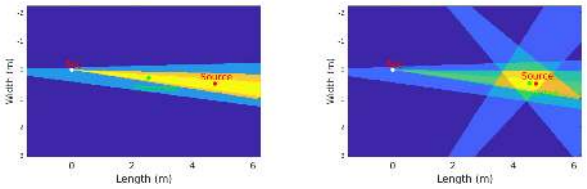
Refine AoA for a regular room: Due to multipath, the MUSIC profile can be noisy, which makes it hard to determine the right peaks to use for distance and AoA estimation of walls. Since most rooms take rectangular shapes, we leverage this information to improve peak selection. Specifically, we select the peaks such that the difference in the azimuth AoA of two consecutive peaks are as close to 90^o as possible. That is, we search for the 4 peaks $\{\theta_0,\theta_1,\theta_2,\theta_3\}$ from the 3D MUSIC profile that minimizes the fitting error with a rectangular room (*i.e.*, $min\sum_i |PhaseDiff(\theta_i-\theta_{i+1})-\pi/2|$, where $PhaseDiff(.)$ is the difference between the two angles by taking into account of the phase wraps every 2π. After finding these peaks, we further adjust the solutions so that the difference between the adjacent AoA is exactly $\pi/2$. This can be done by find θ'_1 that minimizes $\sum_i |PhaseDiff(\theta'_1+\pi/2(i-1)-\theta_i)|$ and the final AoA is set to $(\theta'_1,\theta'_1+\pi/2,\theta'_1+\pi,\theta'_1+3/2\pi)$.

Improve distance estimation by beamforming: Accurate distance estimation requires a large bandwidth and high SNR. Therefore, to improve distance estimation, we send $1kHz$-$10kHz$ FMCW chirps. Among them, we only use $1KHz-3KHz$ for AoA estimation to reduce computational cost since MUSIC requires expensive eigenvalue decomposition, but use the entire FMCW for distance estimation. We increase the SNR using beamforming. We use delay-and-sum (DAS) beamforming algorithm towards the estimated azimuth AoAs. Then we search a peak in the beamformed FMCW profile. We find that the peak magnitude increases significantly and get

more accurate distance estimation after beamforming.

3.3 Constrained Beam Retracing

We can localize the user by retracing the paths using the estimated AoA of the voice signals and room structure. As shown in left figure of Figure 8, we can first find the reflection points on the walls by the propagation path derived from the estimated AoA. Then we trace back the incoming path of voice signals before the wall reflection based on the reflection property. If we have at least two paths, the user is localized at the intersection between the incoming paths. However, the above method is not robust against AoA estimation error. When simulating the retracing algorithm, we find that even when the AoA estimation errors of 2 paths are only 0.5 degrees, it can cause a localization error of more than 60 cm at a distance of 4 meters. A small AoA error can result in a large localization error at a large distance. Moreover, an AoA error in the outgoing path can result in an error in the incoming path, thereby further amplifying this effect. To enhance robustness



(a) Two near parallel paths (b) More paths

Figure 8: Retracing using ray or cone.

against AoA estimation, we employ two strategies. First, instead of treating each propagation path as a ray defined by the estimated AoA, we treat it as a cone where the cone center is determined by the estimated AoA and the cone width is determined by the MUSIC peak width. This allows us to capture the uncertainty in the AoA estimation.

Second, while theoretically two paths are sufficient to perform triangulation, it is challenging to select the right paths for triangulation. Therefore, instead of prematurely selecting incorrect paths, we let the AoA estimation procedure return more paths so that we can incorporate the room structure to make informed decision on which paths to use for localization. Specifically, for each of the K paths returned by our AoA estimation, we trace back using the cone structure as shown in Figure 8. We observe that the azimuth AoA is reliable for the strongest path, which is the direct path in LoS or the path from the user to the ceiling and then to the microphone in NLoS. Therefore, within the cone corresponding to the strongest path we search for a point O such that the circle centered at the point with radius of 0.5m overlaps with the maximum number of cones corresponding to the other $K-1$ paths. We localize the user at the point O. Our evaluation sets $K=4$.

4 Implementation

Setups: We implement our system on a Bela platform [4]. It is connected with a JBL Clip 3 or an echo dot speaker and a circular microphone array with 8 microphones. Figure 9 shows an example setup in a conference room. Each microphone uses a sampling rate of $22.05kHz$. Many commercial smart speakers have similar numbers of speakers and microphones. We test our system using two microphone arrays: a larger array has radius of $9.6cm$ and a smaller one has radius of $5.0cm$. We use the smaller array to compare with VoLoc [37] since its size is similar to their setup. The Bela board uses a 1 GHz ARM Cortex-A8 single-core processor. The Bela is connected to a laptop with Intel I5 processor and $8GB$ memory. We use javaosc protocol to listen and continuously transmit the audio signals in WAV format encapsulated in OSC packets to the laptop through USB in real time and run the processing program in MATLAB on the laptop to derive the AoAs and localize the user. In MAVL , AoA estimation takes 2.35 seconds, room estimation takes 87 seconds, and retracing takes 0.16 seconds. In comparison, VoLoc spends hours in estimating wall parameters and 5 seconds in AoA estimation.

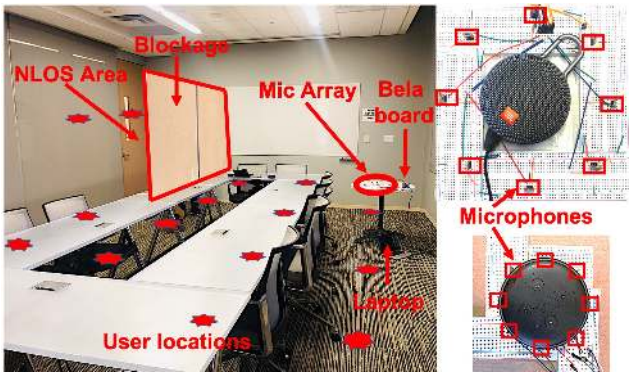


Figure 9: System setups in conference room and mic arrays.

Evaluation environments: We evaluate our system in different environments, including an anechoic chamber, conference room, bedroom and living room. These rooms take different sizes: 2.5m×3.5m, 3.5m×4.0m, and 5.1m×7.5m. We use a wooden board as a blockage in NLoS cases as shown in Figure 9. We let a person speak at $1-6$ meters away from the microphone array in the room. We also vary the distance, users, type of voices (*e.g.*, man, women, child and applause), smartspeaker positions, clutter and noise levels to assess their impacts.

Ground truth: We measure the relative locations of the smartspeaker, user and walls using a measuring tap. We derive the ground truth AoAs of the direct path and 5 reflected paths (*i.e.*, the paths from 4 side walls and ceiling) in LoS scenarios. In NLoS scenarios, we derive the AoAs of the 4 reflected paths and 1 diffraction path.

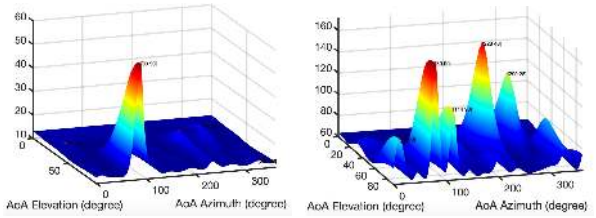
Metrics: We quantify the errors using both AoA estimation error and localization error. The localization error is computed based on the Euclidean distance between the ground truth and estimated positions.

5 Evaluation

In this section, we evaluate our AoA estimation, room contour estimation, and voice localization accuracy.

5.1 Performance of AoA Estimation

Two paths in anechoic chamber. We start from testing our AoA estimation algorithm in the anechoic chamber, where there is no reflection in the room. We put our microphone array on the ground and place an acrylic board to act as a wall to introduce a reflection path. The ground truth of two angles are 81.95^o and 112.68^o. Figure 10 shows the MUSIC power profile. It has a single merged peak around 90^o, which results in 8^o and 22.68^o errors for the two paths. In comparison, our algorithm accurately estimates these two paths within the error of 1.5^o. We can clearly see there are two separate peaks in our MUSIC profile Figure 10. We also change the acrylic board reflector to other places, and find that MUSIC can separate the two paths only when the difference between two ground truth angles is greater than 90^o. This resolution is not sufficient for voice localization since it is quite likely to have reflected paths within 90^o. In comparison, our approach can separate the two paths as long as they are 30^o apart.



(a) MUSIC with merged peak. (b) MAVL with separated peaks.

Figure 10: Comparison of power profiles in anechoic chamber.

AoA accuracy for LoS and NLoS: Next we conduct experiments in three rooms. Figure 11 shows the CDF of LoS AoA estimation error of six methods for the top 3 angles across all experiments. We use a large UCA of radius $9.6cm$, comparable to Amazon Echo Studio, Google Home Max and Apple HomePod. The median error of our approach for the top two paths are 1.49^o and 3.33^o, respectively. This accuracy is sufficient for retracing. In comparison, the corresponding numbers for MUSIC are 2.55^o and 14.54^o , which are significantly worse.

Figure 12 shows the CDF of NLoS AoA estimation error for the top 3 angles across all experiments. The median errors of the top two paths are 2.75^o and 6.49^o. We also plot the CDF for the third angle estimation. In theory, one can retrace the user's location using two paths. However, a median error around 10^o for the third path is too large to be used directly for triangulation. Nevertheless our cone-based retracing algorithm can still leverage the AoA of the paths beyond the top two paths to improve the localization accuracy despite their relatively high errors.

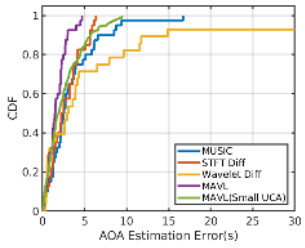
We also evaluate MAVL using a smaller UCA with a radius of $5cm$, comparable to the size of Echo Dot, Amazon Echo and Google home. Figure 13 compares the AoA accuracy of the first path with MAVL using small UCA, VoLoc using ICA algorithm only and VoLoc using joint estimation. Using our approach, the median AoA error of the first path is 1.98^o and the second path is 4.08^o, both of which are larger than the errors from the larger UCA, which are 1.49^o and 3.33^o, respectively. In comparison, Voloc yields median errors of 18.04^o and 5.28^o before and after joint optimization, respectively, much larger than the errors of MAVL.

AoA performance to distance: Figure 14 plots the AoA error versus the distance between the user and smart speaker in a $7.5m \times 5.1m$ conference room. Overall, the accuracy degrades slightly as the user moves away from the microphone array. The SNR of voice is not a serious problem because its frequency is low and it attenuates slowly in the air.

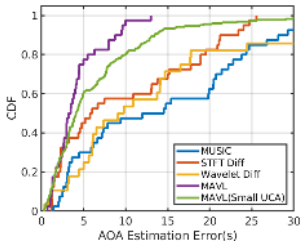
Interestingly, the AoA error of our approach at $4m$ is better than many other distances. This could be due to the specific room structure and user's distance to the nearby wall. Measurements at the distance around $4m$ were collected when the user is near the middle of the room, which makes the propagation delay from the reflected path well separated from the direct path and alleviates the coherence effects. The measurements at a larger distance (*e.g.*, $5m$) were collected when the user was close to the wall and the difference between the direct path and reflect path is smaller, which makes it more challenging to separate in the MUSIC profile.

Performance to different voices: We classify our measurements into four groups: *i.e.*man, woman, child, and applause. Figure 15 shows the sensitivity to different users' voices. The bars are centered at the mean error and their two ends denote the minimum and maximum values across all traces. Our system is fairly robust across the users and the voice they produced. We also evaluate the applause sound, and find the AoA errors of the two paths are about 1.4^o and 3.0^o. The applause sound has smaller AoA error because it is shorter than the human voice, which reduces coherence and improves AoA estimation accuracy.

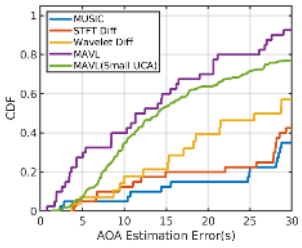
Impact of smartspeaker positions: The relative positions between the microphone array and walls have direct influence on multiple propagation paths. VoLoc requires the microphone array to be close to a wall to ensure that the first two



(a) AoA 1 CDF.



(b) AoA 2 CDF.



(c) AoA 3 CDF.

Figure 11: Comparison of LoS CDF of AoA estimation.

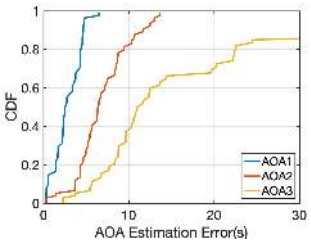


Figure 12: CDF of AoAs error for NLoS.

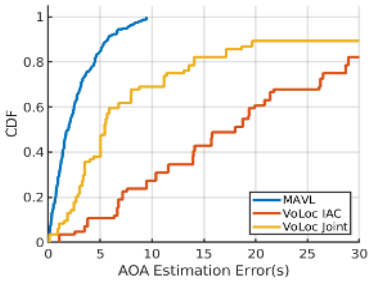


Figure 13: Comparison of AoA estimation for the small UCA.

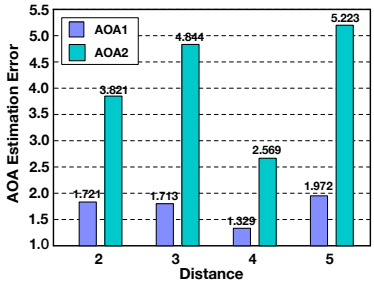


Figure 14: AoA accuracy vs distance.

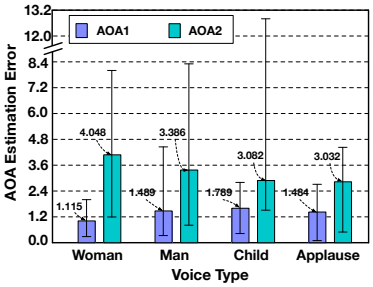


Figure 15: AoA accuracy to voices.

paths come much earlier than other paths. We evaluate the robustness of MAVL against smartspeaker positions. We evaluate UCA setup at three positions:

(1) *center*: 2.35m and 2.92m to the two closest walls;
(2) *close to one wall*: 0.3m and 2.4m to the two closest walls;
(3) *corner*: 0.26m and 0.39m to the two closest walls.

The median AoA errors of MAVL are 1.80^o, 1.97^o, 2.08^o for the direct path, when the smartspeaker is at *center*, *close to one wall* and *corner*, respectively; the corresponding AoA errors are 3.07^o, 4.51^o, 4.37^o for the second path AoA, respectively. MAVL performs best at the *center* and worst near the *corner*. The latter is because the second and third paths have comparable SNR and closer AoAs to the direct path, which increases coherence. But overall it is fairly robust to different placement. In comparison, the median AoA error of VoLoc before its joint optimization is 18.04^o for direct path, when the UCA is placed *close to one wall*. It does not work at the *center* or *corner*. VoLoc only works when the UCA is *close to one wall* and users are not close to any wall.

5.2 Performance of Room Estimation

Next we evaluate our room structure estimation algorithm using different room sizes and microphone placements.

Overall Room estimation Performance: We use room sizes of 2.5m×3.5m, 3.5m×4.0m, and 5.1m×7.5m. The median distance error for all walls is 2.8cm and azimuth error is 1.8^o. We can reduce the azimuth error to 1.4^o by leveraging the knowledge of room shape (*i.e.*, the azimuth angles of walls differ by 90 degrees for rectangular rooms). VoLoc jointly estimates the wall parameters. We follow the VoLoc's setup that the UCA is close to one wall. We speak 5 commands to find the best parameters. The distance error is 2.5cm and azimuth error is 12^o. Its performance is sensitive to the selection of the beginning samples and window size for cancellation.

Impact of smart speaker positions: We also vary the positions of the smart speaker in the rooms to evaluate its impact. We plot the median AoA and distance errors in Figure 16 as we vary the distance between the smart speaker and the wall from 5cm to 20cm. We find an interesting trade-off between the distance error and azimuth error. For the shortest distance range ($< 0.5m$), it has a small distance error of 1.5cm and a larger azimuth error 5.1^o. For the longest distance range ($> 2m$), it has an azimuth error of 1.1^o and a distance error of 5.4cm. The worse distance error for the far away wall has little impact on the final localization error, because the reflected signals from this wall always have a much lower SNR and these results are rarely used for retracing.

5.3 Overall localization results

Localization accuracy: Figure 17 shows the CDF of MAVL localization errors in LoS (blue line) and NLoS (or-

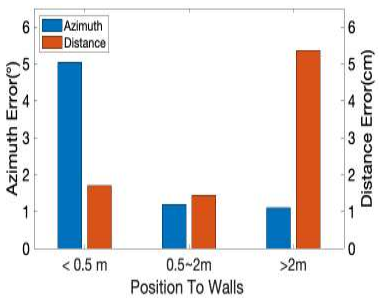


Figure 16: Wall estimation performance over distance.

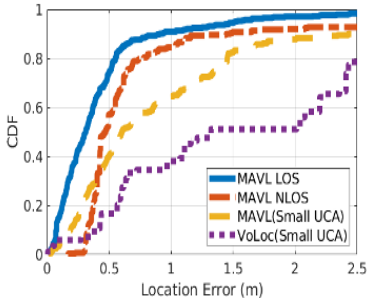


Figure 17: CDF of Localization error for LoS and NLoS, small UCA and Voloc.

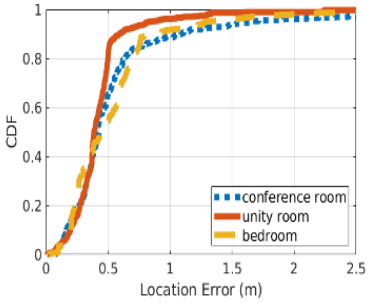


Figure 18: CDF of MAVL Localization error in different rooms.

ange line) scenarios. The median error is 0.31m for LoS and 0.47m for NLoS across all ranges and environments in our evaluation. The accuracy decreases slightly in the NLoS scenario compared to LoS because the diffraction path has lower SNR. The overall localization error for smaller UCA is 0.56m in MAVL . VoLoc [37] reports an overall median error of 0.44m in LoS and a median error of 1.7 m at a large distance (>4m). In our setup, we put the smart speaker *close to one wall*, which is the only setup that VoLoc can work, and find the median error of 1.32 m. This error is larger than the one reported in [37] likely due to different distances and environments.

Performance in different rooms: Figure 18 presents the CDF of localization errors in different rooms. We select three representative environments: a 7.5m x 5.1m conference room with a large desk and many chairs, a 4m x 3.5m bedroom with strong reflectors, such as monitors and wooden furniture, a 3.5m x 2.5m utility room with soft reflectors. We can see that localization error increases with the increasing room size and the number of strong reflectors. A larger room size reduces SNR. For many locations in a large room, the directions of reflected paths are close to each other, which makes it more difficult to separate difference paths. Strong reflection from walls and large furniture may produce merged peaks in the MUSIC profiles. Nevertheless, MAVL still achieves 0.45m median error for the complex bedroom .

Impact of UCA size: As discussed earlier, a smaller UCA size degrades the accuracy of AoAs. The overall localization error for smaller UCA is 0.56m. The yellow line in Figure 17 shows how small UCA works in our system. Although it is worse than that of the larger UCA size, the error can still support many indoor localization applications (*e.g.*, providing useful context information for speech recognition and beamforming to strengthen SNR).

Impact of different positions of UCA: Position of the microphone array have impact on both room contour estimation and source AoA estimation. We place the UCA at three predefined locations, *center*, *close to one wall* and *corner* and evaluate our system. The median localization errors are 0.41m, 0.59m, 0.76m at *center*, *close to one wall*, and *corner*, respectively. Our system works the best when the UCA is placed at the *center*. The accuracy degrades significantly if the UCA is placed at the corner due to increased coherence. VoLoc reports 0.44m overall error and 1.7m error beyond 4m when UCA is placed *close to one wall*. But in our settings with a larger room size and larger distance, VoLoc yields a median error of 1.32 m. VoLoc relies on direct path and reflection path from the close wall in the back. When one retrace using these two paths, a small AoA error may lead to a large localization error. Note that what matters is not the absolute distance to the wall but the ratio between the distance to the wall and the room size. For instance, 0.5m to a wall is considered close for a 5.1$m \times 7.5m$ room and large for a $2m \times 3m$ room. Our system works best in the center position, but also works well for the other setups. Therefore it can support more flexible placement.

Performance to clutter levels: Nearby objects introduce multipath, which makes the AoA estimation more challenging. Figure 20 shows how the clutter level affects the final localization errors across different types of voice. Increasing the clutter level increases the localization errors as we would expect.

(a) Sparse (b) Moderate (c) Dense

Figure 19: Clutter Setups.

Performance to noise level: MAVL is robust to different background noise. Figure 21 shows the influence of various background noise and noise levels. White noise just degrades the accuracy slightly even when SNR is as low as -10dB,

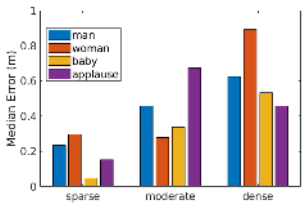


Figure 20: Localization accuracy across clutter levels.

and background music has larger impact than white noise as there are human voices in songs. Our approach is fairly robust against background music unless the SNR is too low (*e.g.*,< -10dB SNR), in which case the error increases to 1.4m.

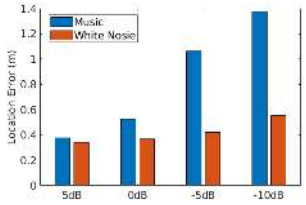


Figure 21: Localization accuracy vs noise levels.

6 Related Work

Acoustic Sensing: A number of systems have been proposed to track a mobile device using acoustic signals [19, 23, 32, 50, 52]. Several recent systems [25, 28, 29, 51] enable device-free tracking using acoustic signals. Many systems generate inaudible acoustic sound for motion tracking. Some use Doppler shift (*e.g.*, AAMouse [50]), time of flight (*e.g.*, BeepBeep [32], or combination (*e.g.*, CAT [23]). Covertband [30] actively sends out OFDM based inaudible signals and builds on top of MUSIC to improve sensing energy. BreathJunior [42] encodes FMCW into white noise to detect motion and breathing of infants. These systems require controlling transmitted acoustic signals and are not suitable for tracking human voice. The most relevant work to ours is VoLoc [37]. Our work advances VoLoc in several important aspects. First, we improve the AoA accuracy from 10 degrees to 1.5 degrees by leveraging multi-resolution analysis in the time-frequency domain. Second, we develop a novel method to automatically estimate the room contour. This significantly eases the deployment effort. Third, we can localize users in both LoS and NLoS whereas they only support LoS.

RF Based Localization: The accuracy of RF based localization approaches are mostly limited by its large wavelength and fast propagation speed for commodity WiFi infrastructure. Chronos [40] can achieve decimeter level localization accuracy by inverting the NDFT. Spotfi [16] incorporates novel filtering and estimation techniques to identify AoA of direct path. Arraytrack [48] designs a novel multipath suppression algorithm to remove reflection between clients and APs. However, they use more than three APs with 16 antennas and require controlling the transmitted signals. Moreover, their approach is focused on eliminating multipath rather than separately estimating each multipath.

Sound Source Localization: There has been a few sound source localization work [26, 34, 46]. [14] builds a real-time system to detect the AoAs of different sound sources. [2] requires a Kinect depth sensor to build a 3D mesh model of an empty room. It estimates multipath AoAs using a cubic microphone array and perform 3D reverse ray-tracing to localize the voice. Its localization error is around 1.12m. [1] considers the diffraction path and applies Uniform Theory of Diffraction for voice localization. Its error is 0.82m. These works either require multiple specialized sensors to get indoor environment or only estimate AoAs instead of localization. They do not address the coherence arising from multipath, so their AoAs are not reliable. MAVL can localize a user using a single smart speaker without extra hardware and explicitly addresses the coherence of multipath.

Audio-Visual Indoor Representation Learning: Recent work combines sound and vision in multimodal learning frameworks to better understand the environment so that they can track audio-visual targets [3, 11, 13], localize pixels relevant to sound in videos [36, 39], and navigate indoor environments [10]. VisualEchoes [12] emits 3ms chirps to combine multipaths and images at different location and learn spatial representation without manual supervision. Soundspaces [7] applies multi-modal deep reinforcement learning on a stream of egocentric audio-visual observations. Our work uses a stand-alone smart speaker, and does not require vision data or pre-training.

7 Conclusion

In this paper, we develop a system, MAVL, to localize users based on their voice using a smartspeaker like device. Our design consists of a novel multi-resolution based AoA estimation algorithm, an easy-to-use acoustic-based room structure estimation approach and a robust retracing to localize the user based on the estimated AoA and room structure. We evaluate MAVL using different sound sources, room sizes, smart speaker setups, noise and clutter levels to demonstrate its effectiveness.

8 ACKNOWLEDGMENTS

This work is supported in part by NSF Grant CNS-1718585 and CNS-2032125. We are grateful to Prof. Shyam Gollakota and anonymous reviewers for their insightful comments and suggestions.

References

[1] Inkyu An, Doheon Lee, Jung-woo Choi, Dinesh Manocha, and Sung-eui Yoon. Diffraction-aware sound localization for a non-line-of-sight source. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4061–4067. IEEE, 2019.

[2] Inkyu An, Myungbae Son, Dinesh Manocha, and Sung-eui Yoon. Reflection-aware sound source localization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 66–73. IEEE, 2018.

[3] Yutong Ban, Xiaofei Li, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Accounting for room acoustics in audio-visual multi-speaker tracking. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6553–6557. IEEE, 2018.

[4] Bela audio expander, 2017. https://github.com/BelaPlatform/Bela/wiki/Using-the-Audio-Expander-Capelet.

[5] Francesco Belfiori, Wim van Rossum, and Peter Hoogeboom. 2d-music technique applied to a coherent fmcw mimo radar. 2012.

[6] Francesco Belfiori, Wim van Rossum, and Peter Hoogeboom. Application of 2d music algorithm to range-azimuth fmcw radar data. In *Radar Conference (EuRAD), 2012 9th European*, pages 242–245. IEEE, 2012.

[7] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc, Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments supplementary materials.

[8] Ross A Clark, Adam L Bryant, Yonghao Pua, Paul McCrory, Kim Bennell, and Michael Hunt. Validity and reliability of the nintendo wii balance board for assessment of standing balance. *Gait & posture*, 31(3):307–310, 2010.

[9] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese. Delay: Robust spatial layout estimation for cluttered indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 616–624, 2016.

[10] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707. IEEE, 2020.

[11] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7053–7062, 2019.

[12] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. *arXiv preprint arXiv:2005.01616*, 2020.

[13] Israel D Gebru, Sileye Ba, Georgios Evangelidis, and Radu Horaud. Tracking the active speaker based on a joint audio-visual observation model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 15–21, 2015.

[14] François Grondin and François Michaud. Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. *Robotics and Autonomous Systems*, 113:63–80, 2019.

[15] Brett Jones, Rajinder Sodhi, Michael Murdock, Ravish Mehra, Hrvoje Benko, Andrew Wilson, Eyal Ofek, Blair MacIntyre, Nikunj Raghuvanshi, and Lior Shapira. Roomalive: magical experiences enabled by scalable, adaptive projector-camera units. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 637–644, 2014.

[16] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. Spotfi: Decimeter level localization using WiFi. In *ACM SIGCOMM Computer Communication Review*, volume 45(4), pages 269–282. ACM, 2015.

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

[18] Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. Roomnet: End-to-end room layout estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4865–4874, 2017.

[19] Qiongzheng Lin, Zhenlin An, and Lei Yang. Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–16, 2019.

[20] David B Lindell, Gordon Wetzstein, and Vladlen Koltun. Acoustic non-line-of-sight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6780–6789, 2019.

[21] Arun Mallya and Svetlana Lazebnik. Learning informative edge maps for indoor scene layout prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 936–944, 2015.

[22] Gleb O Manokhin, Zhargal T Erdyneev, Andrey A Geltser, and Evgeny A Monastyrev. Music-based algorithm for range-azimuth fmcw radar data processing without estimating number of targets. In *Microwave Symposium (MMS), 2015 IEEE 15th Mediterranean*, pages 1–4. IEEE, 2015.

[23] Wenguang Mao, Jian He, and Lili Qiu. CAT: high-precision acoustic motion tracking. In *Proc. of ACM MobiCom*, 2016.

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

[25] Wenguang Mao, Mei Wang, Wei Sun, Lili Qiu, Swadhin Pradhan, and Yi-Chao Chen. Rnn-based room scale hand motion tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–16, 2019.

[26] Kazuhiro Nakadai, Tino Lourens, Hiroshi G Okuno, and Hiroaki Kitano. Active audition for humanoid. In *AAAI/IAAI*, pages 832–839, 2000.

[27] Rajalakshmi Nandakumar, Krishna Kant Chintalapudi, and Venkata N. Padmanabhan. Dhwani : Secure peer-to-peer acoustic nfc. In *Proc. of ACM SIGCOMM*, 2013.

[28] Rajalakshmi Nandakumar, Shyam Gollakota, and Nathaniel Watson. Contactless sleep apnea detection on smartphones. In *Proc. of ACM MobiSys*, 2015.

[29] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. FingerIO: Using active sonar for fine-grained finger tracking. In *Proc. of ACM CHI*, pages 1515–1525, 2016.

[30] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota. Covertband: Activity information leakage using music. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–24, 2017.

[31] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 741–754, 2016.

[32] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. BeepBeep: a high accuracy acoustic ranging system using COTS mobile devices. In *Proc. of ACM SenSys*, 2007.

[33] Swadhin Pradhan, Wei Sun, Ghufran Baig, and Lili Qiu. Combating replay attacks against voice assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–26, 2019.

[34] Caleb Rascon and Ivan Meza. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 96:184–210, 2017.

[35] Ralph Otto Schmidt. A signal subspace approach to multiple emitter location spectral estimation. *Ph. D. Thesis, Stanford University*, 1981.

[36] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

[37] Sheng Shen, Daguan Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury. Voice localization using nearby wall reflections. In *Proc. of ACM MobiCom*, 2020.

[38] Petre Stoica and Arye Nehorai. Music, maximum likelihood, and cramer-rao bound. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(5):720–741, 1989.

[39] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

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

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

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

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

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

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

[46] Xinyu Wu, Haitao Gong, Pei Chen, Zhi Zhong, and Yangsheng Xu. Surveillance robot utilizing video and audio information. *Journal of Intelligent and Robotic Systems*, 55(4-5):403–421, 2009.

[47] Shumian Xin, Sotiris Nousias, Kiriakos N Kutulakos, Aswin C Sankaranarayanan, Srinivasa G Narasimhan, and Ioannis Gkioulekas. A theory of fermat paths for non-line-of-sight shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6800–6809, 2019.

[48] Jie Xiong and Kyle Jamieson. Arraytrack: A fine-grained indoor location system. In *Proc. of NSDI*, pages 71–84, 2013.

[49] Mao Ye, Yu Zhang, Ruigang Yang, and Dinesh Manocha. 3d reconstruction in the presence of glasses by acoustic and stereo fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4885–4893, 2015.

[50] Sangki Yun, Yi chao Chen, and Lili Qiu. Turning a mobile device into a mouse in the air. In *Proc. of ACM MobiSys*, May 2015.

[51] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 15–28. ACM, 2017.

[52] Zengbin Zhang, David Chu, Xiaomeng Chen, and Thomas Moscibroda. Swordfight: Enabling a new class of phone-to-phone action games on commodity phones. In *Proc. of ACM MobiSys*, 2012.

[53] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.