RoSS: Rotation-induced Aliasing for Audio Source Separation

Hyungjoo Seo, Sahil Bhandary Karnoor, and Romit Roy Choudhury

***Abstract*— This paper considers the problem of audio source separation, where the goal is to isolate a target audio signal (say Alice's speech) from a mixture of multiple interfering signals (e.g., when many people are talking). This problem has gained renewed interest mainly due to the significant growth in voice-controlled devices, including robots in homes, offices, and other public facilities. Although a rich body of work exists on the core topic of source separation, we find that rotational motion of the microphones (e.g., a swiveling robot-head) offers complementary gains. We show that rotating the microphone array to the optimal orientation can produce desirable *"delay aliasing"* between two interferers, causing the two interferers to appear as one. In general, *a mixture of* K *signals becomes a mixture of* $(K-1)$ *signals*, a mathematically concrete gain. We show that the gain translates well to practice, provided two rotation-related challenges can be mitigated. This paper is focused on mitigating these challenges and demonstrating the end-to-end performance on a fully functional prototype. We believe that our *Rotational Source Separation* (RoSS) module could be plugged into actual robot heads or into other devices (like Amazon Show) that are also capable of rotation.**

I. INTRODUCTION

As speech recognition and conversational AI matures, voice interactions with robots will become even more popular [1]. Robots in homes, hospitals, restaurants, and airports will interface with humans, with speech serving as the primary medium of interaction [2]–[5]. In such scenarios, separating a user's voice will be essential, especially when these interactions occur in noisy environments. In signal processing, this problem is called "source separation," and has been studied extensively (e.g. ICA, IVA, Adaptive Beamforming) [6]–[13]; today's results are impressive, to the extent that K source signals can be separated using M microphones, even when K is slightly larger than M [14]–[18]. Observe that this $K>M$ problem is particularly challenging not only because the K signals are unknown, but because the K propagation channels – over which the signals arrive to the microphones – are unknown as well. Hence, this problem is specifically called *under-determined blind source separation* (UBSS).

A rich body of work has concentrated on UBSS, and state of the art (SOTA) algorithms range from unsupervised methods (e.g., Nonlinear beamformers, Kernels) and speech specific techniques (e.g., DUET, Bayesian-DUET), to compressed sensing and supervised deep learning approaches [14], [16], [18]–[21]. However, majority of past work must rely on interpolations and regressions since some source information is lost in the (under-determined) mixing process. Therefore, performance degrades, understandably, as K increases for a fixed M. Said differently, any reduction in the $(K-M)$ gap can directly improve the quality of source separation.

All authors are with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA {$hseo17, sahilb5, croy$}@illinois.edu

This paper proposes to utilize robotic rotation to spatially alias interferers, thereby reducing the $(K-M)$ gap. The core idea is simple. Observe that signals arriving from different angles θ_i produce relative delays δ_i at the microphone array. Rotation of the array causes these relative delays to change non-linearly, offering the opportunity to "move" the sources in this relative-delay space. When the microphone rotates to bisect two sources — such as in Fig. 1 where the line joining the microphones bisects the sources A and B — the relative delays of the bisected sources become identical. Hence, in the relative-delay space, K sources manifest as $(K-1)$ sources. This implies that the scenario in Fig. 1, which was originally an under-determined [K=3, M=2] system, has now become determined with [K=2, M=2]. Even when $K > M+1$, the reduction from K to $(K-1)$ offers concrete improvements, both in source separation and localization.

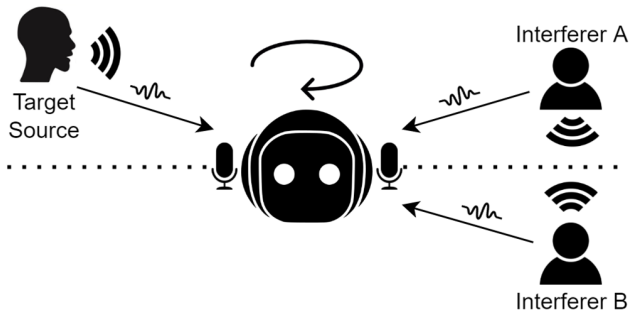

Fig. 1. Rotation of the microphone array to the correct orientation (that bisects the sources A and B) produces a desired "aliasing" in relative delay.

Realizing the above idea presents 2 challenges:

1) Since the angle of arrivals (AoA) of the K signals are not known, the correct microphone orientation θ^* is unknown as well. Estimating all K AoAs is difficult with $M(<K)$ microphones [22]–[26], and worse, AoA estimates are plagued by front-back ambiguities (i.e., it is difficult to tell whether a signal is arriving from a direction θ in front, or $-\theta$ from the back).
2) Even if the K AoAs are estimated, it is not clear which interferers should be bisected to maximize performance. There are $\binom{K-1}{2}$ candidate pairs to bisect, and not all of them help equally in separating the given target signal.

This paper addresses these two problems in Section III through a mobility-guided algorithm that first estimates the source AoAs and, based on the AoAs, decides on the optimal microphone orientation. Once rotated to this orientation, the

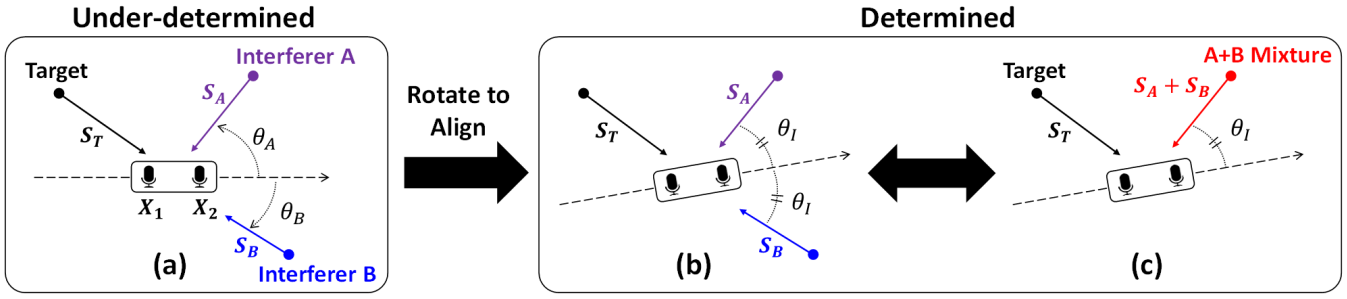

Fig. 2. (a) 2-microphone array faced with 3 sources resulting in a UBSS problem. (b) Rotation causes interferers to arrive over the same absolute AoA angle (θ_I and $-\theta_I$). (c) The steering vector for interferers get aliased (or aligned), resulting in a determined system.

recorded signal is fed to a source separation (SS) algorithm. Our proposed RoSS module is complementary, hence compatible, with most SS algorithms.

We implement RoSS on a rotating microphone prototype, and perform experiments in simulated and uncontrolled (indoor/outdoor) environments (Section IV). Results† show that RoSS achieve around 10-to-15dB of scale-invariant signal distortion ratio (SI-SDR) [27], consistently outperforming existing UBSS/BSS methods by upto 6dB in various scenarios. We believe RoSS could also be effective with smartphones, earbuds, moving video-conference systems, and surveillance cameras, all of which have limited number of microphones but contain actuators or inertial measurement units (IMUs) for angular rotation and sensing.

II. Formulation and Opportunity

A. Signal Model

Let $S_T(t), S_A(t), S_B(t)$ be 3 source signals, of which S_T is the target and others are interference (Fig. 2(a)). A linear 2-microphone array receives the mixture of these signals as $X_1(t)$ and $X_2(t)$ and we designate $X_1(t)$ as the reference for relative delay calculations. The signals travel from the far-field over AoAs θ_k (k=T,A,B). We explain our proposed method with $K = 3$ signals and consider $K > 3$ later.

We make the following Assumptions:

(A1) The sound sources are human speech, widely assumed to be mutually independent, non-Gaussian signals.

(A2) Once a speech has been separated, it is possible to tell if it is from the target user (i.e., a voice fingerprint is available).

(A3) Sources are not moving in the time scale of seconds.

Thus, the received (convolutive) signal mixture is:

$$X_1(t) = \sum_{k\in\{T,A,B\}} S_k(t), X_2(t) = \sum_{k\in\{T,A,B\}} S_k(t+\tau_k) \quad (1)$$

where $\tau_k = \frac{d}{v_p} cos(\theta_k)$, (k=T,A,B), are time-difference-of-arrivals (TDOA) between the microphones (also called relative delay), while v_p and d denote velocity-of-sound and distance-between-microphones, respectively.

Thus, in the time-frequency domain(time index omitted):

$$\vec{X}(f) = \vec{a}_T(f)S_T(f) + \vec{a}_A(f)S_A(f) + \vec{a}_B(f)S_B(f) \quad (2)$$

†More results and demos : https://uiuc-ss.github.io/RoSS

which in the matrix form can be written as:

$$\begin{bmatrix} X_1(f) \\ X_2(f) \end{bmatrix} = \begin{bmatrix} | & | & | \\ \vec{a}_T & \vec{a}_A & \vec{a}_B \\ | & | & | \end{bmatrix} \begin{bmatrix} S_T(f) \\ S_A(f) \\ S_B(f) \end{bmatrix} \quad (3)$$

Here $\vec{a}_k = [1\ \exp(j2\pi f\tau_k)]^T$ (k=T,A,B) is the steering vector. Note that even if all $\vec{a}_k$'s are known, the system is still under-determined.

B. Interference Alignment

What if we rotate the array such that the line joining the microphones bisect the two interferers? While the correct rotation angle needs to be inferred blindly, for now let us assume we know it. Fig.2(b) shows the outcome. *Since the new AoAs of the two interferers are now θ_I and $-\theta_I$, their corresponding TDOAs become equal, or aliased*, as follows:

$$\tau'_A = \frac{d}{v_p} cos(\theta_I) = \frac{d}{v_p} cos(-\theta_I) = \tau'_B$$

Thus, in frequency domain, interferers A and B have identical array vectors $\vec{a}_I(f) = [1\ \exp(j2\pi f\tau_I)]^T$ where $\tau_I = \tau'_A = \tau'_B$. Hence, the new measurement vector $\vec{X}'(f)$ is:

$$\begin{bmatrix} X_1{}'(f) \\ X_2{}'(f) \end{bmatrix} = \begin{bmatrix} | & | \\ \vec{a}_T & \vec{a}_I \\ | & | \end{bmatrix} \begin{bmatrix} S_T(f) \\ S_A(f) + S_B(f) \end{bmatrix} \quad (4)$$

This expression means that the array *would sense two groups of signals, not three*; one is the target and the other is the sum of two interferers. Fig. 2(c) shows these two signals arriving from distinct angles. This produces a determined system of equations except that one of the mixed signals arriving from AoA $\vec{a}_I$ is actually a sum of independent sources. If this sum $(S_A(f) + S_B(f))$ remains independent of the target signal $S_T(f)$ (as shown next), we can apply classical source separation.

C. Sum of Mutually Independence Sources

We briefly show that a mixture of two independent sources remains independent from the third source when all three are mutually independent. Define A, B and T as mutually independent continuous random variables, and $J = A + B$ is a fourth random variable. Let $F_i(\cdot)$ and $f_i(\cdot)$ be cumulative distribution function (CDF) and probability distribution

function (PDF) of variable i, respectively. Then, the joint distribution of T and J can be written as:

$$\begin{aligned} F_{JT}(j,t) &= P(A+B \le j, T \le t) \\ &= \int P(A+B \le j, T \le t | A=a) f_A(a)\, da \\ &= \int P(B \le j-a, T \le t) f_A(a)\, da \\ &= \int P(B \le j-a) f_A(a)\, da \cdot P(T \le t) \\ &= P(A+B \le j) \cdot P(T \le t) = F_J(j) F_T(t) \end{aligned} \tag{5}$$

Therefore, J and T are also mutually independent [28].

III. RoSS: AoA Estimation and Optimal Bisection

Our end-goal now is to rotate the microphone so that the correct interferer-pair gets aligned. For this, we first need to estimate all the AoAs, and using the AoAs, determine the optimal interferer-pair that must be bisected.

A. Estimating AoAs in Under-determined Scenarios

Estimating K AoAs with $M < K$ microphones is known to be a hard problem for general signals. However, literature has shown promise with speech signals, due to what is known as the *W-Disjoint Orthogonality* (WDO) property [14]. Briefly, extensive experiments have shown that speech from two humans have a low probability of collision in a given time-frequency (TF) bin. Thus, if one calculates the TDOA for each TF bin — called *inter-microphone time difference* (ITD) — one can extract information about AoAs. Fig. 3 illustrates this with a toy example of red and blue signals; the calculated ITDs from the red and blue TF bins form 2 clusters. The means of these clusters partly reveals the red/blue signal's AoA.

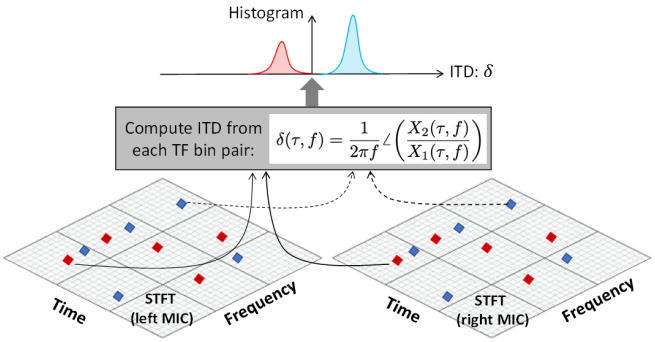

Fig. 3. ITD computed from TF bins produce 2 clusters around two mean ITDs. These mean ITDs are estimates of AoA.

Unfortunately, the mapping between ITD and AoA is not 1:1 because AoAs of both θ and $-\theta$ produce identical ITDs at the microphone array. Said differently, ITD is calculated as $\delta_k = \frac{d}{v_p} cos(\theta_k)$ and both θ_k or $-\theta_k$ produce the same ITD. Fig. 4 shows how 2 ITD clusters map to 4 candidate AoAs (of which 2 AoAs are spurious). This is classically known as the *front back ambiguity*. Worse, if the true AoA's happen to be $\theta_1 = -\theta_2$, then it becomes difficult to even recognize the presence of 2 signals. Rotating the microphone array to the correct orientation θ^*_{final} would obviously require to resolve this ambiguity problem first.

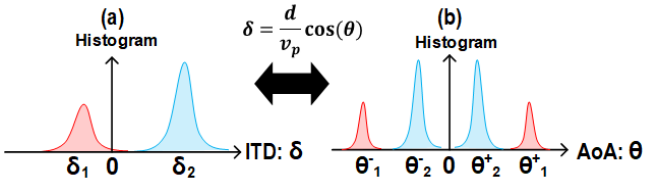

Fig. 4. 2 ITD clusters gets mapped to 4 clusters in $(-\pi, \pi]$ AoA space.

B. Rotation-enabled AoA Disambiguation

We propose to disambiguate AoA using rotation of the microphone array. The idea is simple — as the array rotates, the ITD will change and the *direction* of this change (higher or lower) should reveal the true AoA. Fig. 5(a) illustrates this with a single-source example, where candidate AoAs are θ_k or $-\theta_k$. Fig. 5(b) plots this ITD on a graph with the X-axis showing the rotation angle of the array. Since the microphone has not made any rotation yet, the ITD is plotted for $\theta_{rot} = 0$. As the array rotates counter-clockwise, the ITD should change in one of two ways: if the true AoA $= \theta_k$, then the ITD should increase, while for AoA$= -\theta_k$, the ITD should decrease (Figure 5(d)). Moreover, the trajectory of change should follow the `Cosine` curve since the ITD is a function of $Cos(\theta)$. Thus, in theory, even one small rotation should disambiguate and give us the true AoA.

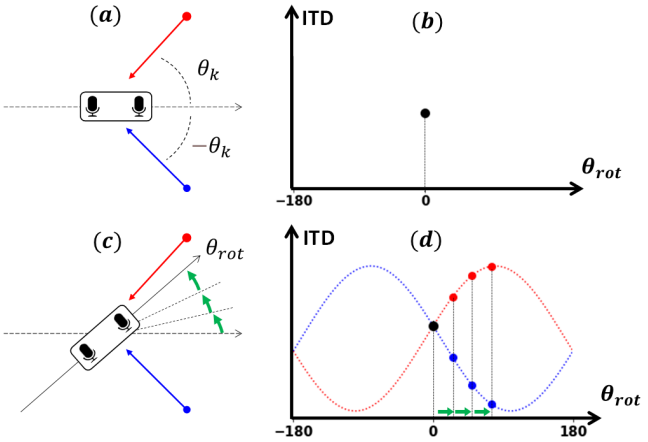

Fig. 5. (a) Ambiguous AoAs in a static scenario, (b) Measured ITD without rotation, (c) Counter-clockwise array rotation, (d) ITD trajectory is Cosine shaped, and the direction reveals the true AoA

Rotation-based disambiguation should be generalizable to K sources. Instead of one ITD value, we now have K ITD values at $\theta_{rot} = 0$. With rotation of the array, each ITD value would move in one of two trajectories – upward `Cosine` or downward `Cosine`, as shown in Figure 6 for $K = 3$. One should be able to fit K distinct `Cosine` functions through all the ITD trajectories, thereby extracting the $K = 3$ true AoAs from 6 candidates.

In practice, disambiguation is far more challenging because the ITD values become noisy. Several reasons contribute:

(1) Background interference arrives from different angles polluting the ITD clusters shown in Figure 4(a). Reverberations add to this pollution.

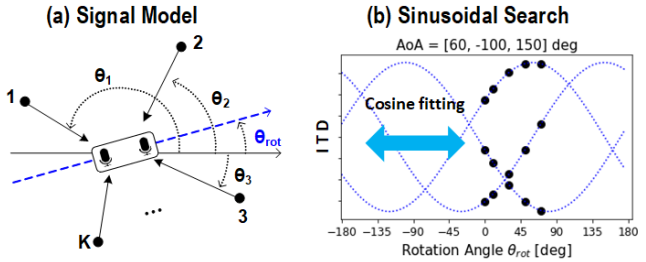

Fig. 6. (a) Two-dimensional rotation under K sources seen from above. (b) $K = 3$ case with 3 fitted Cosine trajectories in the ITD measurements.

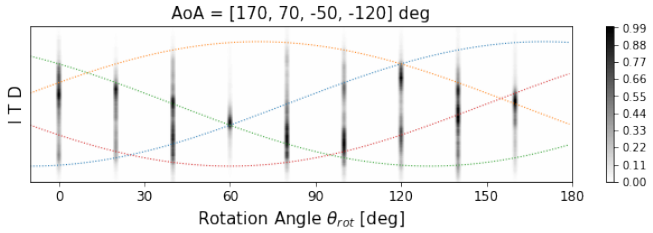

Fig. 7. ITD measurements with $K = 4$ sources for consecutive microphone rotation, performed in a real indoor environment with background noise and reverberation.

(2) With increasing number of sources, $K \geq 3$, the WDO property begins to break down, meaning that sources begin to collide with higher probability in time-frequency bins. Collisions produce incorrect ITD values, shifting the peaks in Figure 4(a).

(3) There is no guarantee that all K ITD peak values would be prominent at every step of rotation; a source pair may have similar (or identical) ITD values, say when their AoAs are 30° and -30°. This smudges the ITD estimates at that rotational step.

(4) Finally, the ITD does not vary linearly with every step of array rotation. The ITD variation is large when θ is near 90° and small when θ is 0° (note that the $\frac{dCos(\theta)}{d\theta}$ is zero when $\theta = 0$). This implies that ITD noise must be treated differently for different regimes of θ_k.

Figure 7 shows measurements from a real indoor scenario where the microphone array is rotated 8 steps, with 20° per step. The smudged ITDs are from $K = 4$ different sources, implying that we have 8 candidate AoAs to disambiguate. Said differently, 4 `Cosine` functions need to be fitted to the measured data, essentially making it a problem in regression.

C. Statistical Approach

Our proposed solution can be intuitively summarized as follows. We compute a likelihood for all AoAs based on the initial ITD measurements. Then, for every rotation of θ_{rot}, we model the expected ITD for each AoA and match it against the new measurement – this gives us an updated likelihood per AoA. With more rotational steps, the likelihood of the true AoAs begin to show sharper peaks, while the ambiguous and the incorrect AoAs die down. We normalize the per-AoA-likelihood and call it the "AoA spectrum" – Figure 8 plots real AoA spectrums after each rotation of the array. The peaks in the AoA spectrum sharpen gradually and

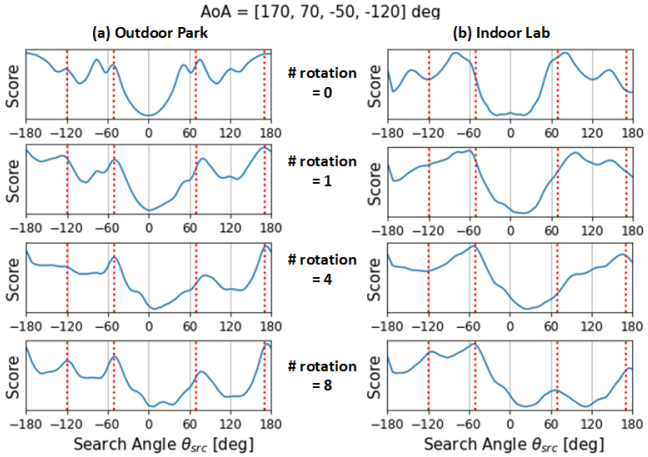

Fig. 8. As microphone takes more rotational measurements, peaks near truth AoAs (dotted red) get clearer, offering better AoA detection.

after several rotations, converge to the $K = 4$ correct peaks. Mathematically, our algorithm can be specified in 3 essential steps as follows:

Step 1: At each rotation angle $\theta_{rot}^{(r)}$, $(r = 0, 1, ..., R)$, use the ITD histogram to estimate probability density function (PDF) as: $\hat{p}(\delta^{(r)})$. Normalize the PDF to not penalize the ITDs that are absent.

Step 2: Calculate likelihood for each AoA, θ_{src}, at the r-th rotation as: $L^{(r)}(\theta_{src}) = \hat{p}\big(\frac{d}{v_p} cos(\theta_{src} - \theta_{rot}^{(r)})\big)$.

Step 3: Compute overall likelihood across R rotations $\prod_{r=0}^{R} L^{(r)}(\theta_{src})$ with normalization. Identify θ_{src} values that do not change more than ϵ for 3 consecutive rotations; announce these as the K source AoAs.

D. Optimal Bisection Angle for Source Separation

Once AoAs are estimated, RoSS needs to rotate the microphone array to bisect two interferers. Given $K-1$ interferers, there are $\binom{K-1}{2}$ candidate pairs. *Which pair should RoSS bisect?*

To answer this question, we need to establish two insights: (1) A target signal can be perfectly isolated when its ITD distribution (as shown in Figure 3) does not overlap with any of the interferer's ITD distributions.

(2) Rotation of the microphone array produces unequal shifts in the ITD distributions. This is because the ITD is proportional to $Cos(AoA)$, hence for a given rotation, AoAs near 0 or 180° experience smaller ITD shifts, compared to AoAs near 90 or 270°.

Given these 2 facts, the optimal rotation becomes the following optimization question: *what final orientation angle θ_{final}^{*} maximizes the minimum ITD separation between the target and the interferers?* The formal optimization is as follows:

$$\theta_{final}^{*} = \underset{\theta_{rot}}{\text{argmax}} \min_{k, k \neq T} |\delta_T - \delta_k| \tag{6}$$

Here δ_T is the mean ITD for the target signal (T) and δ_k is the mean ITD of each interferer. Barring some rare cases, θ_{final}^{*} is indeed an angle that bisects a pair of interferers

(we omit the proof in the interest of space). Hence, the above optimization needs to search only across the $\binom{K-1}{2}$ bisection angles, as opposed to all possible θ_{rot}.

Isolating Any Given Target: In conclusion, given a mixture of K sources, and a target signal T for isolation, RoSS rotates to the θ^*_{final} orientation. The target signal T can be specified either by its AoA (e.g., a robot sees a person in its camera view and isolates that person's voice), or the target signal's voice fingerprint may be given to the robot, in which case it checks which voice signal matches the fingerprint. Once the fingerprint matches, RoSS continues to track that AoA and isolate that voice signal.

Delay: Note that if sources come and go, the problem is easier because K is smaller at any given time. However, if K sources are continuously present, RoSS has the time to rotate and resolve them. Once AoAs are known once, rotation to θ^*_{final} is fast, hence, any given source can be separated so long as they are not moving fast.

IV. Evaluation

A. Experimental Settings

Measurements: RoSS is implemented on a custom-built rotary platform actuated by a NEMA-17 stepper motor (Fig. 9(a)). The open-loop motor uses a TB-6600 driver with peak rotation speed and acceleration of 225 deg/s and 112.5 deg/s^2. A ReSpeaker microphone array [29] connected to a Raspberry Pi is mounted on the rotary platform and 2 adjacent microphones, with $5cm$ spacing, are used to record audio signals. Rotations are performed in 20° increments. The Table I lists some of the environmental parameters. In each environment, K speech signals were played from loud speakers placed radially around the microphone, at distances between 2 to $2.5m$. Each signal is 1-minute-long male/female voice recordings randomly selected out of 11 independent speakers, drawn from the LibriTTS dataset [30] where signal powers are almost identical, i.e., $SIR \approx -10log(K-1)$ for K-sources. Multiple runs were performed per configuration, with various mixtures of voices (males, females, and mixed genders), $K \in [3, 4]$, and K AoA angles chosen uniform randomly between $[0, 360]$. Fig. 9(b,c,d) show example images from our experiment sites.

TABLE I
Evaluation Environments

Settings	Location	SNR [dB]	Room size [$m \times m$]
Lab	Indoor Lab	22	$\approx 8.4 \times 8.2$
Room	Indoor Room	23	$\approx 6 \times 8$
Park	Outdoor Park	15.4	$> 20 \times 20$
Sim	Simulation	15	10×10

The audio recordings are sampled at 16kHz, with STFT frame lengths of 512 or 1024 with 25% overlap with adjacent frames. For comparison, we use three popular source separation algorithms, namely natural gradient-based IVA [7], DUET [14] and MVDR [8].

Simulation: To test RoSS over a wider range of parameters, we simulate the microphone recordings using a room impulse response (RIR) generator [31]. The convolutive mixtures

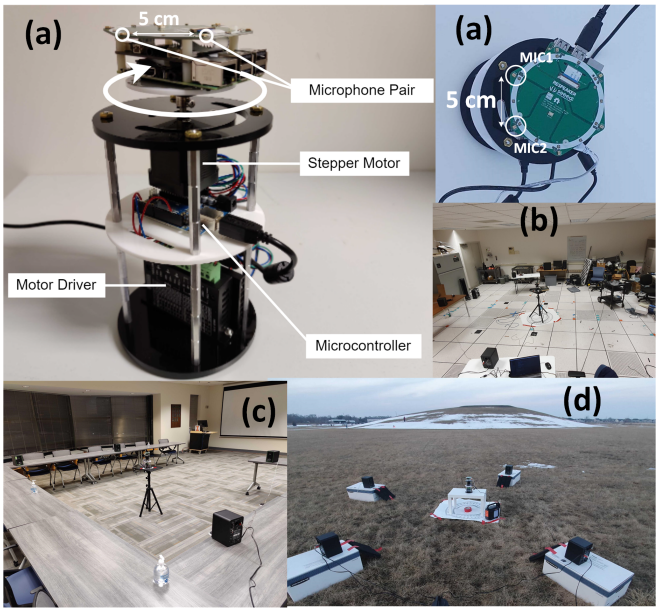

Fig. 9. (a) Custom-built rotary platform with ReSpeaker microphone array. (b) Laboratory. (c) Conference room. (d) Local park.

from K sources are denoted $X_1(t), X_2(t)$. The key parameters of the simulations are:

- Room size: 10m x 10m (2-dimensional space assumed) with reverberation time T_{60} of 0, 450, 700 ms.
- Two omni-directional microphones with $5cm$ spacing are located in the room-center, rotating around their center.
- Gaussian noise is added so that microphone SNR is 15 dB while maintaining SIR of -10log(K-1) dB
- Separated sources are evaluated by comparing with each source alone measured at the reference microphone $X_1(t)$.
- Algorithm settings are similar to the measurement settings except for 24kHz sampling frequency.

B. Performance Metric

AoA Error: Once the AoA estimate $\hat{\theta}$ is available, the AoA error is the smaller angular difference between the ground truth AoA θ^* and the $\hat{\theta}$. However, recall that AoA ambiguity exists, meaning $2K$ AoA candidates appear for K true AoAs. In such settings, we calculate the AoA error as follows. We create K buckets, one for each true AoA. A candidate AoA is assigned to bucket j if that candidate is angularly closest to the j^{th} true AoA. The average AoA error per bucket is then computed – this gives us K AoA errors. If a bucket has no AoA, we assign a maximum possible error as a penalty.

Source Separation: Once a source has been separated as $\hat{s}$ from a mixture m, we report SI-SDR and SI-SDR *improvement* [27], [32] defined as:

$$\text{SI-SDRi} = \text{SI-SDR}(\hat{s},\ s) - \text{SI-SDR}(m,\ s)$$

Here s is the source signal recorded at the microphone without any interference; this serves as ground truth.

C. Results

Comparison between RoSS and Existing Algorithms: Fig. 11 compares RoSS's source separation performance with SOTA algorithms, IVA and DUET. The X-axis shows the

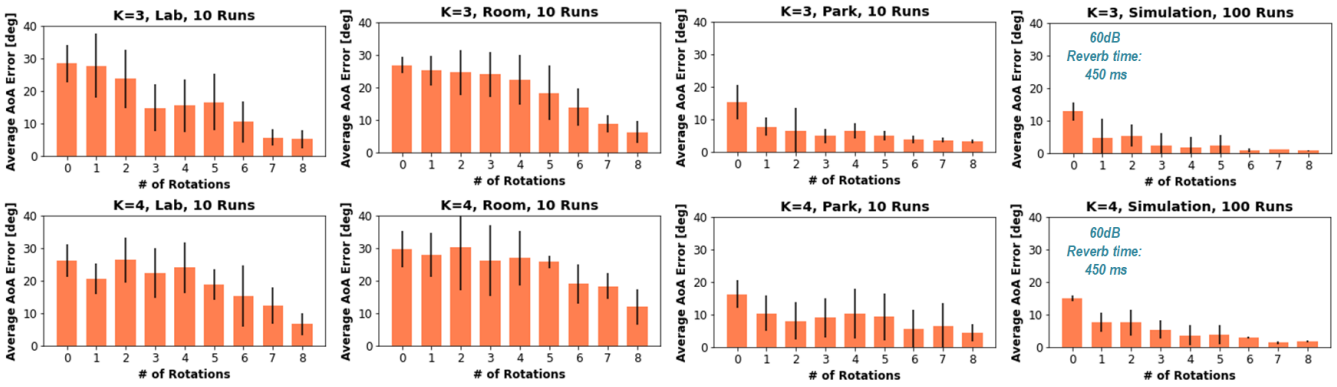

Fig. 10. Average AoA estimation error over consecutive rotational steps in various locations and configurations. Error bars show standard deviations.

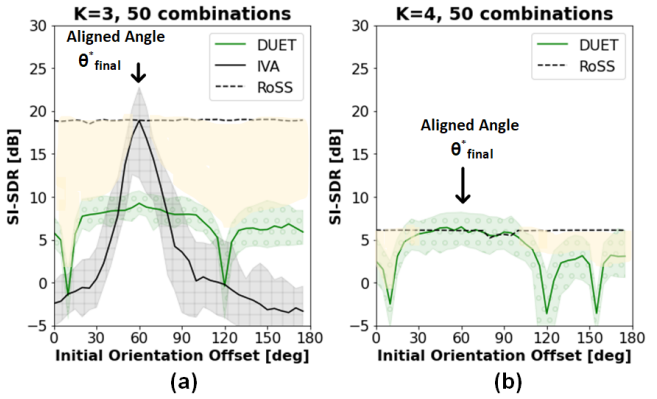

Fig. 11. Source separation performance with different initial orientation of the microphone array, showing non-uniform patterns.

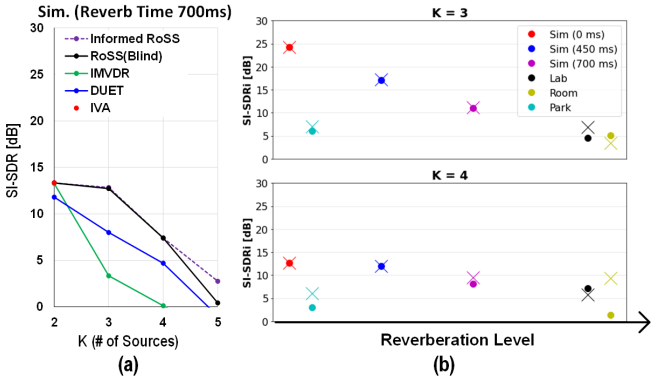

Fig. 12. Average SI-SDR/SI-SDRi of (a) various algorithms, in (b) different setups where X markers are AoA-informed RoSS.

initial orientation of the microphone array – understandably, IVA and DUET's performance vary as a function of this initial orientation. The solid lines show their median performance over 50 different configurations, while the light-color bands are $[80, 20]$ percentiles. Since RoSS rotates to the optimal orientation, its performance remains consistent (and matches IVA when the initial orientation is luckily the optimal). With $K = 3$ sources, IVA outperforms DUET when the array orientation is favorable to it, but for other orientations (and when the sources increase to $K = 4$), DUET gains due to the inherent WDO property of speech. The yellow shaded area depicts the overall gain from RoSS, which is essentially the value of microphone rotation. Since RoSS is complementary to IVA, DUET, and other algorithms, this gain should be always available.

AoA Estimation: Fig. 10 plots the reduction of AoA error against rotation, where each rotation-step is ≈ 1.6 seconds. Each graph shows the average AoA error across all experiments in a given setting (Lab, Room, Park, Sim); the error bars denote standard deviations. As RoSS rotates the microphones, the AoA error reliably converges to the true AoA angle. Simulation and outdoor settings converge faster and more accurately, mainly due to lower reverberation, compared to indoor labs and rooms. Importantly, AoA estimation is a by-product of RoSS and can be leveraged as an independent capability in other applications, such as localization, imaging, and radar-based perception.

Parameterized Simulations: Fig. 12(a) shows how SI-SDR degrades with increasing K but RoSS continues to outperform others. `Informed RoSS` is a variant of RoSS where the AoAs are accurately known, such as in audio-visual systems [33], [34] — the gains are slight, implying RoSS's AoA estimation is reliable. Fig. 12(b) plots SI-SDR against varying reverberation levels – the indoor setting exhibiting the highest reverberation. Performance understandably degrades with reverberation and larger K since AoA errors and TF-collisions are both high. Performance sometimes degrades outdoors from strong winds and background noise, however such degradation affect all source separation methods.

V. Conclusion and Future Work

We show that microphone rotation ushers an opportunity in audio AoA estimation and source separation, especially in under-determined settings. We demonstrate that optimal rotation can align/alias two interferers in the delay space, making them appear as one. This alignment is complementary to existing algorithms, offering promising results in simulations and real reverberant environments.

Further improvements are possible in at least 2 directions: (1) an adaptive rotation policy that converges faster, ideally within a few spoken words, and (2) updating the algorithm to circular microphone arrays. We leave these to future work.

REFERENCES

[1] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5610–5614.

[2] T. Mizumoto, K. Nakadai, T. Yoshida, R. Takeda, T. Otsuka, T. Takahashi, and H. G. Okuno, "Design and implementation of selectable sound separation on the texai telepresence system using hark," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 2130–2137.

[3] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoto, "Robust speech interface based on audio and video information fusion for humanoid hrp-2," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 3, 2004, pp. 2404–2410 vol.3.

[4] A. Deleforge, A. Schmidt, and W. Kellermann, "Chapter 2 - audio-motor integration for robot audition," in *Multimodal Behavior Analysis in the Wild*, ser. Computer Vision and Pattern Recognition, X. Alameda-Pineda, E. Ricci, and N. Sebe, Eds. Academic Press, 2019, pp. 27–51. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128146019000122

[5] A. Magassouba, N. Bertin, and F. Chaumette, "Sound-based control with two microphones," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 5568–5573.

[6] A. Hyvärinen, J. Karhunen, and E. Oja, "Independent component analysis," vol. VII. Wiley, 2001.

[7] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[8] O. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[9] K. Nakadai, H. Nakajima, G. Ince, and Y. Hasegawa, "Sound source separation and automatic speech recognition for moving sources," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 976–981.

[10] M. Heckmann, F. Joublin, and E. Korner, "Sound source separation for a robot based on pitch," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 2197–2202.

[11] K. Noda, N. Hashimoto, K. Nakadai, and T. Ogata, "Sound source separation for robot audition using deep learning," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, 2015, pp. 389–394.

[12] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP*, 2018.

[13] M. Maazaoui, K. Abed-Meraim, and Y. Grenier, "Blind source separation for robot audition using fixed hrtf beamforming," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 58, Mar 2012.

[14] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[15] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International Journal of Neural Systems*, vol. 25, no. 01, p. 1440003, 2015, pMID: 25164245.

[16] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *International Conference on Machine Learning*, 2020.

[17] J. Yang, Y. Guo, Z. Yang, and S. Xie, "Under-determined convolutive blind source separation combining density-based clustering and sparse reconstruction in time-frequency domain," *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp. 3015–3027, 2019.

[18] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," in *IEEE/ACM transactions on audio, speech, and language processing*, 2019.

[19] S. Miyabe, B.-H. Juang, H. Saruwatari, and K. Shikano, "Kernel-based nonlinear independent component analysis for underdetermined blind source separation," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1641–1644.

[20] G. F. Edelmann and C. F. Gaumond, "Beamforming using compressed sensing," *The Journal of Acoustic Society of America*, 2011.

[21] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ilrma originating from ica and nmf," *APSIPA Transactions on Signal and Information Processing*, vol. 8, p. e12, 2019.

[22] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, 2017.

[23] E. Vincent, A. Sini, and F. Charpillet, "Audio source localization by optimal control of a mobile robot," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5630–5634.

[24] K. Nakamura, K. Nakadai, F. Asano, and G. Ince, "Intelligent sound source localization and its application to multimodal human tracking," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 143–148.

[25] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, vol. 1, 2004, pp. 1033–1038 Vol.1.

[26] X. Li, H. Liu, and X. Yang, "Sound source localization for mobile robot based on time difference feature and space grid matching," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 2879–2886.

[27] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr-half-baked or well done?" in *ICASSP*, 2019.

[28] H. Pishro-Nik, *Introduction to Probability, Statistics, and Random Processes*. Kappa Research, LLC, 2014.

[29] "Respeaker 6 mic circular array kit for raspberry pi." [Online]. Available: https://wiki.seeedstudio.com/

[30] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[31] E. Habets, "ehabets/rir-generator: Rir generator," Oct. 2020. [Online]. Available: https://github.com/ehabets/RIR-Generator

[32] M. Maciejewski, J. Shi, S. Watanabe, and S. Khudanpur, "Training noisy single-channel speech separation with noisy oracle sources: A large gap and a small step," in *ICASSP*, 2021.

[33] S. Majumder, Z. Al-Halah, and K. Grauman, "Move2hear: Active audio-visual source separation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 275–285.

[34] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 9701–9707.