A REAL-TIME SYSTEM FOR ASYNCHRONOUS USER LOCALIZATION WITH VOICEBAND PROCESSING

Spencer Russell, Md Tamzeed Islam, Renato Nakagawa, Ravi Pulugurtha,
Paris Argyropoulos, *Wontak Kim*, *Mohamed Mansour*, *Mike Goodwin*, *Romit Roy Choudhury*

Amazon Inc., USA

ABSTRACT

We present a real-time system for computing user location and orientation under constraints of memory, computation, and latency. The system utilizes a set of distributed microphone arrays with independent clocks. It computes angle-of-arrival (AoA) and time-difference-of-arrival (TDoA) information asynchronously, and deploys late fusion to compute the maximum likelihood user position. The user orientation is then computed using a model for the radiation pattern of the human voice. We also describe the Cramer-Rao Bound of the estimation, and a deep-neural-network (DNN) based model for computing the estimation confidence. The effectiveness of the proposed system is established using large datasets of real-life recordings under different room environments.

***Index Terms*—** Acoustic Localization, Time-Difference-of-Arrival, Angle-of-Arrival, User Orientation, Telepresence.

1. INTRODUCTION

External loudspeakers are frequently used to render spatial audio, whether in telepresence applications, gaming, or home theater. These systems generally assume a particular listening position (often the center of the speaker array) and the spatial image typically degrades as the user moves farther from the assumed position. Additionally, the loudspeaker array is assumed to have a well-defined front that is known *a priori* and corresponds to the user facing direction. This paper proposes a system to estimate the user position and orientation (facing direction) based on a single short utterance. This is used by the rendering system to re-calibrate to fit the user pose.

In this system, each loudspeaker is co-located with a microphone array, as found in a smart speaker such as the Amazon Echo or Google Nest. We use this array-of-arrays to estimate the user location and orientation. This process is performed on a heterogeneous set of devices, and does not assume that the devices share a common synchronized clock. We can consider the microphones of all devices to form a single distributed array, comprised of sub-arrays with the microphones from each device. The proposed system leverages far-field angle-of-arrival (AoA) estimation from each sub-array, as well as near-field time-difference-of-arrival (TDoA) position estimation using a subset of microphones across devices. An example of such configuration is shown in Fig. 1, where the map of interest is located within a 3×5 m room (as shown by the x/y coordinates). The map consists of: four devices (D0-3), and a user location U0. Facing directions are shown by the solid-black arrows. The proposed system takes as input the positions and orientations of the devices and estimates the position and orientation of the user. For evaluation, errors are compared against ground-truth locations from real-world data collections.

In the proposed system, the user location and orientation are estimated separately as described in sections 3 and 4, respectively.

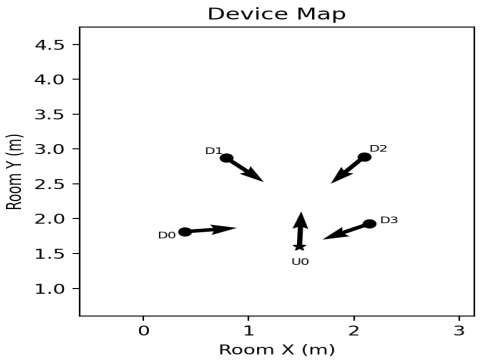


Fig. 1: Device map D0-3 and user location U0 to be identified

Far-field localization methods such as in [1, 2], utilize a small microphone array with omnidirectional elements, and the task is typically to estimate the AoA of the target relative to the array. Conversely, near-field localization, typically uses a larger array, with the source near or within the aperture of the array. This geometry allows the full 2D or 3D position of the target to be estimated using TDoA [3, 4]. Hybrid approaches create a near-field array whose elements are themselves far-field arrays. For example, the sub-arrays can utilize beamforming to improve the signal-to-noise ratio (SNR) prior to time-based processing [5]. Alternatively, the far-field AoA estimates can be combined to triangulate the source position [6]. Some systems (such as the proposed) incorporate AoA from the far-field arrays as well as near-field TDoA-based location [7].

In order to utilize time delays between array elements, they need to be synchronized, and errors in synchronization have a direct effect on the accuracy of the resulting system. This is particularly challenging when the array is a distributed system of separate devices. Prior work generally assumes that this synchronization is performed separately. The proposed system takes advantage of the fact that each device consists of co-located speakers and microphones, and synchronizes the device captures using only the captured audio itself, without assuming a common time base in advance.

The final component of the proposed system is to estimate the user orientation. Estimating orientation, or facing direction, of a human speaker is a less well-studied problem than source localization [8] and typical applications are limited to device arbitration [9, 10, 11](determining which device should respond to a query when multiple are near the user). In the proposed system, orientation is estimated by comparing the received signal power at each sub-array with the power that is expected given a particular emission model. For instance, we expect to receive less power behind the user than in front of them. This emission pattern is computed from acoustic simulation. The user orientation procedure uses the user position estimate as a proxy of the true position.

2. CAPTURE SYNCHRONIZATION

In a distributed system, it is common for each device to operate on its own clock domain, independent from other devices. Acoustic source localization requires tight synchronization which is a non-trivial problem, due to both network and buffering latencies. Such latencies are in general unknown and time-varying.

The process is illustrated in Fig. 2, where it starts with a command from the coordinator to each device, at which point each device starts its capture. However, captures may not start at exactly the same time. This effect is shown in Fig. 3, where 3a shows how the playback sequence actually happened, and 3b shows how it appears in the device captures, due to offsets in their capture start time. In order to compare timestamps across devices, synchronization of the captures (i.e. estimating the offsets) is required.

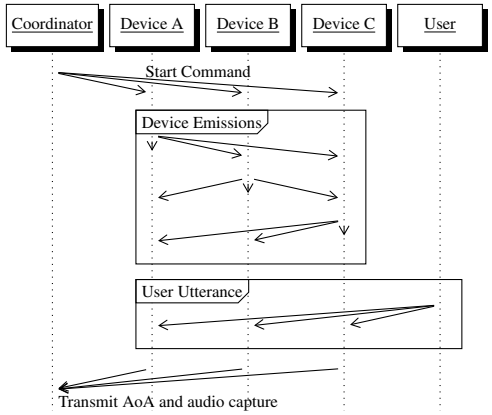


Fig. 2: Sequence diagram of the synchronization and localization

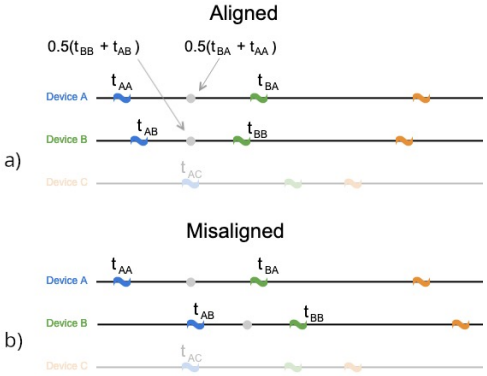


Fig. 3: Playback sequence for time synchronization of four devices as: (a) it occurred relative to a global-clock, and (b) recorded by each local device.

The key assumption that allows these offsets to be estimated is that the acoustic propagation delay from device i to j ($\delta_{ij} \triangleq t_{ij} - t_{ii}$) is the same as that from device j to i (δ_{ji}). Such assumption holds well when each device's speaker-to-microphone distances are small. In Fig. 3a we see that device A emits a sound at time t_{AA} which is received by device B at $t_{AB} = t_{AA} + \delta_{AB}$. Some time later (at t_{BB}), device B emits a sound that is received at $t_{BA} = t_{BB} + \delta_{BA}$. When the clocks are aligned we see that $\frac{1}{2}(t_{BA} + t_{AA}) = \frac{1}{2}(t_{BB} + t_{AB})$. This implies that the alignment problem can be posed as finding the offsets to match up these center points from each emission pair, indicated by the grey dots.

Let Δ_i be the time offset of device i relative to some arbitrary reference clock, so that we may write:

$$\Delta_i + \frac{t_{ji} + t_{ii}}{2} = \Delta_j + \frac{t_{jj} + t_{ij}}{2} \quad (1)$$

$$2(\Delta_i - \Delta_j) = t_{ij} + t_{jj} - t_{ji} - t_{ii} \quad (2)$$

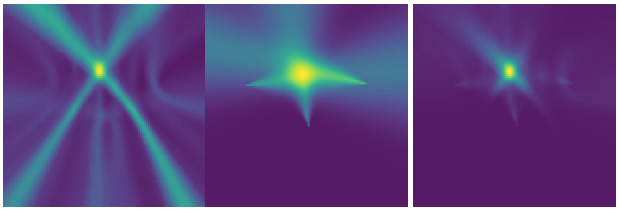
An example system for 3 devices is shown in (3). The local (unaligned) timestamps t_{ij} can be estimated from the captures via cross-correlation [12], and the time offsets may be obtained by the Ordinary Least Squares (OLS) solution. The fourth equation in (3) ensures that the average delay is 0. Without such a constraint there are an infinite number of solutions, because adding an equal offset to all devices is also a solution.

$$\begin{bmatrix} 2 & -2 & 0 \\ 2 & 0 & -2 \\ 0 & 2 & -2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \Delta_A \\ \Delta_B \\ \Delta_C \end{bmatrix} = \begin{bmatrix} t_{AB} + t_{BB} - t_{BA} - t_{AA} \\ t_{AC} + t_{CC} - t_{CA} - t_{AA} \\ t_{BC} + t_{CC} - t_{CB} - t_{BB} \\ 0 \end{bmatrix} \quad (3)$$

The proposed emission sequence for capture-alignment imposes a significant constraint on the overall system. However, such emission sequence can be repurposed to estimate device-to-device range [13], which could be used to estimate or improve the device map[14].

3. USER LOCALIZATION

The user is localized by probabilistic fusion of AoA (with one AoA estimate for each sub-array) and TDoA (using one channel from each sub-array). The boundaries of the user utterance are determined after detecting a keyword spoken by the user. The output of the TDoA and AoA estimators are each a 2D grid of likelihood values, known as a Spatial Likelihood Function (SLF) [15], which approximates the likelihood of each position on the grid given the input measurements. The TDoA estimates and AoA estimates are assumed independent. Hence, the fused likelihood is computed by element-wise multiplication of the individual likelihoods of the grid points. The maximum of the resulting surface gives the maximum-likelihood estimate for the user position. Examples of these likelihoods are shown in Fig. 4.



(a) (b) (c)

Fig. 4: (a) TDoA likelihood map, (b) AoA likelihood map, (c) TDoA-AoA fused likelihood map

3.1. TDoA Estimation

When a signal is emitted from some location (ex. U0 of Fig. 1), it arrives at each microphone at different times. For a given TDoA between a pair of microphones, we can constrain the source location to a hyperbola. Finally, given multiple pairs of microphones, the source location may be estimated by the intersection of these hyperbolas. Fig. 4 shows a number of these hyperbolas as bright lines in the likelihood surface. While we do not need to know the emission time to compute these time differences, we do require that temporal

events can be compared across devices, requiring the synchronization process described in Sec. 2.

The squared pairwise cross-correlations can be interpreted as likelihood functions, and they are normalized into a probability distribution associated with each pairwise TDoA/AoA. Under the assumption that each pairwise probability is independent, we can finally assign a probability to any point in space by computing the time difference for each pair of sensors, and taking the product of all the resulting probability densities. This is known as *direct* TDoA [3], as opposed to *indirect* TDoA which requires first forming an explicit TDoA estimate for each pair. Using this approach, uncertainty in the device locations or time synchronization can be modeled by convolving an uncertainty window to smooth the cross-correlation function [16].

3.2. AoA Estimation

The angles between the user and individual devices represent another key information for user localization. Computing the AoA of a sound signal is a classical estimation problem [17, 18]. In this system, we developed a novel algorithm based on acoustic wave decomposition of the microphone array observations. The algorithm exhibits high robustness to reverberation and user orientation. Furthermore, it is scalable to diverse microphone array geometries and form-factors with a small calibration effort.

The underlying physical model is the generalized acoustic wave decomposition [19]. In this model, the acoustic pressure, $\mathbf{p}(\omega; t)$, is modeled as a superposition of acoustic plane waves at different directions. This can be expressed as

$$\mathbf{p}(\omega; t) = \sum_{l} \alpha_l(\omega; t)\, \boldsymbol{\psi}(\theta_l(t), \phi_l(t); \omega) \tag{4}$$

where θ_l and ϕ_l denote respectively the elevation and azimuth of the direction of propagation of the l-th acoustic wave; and $\boldsymbol{\psi}(\theta, \phi; \omega)$ is the total acoustic pressure at the microphone array when a free-field acoustic plane wave with direction (θ, ϕ) impinges on the device [19]. The total acoustic pressure is the superposition of the incident free-field plane wave and the scattered component at the device surface and it is computed offline using acoustic simulation techniques [20]. The estimation procedure maps the physical model to a statistical framework to extract the line-of-sight component by computing the delay between the individual directional components, and integrate over the duration of the user utterance. The AoA estimate, $\hat{\theta}_i$, is unbiased, and it is modeled with a normal distribution with variance $\hat{\sigma}_i^2$. Hence, the 2D SLF at a point (x, y) is:

$$\mathrm{SLF}_{\mathrm{aoa}}(x, y) = \prod_{i=1}^{N_d} \mathcal{N}\left(\arctan\left(\frac{y - y_i}{x - x_i}\right); \hat{\theta}_i, \hat{\sigma}_i^2\right) \tag{5}$$

where N_d is the number of devices, and (x_i, y_i) is the coordinate of the i-th device. More details are described in [21]

3.3. Confidence Measures of User Localization

3.3.1. Cramer-Rao Bound

The Cramer-Rao Bound (CRB) is a lower bound of the estimation covariance of any estimator of a given estimation problem [22]. If the vector of unknown parameters (to be estimated) is $\mathbf{u}$, and the input observation vector is $\mathbf{b}$, with joint probability $P(\mathbf{b}; \mathbf{u})$; then the CRB matrix is defined as the inverse of the Fisher Information Matrix, $\mathbf{I}(\mathbf{u})$; whose (i, j) entry is defined as

$$[\mathbf{I}(\mathbf{u})]_{i,j} = -E\left\{\frac{\partial^2 \log P(\mathbf{b}; \mathbf{u})}{\partial u_i \partial u_j}\right\} \tag{6}$$

For the user localization problem, the unknown parameters are the user position, i.e., $\mathbf{u} = (x, y)$. Assume K anchor devices are used for localization, with *known* positions $\{(x_k, y_k)\}_{1 \le k \le K}$. If both AoA measurements, $\{\theta_k\}_{1 \le k \le K}$, and TDoA measurements, $\{\tau_{lk}\}_{1 \le l,k \le K}$, are fused to compute the user position, then the observation vector has $K + K(K-1)/2$ observations (after removing repeated terms in TDoA). Assuming all observations are independent, then

$$\begin{aligned} \log P(\mathbf{b}; \mathbf{u}) = &\sum_{k} \log P_\theta(\theta_k; x_k, y_k, x, y) + \\ &\sum_{l,k} \log P_\tau(\tau_{l,k}; x_l, y_l, x_k, y_k, x, y) \end{aligned} \tag{7}$$

Furthermore, we assume that both P_θ and P_τ follow a normal distribution, with mean and variance (μ_k, σ_θ^2) and $(\eta_{l,k}, \sigma_\tau^2)$ respectively. If the estimators are unbiased, then $\mu_k = \tan^{-1}(y - y_k)/(x - x_k)$ and $\eta_{lk} = r_l - r_k$ where $r_k^2 \triangleq (x - x_k)^2 + (y - y_k)^2$. The 2×2 Fisher Information Matrix, $\mathbf{I}(x, y)$, can be computed with straightforward arithmetic from the above components and the joint probability density function in (7). For example,

$$[\mathbf{I}(x, y)]_{1,1} = \sum_{k} \frac{(y_k - y)^2}{\sigma_\theta^2 r_k^4} + \frac{1}{\sigma_\tau^2} \sum_{l,k} \left(\frac{x - x_l}{r_l} - \frac{x - x_k}{r_k}\right)^2 \tag{8}$$

3.3.2. DNN-based Confidence Estimation

The TDoA/AoA likelihood map carries information related to user localization performance. For high quality examples (ex. Fig. 4), peaks are prominent which accurately represent user location with high certainty. Conversely, in low certainty examples, peaks may be smeared or noisy. Instead of using any heuristics, we pose this as a deep neural network (DNN) problem, where the input to the network is the likelihood map and the output is a classification of whether the estimate is reliable.

The network is designed following image recognition networks [23], with N convolutional layers followed by M fully connected layers. The last layer has sigmoid as activation to produce a probability output p of whether the localization error is beyond a predefined threshold, which acts as a proxy for confidence. To train the network we use binary cross-entropy

$$L(y, p) = -(y \log(p) + (1 - y) log(1 - p)) \tag{9}$$

where the label y is a binary indicator of whether user localization error is above a target threshold or not. The detection accuracy of the above model is 88% for detecting high and low confidence case. The true positive rate and true negative rate are 0.87 and 0.88, respectively.

4. USER ORIENTATION

4.1. Radiation Pattern

Due to human head and torso, speech is not omni-directional, a desirable property which allows us to estimate user orientation i.e., facing direction based on the recorded speech signal. In particular, Fig. 5 shows an example of an anechoic radiation pattern of human speech,

computed from simulation of a head and torso simulator (HATS). From Fig. 5 we observe that when the frequency is low (e.g. 125, 250 Hz), the radiation pattern is more omnidirectional and as we move to higher frequency components, the radiation pattern takes a directional shape. Therefore, most of the voice signal energy is captured at the microphone arrays placed in front of the user. Conversely, the back microphone arrays receive lower energy signal especially for high frequency components.

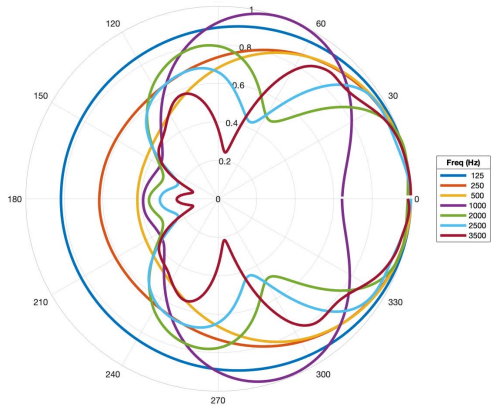


Fig. 5: Radiation pattern of human speech

4.2. Power-based Facing Direction

For facing-direction estimation, we propose an algorithm that utilizes an array of N microphones, the user location, and the recorded speech signals. These steps are outlined below:

Step 1: We calculate the power of the speech signal at each microphone. Then, we use the distance between the user and each microphone to normalize the power of the signal. The normalization is used to eliminate the path loss attenuation. Therefore, we get a N dimensional vector of normalized powers $P = \{P_1, P_2, ..., P_N\}$.

Step 2: Using the normalized power vector, we create a sparse vector of estimated radiation pattern $\hat{R}$ with a dimension of 360 by placing the normalized powers in the indices of their azimuth estimated by the user AoA with respect to the microphone arrays. The correlation between this estimated radiation pattern $\hat{R}$ and ground truth R is maximum at the true user facing direction. Therefore, we rotate the estimated radiation pattern with an increment of five degree and calculate the correlation between the rotated $\hat{R}$ and ground truth R. The angular shift with the maximum correlation is estimated as the user facing direction.

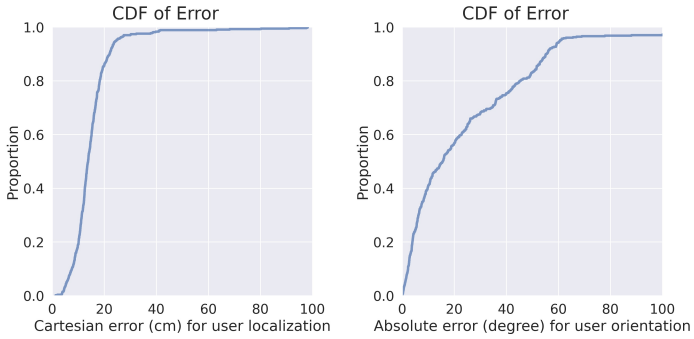
5. EXPERIMENTAL EVALUATION

We evaluate the proposed system on a set of recordings captured with actual devices in a variety of real rooms. Ground truth locations of devices and the user (represented by a HATS dummy head) were measured to within 3 cm accuracy, and orientations were measured within 3°. A total of 76 captures were utilized that represent 532 device configurations. The localization error is reported as:

Location Error: The Cartesian distance (in cm) between the estimated and actual user positions. This is shown in Fig. 6 (a).

Orientation Error: The difference (in degrees) between the estimated and actual user orientation (facing direction). This is shown in Fig. 6 (b).

In Fig. 7, the CRB variance, as described in section 3.3.1, is compared to the measured mean-squared localization error. Note



(a) (b)

Fig. 6: (a) Localization performance of user, (b) Orientation detection performance of user

that, at low-error values , e.g., < 0.5m, the averaged measured MSE (solid-blue-line) approaches the corresponding CRB value (dashed-red-line).

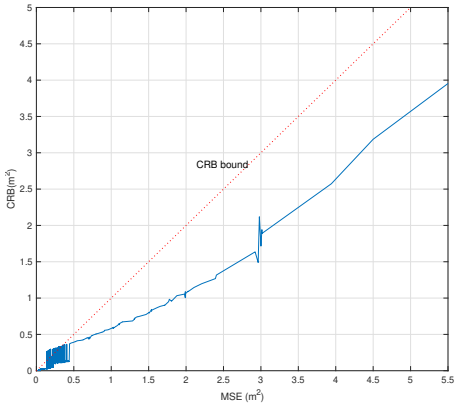


Fig. 7: Localization Performance vs. Cramer-Rao bound

6. CONCLUSION

The proposed system provides a practical solution to the joint user localization and orientation problem with voice-band processing. It utilizes a set of distributed, asynchronous, and heterogeneous microphone arrays; which is encountered in a typical teleconference setup. The TDoA and AoA estimates are combined in a statistical framework to compute the maximum likelihood estimate of the user position; which is then deployed to estimate user orientation. Further, the proposed system has the following practical merits:

1. It automatically calibrates for the clock offset between distributed devices to enable TDoA estimation.
2. It operates in real-time and requires minimal user interaction through a single utterance to perform the localization.
3. The computational and memory requirements are suited for embedded implementation, and it is flexible with respect to microphone array geometry and device form-factor

The discussion in the paper assumes a known device map. However, the proposed system can utilize similar techniques to estimate the device map to further enhance system integration. Future work would extend the framework to identify and localize relevant acoustic events.

7. REFERENCES

[1] Giuseppe Valenzise, Luigi Gerosa, Marco Tagliasacchi, Fabio Antonacci, and Augusto Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*. pp. 21–26, IEEE.

[2] S. Sumitani, R. Suzuki, N. Chiba, S. Matsubayashi, T. Arita, K. Nakadai, and H. G. Okuno, "An integrated framework for field recording, localization, classification and annotation of birdsongs using robot audition techniques — harkbird 2.0," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8246–8250, ISSN: 2379-190X.

[3] Joseph Hector DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," .

[4] Daniel J. Mennill, John M. Burt, Kurt M. Fristrup, and Sandra L. Vehrencamp, "Accuracy of an acoustic location system for monitoring the position of duetting songbirds in tropical forest," vol. 119, no. 5, pp. 2832–2839.

[5] Matthew Wijers, Andrew Loveridge, David W. Macdonald, and Andrew Markham, "CARACAL: a versatile passive acoustic monitoring tool for wildlife research and conservation," vol. 0, no. 0, pp. 1–17.

[6] Daniel Gabriel, Ryosuke Kojima, Kotaro Hoshiba, Katsutoshi Itoyama, Kenji Nishida, and Kazuhiro Nakadai, "2d sound source position estimation using microphone arrays and its application to a VR-based bird song analysis system," vol. 33, no. 7, pp. 403–414.

[7] Robert Showen, "Operational gunshot location system," in *Surveillance and Assessment Technologies for Law Enforcement*. vol. 2935, pp. 130–139, International Society for Optics and Photonics.

[8] Yu-Lin Wei, Rui Li, Abhinav Mehrotra, Romit Roy Choudhury, and Nic Lane, "Inferring facing direction from voice signals," arXiv, 2021.

[9] Jarred Barber, Yifeng Fan, and Tao Zhang, "End-to-end alexa device arbitration," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 926–930, ISSN: 2379-190X.

[10] Jackie (Junrui) Yang, Gaurab Banerjee, Vishesh Gupta, Monica S. Lam, and James A. Landay, "Soundr: Head position and orientation prediction using a microphone array," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20, pp. 1–12, Association for Computing Machinery.

[11] Karan Ahuja, Andy Kong, Mayank Goel, and Chris Harrison, "Direction-of-voice (DoV) estimation for intuitive speech interaction with smart devices ecosystems," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. UIST '20, pp. 1121–1131, Association for Computing Machinery.

[12] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," vol. 24, no. 4, pp. 320–327.

[13] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan, "BeepBeep: a high accuracy acoustic ranging system using COTS mobile devices," in *Proceedings of the 5th international conference on Embedded networked sensor systems - SenSys '07*. p. 1, ACM Press.

[14] Yi Shang, Wheeler Ruml, Ying Zhang, and Markus P. J. Fromherz, "Localization from mere connectivity," in *Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing*. MobiHoc '03, pp. 201–212, Association for Computing Machinery.

[15] Parham Aarabi, "The fusion of distributed microphone arrays for sound localization," vol. 2003, pp. 338–347.

[16] Spencer Russell, "Resynthesizing volumetric soundscapes : low-rank subspace methods for soundfield estimation and reconstruction," Accepted: 2020-09-15T22:01:09Z ISBN: 9781193026695.

[17] Joseph H DiBiase, Harvey F Silverman, and Michael S Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays*, pp. 157–180. Springer, 2001.

[18] Pierre-Amaury Grumiaux, Sran Kitić, Laurent Girin, and Alexandre Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[19] Mohamed Mansour, "Sparse recovery of acoustic waves," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 5418–5422.

[20] Amit Chhetri, Mohamed Mansour, Wontak Kim, and Guangdong Pan, "On acoustic modeling for broadband beamforming," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[21] Mohamed Mansour, "Maximum likelihood estimation of the direction of sound in a reverberant noisy environment," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 16–20.

[22] Steven M Kay, *Fundamentals of statistical signal processing: estimation theory*, Prentice-Hall, Inc., 1993.

[23] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.