TOWARDS WIRELESS ACOUSTIC SENSOR NETWORKS FOR LOCATION ESTIMATION AND COUNTING OF MULTIPLE SPEAKERS IN REAL-LIFE CONDITIONS

Anastasios Alexandridis[⋆†], *Nikolaos Stefanakis*[⋆], *Athanasios Mouchtaris*[⋆†]

[⋆] FORTH-ICS, Signal Processing Laboratory, Heraklion, Crete, Greece, GR-70013
[†] University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-70013
{analexan, nstefana, mouchtar}@ics.forth.gr

ABSTRACT

Speaker localization and counting in real-life conditions remains a challenging task. The computational burden, transmission usage and synchronization issues pose several limitations. Moreover, the physical characteristics of real speakers in terms of directivity pattern and orientation, as well as restrictions in the microphone array positioning, which commonly have to be placed close to walls, deteriorate the localization performance. In this paper, we propose a localization and counting method that accounts for the adjacent wall reflections and evaluate it using a dataset of real recorded signals of actual speakers that we collected. Our dataset is publicly available to foster further investigation towards localization in real-life scenarios.

Index Terms— location estimation, source counting, reflections, real recordings, wireless acoustic sensor networks

1. INTRODUCTION

Speaker localization in a wireless acoustic sensor network (WASN) where each sensor consists of a microphone array has been an emerging field of interest. Usually, a centralized scheme is adopted, where a dedicated node (the “fusion center”) performs the localization based on information transmitted from the arrays. Among the variety of localization approaches, different methods make different assumptions about the acoustic environment and the signal model and come with specific advantages and limitations. However, few of them have been tested in real-life conditions.

The practical, real-life deployment of a WASN for localization poses many new challenges. Some of them arise from computational complexity, bandwidth usage, and synchronization issues [1]: the computational complexity becomes an important issue for real-time implementations, while the wireless nature of the sensors limits the amount of information that can be transmitted. Also, as each array has its own clock, the signals across the arrays are not synchronized.

Another important issue arises from limitations on the positioning of the microphone arrays in the room, which in real-life need to be placed close to walls in order not to pose restrictions on the activities of the speakers inside the room. The presence of such reflecting surfaces so close to the microphone arrays is generally known to have an adverse effect on the performance of the array, which can degrade the localization performance of the entire system. Also, the characteristics of real speakers in terms of their directivity pattern, spatial volume, and orientation are far more complicated than omnidirectional point sources which are usually used for simulations. Lastly, in realistic scenarios with multiple active speakers, it is reasonable to assume that their number is also unknown and may vary arbitrarily in time, which constitutes the estimation of the number of sources also important.

A well-known class of localization methods is based on constructing a likelihood map of candidate source locations, usually with the use of the steered response power with phase transform (SRP-PHAT) [2], and estimating the speaker locations from the peaks of the likelihood map [3]. Although the computational burden of SRP-PHAT methods has been optimized (e.g., [4–6]), the transmitted information is still high, since to compute the SRP each node must transmit its entire Generalized Cross-Correlation function [2]. SRP-PHAT has also been used for multiple sources (e.g., [7]), however due to the many local extrema of the SRP space, the performance deteriorates with an increasing number of speakers.

Other approaches rely on time-differences of arrival (TDOA) or direction of arrival (DOA) estimates from the sensors. The location can be found as the intersection of DOA lines [8–11], or the intersection of hyperbolas defined by the estimated TDOAs [12, 13]. Such methods, maintain low transmission requirements—as only the TDOAs/DOAs need to be transmitted—but can become quite complicated for multiple sources: the number of estimated TDOAs/DOAs in each time instant can vary across the sensors due to missed detections or overestimation of the number of sources and an association procedure is needed to find the TDOAs/DOAs combinations that correspond to the same source [14]. Finally, other methods utilize statistical approaches to model spatial features extracted from the arrays. An Expectation-Maximization algorithm for speaker localization is presented in [15], while distributed approaches are discussed in [16,17].

In this paper, we present the implementation of a source counting and location estimation system in the challenging conditions of a real environment. Our previous work of [18] is employed at the fusion center to perform the source counting and localization, based on clustering of per-frequency location estimates. It offers reduced transmission requirements, while it does not require perfect synchronization, facilitating its use in real-life applications. We incorporate our system with our recently proposed method of [19] to explicitly take into account the reflections that occur when the arrays are close to walls. To investigate the challenges occurred in real-life conditions, we collected and present a dataset[1] of real recordings in a typical office room. We evaluate the performance of our system in this challenging dataset, which is publicly available in order to assist the research community move a step towards accurate localization in real-life scenarios.

2. “REFLECTION-AWARE” DOA ESTIMATION

Following [19], the DOA estimation accuracy for a circular array of M sensors and radius R, placed in front of a wall can be significantly improved by designing a propagation model which is aware of the earliest reflection introduced by the adjacent wall. Let us first review the typical propagation model, expressing the relative sound pressure at the mth microphone as a function of frequency ω and incident angle θ as

$$a_m(\omega, \theta) = e^{jkR\cos(\phi_m - \theta)}. \tag{1}$$

Here, $k = \omega/c$ is the wavenumber, c is the speed of sound and ϕ_m is the angle of the mth sensor which similar to θ, is defined with respect to the center of the circular disk. This typical propagation model accounts for the direct path of the sound only and ignores any distinct reflections that may occur.

A so-called half-space version of the propagation model associated to the same circular array can be designed with the model for the mth microphone defined as [19]

$$\begin{aligned}\hat{a}_m(\omega, \theta) = {} & e^{jkR\cos(\phi_m - \theta)} e^{jk\epsilon\cos\theta} \\ & + h e^{jkR\cos(\phi_m - \pi + \theta)} e^{-jk\epsilon\cos\theta},\end{aligned} \tag{2}$$

where ϵ is the distance of the array center from the adjacent wall, θ is the incident angle defined so that $\theta = 0^\circ$ is normal to the wall (Fig. 1), and $h \in [0, 1]$ is the so-called Image Source Relative Gain (ISRG) which encodes the reflective properties of the wall. Assuming h to be real and constant with frequency is an affordable simplification, although in practice the wall reflectivity would be more accurately represented by a complex and frequency-varying ISRG [19]. Letting now $\hat{\mathbf{a}}(\omega, \theta) = [\hat{a}_1(\omega, \theta), \cdots, \hat{a}_M(\omega, \theta)]^T$ be the vector concatenating all the M terms from Eq. (2), the half-space steering vector is derived as

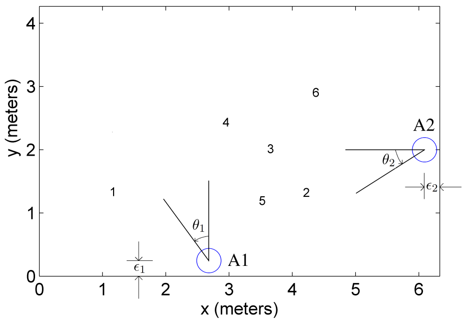


Fig. 1. Recording setup used for evaluation.

$$\mathbf{a}(\omega, \theta) = \hat{\mathbf{a}}(\omega, \theta) / \left\| \hat{\mathbf{a}}(\omega, \theta) \right\|_2, \tag{3}$$

where $\|\cdot\|_2$ denotes the Euclidean norm.

Using any of the two propagation models, different approaches to DOA estimation can be employed. Here, we utilize a Minimum Variance Distortionless Response (MVDR) beamformer [20]. Performed in a time-frequency (TF) basis, with τ denoting the time-frame index, we find the local DOA $\hat{\theta}(\tau, \omega)$ where the MVDR beamformer response is maximized by searching across the entire range of potential directions from 0 to 360 degrees. The DOA estimates, $\hat{\theta}(\tau, \omega)$ up to a maximum cutoff frequency ω_c from each array are then transmitted to the fusion center.

3. LOCATION ESTIMATION AND COUNTING

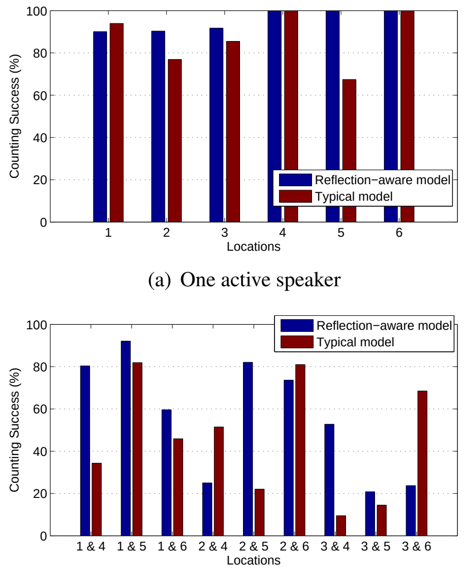
Assuming a WASN with N arrays at known locations the fusion center estimates the locations of an unknown number of K sources which are present in the environment.

3.1. Per-frequency location estimation

The estimated DOAs from each array at a given frequency are used to infer a location estimate for that frequency. To do so, we utilize our single-source grid-based (GB) location estimator described in [11,21]. The method constructs a grid over the area where localization is performed and divides it into a set of grid points. The location for each frequency is estimated as the grid point whose DOAs most closely match the estimated DOAs from the arrays at that frequency.

Based on the geometry of the space and the array locations, each array has a specific range of “allowable” DOAs, i.e, DOAs that can result in locations inside the localization area defined by the grid. Thus, before we apply our single-source GB method we check that the DOA of each array is

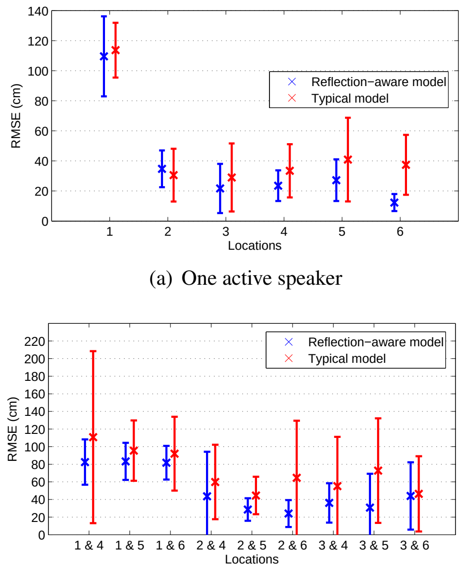
[1] Our dataset is publicly available at `https://github.com/spl-icsforth/WASN-Recordings-OfficeRoom`.



(a) One active speaker

(b) Two active speakers

Fig. 2. Counting success rates for our dataset, using our "reflection-aware" and the typical model for DOA estimation.



(a) One active speaker

(b) Two active speakers

Fig. 3. Localization error for our dataset, using our "reflection-aware" and the typical model for DOA estimation.

in its "allowable" range. If not, we don't estimate a location and move to the DOAs for the next frequency. Finally, for each time-frame we create a block of per-frequency location estimates that contains the estimates of the current frame and B previous frames (history length).

3.2. Outlier rejection

To remove erroneous location estimates that occurred from the previous step, we construct the two-dimensional histogram of location estimates and smooth it by applying an averaging filter with a rectangular window of length h_X and h_Y in the $x-$ and $y-$ dimension, respectively. Finally, we remove the per-frequency location estimates whose cardinality in the histogram is less than q times the maximum cardinality, where $q \in [0, 1]$ is a pre-defined threshold. The underlying assumption behind this scheme is that erroneous estimates are expected to be of low cardinality in the histogram.

3.3. Final location estimation and counting

The remaining per-frequency location estimates are expected to form K clusters around the K sources' locations. The final location estimation and counting is performed by clustering the per-frequency location estimates using the bayesian K-means algorithm [22], where the number of clusters is also unknown. For more details the reader is referred to [18].

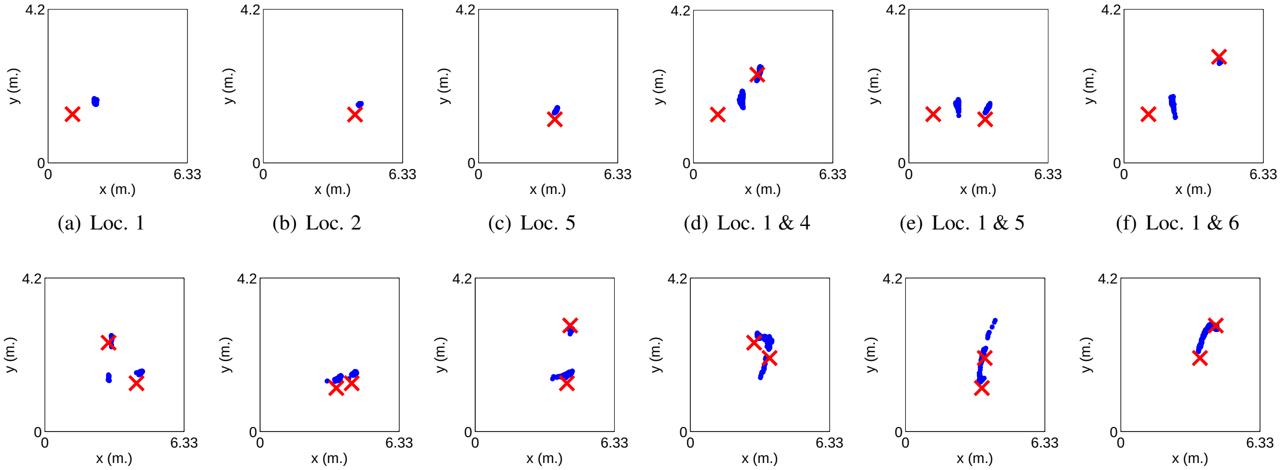
4. DATASET OF REAL RECORDINGS

The dataset contains real recordings of speech sources in a typical office room of dimensions $L_x = 6.33$ and $L_y = 4.2$ meters with reverberation time of approximately equal to 400 ms. The recording setup is depicted in Fig. 1. The recordings were made at predefined locations (1–6 in Fig. 1), by two male speakers. The first speaker (M01) was recorded at locations 1–3, and the second one (M02) at locations 4–6. The speakers were asked to stand in the predefined locations with an orientation towards the center of the room, without further advising them about where to look at or how loud to speak.

We used two uniform circular microphones arrays. The array locations were measured to be $(2.68, 0.086, 1.20)$ meters for the first array (A01) and $(6.248, 2, 1.20)$ meters for the second array (A02). Both arrays were placed very close to walls: A01 is 8.6 cm and A02 is 8.2 cm away from the corresponding walls. Both arrays consisted of 8 Shure SM95 omnidirectional microphones and a radius of 5 cm. They operated individually (i.e., they were connected to different host PCs). Utterances for each speaker and each location were segmented from the original recordings and synchronized by eye-inspection. The signals were recorded at 48 kHz sampling rate.

5. RESULTS AND DISCUSSION

We used the recordings (downsampled at 12 kHz) of our dataset to evaluate our method in real-life conditions. For processing, we used an FFT of 512 samples length, with 50% overlap windowed with a square root hanning window and



(a) Loc. 1 (b) Loc. 2 (c) Loc. 5 (d) Loc. 1 & 4 (e) Loc. 1 & 5 (f) Loc. 1 & 6

(g) Loc. 2 & 4 (h) Loc. 2 & 5 (i) Loc. 2 & 6 (j) Loc. 3 & 4 (k) Loc. 3 & 5 (l) Loc. 3 & 6

Fig. 4. Location estimates (the blue clouds) for the real recordings of one [(a)–(c)] and two [(d)–(l)] speakers (the red X's).

$\omega_c = 4$ kHz. For the per-frequency location estimates we used 1 second history length ($B = 46$ frames). The parameters for outlier rejection were set to $h_X = h_Y = 20$ cm and $q = 0.25$, as these parameters were found to perform best in most of the cases. The "allowable" range of DOAs was set to $[0°, 70°] \cup [290°, 360°)$ for A01 and $[0°, 45°] \cup [315°, 360°)$ for A02. Finally, ISRG was set to $h = 0.9$ and all other parameters were set according to [18, 19]. To consider scenarios of two sources, we artificially added the microphone array signals at different locations and from different speakers. This resulted in 9 different cases, which represent all combinations of mixing speaker M01 at locations 1–3 with speaker M02 at locations 4–6. The energy of the speakers was not equalized to better model real-life conditions where energies will not be equal.

Fig. 2 depicts the counting success rate as the percentage of time-frames where the correct number of sources was found for the single-source and two-source case. The corresponding Root Mean Squared Error (RMSE) with error bars representing one standard deviation is shown in Fig. 3. For comparison, we include the results when the sensors use the same DOA estimator but with the typical steering vector (i.e, not accounting for reflections). It is evident that when the sensors utilize our "reflection-aware" DOA estimator, we can achieve better performance: while the counting success rate is not always improved, the location estimation error is significantly reduced for all tested locations, especially in the two sources case.

Generally, our system operates in a functional range of values for these challenging scenarios, both in terms of counting and localization performance. It is also important to note how the performance varies in different locations. A large error is observed at location 1 in the single source case, which can be explained by the fact that this location is much further from the arrays (especially array A02). A performance degradation especially in terms of counting success rate is evident also in some cases of the two sources scenario (e.g, location pairs 2 & 4, 3 & 5), which can be attributed to the small distance between the sources, as well as to the small angular separation of the sources with respect to one (or both) the arrays. Such location pairs could benefit from the deployment of more microphone arrays. These results also highlight the importance of evaluation across the entire localization cell, a direction towards which little effort has been made so far.

Fig. 4 shows the location estimates using our "reflection-aware" DOA estimation, for 12 out of 15 tested source locations (due to space limitations the three single-source cases with the smallest error have been omitted). The blue dots show the cloud of estimates over the entire duration of the signals for the time-frames where the correct number of speakers was found, revealing again a quite accurate localization. Finally, given that the locations and orientations of the arrays were not finely calibrated and had unintended offsets of a few centimetres and degrees, the conditions were far from ideal, making our results quite encouraging.

6. CONCLUSIONS

In this paper, we considered the location estimation and counting problem in a real WASN and real-life conditions. We extended our previously proposed method of [18] by incorporating a model that takes into account the early reflections for DOA estimation [19], resulting in improved localization and counting when the microphone arrays are close to walls, a setup which is quite ordinary in real-life situations. We evaluated our approach using real recorded signals that reflect the challenges that occur in real-life scenarios, such as the directivity pattern, spatial volume, and orientation of real speakers. Our dataset of real recordings is publicly available to allow the evaluation of localization approaches in real-life conditions. In the future, we plan to extend our dataset with more speakers, source locations, and microphone arrays.

7. REFERENCES

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *IEEE Symp. on Communications and Vehicular Technology in the Benelux*, 2011, pp. 1–6.

[2] J. H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in reverberant environments using microphone arrays*, Ph.D. thesis, Brown University, Providence, RI, 2000.

[3] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 338–347, Jan. 2003.

[4] H. Do and H. F. Silverman, "A Fast Microphone Array SRP-PHAT Source Location Implementation using Coarse-To-Fine Region Contraction (CFRC)," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2007, pp. 295–298.

[5] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2007, vol. 1, pp. I–121–I–124.

[6] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 71–74, Jan 2011.

[7] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–17, 2010.

[8] M. Gavish and A. J. Weiss, "Performance analysis of bearing-only target location algorithms," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 28, no. 3, pp. 817–828, 1992.

[9] L. M. Kaplan, Q. Le, and N. Molnar, "Maximum likelihood methods for bearings-only target localization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, vol. 5, pp. 3001–3004.

[10] A. Griffin and A. Mouchtaris, "Localizing multiple audio sources from DOA estimates in a wireless acoustic sensor network," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2013, pp. 1–4.

[11] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, "Localizing multiple audio sources in a wireless acoustic sensor network," *Signal Processing*, vol. 107, pp. 54 – 67, 2015, Special Issue on ad hoc microphone arrays and wireless acoustic sensor networks.

[12] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 439–443, Feb 2013.

[13] M. Compagnoni, P. Bestagini, E. Antonacci, A. Sarti, and S. Tubaro, "Localization of acoustic sources through the fitting of propagation cones using multiple independent arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1964–1975, 2012.

[14] A. Alexandridis, G. Borboudakis, and A. Mouchtaris, "Addressing the data-association problem for multiple sound source localization using DOA estimates," in *European Signal Processing Conference (EUSIPCO)*, Aug 2015, pp. 1551–1555.

[15] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 392–402, Feb 2014.

[16] Y. Dorfan, G. Hazan, and S. Gannot, "Multiple acoustic sources localization using distributed expectation-maximization algorithm," in *Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2014, pp. 72–76.

[17] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1692–1703, Oct 2015.

[18] A. Alexandridis and A. Mouchtaris, "Multiple sound source location estimation and counting in a wireless acoustic sensor network," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2015, pp. 1–5.

[19] N. Stefanakis and A. Mouchtaris, "Direction of arrival estimation in front of a reflective plane using a circular microphone array," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2016.

[20] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996.

[21] A. Griffin, A. Alexandridis, D. Pavlidi, and A. Mouchtaris, "Real-time localization of multiple audio sources in a wireless acoustic sensor network," in *European Signal Processing Conference (EUSIPCO)*, Sep 2014, pp. 306–310.

[22] K. Kurihara and M. Welling, "Bayesian k-means as a "maximization-expectation" algorithm," *Neural Computation*, vol. 21, no. 4, pp. 1145–1172, Apr 2009.