Microphone Pair Selection for Sound Source Localization in Massive Arrays of Spatially Distributed Microphones

Bilgesu Çakmak*, Thomas Dietzen*, Randall Ali†, Patrick Naylor‡, Toon van Waterschoot*

* *Dept. of Electrical Engineering (ESAT), STADIUS, KU Leuven*, Leuven, Belgium
† *Institute of Sound Recording Dept. of Music & Media, University of Surrey*, Guildford, UK
‡*Dept. of Electrical and Electronic Engineering, Imperial College London*, London, UK

***Abstract*—In massive distributed microphone arrays, such as in conference systems, the use of all sensors leads to unnecessary energy consumption and limited network life, as some sensors may have a limited contribution to specific estimation tasks, such as sound source localization (SSL). In this work, we propose two pre-processing methods for microphone pair selection in steered response power (SRP) based SSL using massive arrays of spatially distributed microphones. The first method is based on thresholding which results in a selection of broadside microphone pairs, that is, microphone pairs that are oriented broadside to the source direction, and only the signals from those microphones are collected for use in the SRP algorithm. The second method is based on sparseness of the cross-correlation in which the pairs which have a high sparsity factor are selected for use in localization. These methods are further improved by leveraging orientational diversity. Simulations show that we achieve a localization performance only using the selected microphone pairs that is comparable to the performance when all microphones are employed.**

***Index Terms*—source localization, steered response power, wireless acoustic sensor network, microphone selection**

I. Introduction

Lately, there has been a surge in the production of small microphone sensors intended for use in wireless acoustic sensor networks (WASNs), and consequently source localization using WASNs has emerged as a trending topic [1]. The steered response power (SRP) approach is a widely used source localization technique [2]–[5].

In SRP, using the generalized cross-correlation (GCC) method [6], usually with a phase-transform (PHAT) weighting [7]–[9], a power map of the acoustic scene is created by a delay-and-sum beamformer (DSB) that is steered towards a grid of candidate locations based on their corresponding time differences of arrivals (TDOAs). The source's position is then estimated as the position on this spatial grid that has the highest power.

This research work was carried out at the ESAT Laboratory of KU Leuven and the EEE department of Imperial College London, and has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 956369 'SOUNDS' and from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

In SRP applied to WASNs, every node must normally send its microphone signal(s) to a fusion center. The conventional SRP requires that the processing takes place in a centralized manner, where it is assumed that all microphone signals are available in a fusion center, which is not usually the case for a WASN. There exist some studies focusing on the adaptation of the SRP to arrays of spatially distributed sensors [10], [11]. Since the exchange of information between nodes, or between nodes and the fusion center, can significantly reduce the battery life of the sensors [12], the power consumption due to exchange of information is a major limiting factor in WASNs [13].

In situations where a large number of spatially distributed microphones are present with a central fusion center, certain pairs will provide reliable spatial information, while others may yield inaccurate or redundant data. Therefore, only a handful may be required to achieve sufficient accuracy. As opposed to using all sensors, selecting a subset of microphones that gives sufficient information for the considered localization task reduces the amount of data to be processed, thus minimizing the transmission costs. Several recent studies [14]–[16] have suggested that microphone subset selection is fundamental for realizing energy-efficient WASNs. Moreover, the computational complexity of the SRP scales linearly with the number of microphone pairs [17]; therefore, microphone pair selection brings a significant computational advantage if the network consists of a massive number of microphones.

To address these issues specifically for source localization using the SRP method in a WASN setting with a fusion center, we propose two methods to select pairs of microphones to ensure robust and efficient localization estimation. The proposed microphone pair selection methods are based on two different criteria. On the one hand, we aim to select microphone pairs that yield a high effective resolution of the SRP map near the source location. Since the TDOA is most sensitive to direction of arrival (DOA) if the source is in the broadside direction of a microphone pair [18], we can therefore argue that preference should be given to such broadside microphone pairs. We propose to estimate a selection of these pairs based on the time-domain cross-correlation (TD-CC) at lag zero. On the other hand, we aim to select microphone pairs that are

least impacted by reverberation. Since reverberation reduces the sparsity of the TD-CC due to reflections, we propose to estimate a selection of such microphone pairs based on the sparsity of the TD-CCs. In the first approach, a selection of microphone pairs is made possible by thresholding their cross-correlations at time lag zero, and using only the pairs that are above a threshold, which often has been found to correspond to the broadside-oriented microphone pairs. The other approach involves exploiting the sparseness of TD-CCs in which the microphone pairs that have the most sparse cross-correlations are chosen for use in localization.

A major advantage of the aforementioned approaches is that, microphone signals do not necessarily need to be centralized. If the TD-CCs for lag zero can be computed by, for example, only exchanging microphone signals between microphone pairs, then only the TD-CC values for lag zero and microphone signals of the selected pairs must be collected in a fusion center. Likewise, if the sparsity values can be computed similarly then only these sparsity values and microphone signals of the selected pairs must be collected in a fusion center.

These methods can further be improved by taking orientational diversity into account. That is, microphone pairs with axes perpendicular to each other can localize in orthogonal directions. This configuration allows processing of different signals from various angles simultaneously. However, when pairs are oriented with parallel axes, they provide less orientational diversity hence less information for source localization.

The remainder of the paper is organized as follows. Section II describes the signal model along with the computation of the GCC and the conventional SRP algorithm. In Section III, we introduce the proposed microphone pair selection methods. In Section IV, we present the results of simulations comparing the proposed approaches to the conventional SRP method. Section V concludes the paper.

II. Signal Model and Steered Response Power (SRP) Computation

A. Signal Model

Let $M \in \mathbb{N}$ be the number of microphones and let $y_m(t)$ with $m = 1, \ldots, M$ denote the mth microphone signal in the time domain, where t is the discrete time index.

In the frequency domain, with ω the radial frequency, the GCC matrix $\mathbf{\Psi}(\omega) \in \mathbb{C}^{M \times M}$ is defined as

$$[\mathbf{\Psi}(\omega)]_{m,m'} = \psi_{m,m'}(\omega) = \frac{Y_m(\omega) Y^*_{m'}(\omega)}{|Y_m(\omega) Y^*_{m'}(\omega)|}, \tag{1}$$

where $Y_m(\omega)$ and $Y_{m'}(\omega)$ are the discrete-time Fourier transforms of a pair of microphone signals and * is the complex conjugation operator. $[\mathbf{\Psi}(\omega)]_{m,m'}$ denotes the element of the GCC matrix with index m, m'. Here, a PHAT weighting function $1/|Y_m(\omega)Y^*_{m'}(\omega)|$ is used, so only phase information remains which is known to improve spatial resolution [7], [8].

B. Steered Response Power

Let $\mathcal{P} = \{(m, m') | m > m'\}$ be the set of microphone pairs and let p index the elements of this set. Then, the frequency-dependent SRP map is defined as [17]

$$\begin{aligned} \text{SRP}(\omega, i) &= \mathbf{h}^H(\omega, i) \mathbf{\Psi}(\omega) \mathbf{h}(\omega, i) - \text{tr}[\mathbf{\Psi}(\omega)] \\ &= 2 \sum_{p \in \mathcal{P}} \Re[\psi_p(\omega) e^{j\omega \Delta t_p(i)}]. \end{aligned} \tag{2}$$

Here, $()^H$, $\Re()$, tr[ ], $\psi_p(\omega)$ denote the Hermitian transpose, the real part, the trace, and the frequency-domain cross-correlation (FD-GCC) respectively. Let J be the number of candidate locations. Then, the DSB steering vector towards the candidate location i is $\mathbf{h}(\omega, i)$ with $i = 1, \ldots, J$ defined relative to the first microphone [19] and

$$\mathbf{h}(\omega, i) = [1 \; e^{j\omega \Delta t_{2,1}(i)} \ldots \; e^{j\omega \Delta t_{M,1}(i)}], \tag{3}$$

where $\Delta t_p(i)$ is the TDOA observed at the microphone pair p from a sound source located at $\mathbf{q}_i$ and is given by

$$\Delta t_p(i) = \frac{(\|\mathbf{q}_i - \mathbf{p}_m\| - \|\mathbf{q}_i - \mathbf{p}_{m'}\|)}{c}. \tag{4}$$

Here $\mathbf{q}_i$ is the Cartesian coordinate vector of the ith candidate location and $\mathbf{p}_m$ is the Cartesian coordinate vector of the mth microphone position and c is the speed of sound. Generally, tr$[\mathbf{\Psi}(\omega)]$ is subtracted in (2), as it creates an offset and does not contribute to the purpose of spatial mapping [7], [17]. The total number of microphone pairs is given by

$$P = M(M-1)/2. \tag{5}$$

By integrating SRP(ω, i) over frequency, one can obtain the broadband SRP value

$$\text{SRP}(i) = \int_{-\omega_0}^{\omega_0} \text{SRP}(\omega, i) \, d\omega, \tag{6}$$

where ω_0 is the bandwidth of the signals, and the source location estimate $\hat{i}$ can be found as the location index with the maximum SRP, such that

$$\hat{i} = \underset{i}{\text{argmax}} \; \text{SRP}(i). \tag{7}$$

This method will be referred to as the 'conventional SRP' method.

III. Proposed Microphone Pair Selection Methods

In this section, we propose some alternative methods for microphone pair selection which will be subsequently evaluated in Sec. IV.

A. Selecting broadside pairs

For microphone pairs in the broadside direction, the TDOA is more sensitive to changes in the DOA as compared to when they are in the endfire direction. Consequently, the effective resolution of the localization map obtained from SRP is highest in the broadside direction. Therefore, selecting broadside microphone pairs would be a reasonable strategy to reduce the number of microphone pairs used in SRP. To achieve this, we compute the (unweighted) TD-CCs at lag 0 directly in the time domain, as

$$R_p(0) = \sum_{t=-\infty}^{\infty} y_m(t) y_{m'}(t). \tag{8}$$

Next, we define the set consisting of microphone pairs where the TD-CCs at lag 0 exceed a specified threshold that corresponds to 98% percentile of the data as $\mathcal{P}_{\text{select}} = \{p | R_p(0) > \text{threshold}\}$ and use only $p \in \mathcal{P}_{\text{select}}$ pairs in computation of the SRP in (2).

B. Sparsity

A sparse signal in time is one in which a large portion of energy is concentrated in only a few elements. Following this definition, the pairs that have the most sparse TD-CCs will have one single peak that will correspond to the true source location and hence contribute more accurately to the localization task. The reverberation causes the TD-CCs to be less sparse, so if a selection is made for the pairs that have sparse TD-CCs, the impact of reverberation can be mitigated. In Fig. 1a, a sparse TD-CC between a particular pair is shown, whereas Fig. 1b demonstrates a less sparse TD-CC between another pair. Various measures to quantify sparseness have been proposed in the literature [20]. A common measure is [21], [22]

$$\epsilon_p = \frac{L}{L - \sqrt{L}} \left[1 - \frac{\sum_{n=1}^{L} |R_p(n)|}{\sqrt{L \sum_{n=1}^{L} R_p(n)^2}} \right]. \tag{9}$$

where R_p is the TD-CC between the signals in pair p and is defined as

$$R_p(n) = \sum_{t=-\infty}^{\infty} y_m(t) y_{m'}(t+n), \tag{10}$$

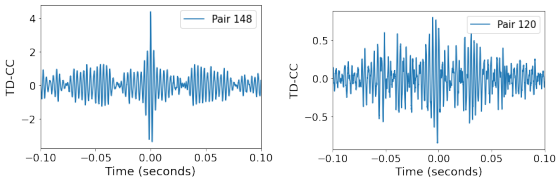
and L is the length of the TD-CC. This measure will be referred to as 'sparsity measure 1' in simulations.

Another measure is kurtosis which measures the peakedness [23] and is defined as

$$\rho_p = \frac{1}{L} \sum_{n=1}^{L} \frac{(R_p(n) - \bar{R}_p)^4}{\left((R_p(n) - \bar{R}_p)^2\right)^2}. \tag{11}$$

where $\bar{R}_p$ represents the average TD-CC value for that pair. This measure will be referred to as 'sparsity measure 2'.

Similar to the method in III-A, we apply a threshold to these sparsity values and use only the pairs that have sparsity values exceeding the threshold in computation of the SRP in (2).



(a) (b)

Fig. 1: (a) sparse TD-CC, (b) less sparse TD-CC



Fig. 2: Conference room

C. Orientational Diversity

The approaches in III-A and III-B can be further improved if orientational diversity is leveraged. That is, assigning weights to the selected microphone pairs based on their spatial orientation maximizes the directional information contribution of each microphone pair. Therefore, a further reduction of the number of selected pairs can be possible. One extreme case would be when the selected microphone pairs have parallel axes hence provide less information for source localization. The other extreme, providing most information would be when the selected microphone pairs are perpendicular to each other. This configuration is extremely useful as the pairs can capture sound from different angles.

Let $\mathbf{v}_p = \mathbf{p}_m - \mathbf{p}_{m'}$ represent the difference vector of Cartesian coordinate vectors formed by the two microphones in pair p. For two different pairs p and p', the angle $\theta_{p,p'}$ between vectors $\mathbf{v}_p$ and $\mathbf{v}_{p'}$ can be computed by

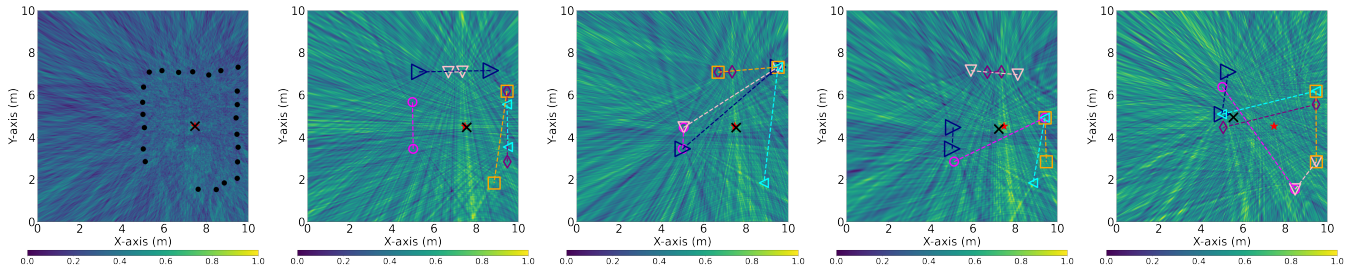
$$\theta_{p,p'} = \arccos\left(\frac{\mathbf{v}_p \cdot \mathbf{v}_{p'}}{\|\mathbf{v}_p\| \cdot \|\mathbf{v}_{p'}\|} \right). \tag{12}$$

To exploit orientational diversity, we propose to assign weigths to each combination of pairs p and p' in $\mathcal{P}_{\text{select}}$ as $w_{p,p'} = \cos(\theta_{p,p'} - \frac{\pi}{2})$, which equals 1 if $\theta_{p,p'} = \frac{\pi}{2}$. A subset of $\mathcal{P}_{\text{select}}$ can then be defined as $\mathcal{P}_{\text{subselect}} = \{p, p' | w_{p,p'} > \text{threshold}\}$ and be used in computation of the SRP in (2).

IV. Simulations

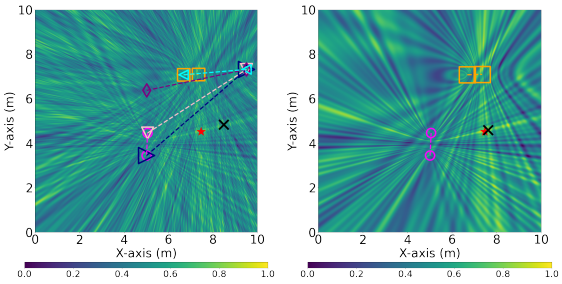
A. Simulation Framework

In this section, we consider a scenario based closely on a conference room that is used in real-world scenarios in Televic Conference Systems as in Fig. 2. We simulated this room of



(a) (b) (c) (d) (e)

Fig. 3: SRP maps (an example with SNR = 6 dB) using (a) Conventional SRP, red star: source position, black cross: estimated source position, black dots: microphones (b) 6 selected microphone pairs by means of broadside pairs selection, (c) 6 selected microphone pairs by means of sparsity measure 1 (d) 6 selected microphone pairs by means of sparsity measure 2, (e) 6 randomly selected microphone pairs, various shapes and colors along with the dashed lines show the selected microphone pairs for each measure.



(a) (b)

Fig. 4: SRP maps (a) using the microphones selected by sparsity measure 1, (b) applying orientational weighting after selection, red star: source position, black cross: estimated source position, various shapes and colors along with the dashed lines show the selected microphone pairs for each measure.

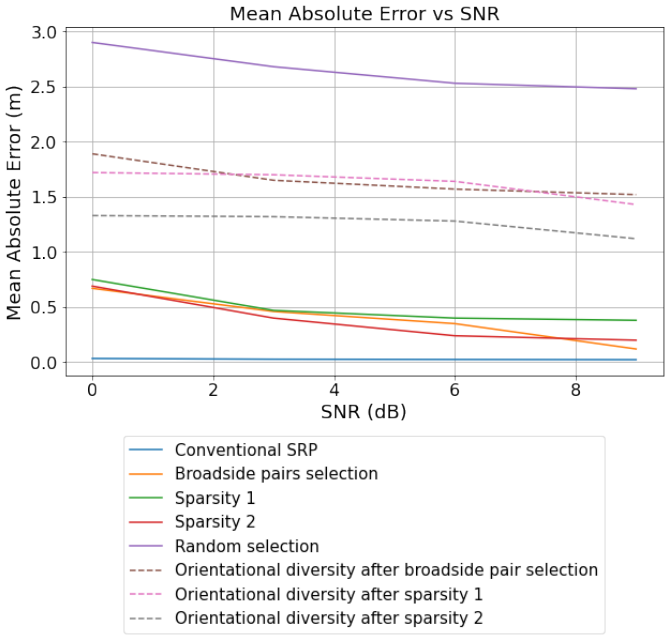


Fig. 5: Mean Absolute Localization Errors

size $10,30$ x $10,13$ x $3,48$ m with 600 ms reverberation time, wherein 23 microphones, as well as a single female speech source at the center [24] are placed. Additive white Gaussian noise (AWGN) is added at signal-to-noise ratios (SNRs) of SNR = $[0, 3, 6, 9]$ dB. The number of microphone pair combinations P is $\binom{23}{2} = 253$ in this case. The acoustic scenario is modelled using the randomized image method [25] at a sampling rate of 16 kHz.

B. Application of Microphone Pair Selection Methods

We now present the SRP maps and localization results after the application of the three aforementioned microphone pair selection methods, and a randomized control which is the result after making a random microphone pair selection along with the conventional SRP result.

Fig. 3a shows the conventional SRP map in which all of the 23 microphones are used. Fig. 3b shows the resulting SRP map obtained from using only 6 selected microphone pairs by thresholding the TD-CCs. Next, we investigate using sparseness of the TD-CCs as a pre-processing method. By using Eq. 9, we compute the sparsity of each microphone combination's TD-CC and use only the most sparse pairs in the computation of the SRP. Fig. 3c shows the corresponding SRP map using only those chosen 6 pairs. Following the same procedure for sparsity measure 2 in Eq. 11, we obtain the SRP map in Fig. 3d which successfully locates the source. Fig. 3e shows the resulting SRP map using randomly selected 6 pairs without making a strategic selection.

In cases where one of the proposed measures give a less satisfactory localization result, orientational diversity in Sec. III-C can be employed to further improve the microphone selection. For instance, in Fig. 4a an SRP map after microphone pair selection using sparsity measure 1 is shown which misses the correct source position by almost 1 meter. Then, orientational diversity is employed to the selected microphone pairs and a better localization result is obtained using 2 pairs as seen in Fig. 4b.

Fig. 5 shows the average localization errors for 30 realizations of AWGN added to the female speech with SNRs of $[0, 3, 6, 9]$ dB, for the conventional SRP, SRP by using 6

selected microphone pairs according to the broadside pairs selection method, sparsity measure 1 and 2, and 6 randomly selected microphone pairs as control. The average errors after applying orientational diversity to the pairs obtained by the proposed methods and selecting the best 2 pairs accordingly are also presented. Here, it should be highlighted that for the simulations in which broadside pair selection or sparsity measures already give good results, using this approach make them worse as we now use only 2 pairs. Yet still we obtain a better localization performance than random pair selection. The mean absolute error between estimated and true locations is used as a performance metric. As expected, the conventional SRP approach gives the best localization performance as it uses all available microphones. Compared with the conventional SRP approach the proposed approaches show acceptable errors with the potential for significantly reduced data transmission and distributed computation. The computation of the SRP map is ~ 46 times faster with 6 selected microphone pairs than the computation of the conventional SRP while using all possible microphone pair combinations. All three proposed measures demonstrate a similar level of performance using only 6 microphone pairs, while sparsity measure 2 slightly outperforms the others overall. Broadside pairs selection performs the best in higher SNRs. Substantial localization error observed in the randomized control clearly demonstrates that utilizing meaningful microphone pair selection methods such as the ones proposed is essential.

V. Conclusion

In this work, we proposed several microphone pair selection methods that facilitate the selection of microphone pairs for SSL in a conference room setup in which there are many possible microphone pairs. These methods can be used separately or in conjunction with each other. Our simulation results show that several sensibly designed methods for microphone pair selection achieve a localization performance close to the solution that uses all the microphones, but with the potential for significantly reduced data transmission and distributed computation. All of the proposed methods outperform the randomized microphone pair selection. Microphone pair selection techniques like the ones we propose open up the opportunity for reduced complexity schemes and are particularly relevant for systems with a large number of microphones.

VI. Acknowledgment

The authors would like to extend their acknowledgements to Televic for their time and energy spent on this project.

References

[1] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, pp. 1–24, 2017.

[2] A. Marti, M. Cobos, J. J. Lopez, and J. Escolano, "A steered response power iterative method for high-accuracy acoustic source localization," *J. Acoust. Soc. of Amer.*, vol. 134, no. 4, pp. 2627–2630, 2013.

[3] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. on Speech Audio Process.*, vol. 12, no. 5, pp. 499–508, 2004.

[4] 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 *Proc. 2007 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2007), Honolulu, HI, USA*, vol. 1. IEEE, 2007, pp. I–121.

[5] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, 2007.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoust., Speech, and Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[7] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Ph.D. thesis, Brown University, Providence, Rhode Island, USA, 2000.

[8] C. Zhang, D. Florêncio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?" in *Proc. 2008 IEEE Int. Conf. on Acoust., Speech Signal Process. Las Vegas, NV, USA*, 2008, pp. 2565–2568.

[9] H. Do and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *Proc. 2007 IEEE Workshop Appl. Signal Process. Audio, Acoust. New Paltz, NY, USA*, 2007, pp. 295–298.

[10] B. Çakmak, T. Dietzen, R. Ali, P. Naylor, and T. van Waterschoot, "A distributed steered response power approach to source localization in wireless acoustic sensor networks," in *Proc. 2022 Int. Workshop Acoust. Signal Enhancement (IWAENC), Bamberg, Germany*, 2022, pp. 1–5.

[11] Y. Huang, J. Tong, X. Hu, and M. Bao, "A robust steered response power localization method for wireless acoustic sensor networks in an outdoor environment," *Sensors*, vol. 21, no. 5, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/5/1591

[12] G. Dekkers, F. Rosas, S. Lauwereins, S. Rajendran, S. Pollin, B. Vanrumste, T. van Waterschoot, M. Verhelst, and P. Karsmakers, "A multi-layered energy consumption model for smart wireless acoustic sensor networks," 2018.

[13] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. IEEE 18th Symp. Comm. Veh. Techn. (SCVT), Ghent, Belgium*, 2011, pp. 1–6.

[14] J. Zhang, S. P. Chepuri, R. Hendriks, and R. Heusdens, "Microphone subset selection for MVDR beamformer based noise reduction," *IEEE/ACM Trans. on Audio, Speech, Language Process.*, 2017.

[15] J. Szurley, A. Bertrand, M. Moonen, P. Ruckebusch, and I. Moerman, "Energy aware greedy subset selection for speech enhancement in wireless acoustic sensor networks," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO), Bucharest, Romania*, 2012, pp. 789–793.

[16] T. C. Lawin-Ore and S. Doclo, "Reference microphone selection for MWF-based noise reduction using distributed microphone arrays," in *Proc. 10th ITG Symp. Speech Communic., Braunschweig, Germany*, 2012, pp. 1–4.

[17] T. Dietzen, E. De Sena, and T. van Waterschoot, "Low-complexity steered response power mapping based on nyquist-shannon sampling," in *Proc. 2021 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA),New Paltz, New York, USA*, 2021, pp. 206–210.

[18] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting a geometrically sampled grid in the steered response power algorithm for localization improvement," *J. Acoust. Soc. of Amer.*, vol. 141, no. 1, p. 586–601, 2017.

[19] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2001.

[20] N. Hurley and S. Rickard, "Comparing measures of sparsity," in *2008 IEEE Workshop on Machine Learning for Signal Processing*, 2008, pp. 55–60.

[21] A. W. H. Khong and P. A. Naylor, "Efficient use of sparse adaptive filters," in *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, 2006, pp. 1375–1379.

[22] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

[23] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs," *Current Opinion in Neurobiology*, vol. 14, no. 4, pp. 481–487, 2004.

[24] Bang and Olufsen, "Music for archimedes," Compact Disc B&O, 1992.

[25] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Trans. on Audio, Speech, Language Process.*, vol. 23, no. 4, pp. 774–786, 2015.