TOWARDS FULLY UNCALIBRATED ROOM RECONSTRUCTION WITH SOUND

Marco Crocco, Andrea Trucco, Vittorio Murino and Alessio Del Bue

Pattern Analysis and Computer Vision (PAVIS)
Istituto Italiano di Tecnologia (IIT)
Via Morego 30, 16163 Genova, Italy

ABSTRACT

This paper presents a novel approach for room reconstruction using unknown sound signals generated in different locations of the environment. The approach is very general, that is fully uncalibrated, i.e. the locations of microphones, sound events and room reflectors are not known a priori. We show that, even if this problem implies a highly non-linear cost function, it is still possible to provide a solution close to the global minimum. Synthetic experiments show the proposed optimization framework can achieve reasonable results even in the presence of signal noise.

Index Terms— Room reconstruction, microphone calibration, source localization, simulated annealing

1. INTRODUCTION

Sensing the shape of a room is a problem that has attracted increasing attention from the research community. In part this is due to the complexity of the task in which the position of the walls has to be found by analysing only a set of acoustic events registered by microphones. These being, in the most blind scenario, unknown signals generated from unknown sources with an arbitrary position. The other practical aspect is that room reconstruction is an enabling technology for ubiquitous localisation with the simple use of microphones such as the ones in mobile phones and other consumer products. Such attractiveness has although clashed against the intrinsic complexity of the problem. Current solutions often need custom hardware requirements or even constraints that make the applicability of each method subject to the specific setup or to limiting assumptions.

On the contrary, we are dealing with the general optimization problem for room reconstruction where each source generates an unknown sound (not impulsive) which is acquired by a set of microphones deployed randomly in an unknown indoor area. The solution of such optimization is the 3D metric positions of the microphones, sources and the room wall positions. Crucially, the cost function derived from this problem is non-convex, highly non-linear and with several ambiguous solutions. Moreover, the problem in such general form requires the solution of four different problems that formally have been treated separately: dereverberation, microphone positions estimation, source localization and room calibration. For this reason, most of the approaches were devised to solve, or to consider solved, a subset of these four problems. Actually, almost all the works dealing with room geometry reconstruction [1–5] use a priori known emitted signals, typically impulsive or with a high bandwidth-time product (e.g., linear frequency sine sweeps) and matched filters in reception, so allowing an easy estimation of the delays by simply looking at the temporal peaks of the received signals. This of course implies the use of additional equipment and the need to accurately measure the microphone and loudspeaker impulse responses. Another common assumption is knowing a priori the positions of the microphones and/or the acoustic sources [1–6]. In addition, many methods require specific microphone arrays or source arrangements to avoid ambiguities in the order of arrival of the reflections [1–4]. The work of Tervo and Korhonen [6] can be considered as the closest approach to ours since it employs continuous and unknown signals. However, they limit the approach to a single reflective surface with known microphone and sound source positions.

Differently, this paper shows that it is possible to obtain a solution even if the resulting cost function is strongly non-linear and characterized by a high number of variables. The devised strategy is based on a bootstrap approach. First, the estimation of the delays of arrival related to the walls is decoupled from the geometry of the problem. This provides a double advantage: the non-linearity of the problem is reduced and the delay search results in a set of smaller independent problems, one for each real source. This allows to also provide a solution to the dereverberation problem and to reconstruct the original signals. Second, the shorter delays, corresponding to the direct paths from the real sources to the microphones, are used to initialise the microphone and real source positions. Then, walls positions are estimated, taking into account a subset of delays found at the first stage, specifically the ones not subject to ambiguities of reconstruction. Finally, once all the ambiguities are solved, a refinement procedure reconsider all the delays estimated initially.

The remaining of the paper first presents in Sec. 2 the room calibration model and define the general optimization problem. Then we explain in detail the step of the optimiza-

tion algorithms for the signal in Sec. 3 and the geometric related parts in Sec. 4. Experimental validation is given in Sec. 5 while future work is discussed in Sec. 6.

2. PROBLEM STATEMENT

Let us first consider the setup for the uncalibrated room reconstruction problem. As previously stated, the only input available are a set of N sound signals (not impulsive) recorded by a set of M microphones in a room with K reflective planes (i.e. walls, floor and ceiling). Each sound source n with $n = 1 \dots N$ generates a signal $y_m^n(tT_s)$ (T_s being the sampling period) that is received by a microphone m with $m = 1 \dots M$ positioned in a 3D space. The signal $y_m^n(tT_s)$ is given by the sum of $K+1$ propagating signals generated by a single source n where K are given by the image sources corresponding to the reflections from the K planar surfaces. No additional knowledge is assumed on the sound sources: they are in general different from each other, not synchronized with the acquisition system (the time of emission is unknown), and nothing is known about the signal statistics. The only hypothesis is that the real sound sources are not overlapped in time, i.e. overlaps occur only between each real source signal and the corresponding reflections [1].

In more detail, define $x_{mk}^n(tT_s)$ for $t = 1 \dots T$ and $m = 1 \dots M$ as the sampled signal received by the m-th sensor and generated by the k-th source (regardless of being real or image sources). Each component $x_{mk}^n(tT_s)$ received by a sensor can be seen as a delayed and scaled version of the signal generated by the real source with a delay depending on the reflector and sensor position $x_{mk}^n(tT_s) = a_{m,k}^n x^n(tT_s - \tau_{mk}^n)$ where $x^n(tT_s)$ is the signal generated by the real source n, $a_{m,k}^n$ and $\tau_{m,k}^n$ are respectively the relative amplitude and the delay of the n-th signal received by the m-th sensor and coming from the k-th source. We consider the planar surfaces as perfect acoustic mirrors with a frequency independent attenuation coefficient. The signal received at each sensor can be expressed as the sum of the $K+1$ components such that:

$$y_m^n(tT_s) = \sum_k a_{m,k}^n x^n(tT_s - \tau_{mk}^n). \quad (1)$$

In practice Eq. (1) is the expression for the convolution of the transmitted signal with the $K+1$-sparse room impulse response (RIR) and recovering the original signal $x^n(t)$ corresponds in solving a dereverberation problem. However, each delay τ_{mk}^n is given by the specific room configuration and the positions of microphones and sound events. Notice that each planar reflector has its own image source $\mathbf{p}_k^n$ defined as:

$$\mathbf{p}_k^n = \mathbf{b}^n + 2\left(1 - \frac{\mathbf{r}_k^T \mathbf{b}^n}{\|\mathbf{r}_k\|^2}\right)\mathbf{r}_k \quad (2)$$

[1] This hypothesis is assumed by all the room reconstruction methods operating with audio signals and can be easily fulfilled in a real situation (e.g. sources can be realized by different people speaking or hand-clapping in sequence, or even a single person moving around in the room)

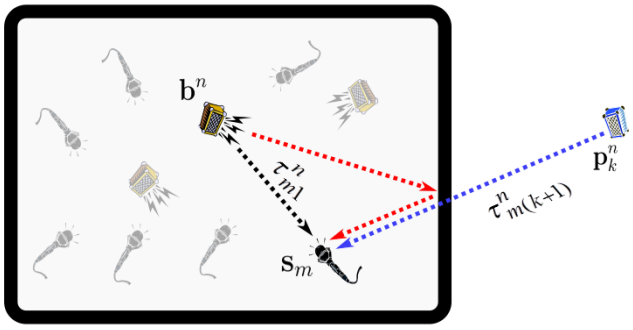

Fig. 1. The figure shows the real source $\mathbf{b}^n$ together with the image source $\mathbf{p}_k^n$. The image source is positioned such that the sound path in red equal the blue path coming from $\mathbf{p}_k^n$

where each of the vectors $\mathbf{b}^n$, of size 3, represents the 3D position of the $n-th$ source, and the normal vector $\mathbf{r}_k$ defines the orientation and distance from the origin $\mathbf{0}$ of the k-th planar reflector (as shown in Fig. 1). Thus we have the real propagation delay τ_{m1}^n and the image propagation delay τ_{mk}^n with $k = 2 \dots K+1$ defined as:

$$\tau_{m1}^n = \|\mathbf{b}^n - \mathbf{s}_m\|/c \qquad \tau_{m(k+1)}^n = \|\mathbf{p}_k^n - \mathbf{s}_m\|/c \quad (3)$$

where the vector $\mathbf{s}_m$ is the 3D position of the m-th microphone; c is the sound propagation speed.

By combining Eqs. (1), (2) and (3) we obtain the final cost function which combines both the signal based terms (convolution of signals) and the geometrical one given by the microphones, events and room configuration. Such function is highly non-linear and finding the right solution among many local minima is not possible if not close to the basin of attraction of the global minimum. However, we show here that is possible to achieve a reasonable solution by dividing the problem in different but yet linked components. Figure 2 shows a schematic representation of our approach. First, a signal processing stage estimates the original signal together with the delays and amplitudes given by the reflected components. Such delays are then used to bootstrap a geometrical optimization procedure that does delays sorting and associations together with a local initialization of the microphones and events positions.

3. SIGNAL AND DELAYS OPTIMIZATION

We now describe the signal and delays optimization that will estimate $x^n(t)$, a_{mk}^n and τ_{mk}^n from the input signals $y_m^n(t)$, generated by the microphones. Equation (1) can be reformulated in the frequency domain as follows by applying a DFT such that:

$$Y_m^n(f) = \sum_k a_{mk}^n X^n(f) e^{-j2\pi f \tau_{mk}^n f_s} \quad (4)$$

where $Y_m^n(f)$ and $X^n(f)$ are the DFTs of $y_m^n(tT_s)$ and $x^n(tT_s)$ respectively, f_s is the sampling frequency normalized by the number of temporal samples T and $f =$

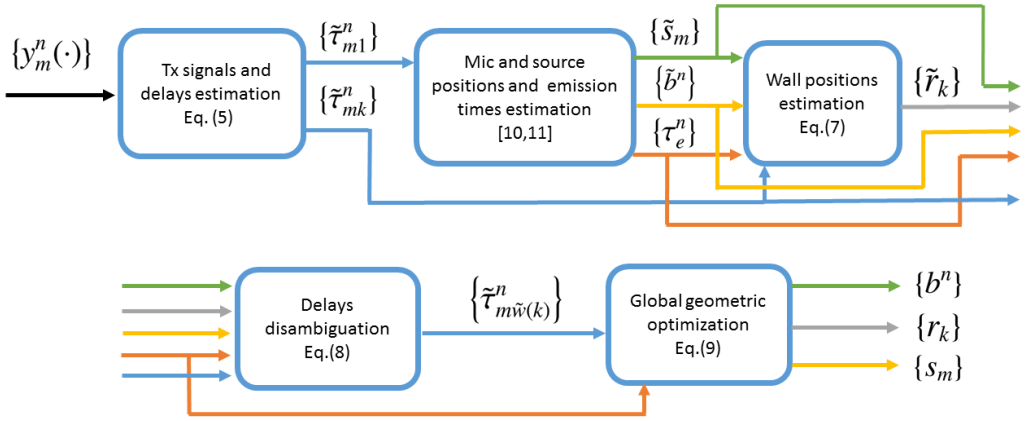

Fig. 2. Diagram of the overall method. Each box corresponds to a computational step as implemented by our approach. The figure also presents corresponding input (audio signals) and outputs (mic/sound source position, the transmitted signal, delays and room walls position).

$0 \ldots T-1$. The frequency domain representation of the model brings two advantages. Firstly, the delayed versions of the signal $x^n(t)$ are exactly represented by its spectrum plus a scalar parameter τ_{mk}^n, whereas in time domain each delayed signal has different values of the time samples (unless τ_{mk}^n is an exact multiple of the sampling period) so resulting in an increased number of variables whose mutual dependencies are not simple to model. Secondly, it is possible to retain only a fraction of the frequency bins, e.g. the ones with the highest SNR. Therefore, given the MN spectra $Y_m^n(f)$, the problem of recovering all the delays of arrival related to the virtual and real sources can be recast as N indepedent nonlinear least squares problem as follows:

$$\underset{\tau_{mk}^n, a_{mk}^n, X^n(f)}{\text{minimise}} \sum_{m,f} \left(Y_m^n(f) - \sum_k a_{mk}^n X^n(f) e^{-j2\pi f \tau_{mk}^n f_s} \right)^2 \tag{5}$$

For known delays and amplitudes, the spectrum can be found in closed-form. Defining $\mathbf{y}^n(f)$ and $\mathbf{z}^n(f)$ as the M-vectors given by $Y_m^n(f)$ and $\sum_k a_{mk}^n exp\,(-j2\pi f \tau_{mk}^n f_s)$ respectively, the estimated frequency bins of $X^n(f)$ are given by:

$$X_{est}^n(f) = \mathbf{z}^+(f)\mathbf{y}^n(f), \tag{6}$$

where $\mathbf{z}^+(f)$ is the pseudo-inverse of $\mathbf{z}^n(f)$. If only propagation delays are known, the cost function in (5) implies a bilinear form and, if also the delays are unknown, the problem becomes even harder. As delays appear in the argument of complex exponentials, the cost function is strongly nonlinear, yielding to many local minima. For this reason, gradient descent alike methods might easily be trapped in local minima thus it is necessary to adopt stochastic minimization procedures such as Simulated Annealing (SA) [7]. Besides its general ability in dealing with local minima, SA has already demonstrated its convergence performance in other problems where variables are arguments of complex exponentials [8,9]. In brief, SA is an iterative procedure aimed at minimizing the energy function $J(\mathbf{v})$, where $\mathbf{v}$ is the vector of the state variables. At each iteration, a small random perturbation is induced in the current state configuration $\mathbf{v}_i$, where i is the iteration. If the new configuration, $\mathbf{v}^*$, causes the value of the energy function to decrease, then it is accepted. If, instead, $\mathbf{v}^*$ causes the value of the energy function to increase, it is accepted with a probability dependent on the system temperature, a parameter that is gradually lowered along with the iterations. In our case we adapted SA using delays and amplitudes as the state variable vector, whereas the transmitted signal spectrum is computed at each iteration in closed form, according to (6).

4. GEOMETRIC OPTIMIZATION

The estimated delays are used to infer the room and sensors geometry. However the geometric optimization problem has to face three main issues. First, the delays are estimated given an unknown time offset representing the time of emission. Second, the order of arrival of the reflections is different at each microphone and for each source, making difficult to match the delays with the corresponding walls. Finally, the estimated delays can be subject to ambiguities, whenever two or more delays are equal[2]. For instance, consider the case $K+1=3$ in which $\tau_1 = \tau_2 \neq \tau_3$ (n and m indexes dropped for simplicity): if the estimated delays and amplitudes (denoted with $\tilde{}$) are set as $\tilde{\tau}_1 = \tau_1$, $\tilde{\tau}_2 = \tau_3$, $\tilde{\tau}_3 = \tau_3$, $\tilde{a}_1 = a_1 + a_2$, $\tilde{a}_2 = \tilde{a}_3 = a_2/2$ the value of the cost function

[2] Although it might seem rare, the coincidence of delays is a quite common effect in most room reconstruction scenarios.

(5) does not change despite the delay estimation is clearly wrong. The problem is present also if two delays are close each other (not exactly equal) because (5) will have local minima, very similar to the correct mininimum, for each arbitrary couple of close delays. To solve for this crucial problem we propose the following strategy.

First, we sort in ascending order the estimated delays for each microphone and source and pick up the lower one which corresponds to the direct path delay. Having collected this MN delays, we apply the sensor localization algorithms [10, 11] which allow to recover an initial microphone and sources 3D positions ($\tilde{\mathbf{s}}_m$ and $\tilde{\mathbf{b}}^n$ respectively) and the times of emission τ_e^n related to each source n. Given such initialization, the walls position can be found exploiting a subset of delays from the previous stage and in particular the ones which do not hold ambiguities problems. This is done by a pruning strategy that removes a signal (n, m) if a pair of the delays associated to that signal is closer than a given threshold. Notice that, thanks to the multiple sources employed, after the pruning stage the data are sufficient for solving the following stages. In any case, if a data starving situation appears, it is possible to lower the threshold to fetch back delays that were previously removed. A non-linear Least Squares cost function is then defined as:

$$\min_{\mathbf{r}_k} \sum_{nm} I\left(n,m\right) \sum_{k} \left(\tilde{\tau}^n_{mh_1(k)} - \tau^n_e - \tau\left(\tilde{\mathbf{b}}^n, \tilde{\mathbf{s}}_m, \mathbf{r}_{h_2(k)}\right)\right)^2 \tag{7}$$

where the indicator function $I(n, m)$ is zero or one according to the fact that the corresponding set of delays τ^n_{mk} has been pruned or not. Instead, the index functions $h_1(k)$ and $h_2(k)$ sort the two set of delays $\tilde{\tau}$ and τ in ascending in order (for n and m fixed). This solves the matching problem between walls and delays since, for the right configuration of walls, the two sets of delays are equal. Given the non-linearity of the cost function in respect to the walls position, SA is used again to solve for (7).

At this stage we have two sets of estimated delays: the ones obtained from the signal optimization stage $\tilde{\tau}^n_{mk}$, close to the correct ones τ^n_{mk}, but suffering from overlap ambiguities, and the ones produced by the current geometric solution $\tau(\tilde{\mathbf{b}}^n, \tilde{\mathbf{s}}_m, \tilde{\mathbf{r}}_k)$ free of ambiguities but more unprecise due to the errors added in the geometric reconstruction step carried out by SA. We can now solve for the ambiguities by a nearest neighbour approach between the two sets of delays. In detail, by defining the following relation $\tilde{w}(k)$ between indexes:

$$\tilde{w}(k) = arg \min_{w} \left(\left| \tilde{\tau}^n_{mw} - \tau^n_e - \tau(\tilde{\mathbf{b}}^n, \tilde{\mathbf{s}}_m, \tilde{\mathbf{r}}_k) \right| \right) \tag{8}$$

we can write the set of ambiguity-free delays as $\tilde{\tau}^n_{m\tilde{w}(k)}$ for $k = 1 \ldots K+1$. The procedure can by further clarified looking at Fig. 3 Once ambiguities have been removed we can now employ the whole set of estimated delays for a final geometric refinement that jointly optimizes walls, microphones and sources positions. Given such initialization, we are likely to be in the basin of attraction of the global minimum, a gradient based descent can be adopted to optimize the following cost function:

$$\underset{\mathbf{b}^n, \mathbf{s}_m, \mathbf{r}_k}{\text{minimise}} \sum_{nmk} \left(\tilde{\tau}^n_{m\tilde{w}(k)} - \tau^n_e - \tau(\mathbf{b}^n, \mathbf{s}_m, \mathbf{r}_k) \right)^2 \tag{9}$$

where the quantities $\tilde{\mathbf{b}}^n$,$\tilde{\mathbf{s}}_m$, $\tilde{\mathbf{r}}_k$ are set as initialization values. The overall procedure is resumed in the scheme as shown in Fig. 2 with the inputs and outputs of each stage in our method. Considering the difficulty of the problem, in our model we neglected higher order reflections from the walls. Nevertheless, if necessary, the proposed method can be extended in order to handle such reflections, simply increasing the number of delays to be estimated in (4) and modifying the image source model (2), employed in the geometric reconstruction, according to equations given in [3].

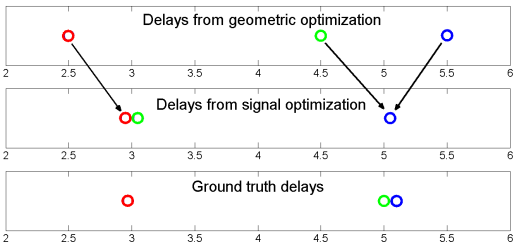

Fig. 3. Example of delay disambiguation by a NN approach. Each circle denotes a delay whose value is given by the x-axis. Each ambiguity free delay calculated from geometry is associated to the closest delay estimated from the signals.

5. EXPERIMENTS

To assess the proposed method we run a set of synthetic experiments. Since, to the best of our knowledge, no literature method is able to work in the unconstrained conditions set above, no comparative analysis is possible. Experiments are mainly aimed at verifying the overall feasibility of the problem. A rectangular room with sides of 7.5 m, 6.5 m and height 5.5 m has been filled with 10 microphones and 12 sources deployed in random positions. Each source has been generated filtering a white noise of 0.1 s between 50 and 1000 Hz. An example of source signal and corresponding signal acquired from a microphone is shown in Fig. 4. It can be seen that the replicas are completely overlapped with the original signal. Reflection amplitudes have been randomly generated according to a uniform distribution between 0.1 and 1. An explicit modeling of amplitudes depending on wall reflection coefficients and wall areas is not trivial and was therefore left for future investigations. The threshold for considering two delays as overlapped is set to 0.0002 s corresponding to 6.8 cm.

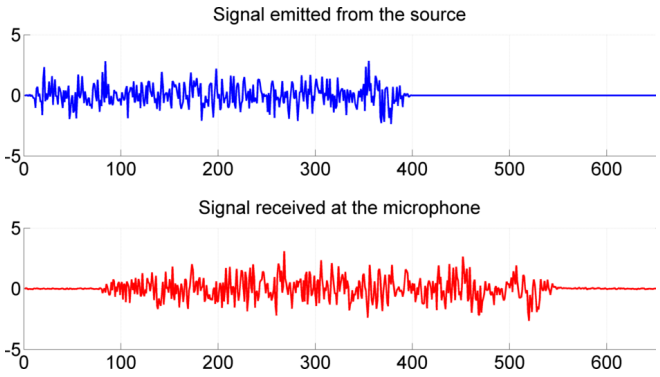

Fig. 4. Example of emitted signal from a generic source (left), and related acquired signals at a generic microphone (right).

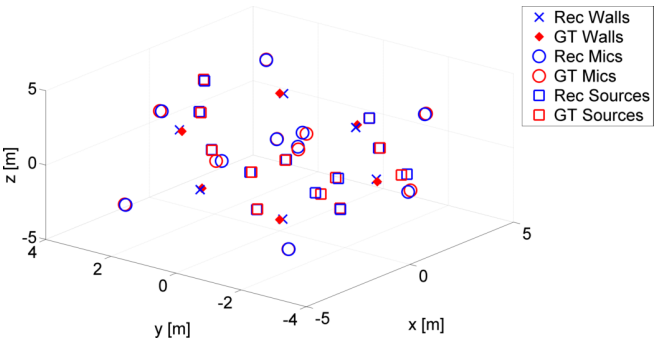

Fig. 5. Ground truth and reconstruction of walls, microphones and sources positions yielded by the proposed algorithm.

A Gaussian noise with a StD of 0.01 and 0.025, corresponding respectively to an SNR of 40 dB and 32 dB, has been added to the signals acquired by the microphones. The whole computation took a few hours on a common PC working with Matlab code.In Fig. 5 the ground truth and reconstructed walls, microphones and sources are displayed for the case with 0.01. One can see to qualitatively good reconstruction of the whole structure. To quantify the error of reconstruction we apply Procrustes analysis and evaluate the RMS value of the distance between the ground truth and the estimated values, obtaining an error of 0.15 m for the walls, 0.086 m for the microphones and 0.082 m for the sources. If normalized with respect to the maximum room side they correspond to about 2% (walls) and 1.1% (microphones and sources). The higher error for the walls is probably due to the outer positions of the virtual sources with respect to microphones and real sources. An analogous reconstruction with 0.025 noise gives the following RMS errors: 0.42 m for the walls, 0.23 m for the microphones and 0.33 m for the sources.

6. CONCLUSIONS

We have presented an approach for uncalibrated room reconstruction that provides a solution when no information a priori is known about sensors, events and room walls position, shape and time of emission of the transmitted signals and amplitudes of walls reflections. Despite the complexity of the problem, involving huge number of variables and thorny cost functions with several local minima, we demonstrated its feasibility. The use of natural sounds allows to multiply the number of sources without expense, making possible the pruning strategy aimed at solving delay ambiguities. The results are qualitatively correct even if there is space for improving the precision of reconstruction. Toward this direction, the obtained solution could be used as starting guess for a final global minimization involving both signal and geometry parameters. Future work will explore the extension of the method to higher order wall reflections and will extend experimentation to real environments.

7. REFERENCES

[1] J. Filos, A. Canclini, F. Antonacci, A. Sarti, and P. A. Naylor, "Localization of planar acoustic reflectors from the combination of linear estimates," in *Sig. Proc. Conf. (EUSIPCO), 2012 Proc. of the 20th European*. IEEE, 2012, pp. 1019–1023.

[2] Sakari Tervo and Timo Tossavainen, "3d room geometry estimation from measured impulse responses," in *Acoustic, Speech and Sig. Proc. (ICASSP), 2012 IEEE Int. Conf. on*. IEEE, 2012, pp. 513–516.

[3] F. Ribeiro, D. Florêncio, D. Ba, and C. Zhang, "Geometrically constrained room modeling with compact microphone arrays," *Audio, Speech, and Lang. Proc., IEEE Trans. on*, vol. 20, no. 5, pp. 1449–1460, 2012.

[4] F. Antonacci, A. Sarti, and S. Tubaro, "Geometric reconstruction of the environment from its response to multiple acoustic emissions," in *Acoust. Speech and Sig. Proc. (ICASSP), 2010 IEEE Int. Conf. on*. IEEE, 2010, pp. 2822–2825.

[5] I. Dokmanić, R. Parhizkar, A. Walther, and M. Lu, Y. M .and Vetterli, "Acoustic echoes reveal room shape," *Proc. of the Nat. Academy of Sciences*, 2013.

[6] S. Tervo and T. Korhonen, "Estimation of reflective surfaces from continuous signals," in *Acoust. Speech and Sig. Proc. (ICASSP), 2010 IEEE Int. Conf. on*. IEEE, 2010, pp. 153–156.

[7] S. Kirkpatrick, D. Gelatt Jr., and M. P. Vecchi, "Optimization by simulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[8] V. Murino, A. Trucco, and C.S. Regazzoni, "Synthesis of unequally spaced arrays by simulated annealing," *Signal Proc., IEEE Trans. on*, vol. 44, no. 1, pp. 119–122, 1996.

[9] M. Crocco. and A. Trucco, "Stochastic and analytic optimization of sparse aperiodic arrays and broadband beamformers with robust superdirective patterns," *Audio, Speech, and Lang. Proc., IEEE Trans. on*, vol. 20, no. 9, pp. 2433–2447, 2012.

[10] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *Acoust., Speech and Sig. Proc. (ICASSP), 2013 IEEE Int. Conf. on*, 2013, pp. 106–110.

[11] M. Crocco, A. Del Bue, and V. Murino, "A bilinear approach to the position self-calibration of multiple sensors," *IEEE Trans. on Sig. Proc.*, vol. 60, pp. 660–673, 2012.