ROOM REFLECTORS ESTIMATION FROM SOUND BY GREEDY ITERATIVE APPROACH

Marco Crocco[†]*, Andrea Trucco*[*] *and Alessio Del Bue*[†]

[†]Visual Geometry and Modelling (VGM) Lab, Istituto Italiano di Tecnologia (IIT)
Via Morego 30, 16163 Genova, Italy
[*]DITEN - University of Genova, Via Opera Pia 11, 16145 Genova, Italy

ABSTRACT

Room reconstruction from sound is the problem of estimating indoor space boundaries given only a set of sound events acquired by an array of microphones. In order to find a solution in realistic scenarios, it is necessary a robust and practical method that can solve the echo labelling problem, i.e. assigning at each signal delay the correct reflector that has generated it. Although being an NP-hard problem, in this paper we demonstrate that it is possible to solve the echo labelling problem in a reasonable computational time without the need of additional hypotheses on the echoes order of arrival.

Index Terms— Echo Labelling problem, NP-hard, Room Reconstruction, Structure from Sound.

I. INTRODUCTION

Whenever a sound is emitted in an enclosed space, the signal received at the microphones is a sum of a set of attenuated time-delayed replicas given by the room walls reflections. The problem at the basis of room reconstruction can be briefly stated as the estimation of the position and orientation of planar room reflectors only from the time-delays in the arrival of the emitted signal at a set of microphones. Such time-delays, once extracted from the signal [1], [2], [3], [4], [5], [6], have to be associated to the correct "source" of the delay: The first being related to the direct path of the sound wave, the other related to the reflections by the walls.

However, the identity of the wall that has generated the delay is unknown at the receivers so leading to a NP-hard assignment problem, called the Echo Labelling Problem (ELP). This task is aggravated by the presence of missing time delays, noise in the temporal localisation of the delays and spurious delays extracted from the signal, due for example to scattering from objects and non-rigid surfaces [7]. To this end, a robust solution to ELP is essential to obtain correct results in any room reconstruction method. ELP has been studied extensively in the literature but few approaches [8], [9], [10] are able to provide results in realistic scenarios as tested in this work.

In practice, planar surface positions can be estimated by an exhaustive search over a discretized grid in the space of all possible configurations, evaluating a proper cost function parametrized by microphone and source positions and signal peaks. In this way, the intrinsic ambiguity problem arising from the unknown matching between signal peaks and planar surfaces in ELP is bypassed. However, as the 3D space of configurations is of dimension three multiplied by the number of planar surfaces, a naive exhaustive search would be computationally hard.

For this reason we adopt a greedy iterative procedure, the core of our approach, in which the search is decoupled for each planar surface, decreasing the space of solution to 3 for each wall search. To do this, we initially did not consider second order reflections since they imply dependencies between different planar surfaces. Moreover, at each iteration, we prune out peaks matched to an already estimated planar surface to simplify the search over subsequent surfaces. Finally we adopt a robust cost function that allows to cope with missing peaks or spurious peaks due to not perfect functioning of the peak finder stage. Through this procedure it is possible to make feasible the solution of this problem in realistic environments which are currently impractical for most approaches in the literature.

II. RELATED WORKS

Several works have attempted to solve ELP but most of them related to synthetic scenarios only. In [9] the signals at the microphones are modelled as a sparse combination of signals related to every possible plane. The method requires to measure or simulate such dictionary of signals, a time consuming procedure. The ELP boils down by the requirement of microphone array compactness but this limits the method applicability. In some works [11], [12], [13], [10], [14], [15], [16], [17] reflectors are modelled as planes tangent to the ellipsoids with foci given by each pair of microphone/source.

Antonacci et al. [11] use a clustering procedure of Times Of Flights (TOFs) based on the Hough transform to solve ELP. However, this method requires a very specific setup with just one microphone and a source moving on a perfect circle around the microphone. In [10], the source is moved very close to a reflector at each acquisition, such that the second TOF is surely related to the same reflector for

all the microphones. In [12] a brute force search is run over all the possible TOF combinations. Once a reflector is estimated, the corresponding TOFs are iteratively discarded: The approach becomes feasible given the low number of microphones/sources and by the fact that just first order TOFs are assumed to be detected. In [17] the use of a compact microphone array strongly limits the possible permutations in echoes order of arrival. In [8] a rank criterion is applied to a matrix derived from pairwise distances between microphones and between microphones and virtual sources, in order to check if the current set of selected TOFs belong to a single reflector. The method results in a combinatorial search, unless heuristics on microphone array compactness are applied. Moreover the method is sensitive to outliers.

Other works [18], [19] do not deal efficiently with ELP, requiring an exhaustive combinatorial search. In a previous work of the authors [20], a stochastic approach, based on Simulated Annealing was able to bypass ELP by associating at each iteration the TOF computed from the tentative reflector configuration to the TOF from the signals using a nearest neighbour procedure. Even if a pruning stage was devised to discard ambiguous TOFs, the method did not have a specific strategy to deal with outliers and missing data.

III. ELP BY GREEDY ITERATIVE APPROACH

Consider N sources and M microphones, with positions given by the 3-vectors $\mathbf{b}^n$, $n = 1, \ldots, N$ and $\mathbf{s}_m$ with $m = 1, \ldots, M$, enclosed in a room defined by a convex polyhedron with a known number of faces K. The 3D positions of these K planar surfaces are defined by the 3-vectors $\mathbf{r}_k$, $k = 1, \ldots, K$, normal to the surfaces and with modulus equal to their distance from the 3D coordinate center. According to the image model [21] the reflection from a planar surface k can by approximated by a virtual source emitting the same signal as the real source n, whose position $\mathbf{p}_k^n$ is specular with respect to the plane $\mathbf{r}_k$ as:

$$\mathbf{p}_k^n = IM(\mathbf{b}^n, \mathbf{r}_k) = \mathbf{b}^n + 2\left(1 - \frac{\mathbf{r}_k^\top \mathbf{b}^n}{\|\mathbf{r}_k\|_2^2}\right)\mathbf{r}_k. \quad (1)$$

The signal reflected from a surface can be reflected by other surfaces, yielding second order reflections At most $K(K-1)$ second order virtual sources can be identified. Their location $\mathbf{p}_{k_1k_2}^n$ for $k_1, k_2 = 1, \ldots, K$ and $k_1 \neq k_2$ is given by $\mathbf{p}_{k_1k_2}^n = IM(\mathbf{p}_{k_1}^n, \mathbf{r}_{k_2})$. Higher order reflections are here considered as noise, together with scattering from objects in the room and self noise of the acquisition chain. The TOFs between real or virtual sources and microphones can be expressed as:

$$\tau_{m0}^n = \|\mathbf{b}^n - \mathbf{s}_m\|_2 / c. \quad (2)$$

$$\tau_{mk}^n = \|\mathbf{p}_k^n - \mathbf{s}_m\|_2 / c, \qquad \tau_{mk_1k_2}^n = \|\mathbf{p}_{k_1k_2}^n - \mathbf{s}_m\|_2 / c. \quad (3)$$

with c being the sound velocity.

From the set of NM acquired signals we have that the times of arrival (TOA), given by the sum of TOF plus emission times, can be estimated by a peak finding procedure, as described in [22]. From the TOA corresponding to direct path, i.e. the first peaks in time, an iterative method by Gaubitch et al. [23], based on bilinear decomposition, can be adopted to estimate the positions of microphones and real sources $\tilde{\mathbf{b}}^n$, $\tilde{\mathbf{s}}_m$ as well as the emission times. Hence, the estimated TOFs can be recovered for both real and virtual sources, by subtracting emission times to TOAs. The TOFs for each microphone n and each source m can be defined as $\mathcal{T}_{m0}^n = \left\{\tilde{\tau}_{m1}^n, \tilde{\tau}_{m2}^n, \ldots, \tilde{\tau}_{mL(n,m)}^n\right\}$ of length $L(n,m)$. Notice that $L(n,m)$ does not match exactly the number of first and second order virtual sources since the peak finding algorithm could detected spurious peaks or even miss true peaks.

A straightforward method to find the correct reflector position is to build a cost function given by the sum of the absolute differences of each TOF, function of the position reflector through Eq. (1), (2), (3) and each TOF estimated from the signal. In the following we will call these two TOFs as the GTOF (geometry TOF) and STOF (signal TOF). Unfortunately, there are three important issues: 1) The nonlinear nature of Eq. (1) leading to possible local minima; 2) the presence of spurious peaks or missing peaks in the estimated TOF from signal; 3) ELP, i.e. the unknown association between TOFs from signals and related reflectors.

A brute force approach to bypass the above three issues would be to discretize the space of reflector positions and perform a systematic search, checking if the resulting GTOFs correctly match the STOFs. In this case the problems of labeling, missing data and spurious peaks is solved simply by considering, for each GTOF, the nearest STOF, and applying opportune robust cost functions to limit the effect of missing and spurious peaks. Moreover, the problem of local minima is implicitly solved by exhaustive grid search over all the possible solutions. Unfortunately, this procedure is computationally infeasible due to dimensionality of the search space equal to $3K$.

However, if we initially consider in the model only the first order reflections, treating higher order ones in the data as noise, we can decouple the search for each single reflector, decreasing in this way the dimensionality of the search space to 3. Then, once the first reflector position is estimated, the image sources associated to it can be exploited to infer the subsequent reflectors positions. This results in a **greedy iterative algorithm**, where at each iteration a new reflector position is estimated and second order reflections are incrementally added to the data.

In detail, we discretize the 3D euclidean space of possible locations of the planar surfaces in a 3D cartesian grid $\mathbf{r}_i^{grid} = (x_i, y_i, z_i)$ with $i = 1, \ldots, I$. The grid boundaries are set according to a coarse guess of the dimension of the room and the grid spacing is given by the required precision

and the computational resources available. Now, rename for convenience the estimated real source positions $\tilde{\mathbf{b}}^n$ with $\tilde{\mathbf{p}}_0^n$. Moreover, consider the index $j = 1, \ldots, K$ as the iteration index of the algorithm and $k = 0, \ldots, j-1$ as the index of image sources (and real ones for $k = 0$) that will be be progressively added along with the iterations. Consider the first iteration for which $j = 1$ and $k = 0$. For a given real source n, a microphone m and a tentative reflector position $\mathbf{r}_i^{grid}$, the time of flight $\tau_{mk}^{geom,n}(i)$ computed from the geometry of the problem is given by:

$$\tau_{mk}^{geom,n}(i) = \|\tilde{\mathbf{s}}_m - IM(\tilde{\mathbf{p}}_k^n, \mathbf{r}_i^{grid})\|_2 / c. \tag{4}$$

For such TOF, we search for the index $\tilde{l}$ of the closest TOF estimated from the signal, with a nearest neighbours (NN) approach:

$$\tilde{l}(n, m, k, i) = NN(\mathcal{T}_{mj}^n, \tau_{mk}^{geom,n}(i)). \tag{5}$$

Now, we need a score function to evaluate the goodness of the fitting between the two delays $\tau_{mk}^{geom,n}(i)$ and $\tilde{\tau}_{m,\tilde{l}(n,m,k,i)}^n$. Such score should be rapidly decaying whenever the two delays are too distant, since in this case the guessed reflector position is probably wrong. At the same time the score should be robust to missing TOFs: if the peak finder fails to detect a peak from the signal, the nearest neighbour procedure identifies a matching within arbitrarily distant TOFs. An empirically suitable score function $S_{mk}^n(i)$ is the following:

$$S_{mk}^n(i) = exp\left(-\left(\tau_{mk}^{geom,n} - \tilde{\tau}_{m,\tilde{l}(n,m,k,i)}^n\right)^2 / (2\sigma^2)\right) + \epsilon. \tag{6}$$

where σ rules the rate of decay of the score function and should be set according to the accuracy of the peak finder error; ϵ imposes a lower bound on the score function, making it robust to missing TOFs. The total score function $S^{tot}(i)$ is given, empirically, by the product of all the score functions related to the couples of microphones and real sources as:

$$S^{tot}(i) = \prod_{n=1}^{N} \prod_{m=1}^{M} \prod_{k=0}^{j-1} S_{mk}^n(i). \tag{7}$$

Recall that at first iteration $j = 1$ and $k = 0$ the last product structure is not relevant. Finally we search by a brute force approach for the reflector position yielding the maximum score:

$$\tilde{i} = \arg\min_i S^{tot}(i), \qquad \tilde{\mathbf{r}}_j = \mathbf{r}_{\tilde{i}}^{grid} \tag{8}$$

with $j = 1$. In this way we have found the first estimated reflector position $\tilde{\mathbf{r}}_1$. In order to perform the search of the second reflector, starting iteration $j = 2$, we remove from the set of STOFs all the STOFs that are likely related to the first reflector, otherwise the search procedure would fall again in the same global maximum already found for the first reflector. To do this we subtract to each set of STOFs $\mathcal{T}_{m,j}^n$ the set of STOFs already matched with the GTOFs calculated from the first reflector $\mathcal{T}_m^{matched,n}$, as follows:

$$\mathcal{T}_{m,j+1}^n = \mathcal{T}_{m,j}^n \setminus \mathcal{T}_m^{matched,n}, \tag{9}$$

where $\mathcal{T}_m^{matched,n}$ is defined as the set of STOFs for which at least a GTOF is distant less than a predefined threshold thr[1]:

$$\mathcal{T}_m^{matched,n} = \left\{\tilde{\tau}_{ml}^n \mid \left(\exists k \mid |\tau_{m,k}^{geom,n}(\tilde{i}) - \tilde{\tau}_{ml}^n| < thr\right)\right\} \tag{10}$$

The threshold thr is necessary to avoid the undesired removal of STOFs $\tilde{\tau}_{ml}^n$ too distant from the corresponding GTOFs, likely due to missing data. Finally, once the first reflector $\tilde{\mathbf{r}}_1$ has been estimated, we can consider the corresponding n image sources $\tilde{\mathbf{p}}_1^n$ for the search of the second reflector. This means to consider the second order reflections between the already estimated reflector and the next one. Thus, for each of the pair n, m, one match will be sought for the real source $\tilde{\mathbf{p}}_0^n$ and one for the virtual source $\tilde{\mathbf{p}}_1^n$. Notice that all the previous formulas are still valid by simply increasing the iteration index j from 1 to 2. By repeating the procedure, the second reflector is estimated and the new image sources are added, making the estimation more and more robust. Actually, the decrease in STOFs quality with the iterations[2] is compensated by the increased amount of TOFs available. Finally, the set of estimated planar reflectors is used as a starting guess for an off-the-grid refined solution, by solving a nonlinear Least Squares problem[3], minimizing the sum of squared differences of geometry TOF and signal TOF, function of microphones, sources and walls positions[4].

IV. REAL EXPERIMENTS

Real experiments were performed in a rectangular room with a vaulted ceiling of size 8.5m ×7.5m ×7m, located in a 16^{th} century mansion in Genova (Fig.1(a)). We displaced 12 omnidirectional Lavalier microphones in an area of about 3m ×3m at the center of the room, with heights ranging from from 20cm to 2m. Then we used a speaker of about 3cm diameter (VEHO360), moved in 17 different locations and the transmitted signal was a chirp of length 5s and frequency sweep from 3 kHz to 6 kHz. An average SNR of 10 dB was measured, due mainly to traffic in the street nearby.

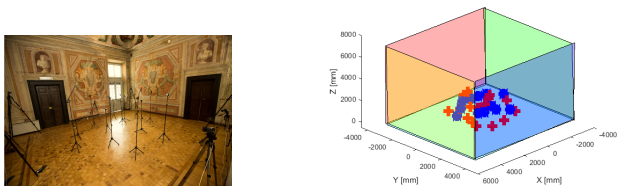
In order to acquire precise ground truth of room boundaries, we employed a Leica C10 laser scanner. Notice that the vaulted ceiling and the niches corresponding to the windows present relevant differences from the assumed piecewise planar model (Fig.2). Moreover a VICON system provided

[1]Notice that more than one STOFs could be included in $\mathcal{T}_m^{matched,n}$ due to spurious peaks

[2]The reflectors characterized by less missing TOFs are estimated at the first iterations and their sTOFs are removed from the sTOFs set.

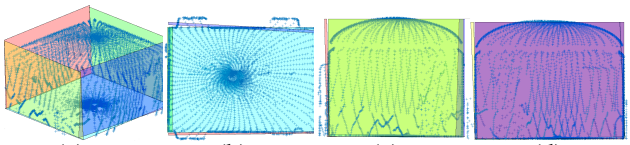
[3]For details please refer to [22].

[4]Notice that ELP has been already solved and this allows to discard spurious signal TOF, and geometry TOF for which the corresponding signal TOF are missing.



(a) (b)

Fig. 1. (a) View of the room for experimental tests; (b) ground truth data: 3D displacement of microphones (stars) and sources (crosses) from VICON system, and planar surfaces (fitted from point cloud given by the laser scanner).



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

Fig. 2. Four views of subsampled 3D point cloud from laser scanner and fitted ground truth planes related to the four walls, ceiling and floor. (a): assonometric view of the room; (b), (c) and (d): top and two lateral views.

the ground truth for microphone and sound events positions. Ground truth data[5], enclosing microphones and sources 3D positions and planes fitted to the 3D subsampled point cloud are displayed in Fig.1(b). Microphones, real sources and emission times were estimated by the method of [23] and used as input to the proposed method. We set parameters σ, ϵ and thr to $1.47 \cdot 10^{-4}$ s, 0.1 s and $5.88 \cdot 10^{-4}$ respectively. The grid step was set to 0.2 m. These values were selected according to physical considerations, such as the transmitted wavelengths, and trade off between computational load and desired accuracy. We compared our method with two recent approaches for room geometry estimation [8], [19]. Differently from our method, both algorithms use the information of just one source. In addition, their combinatorial nature does not allow to exploit 12 microphones in reasonable computation times. To allow a comparison as fair as possible, we run both the algorithms 17 times, one for each source. For each of the run, we selected the subset of five [19][6] or six [8] microphones that gave the best match between estimated and ground truth delays. Obviously, in a real situation ground truth delays are not known, however our aim was to test the two competing algorithms at their best. Finally, we collected all the estimated planes from the 17 trials and clustered them in six groups representing the reflectors. We then pruned out all the reflectors that were too distant from ground truth planes given by the dataset. Notice that also this step would require a preliminary knowledge of the room. As it can be seen from Table I both methods are not able to estimate the ceiling, while the method of [19] misses also the third wall. Moreover the accuracy of estimation is on average significantly worse with respect to our method In particular, the average distance error is 24.6 mm for our method, 87.5 mm for [19] and 80.6 mm for [8], while the average angle error is 1.01 degrees for our method, 1.41 degrees for [19] and 2.13 degrees for [8]. Overall, our method shows remarkable accuracy for floor and the four walls and a qualitatively correct estimation of the ceiling. The considerable departure of the vaulted ceiling from the planar reflector model accounts for the difference in estimation accuracy. Table II reports the actual computation time, the number of microphones adopted and the computational complexity of the walls estimation procedure with respect to number of sources, microphones and average number of detected peaks per acquisition. In terms of computational complexity, the proposed method grows linearly with the number of microphones, sources and average estimated peaks, due in particular to nearest neighbour search in Eq.5.

[5]Dataset with ground truth can be freely downloaded at: vgm.iit.it/datasets/3d-room-reconstruction-with-sound

[6][19] is actually formulated to work with 5 microphones only.

	DE Our	**DE [19]**	**DE [8]**
wall1	**22 mm**	150 mm	100 mm
wall2	**34 mm**	100 mm	60 mm
wall3	**62 mm**	-	143 mm
wall4	**4 mm**	70 mm	80 mm
floor	**1 mm**	30 mm	20 mm
ceiling	**1824 mm**	-	-
	AE Our	**AE [19]**	**AE [8]**
wall1	**0.89 deg**	1.92 deg	1.60 deg
wall2	**1.72 deg**	1.91 deg	2.13 deg
wall3	**0.63 deg**	-	4.32 deg
wall4	**0.43 deg**	0.76 deg	1.23 deg
floor	1.40 deg	1.18 deg	**1.01 deg**
ceiling	**2.98** deg	-	-

Table I. Distance error (DE) and Angle error (AE) between ground truth and estimated planes according to the proposed method (Our), the method of [19] and the method of [8].

Our	**[19]**	**[8]**
300 seconds	120 seconds	5 hours
12 microphones	5 microphones	6 microphones
$O(NML)$	$> O(NL^{(M-1)})$	$> O(NL^{M})$

Table II. First row: computation time for the real experiment. Second row: actual number of microphones employed. Third row: computational complexity, where N, M, and L are respectively the number of sources, microphones and the average number of extracted peaks per acquisition.

V. CONCLUSION

We presented a robust room geometry estimation method able to efficiently solve ELP, thus allowing to increase the number of microphones and sources involved used in thi problem and consequently improving the accuracy of reconstruction, while keeping reasonable computation times. Experiments on a real environment witness the advantages of the method with respect to recent state-of-art.

VI. REFERENCES

[1] L. T., G. X., and T. Kailath, "Blind identification and equalization based on second-order statistics: a time domain approach," *Inf. Theory, IEEE Trans. on*, vol. 40, no. 2, pp. 340–349, Mar 1994.

[2] M. Crocco and A. Del Bue, "Room impulse response estimation by iterative weighted l1-norm," in *EUSIPCO'15*, 2015.

[3] ——, "Estimation of TDOA for room reflections by iterative weighted l1," in *ICASSP16*, 2016.

[4] Y. Lin, J. Chen, Y. Kim, and D. D. Lee, "Blind channel identification for speech dereverberation using l1-norm sparse learning," in *Advances in Neural Information Processing Systems*, 2007, pp. 921–928.

[5] L. Yuanqing, C. Jingdong, K. Youngmoo, and D. Lee, "Blind sparse-nonnegative (bsn) channel identification for acoustic time-difference-of-arrival estimation," in *WASPAA'07*, Oct 2007, pp. 106–109.

[6] K. Kowalczyk, E. Habets, W. Kellermann, and P. Naylor, "Blind system identification using sparse learning for tdoa estimation of room reflections," *Sig. Proc. Letters, IEEE*, vol. 20, no. 7, pp. 653–656, July 2013.

[7] T. Nowakowski, N. Bertin, R. Gribonval, J. De Rosny, and L. Daudet, "Membrane shape and boundary conditions estimation using eigenmode decomposition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE Int. Conf. on*. IEEE, 2016, pp. 3336–3340.

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

[9] F. Ribeiro, D. Florêncio, D. Ba, and C. Zhang, "Geometrically constrained room modeling with compact microphone arrays," *IEEE TASLP*, vol. 20, no. 5, pp. 1449–1460, 2012.

[10] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE TASLP*, vol. 20, no. 10, pp. 2683–2695, 2012.

[11] F. Antonacci, A. Sarti, and S. Tubaro, "Geometric reconstruction of the environment from its response to multiple acoustic emissions," in *ICASSP'10*. IEEE, 2010, pp. 2822–2825.

[12] J. Filos, E. Habets, and P. Naylor, "A two-step approach to blindly infer room geometries," in *Proc. Int.. Workshop Acoust. Echo Noise Control (IWAENC), Tel Aviv, Israel*, 2010.

[13] J. Filos, A. Canclini, M. Thomas, F. Antonacci, A. Sarti, and P. A. Naylor, "Robust inference of room geometry from acoustic measurements using the hough transform," in *EUSIPCO'11*. IEEE, 2011, pp. 161–165.

[14] L. Remaggi, P. J. Jackson, P. Coleman, and W. Wang, "Room boundary estimation from acoustic room impulse responses," in *Sensor Sig. Proc. for Defense (SSPD), 2014*. IEEE, 2014, pp. 1–5.

[15] L. Remaggi, P. J. B. Jackson, W. Wang, and J. A. Chambers, "A 3d model for room boundary estimation," in *ICASSP'15*, April 2015, pp. 514–518.

[16] L. Remaggi, P. J. Jackson, and P. Coleman, "Source, sensor and reflector position estimation from acoustical room impulse responses," *ICSV22 - The 22th Int. Congress on Sound and Vibration*, 2015.

[17] L. Remaggi, P. J. Jackson, P. Coleman, and W. Wang, "Acoustic reflector localization: Novel image source reversion and direct localization methods," *IEEE/ACM TASLP*, vol. 25, no. 2, pp. 296–309, 2017.

[18] T. Wang, F. Peng, and B. Chen, "First order echo based room shape recovery using a single mobile device," in *ICASSP'16*, March 2016, pp. 21–25.

[19] T. Rajapaksha, X. Qiu, E. Cheng, and I. Burnett, "Geometrical room geometry estimation from room impulse responses," in *ICASSP'16*, March 2016, pp. 331–335.

[20] M. Crocco, A. Trucco, V. Murino, and A. Del Bue, "Towards fully uncalibrated room reconstruction with sound," in *EUSIPCO'14*, 2014.

[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *JASA*, vol. 65, no. 4, pp. 943–950, 1979.

[22] M. Crocco, A. Trucco, and A. Del Bue, "Uncalibrated 3d room geometry estimation from sound impulse responses," *Journal of the Franklin Institute*, vol. 354, no. 18, pp. 8678–8709, 2017.

[23] N. Gaubitch, B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *ICASSP'13*, 2013, pp. 106–110.