L1 REGULARIZED ROOM MODELING WITH COMPACT MICROPHONE ARRAYS

Demba Ba[1]*, Flávio Ribeiro*[2]*, Cha Zhang*[3]*, Dinei Florêncio*[3]

[1] Dept. of Electrical Eng. and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 02139
[2] Electronic Systems Engineering Department, Universidade de São Paulo, Brazil
[3] Microsoft Research, One Microsoft Way, Redmond, WA, 98052

ABSTRACT

Acoustic room modeling has several applications. Recent results using large microphone arrays show good performance, and are helpful in many applications. For example, when designing a better acoustic treatment for a concert hall, these large arrays can be used to help map the acoustic environment and aid in the design. However, in real-time applications – including de-reverberation, sound source localization, speech enhancement and 3D audio – it is desirable to model the room with existing small arrays and existing loudspeakers. In this paper we propose a novel room modeling algorithm, which uses a constrained room model and ℓ_1-regularized least-squares to achieve good estimation of room geometry. We present experimental results on both real and synthetic data.

Index Terms— Shoebox room modeling, wall discrimination, circular microphone array, l1-constrained least squares.

1. INTRODUCTION

The problem of extracting 3D models from real world measurements has been an active area of research for decades, particularly in the areas of machine vision, remote sensing and robotics [1]. Popular and effective methods involve using passive or active sensors to obtain high resolution maps from which 3D models can be extracted. Passive techniques can infer spatial information from shading, edges, texture, or other features in one or more images. Active methods work by illuminating a given region with structured light or laser light. While these techniques are quite effective for extracting visual information, they don't offer any information regarding sound reflection characteristics. To determine reflection coefficients, one must measure audio, and little has been published about audio 3D modeling. This is understandable, since sound possesses much longer wavelengths than light, which limits its resolution, and brings about near field effects which degrade performance even further.

Due to the difficulties associated with sound, room acoustics analysis and design is often made by physical measurements, followed by material and propagation modeling [2]. Nevertheless, interest in such problems has apparently increased in recent years. [3] uses MVDR beamforming with a single ultrasound transmitter/receiver pair mounted on a precision 2D positioning system to perform ultrasound imaging in air, with which the position and outline of obstacles can be determined. [4] uses a 32-microphone spherical array to visualize the location of sound reflections in concert halls. [5, 6] use a single microphone and either a moving source on a circular trajectory or multiple sources to estimate the coordinates of reflectors.

In this paper we consider the problem of fitting a six-wall room model to a 3D enclosure based on data recorded by an array of M microphones, by reproducing a known signal from a source at the center of the array. This approach is quite convenient, since it is compact, self contained, does not have moving parts, does not require multiple sources and estimates reflection coefficients for frequencies in the audible range, which allows them to be used in applications involving speech capture and enhancement. In essence, our proposal involves estimating the impulse responses from the array loudspeaker to each of the array's microphones, and then extracting the wall positions and distances from this set of impulse responses.

There are numerous applications even for such a simple room model. It can increase robustness in MVDR arrays by improving the desired signal manifold estimates (instead of directly estimating room transfer functions as in [7]), improve 3D sound spatialization by incorporating more accurate room models [8], help initialize acoustic echo cancelation algorithms, assist in tracking environment changes, and help alleviate the drawbacks of reverberation in many algorithms. More impressively, in [9] we show that applying this model to sound source localization can yield better results than for state-of-the-art algorithms [10, 11] in non-reverberant rooms.

This paper is organized as follows: Section 2 gives an overview of the problem and the main assumptions under consideration. Section 3 presents the mathematical details and approximations behind the room estimation method. Section 4 shows experimental results on both real and synthetic data, and Section 5 presents some of our conclusions and future work.

2. PROBLEM STATEMENT

We want to obtain a room model which could be used to predict the way sound propagates inside a room. We do not need to perfectly predict room propagation, as long as we can help explain at least part of the sound behavior. Indeed, real rooms are potentially complex environments. Yet, in sampling a few conference rooms in corporate environments, we find that almost every room has four walls, a ceiling and a floor; the floor is leveled and the ceiling parallel to the floor; walls are vertical, straight, and extend from floor to ceiling and from adjoining wall to adjoining wall. Carpet is common, and almost invariably there is a conference table in the center of the room. Furthermore, many objects that seem visually important are small enough that may actually be acoustically transparent for most frequencies of interest. Based on these observations, we adopt a simple room model: four walls and a ceiling.

Even with such a simplified room model, it would be hard to passively estimate the components of the model based solely on unknown signals already existing in the room. Instead, we follow the same approach as [4, 5, 6] and actively probe the room by emitting a known signal (e.g., a sweep) from a known location (e.g., a loudspeaker co-located with the array). For the purposes of this discussion, we consider a uniform circular array with a speaker rigidly mounted in its center. This is the geometry used by the RoundTable

To appear in *Proc. ICASSP'2010, March 14-19, 2010, Dallas, TX, USA*

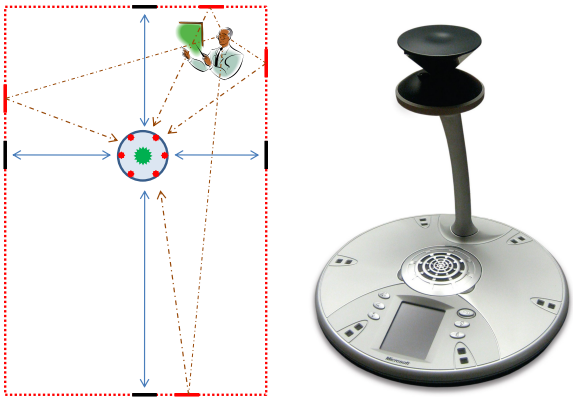

Fig. 1. Room model and RoundTable device

device depicted in Figure 1.

Note that, in contrast to previous work, we use a single sound source, fixed, and close to the microphones. This implies that we only sample each wall at one point: the point where the wall's normal vector points to the array. Depending on the application, we need to assume that the walls extend beyond the location at which they will be detected. Figure 1 illustrates the concept when using the proposed room model to do speech enhancement or sound source localization. The circular device in the center of the room (i.e., the RoundTable) will detect the reflections from the walls, indicated by the black segments in each of the four walls. However, the locations of interest for the walls are in fact the ones indicated by the red segments. The underlying assumption is that the walls extend linearly and with similar acoustic characteristics.

We consider the problem of fitting a five wall model to a 3-D enclosure based on data recorded by an array of M microphones, by reproducing a known signal such as a sine sweep from a source positioned at the center of the array. The room model is denoted $\mathcal{R} = \{(a_i, d_i, \theta_i, \phi_i)\}_{i=1}^{5}$, where the vector $(a_i, d_i, \theta_i, \phi_i)$ specifies respectively the reflection coefficient, distance, azimuth and elevation of the i^{th} wall with relation to a known coordinate system. We assume that the geometry of the array is fixed and known a priori.

The optimal manner in which to solve this problem would be a completely parametric approach, where $\mathcal{R}$ is estimated directly. However, there are two issues with this approach: (a) there is no straightforward functional relationship between $\mathcal{R}$ and the room impulse responses; (b) the estimation problem is a highly nonlinear one which suffers from the presence of multiple local extrema. We therefore resort to a non-parametric approach which assumes that early segments of impulse responses can be decomposed into a sum of isolated wall reflections.

3. ROOM MODELING

Without loss of generality, a spherical coordinate system (r, θ, ϕ) is defined such that r is the range, θ is the azimuth, ϕ is the elevation and $(0, 0, 0)$ is at the phase center of the array. We assume that the geometry of the array and loudspeaker is fixed and known a priori.

Define $h_m^{(r,\theta,\phi)}(n)$ as the discrete time impulse response from the loudspeaker to the m^{th} microphone, considering that: (1) the direct path from loudspeaker to the microphone has been removed and (2) the array is mounted on free space, except for the presence of a lossless, infinite wall with normal vector $\mathbf{n} = (r, \theta, \phi)$ and which contains the point (r, θ, ϕ). Let r be sufficiently large so that the wall does not intersect the array or offer significant near-field effects. We denote $h_m^{(r,\theta,\phi)}(n)$ a single wall impulse response (SWIR).

Our discrete time observation model is

$$y_m(n) = h_m(n) * s(n) + u_m(n), \tag{1}$$

where n is the sample index, m is the microphone index, $h_m(n)$ is the room's impulse response from the array center to the m^{th} microphone, $s(n)$ is the reproduced signal, and $u_m(n)$ is measurement noise. Given a persistently exciting signal $s(n)$, one can estimate the room impulse responses (RIRs) from the observations $y_m(n)$. It is from these estimates that we infer the geometry of the room.

We assume that the early reflections from an arbitrary RIR $h_m(n)$ may be approximately decomposed into a linear combination of the direct path and individual reflections, such that

$$h_m(n) \approx h_m^{(dp)}(n) + \sum_{i=1}^{R} \rho^{(i)} h_m^{(r_i,\theta_i,\phi_i)}(n) + v_m(n), \tag{2}$$

where $h_m^{(dp)}(n)$ is the direct path; R is the total number of modeled reflections; the superscript i is the reflection index; $h_m^{(r_i,\theta_i,\phi_i)}(n)$ is the SWIR from a perfectly reflective wall at position (r_i, θ_i, ϕ_i), and from which the direct path from the loudspeaker to the microphone has been removed; $\rho^{(i)}$ is the reflection coefficient (which we assume to be frequency invariant); $v_m(n)$ is noise and residual reflections not accounted in the summation.

Note that we assume that $\rho^{(i)}$ does not depend on m, and this claim deserves justification. While the reflection coefficient obviously depends on a wall and not on the array, it is conceivable (albeit unlikely) that the sound impinging on a pair of microphones could have reflected off different walls. However, for reasonably small arrays the sound will take approximately the same path from the source to each of the microphones, which implies that it should with high probability reflect off the same walls before reaching each microphone, such that the reflection coefficients will be the same for every microphone.

Now define

$$\begin{aligned}
\mathbf{x}_m &= \left[\, x_m(0) \;\cdots\; x_m(N) \,\right]^T \\
\mathbf{x} &= \left[\, \mathbf{x}_1^T \;\cdots\; \mathbf{x}_M^T \,\right]^T \\
\mathbf{x}_{m,\tau} &= \left[\, x_m(\tau) \;\cdots\; x_m(N+\tau) \,\right]^T \\
\mathbf{x}_\tau &= \left[\, \mathbf{x}_{1,\tau}^T \;\cdots\; \mathbf{x}_{M,\tau}^T \,\right]^T
\end{aligned}$$

for any signal $x_m(n)$ associated with the m^{th} microphone.

We can then rewrite (2) in truncated vector form as

$$\mathbf{h} \approx \mathbf{h}^{(dp)}(n) + \sum_{i=1}^{R} \rho^{(i)} \mathbf{h}^{(r_i,\theta_i,\phi_i)} + \mathbf{v}, \tag{3}$$

where we have selected a vector length N that is just large enough to contain the first order reflections, but that cuts off the higher order reflections and the reverberation tail. Therefore, given a measured $\mathbf{h}$, our problem is to estimate $\rho^{(i)}$ and (r_i, θ_i, ϕ_i) for the dominant 1st order reflections, which in turn should reveal the position of the closest walls and their reflection coefficients.

Our proposed method for room modeling first consists of obtaining synthetically and/or experimentally for the array of interest: (1) a set $\left\{\mathbf{h}^{(r_0,\theta,0)}\right\}_{\theta\in\mathcal{A}}$ of SWIRs, each measured at fixed range $r = r_0$ over a grid $\mathcal{A}$ of azimuth angles, and (2) the SWIR $\mathbf{h}^{(r_0,0,\pi/2)}$ con-

taining only the reflection from a ceiling at the same fixed range. We define

$$\mathcal{H} = \left\{ \mathbf{h}^{(r_0,\theta,0)} \right\}_{\theta \in \mathcal{A}} \cup \left\{ \mathbf{h}^{(r_0,0,\pi/2)} \right\}. \quad (4)$$

In essence, $\mathcal{H}$ carries a time-domain description of the array manifold vector for multiple directions of arrival. If we assume a far field approximation and a sufficiently high sampling rate, given an arbitrary $\mathbf{h}^{(r_*,\theta_*,\phi_*)}$ with $r_* > r_0$ we have that

$$\mathbf{h}^{(r_*,\theta_*,\phi_*)} \approx \frac{r_0}{r_*} \mathbf{h}_{\tau_*}^{(r_0,\theta_*,\phi_*)}, \quad (5)$$

for $\tau_* = [2\,(r_* - r_0)\,/c]$, where $[\cdot]$ denotes the nearest integer, and c is the speed of sound. Thus, $\mathbf{h}^{(r_0,\theta_*,\phi_*)}$ generates a family of reflections for a given direction. Since a room is essentially a linear system, if we assume that reflection coefficients are frequency-independent and neglect the direct path from loudspeaker to microphone, the 1st order reflections can always be expressed as a linear combination of time-shifted and attenuated SWIRs. Furthermore, if $\mathcal{A}$ is sufficiently fine, for a set of walls $\mathcal{W} = \{(r_i, \theta_i, \phi_i)\}_{i\in[1,W]}$ there are coefficients $\{c_i\}_{i\in[1,W]}$ such that given an impulse response $\mathbf{h}_{room}$, which had the direct path removed and was truncated as to only contain early reflections,

$$\mathbf{h}_{room} \approx \sum_{i\in[1,W]} c_i \mathbf{h}^{(r_0,\theta_i,\phi_i)}. \quad (6)$$

Thus, under the approximations above we can claim that the set of all delayed SWIRs approximately generates the space of truncated impulse responses over which we will make estimations. Define $\mathcal{H}_* = \{\mathbf{h}_\tau : \mathbf{h} \in \mathcal{H} \wedge 0 \le \tau \le T\}$, where T is the maximum delay we wish to model for a reflection. Our problem is then to fit elements $\mathcal{H}_*$ to the measured impulse response, adjusting for attenuation. A sparse solution is also required, given that we are interested in only a few major 1st order reflections, and that $\mathcal{H}_*$ will contain a very large number of candidate reflections.

Consider an enumeration of $\mathcal{H}$ such that $\mathcal{H} = \left\{ \mathbf{h}^{(1)}, ..., \mathbf{h}^{(K)} \right\}$, with $K = |\mathcal{H}|$. We define

$$\mathbf{H} = \left[\mathbf{h}_{\tau=0}^{(1)} \cdots \mathbf{h}_{\tau=T}^{(1)} \cdots \mathbf{h}_{\tau=0}^{(K)} \cdots \mathbf{h}_{\tau=T}^{(K)} \right], \quad (7)$$

where each single wall impulse response appears for each integer delay τ such that $0 \le \tau \le T$. We then solve the following ℓ_1-regularized least-squares problem [12]:

$$\min_{\mathbf{a}} \|\mathbf{h}_{room} - \mathbf{H}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1, \quad (8)$$

where λ controls the sparsity of the desired solution. Each coefficient in the solution indicates a reflection, and we must assume each reflection is from a different wall. Thus the need to use a sparsity-inducing penalty as the ℓ_1 norm. Without it, a typical minimum mean square solution will provide hundreds or thousands of small-valued reflections, instead of the few strong reflections corresponding to the wall candidates.

If we consider only SWIRs with coefficients $[\mathbf{a}]_i$ larger than a given threshold, then we have a set of candidate walls. A post-processing stage is necessary in order to only accept solutions which contain walls which make 90° angles to each other, and reject impossible solutions such as more than one ceiling or multiple walls at approximately the same direction.

A practical consideration involves the computational tractability of solving (8). It is desirable to have spatial resolutions on the order of 2 cm or better. Given the restriction of integer delays, this translates to having a sampling rate of 16 kHz or higher. If one wishes to identify walls located at 4 meters or less, one must plan for a round-trip time of around 350 samples, which implies allowing $0 \le \tau \le 350 = T$. The grid of single wall reflections should be sufficiently fine, otherwise walls will not be detected. We have sampled in azimuth with 4° resolution, resulting in 90 SWIRs. One SWIR for the ceiling is also necessary, giving $K = 90 + 1$. Therefore, $\mathbf{H}$ has $T \cdot K = 31850$ columns. Since impulse responses can be long, computational requirements for operating explicitly with $\mathbf{H}$ will typically be prohibitive.

In order to solve (8) following [12] one must implement the $\mathbf{H}\mathbf{x}$ and $\mathbf{H}^T\mathbf{y}$ operations for arbitrary vectors $\mathbf{x}$ and $\mathbf{y}$. Fortunately, it is possible to exploit $\mathbf{H}$'s block matrix nature in order to avoid representing $\mathbf{H}$ explicitly, and also to accelerate the matrix-vector product operations. Indeed, $\mathbf{H}$ has a block structure such that

$$\mathbf{H} = \left[\mathbf{H}^{(1)} \; \mathbf{H}^{(2)} \; \cdots \; \mathbf{H}^{(K)} \right], \quad (9)$$

where

$$\mathbf{H}^{(i)} = \left[\mathbf{h}_{\tau=0}^{(i)} \; \mathbf{h}_{\tau=1}^{(i)} \; \cdots \; \mathbf{h}_{\tau=T}^{(i)} \right]. \quad (10)$$

It is easy to see that for all i, $\mathbf{H}^{(i)}$ is Toeplitz. Therefore, $\mathbf{H}^{(i)}\mathbf{x} = \mathbf{h}_{\tau=0}^{(i)} * \mathbf{x}$, which can be implemented with a fast FFT-based convolution. It is easy to show that $\left[\mathbf{H}^{(i)}\right]^T \mathbf{y} = \mathbf{h}_{\tau=0}^{(i)} \star \mathbf{y}$ (where $\star$ denotes cross-correlation), which can also be evaluated with FFTs. Using this method, both matrix-vector products can be performed using K fast convolutions or fast correlations.

After solving (8) and post processing to reject invalid walls, one is left with a handful of wall coordinates and their associated coefficients $[\mathbf{a}]_i = \rho^{(i)} \cdot \frac{r_0}{r^{(i)}}$. It turns out that

$$r^{(i)} = r_0 + \mathrm{mod}\,(i-1, T)\,/\,(2f_s), \quad (11)$$

where f_s is the sampling rate. Thus we are able to estimate $\rho^{(i)}$. However, one must consider that the ℓ_1-regularized least-squares procedure is designed for producing sparse solutions. As such, it tends to underestimate coefficients, such that reflection coefficients obtained directly from solving (8) can be too small. To get better estimates of reflection coefficients, we gather only the $\mathbf{h}_{\tau=\tau_i}^{(i)}$ single wall responses corresponding to the identified walls and fit them to the measured impulse response using conventional least squares.

One final consideration must be made concerning how to pre-process impulse responses before solving (8). Individual single wall reflections tend to be very short, while the impulse response $\mathbf{h}_{room}$ is usually long, and contains many features other than the first reflections that one would wish to identify with greater precision. These features can be due to clutter, multiple reflections, bandpass responses from microphones or reflections from the table over which the array is set. In order to reduce these extraneous features, we perform soft thresholding on SWIRs and room RIRs, according to

$$\mathbf{h}_{thresh} = \mathrm{sign}\,(\mathbf{h}) \cdot \max\,(|\mathbf{h}| - \sigma, 0), \quad (12)$$

where σ determines the thresholding level and should be adjusted as a fraction of the signal's level. With soft thresholding, the RIR gains the appearance of a synthetic impulse response generated using the image method. The sparsity of the thresholded RIR lends well to the ℓ_1-constrained least squares procedure, both in running time and estimation precision.

Ground Truth				Estimates			
r	θ	ϕ	ρ	r	θ	ϕ	ρ
-1.00	0.0	90.0	0.77	1.00	0.0	90.0	0.73
2.00	0.0	90.0	0.77	2.00	0.0	90.0	0.65
4.00	0.0	0.0	0.77	4.00	0.0	0.0	0.68
1.50	90.0	0.0	0.77	1.50	92.0	0.0	0.71
3.00	180.0	0.0	0.77	3.00	-180.0	0.0	0.69
4.50	270.0	0.0	0.77	–	–	–	–

Table 1. Estimated walls for the synthetic room

4. EXPERIMENTAL RESULTS

Using the image model, we obtained 90 SWIRs for vertical walls at 4^o azimuth intervals and 1 ceiling SWIR, and zero padded them to allow for up to $T = 350$ integer sample delays. We simulated an array with dimensions matching the RoundTable array (see Figure 1), which is a 6 directional microphone, uniform circular array with a radius of 13.5 cm with a fixed sampling rate $f_s = 16kHz$.

A virtual room with dimensions $6 \times 7 \times 3$ m and $R_{60} = 250\,ms$ was simulated using the image method [13]. RIRs from the center of the array to all channels were extracted, and truncated to 450 samples. A ℓ_1-regularized least-squares problem with $\lambda = 10^{-2}$ was solved to determine candidate wall locations, and a post-processing stage was used to discard false candidates. The wall positions were estimated within 1 cm of their true position, and when the postprocessing stage was set to select the 5 dominant walls, the estimated reflection coefficients fell within 0.12 of their true value, which was 0.77 for all walls. When it was set to select the 3 dominant walls, the estimated reflection coefficients were exactly 0.77. The array's coordinates and estimation results are shown in Table 1.

Using the anechoic chamber at Microsoft Research and a real RoundTable device, we obtained 90 SWIRs for vertical walls and one ceiling SWIR by using a circular acrylic barrier measuring about 1 meter in diameter. Real impulse responses were collected in a conference room in the Microsoft campus with dimensions $5.30 \times 7.01 \times 2.77$ m. The array was placed on top of a conference room table which was about 0.8 m from the ground. Therefore, the distance to the ground could not be estimated. A 3-second linear sine sweep from 30 Hz to 8 kHz was played through the RoundTable's internal speaker, and recorded simultaneously by all 6 microphones. Impulse responses were then estimated by frequency domain division.

After inspecting the impulse responses, it became apparent that the RoundTable is not the ideal device to capture reflections coming from side walls. Indeed, its microphone enclosures give highest gain to signals arriving from the ceiling, and lowest gain to signals arriving directly from the sides. Additionally, the RoundTable loudspeaker is mounted facing upwards, such that its directivity is low to the sides. As a matter of fact, some secondary reflections from the ceiling and walls were being detected with better clarity than the primary reflections off the side walls. Unfortunately, detecting secondary reflections is less reliable, because they tend to appear together with many other reflections. Regardless, we could determine the location of the closest walls with good accuracy, which is sufficient to enhance algorithms such as SSL with an image model of the room. Real distances and estimates are presented in Table 2.

Note that the wall at $\theta = 0^\circ$ could not be estimated, while the wall at $\theta = 90^\circ$ was found at its exact distance. It turns out that the wall at $\theta = 90^\circ$ was completely covered by a whiteboard, which is quite reflective. Since both walls are approximately at the same distance to the RoundTable, their reflections arrived at approximately the same time, and the impulse responses were dominated by the reflection from the whiteboard. Finally, the wall at $\theta = 180^\circ$ could not be detected simply because it is too far away.

Ground Truth				Estimates			
r	θ	ϕ	ρ	r	θ	ϕ	ρ
1.98	0.0	90.0	?	1.98	0.0	90.0	0.70
2.52	0.0	0.0	?	–	–	–	–
2.49	90.0	0.0	?	2.49	88.0	0.0	0.99
4.49	180.0	0.0	?	–	–	–	–
2.81	270.0	0.0	?	2.78	272.0	0.0	0.72

Table 2. Estimated walls for conference room 1

5. CONCLUSION

We have presented a method capable of identifying wall distances, positions and reflection coefficients with a small microphone array and loudspeaker. This information has already shown useful in enhancing SSL [9] and 3D audio spatialization [8], and it can be expected to be useful in many acoustic signal processing applications, including beamforming, speech enhancement, and others.

Future enhancements of the room estimation algorithm involve better identifying higher order reflections, in order to work around device limitations such as seen with the RoundTable. In particular, we are currently acquiring a more complete dataset of reflection basis functions, which incorporate different elevations, in addition to the 0° and 90° we currently use.

6. REFERENCES

[1] F. Remondino and S. El-Hakim, "Image-based 3D modeling: a review," *The Photogrammetric Record*, vol. 21, no. 115, pp. 269–291, 2006.

[2] Y. Jing and N. Xiang, "On boundary conditions for the diffusion equation in room-acoustic prediction: theory, simulations, and experiments," *J. Acoust. Soc. Am*, vol. 123, no. 1, pp. 145–153, 2008.

[3] M. Moebus and A. Zoubir, "Three-dimensional ultrasound imaging in air using a 2D array on a fixed platform," in *Proc. of ICASSP*, 2007.

[4] A. O'Donovan, R. Duraiswami, and D. Zotkin, "Imaging concert hall acoustics using visual and audio cameras," in *Proc. of ICASSP*, 2008.

[5] F. Antonacci, A. Sarti, and S. Tubaro, "Geometric reconstruction of the environment from its response to multiple acoustic emissions," in *Proc. of WASPAA*, 2009.

[6] D. Aprea, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic reconstruction of the geometry of an environment through acquisition of a controlled emission," in *Proc. of EUSIPCO*, 2009.

[7] I. Papp, Z. Saric, and A Jovicic, "Adaptive microphone array for unknown desired speaker's transfer function," *J. Acoust. Soc. Am*, vol. 122, no. 2, pp. EL44–EL49, 2007.

[8] M. Song, C. Zhang, and D. Florencio, "Enhanced binaural loudspeaker audio system with room modeling," submitted.

[9] F. Ribeiro, D. Ba, C. Zhang, and D. Florencio, "Turning enemies into friends: using reflections to improve sound source localization," submitted.

[10] C. Zhang, Z. Zhang, and D. Florencio, "Maximum likelihood sound source localization for multiple directional microphones," in *Proc. of ICASSP*, 2007.

[11] Y. Rui, D. Florencio, W. Lam, and J. Su, "Sound source localization for circular arrays of directional microphones," in *Proc. of ICASSP*, 2005.

[12] S.J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l1-regularized least squares," *IEEE Journal of Selected Topics in Sig. Proc.*, vol. 1, no. 4, pp. 606–617, 2007.

[13] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am*, vol. 65, pp. 943–950, 1979.