L1 REGULARIZED ROOM MODELING WITH COMPACT MICROPHONE ARRAYS

Demba Ba1, Flávio Ribeiro2, Cha Zhang3, Dinei Florêncio3

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 $\ell_1$-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, 11-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...
does not intersect the array or offer significant near-field effects. We denote \( h_{m,n}^{(r,\theta,\phi)} (n) \) a single wall impulse response (SWIR).

Our discrete time observation model is

\[
y_m (n) = h_m (n) * s(n) + u_m (n),
\]

where \( n \) is the sample index, \( m \) is the microphone index, \( h_m (n) \) is \( n \)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

\[
h_m (n) \approx h_m^{(dp)} (n) + \sum_{i=1}^{R} \rho_i h_{n}^{(r_i,\theta_i,\phi_i)} (n) + v_m (n),
\]

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 \((r_i,\theta_i,\phi_i)\), 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

\[
\mathbf{x}_m = \begin{bmatrix} x_{m}(0) & \cdots & x_{m}(N) \end{bmatrix}^T
\]

\[
\mathbf{x} = \begin{bmatrix} \mathbf{x}_1^T & \cdots & \mathbf{x}_M^T \end{bmatrix}^T
\]

\[
\mathbf{x}_{m,r} = \begin{bmatrix} x_{m(r)} & \cdots & x_{m(N+\tau)} \end{bmatrix}^T
\]

\[
\mathbf{x}_r = \begin{bmatrix} \mathbf{x}_{1,r}^T & \cdots & \mathbf{x}_{R,r}^T \end{bmatrix}^T
\]

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

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

\[
\mathbf{h} \approx \mathbf{h}^{(dp)} (n) + \sum_{i=1}^{R} \rho_i \mathbf{h}^{(r_i,\theta_i,\phi_i)} + \mathbf{v},
\]

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 \((r_i,\theta_i,\phi_i)\) 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 \( \{ h_{(\theta,\phi)} \}_{\theta \in \mathcal{A}} \) of SWIRs, each measured at fixed range \( r = r_0 \) over a grid \( \mathcal{A} \) of azimuth angles, and (2) the SWIR \( h_{(\theta,\phi)} \) con-
containing only the reflection from a ceiling at the same fixed range. We define
\[
H = \left\{ h^{(r_0, \theta, \phi)} \right\}_{\theta \in \mathbb{A}} \cup \left\{ h^{(r_0, \frac{\pi}{2})} \right\},
\]
In essence, \(H\) 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 \(h^{(r, \theta, \phi)}\) with \(r > r_0\) we have that
\[
h^{(r, \theta, \phi)} = \frac{r_0}{r_h} h^{(r_0, \theta, \phi)},
\]
for \(r_h = \left[ (r - r_0) / c \right]\), where \([\cdot]\) denotes the nearest integer, and \(c\) is the speed of sound. Thus, \(h^{(r_0, \theta, \phi)}\) 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 \(\mathbb{A}\) is sufficiently fine, for a set of walls \(\mathcal{W} = \{(r_i, \theta_i, \phi_i)\}_{i=1}^{|\mathcal{W}|}\) there are coefficients \(c_{i}\) such that given an impulse response \(h_{\text{room}}\), which had the direct path removed and was truncated as to only contain early reflections,
\[
h_{\text{room}} \approx \sum_{i=1}^{|\mathcal{W}|} c_{i} h^{(r_0, \theta_i, \phi_i)}.
\]
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 \(\mathcal{H}_t = \{ h : h \in H \land 0 \leq \tau \leq T \}\), where \(T\) is the maximum delay we wish to model for a reflection. Our problem is then to fit elements \(h_{\mathcal{H}_t}\) 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 \(H_{\mathcal{H}}\) will contain a very large number of candidate reflections.

Consider an enumeration of \(\mathcal{H}\) such that \(\mathcal{H} = \{ h^{(1)}, ..., h^{(K)} \}\), with \(K = |\mathcal{H}|\). We define
\[
H = \left[ h^{(1)}_{\tau=0} \cdots h^{(1)}_{\tau=T} \cdots h^{(K)}_{\tau=0} \cdots h^{(K)}_{\tau=T} \right],
\]
where each single wall impulse response appears for each integer delay \(\tau\) such that \(0 \leq \tau \leq T\). We then solve the following \(\ell_1\)-regularized least-squares problem [12]:
\[
\min_{a} \| h_{\text{room}} - Ha \|_2^2 + \lambda \|a\|_1,
\]
where \(\lambda\) 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 \(\ell_1\) 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 \(|a|\) 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 \(90^\circ\) 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 \(0 \leq \tau \leq 350 = T\). 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, \(H\) has \(T \cdot K = 31850\) columns. Since impulse responses can be long, computational requirements for operating explicitly with \(H\) will typically be prohibitive.

In order to solve (8) following [12] one must implement the \(Hx\) and \(H^T y\) operations for arbitrary vectors \(x\) and \(y\). Fortunately, it is possible to exploit \(H\)‘s block matrix nature in order to avoid representing \(H\) explicitly, and also to accelerate the matrix-vector product operations. Indeed, \(H\) has a block structure such that
\[
H = \left[ H^{(1)} \ H^{(2)} \cdots H^{(K)} \right],
\]
where
\[
H^{(i)} = \left[ h_{\tau=0}^{(i)} \ h_{\tau=1}^{(i)} \cdots h_{\tau=T}^{(i)} \right].
\]
It is easy to see that for all \(i\), \(H^{(i)}\) is Toeplitz. Therefore, \(H^{(i)} x = h_{\tau=0}^{(i)} x\), which can be implemented with a fast FFT-based convolution. It is easy to show that \(H^{(i)} T y = h_{\tau=0}^{(i)} \ast y\) (where \(\ast\) 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 \(|a_i| = \rho^{(i)} \cdot r_{\pi, i}\). It turns out that
\[
r^{(i)} = r_0 + \text{mod}(i - 1, T) / (2f_s),
\]
where \(f_s\) is the sampling rate. Thus we are able to estimate \(\rho^{(i)}\). However, one must consider that the \(\ell_1\)-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 \(h^{(i)}_{\tau=0}\) 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 preprocess impulse responses before solving (8). Individual single wall reflections tend to be very short, while the impulse response \(h_{\text{room}}\) 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
\[
h_{\text{thresh}} = \text{sign}(h) \cdot \max(|h| - \sigma, 0),
\]
where \(\sigma\) 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 \(\ell_1\)-constrained least squares procedure, both in running time and estimation precision.
4. EXPERIMENTAL RESULTS

Using the image model, we obtained 90 SWIRs for vertical walls at 4° 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 \( \ell_1 \)-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 post-processing 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 m in diameter. Real impulse responses were collected in a conference room in the Microsoft campus with dimensions 5.30 × 7.01 × 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.

Table 1. Estimated walls for the synthetic room

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r )</td>
<td>( \theta )</td>
</tr>
<tr>
<td>-1.00</td>
<td>0.0</td>
</tr>
<tr>
<td>2.00</td>
<td>0.0</td>
</tr>
<tr>
<td>4.00</td>
<td>0.0</td>
</tr>
<tr>
<td>1.50</td>
<td>90.0</td>
</tr>
<tr>
<td>3.00</td>
<td>180.0</td>
</tr>
<tr>
<td>4.50</td>
<td>270.0</td>
</tr>
</tbody>
</table>

Table 2. Estimated walls for conference room 1

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r )</td>
<td>( \theta )</td>
</tr>
<tr>
<td>1.98</td>
<td>0.0</td>
</tr>
<tr>
<td>2.52</td>
<td>0.0</td>
</tr>
<tr>
<td>2.49</td>
<td>90.0</td>
</tr>
<tr>
<td>4.49</td>
<td>180.0</td>
</tr>
<tr>
<td>2.81</td>
<td>270.0</td>
</tr>
</tbody>
</table>

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