Using Reverberation to Improve Range and Elevation Discrimination for Small Array Sound Source Localization

Flavio Ribeiro, *Student Member, IEEE*, Cha Zhang, *Senior Member, IEEE*, Dinei A. Florêncio, *Senior Member, IEEE*, and Demba Elimane Ba

***Abstract*—Sound source localization (SSL) is an essential task in many applications involving speech capture and enhancement. As such, speaker localization with microphone arrays has received significant research attention. Nevertheless, existing SSL algorithms for small arrays still have two significant limitations: lack of range resolution, and accuracy degradation with increasing reverberation. The latter is natural and expected, given that strong reflections can have amplitudes similar to that of the direct signal, but different directions of arrival. Therefore, correctly modeling the room and compensating for the reflections should reduce the degradation due to reverberation. In this paper, we show a stronger result. If modeled correctly, early reflections can be used to provide more information about the source location than would have been available in an anechoic scenario. The modeling not only compensates for the reverberation, but also significantly increases resolution for range and elevation. Thus, we show that under certain conditions and limitations, reverberation can be used to *improve* SSL performance. Prior attempts to compensate for reverberation tried to model the room impulse response (RIR). However, RIRs change quickly with speaker position, and are nearly impossible to track accurately. Instead, we build a 3-D model of the room, which we use to predict early reflections, which are then incorporated into the SSL estimation. Simulation results with real and synthetic data show that even a simplistic room model is sufficient to produce significant improvements in range and elevation estimation, tasks which would be very difficult when relying only on direct path signal components.**

***Index Terms*—Array processing, circular microphone array, distance discrimination, image method, range estimation, sound source localization (SSL).**

I. Introduction

A major goal in speech research is the acquisition of high-quality audio without constraining users with devices such as close-talking microphones. Microphone arrays can be used in this regard, and are progressively gaining popularity in applications such as videoconferencing [1], smart rooms [2]–[4], and human–computer interaction [5], [6]. Unlike a single microphone, a microphone array can be electronically steered to emphasize a signal coming from a direction of interest and reject noise coming from other locations. Such spatial filtering techniques require knowledge of the location of the speaker, which must be known *a priori* or estimated.

A significant trend in human–computer interaction is the use of joint audio and video sensor arrays to acquire the user's environment. For example, a combination of video cameras can be used to record a panoramic view of a scene, capturing more detail than a single camera possibly could. Once again, it is typically necessary to identify regions of interest—for instance, the location of individuals in a conference room. For videoconferencing applications, speaker localization can be used to automatically determine which sections of the acquired panoramic frame should be transmitted to a remote location. Furthermore, the knowledge of the range to the speaker can be used to identify him, given a choice between two individuals located at approximately the same direction of arrival, but at different distances to the device. This information can then be used to zoom, focus, and align individual cameras.

The general problem of sound source localization (SSL) has been an active area of research for many years, and finds applications in most array processing algorithms. Several methods have been proposed over the previous decades with varying degrees of accuracy, noise robustness, and computational complexity. Most algorithms can be classified into four categories: beamformer steering [7], energy ratio estimation [8], subspace characterization [9], [10], and time difference of arrival (TDOA) estimation [1], [11]–[15]. Common to these techniques is the fact that performance decreases with increasing reverberation [16]. This can be readily explained, given that in typical indoor environments, early reflections can have amplitudes similar to that of the direct signal, but different directions of arrival. If not accounted for explicitly, they will interfere with the estimation.

Another characteristic of these algorithms when applied to small arrays is their emphasis on estimating only azimuth. Indeed, a practical array designed for offices or conference rooms can be expected to have a limited number of low cost microphones (typically between 4 and 8), relatively small dimensions (probably featuring an inter-element spacing of at most 15 cm) and a simple circular or linear geometry. Under these constraints, estimating elevation may be difficult, and estimating range with traditional methods is an almost impossible task (see Fig. 1).

Manuscript received October 29, 2009; revised March 01, 2010 and May 15, 2010; accepted May 15, 2010. Date of current version August 13, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tomohiro Nakatani.

F. Ribeiro is with the Electronic Systems Engineering Department, Escola Politécnica, Universidade de São Paulo, São Paulo-SP, 05508-900, Brazil (e-mail: fr@lps.usp.br).

C. Zhang and D. Florêncio are with Microsoft Research, Redmond, WA 98052 USA (e-mail: chazhang@microsoft.com; dinei@microsoft.com).

D. Ba is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: demba@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2052250

1558-7916/$26.00 © 2010 IEEE

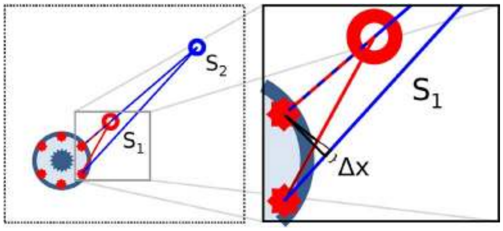


Fig. 1. Range discrimination problem for a six-element circular array. The ranges to sources S_1 and S_2 can be discriminated only by implicitly or explicitly estimating Δx, which corresponds to the difference between TDOAs. For compact arrays, Δx will be very small and its estimation will be very sensitive to noise and reverberation.

Given the small array constraint and a reverberant environment, the choices for SSL algorithms are very limited. For instance, many subspace methods were not developed for acoustic environments, and perform poorly in the presence of correlated signals resulting from reflections. SSL algorithms that rely on sensing the difference in source energy among different microphones cannot be applied reliably due to the close distance between microphones. Also, for any commercial device, it is not cost-effective to use microphones with matched directivity patterns, frequency responses and gains. Therefore, the algorithm should estimate these quantities wherever possible, and should be robust to estimation errors.

In this paper, we propose a novel approach to significantly improve the resolution and accuracy of range and elevation estimation: we use a room model to help extract the indirect source location information contained in the early reflections. We extend the TDOA method introduced in [15], [17] by explicitly accounting for the attenuation and path of dominant early reflections, in a method that reduces gracefully to the original algorithm in an anechoic scenario, and shows increased accuracy and robustness in the presence of reverberation.

Previous research has tried to improve robustness to reverberation by incorporating models to account for room reverberation [14], [15], or directly trying to estimate room impulse responses (RIRs) [18]–[20]. However, both approaches have limited effectiveness: generic reverberation models will only reduce the interference caused by reverberation, and estimating RIRs is a hard task. Furthermore, RIRs change rapidly and significantly with the position and orientation of the source. We choose an indirect approach: instead of trying to directly estimate RIRs, we build a 3-D model of the room to help estimate the position of the main reflectors (e.g., the closest walls and the ceiling). Using this room model, we analytically compute the strongest reflections and incorporate them into the SSL. Although more complex 3-D models could be used, in our simulations we used a simple model: four walls and a ceiling, with distances estimated with the method proposed in [21]. As we show in Section IV, this significantly improves range and elevation estimates, even with imperfect estimation and modeling of the reflectors.

The remainder of this paper is organized as follows. Section II gives an overview of room estimation methods and their requirements. Section III derives a maximum-likelihood SSL algorithm that incorporates the room model's early reflections. Section IV shows experimental results on both real and synthetic data, and Section V presents our conclusions.

II. Room Estimation

The proposed SSL algorithm is based on using a room model to estimate and predict early reflections. Thus, the first step is to obtain such a model. The most obvious way would be to measure the size, distance, and reflection coefficient of every major surface in the room. While cumbersome, this solution may be practical for large auditoriums, amphitheaters, and other large, instrumented rooms. These usually require a detailed and expensive setup, and adding a few measurements could be the most effective approach. Indeed, this is the method we used for one of the rooms reported in Section IV-C. Nevertheless, requiring professional measurement during setup is not practical for SSL in meeting rooms or homes, which are two of the important applications of the proposed technology. Thus, for many applications, we need to automatically generate the room model.

Extensive research exists for obtaining 3-D models based on video and images. Common passive methods include depth from focus, depth from shading, and stereo edge matching. Active methods include illuminating the scene with laser, or with structured or patterned infrared light. Most of this research is targeted at estimating 3-D objects, but could be readily applied to obtain room models (see, for example [22]). These image based methods may provide very precise spatial models, but have the disadvantage of not estimating reflection coefficients. However, as will be shown in Section IV, the estimation of reflection coefficients is not strictly required.

To obtain estimates of reflection coefficients, acoustic measurements have to be performed. Again, several algorithms have been proposed for automatic acoustic room measurements. O'Donovan [23] uses a 32-microphone spherical array to visualize the location of sound reflections in concert halls. Antonacci and Aprea [24], [25] use a single microphone and either a moving source on a circular trajectory or multiple sources to estimate the coordinates of reflectors. Moebus [26] 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. Similarly to the video solutions, this is particularly intriguing, because the use of ultrasound allows measurements to be performed during operation.

To avoid the need for physical measurements, as well as any additional hardware or moving parts, a reasonable method is the one we recently proposed in [21]. Instead of finding a full, detailed 3-D model of the room, the estimation is restricted to finding the location of the nearest major reflectors, which are usually the walls and the ceiling. To obtain these estimates, a test signal is reproduced through an existing single speaker integrated into a teleconferencing array, recorded by the array microphones and processed to extract the room model. This method does not require ultrasound hardware, moving parts, or multiple speakers, and was used to estimate the parameters of one of the real rooms, and of the synthetic room in Section IV.

The estimation method and simple model we used produce reasonable results. Note, however, that the optimum solution would be a more complex 3-D model, and a combination of acoustic and visual measurements. Acoustic measurements could be performed during setup, estimating the general room geometry and reflection coefficients. Visual information could be used during a meeting to account for people moving, rotation or movement of the physical device, etc.

III. ML SSL Framework With Room Models

A. Signal Propagation Model

Consider an array of M microphones in a reverberant environment. Given a signal of interest $s(n)$ with frequency representation $S(\omega)$, a simplified model for the signal arriving at each microphone is [14]

$$X_i(\omega) = \alpha_i(\omega) e^{-j\omega\tau_i} S(\omega) + H_i(\omega)S(\omega) + N_i(\omega) \quad (1)$$

where $i \in \{1, \ldots, M\}$ is the microphone index, τ_i is the time delay from the source to the ith microphone, $\alpha_i(\omega)$ is a microphone dependent gain factor, which is a product of the ith microphone's directivity, the source gain and directivity, and the attenuation due to the distance to the source, $H_i(\omega)S(\omega)$ is a reverberation term corresponding to the room's impulse response minus the direct path, convolved with the signal of interest, and $N_i(\omega)$ is the noise captured by the ith microphone.

A more elaborate version of (1) can be obtained by explicitly considering R early reflections. In this case, $H_i(\omega)S(\omega)$ only models reflections which were not explicitly accounted for. The microphone signals can then be represented by

$$X_i(\omega) = \sum_{r=0}^{R} \alpha_i^{(r)}(\omega) e^{-j\omega\tau_i^{(r)}} S(\omega) + H_i(\omega)S(\omega) + N_i(\omega) \quad (2)$$

where $\alpha_i^{(r)}(\omega)$ is a gain factor which is a product of the ith microphone's directivity in the direction of the rth reflection, the source gain and directivity in the direction of the rth reflection, the reflection coefficients for all walls involved in the rth reflection, and the attenuation due to the distance to the source, and $\tau_i^{(r)}$ is the time delay for the rth reflection. We also define $\alpha_i^{(0)}(\omega) = \alpha_i(\omega)$ and $\tau_i^{(0)} = \tau_i$, which correspond to the direct path signal.

When early reflections are modeled, traditional SSL algorithms such as [15] cannot be applied any more. In the following, we present a scheme that models early reflections as a whole, which results in a maximum likelihood algorithm that is both accurate and efficient.

Let $G_i(\omega) = \sum_{r=0}^{R} \alpha_i^{(r)}(\omega) e^{-j\omega\tau_i^{(r)}}$, which is further decomposed into gain and phase shift components $G_i(\omega) = g_i(\omega) e^{-j\varphi_i(\omega)}$, where

$$g_i(\omega) = \left| \sum_{r=0}^{R} \alpha_i^{(r)}(\omega) e^{-j\omega\tau_i^{(r)}} \right| \quad (3)$$

$$e^{-j\varphi_i(\omega)} = \frac{\sum_{r=0}^{R} \alpha_i^{(r)}(\omega) e^{-j\omega\tau_i^{(r)}}}{\left| \sum_{r=0}^{R} \alpha_i^{(r)}(\omega) e^{-j\omega\tau_i^{(r)}} \right|}. \quad (4)$$

We further approximate the phase shift components by modeling each $\alpha_i^{(r)}(\omega)$ with only attenuations due to reflections and path lengths, such that

$$e^{-j\varphi_i(\omega)} \approx \frac{\sum_{r=0}^{R} \frac{\rho_i^{(r)}}{r_i^{(r)}} e^{-j\omega\tau_i^{(r)}}}{\left| \sum_{r=0}^{R} \frac{\rho_i^{(r)}}{r_i^{(r)}} e^{-j\omega\tau_i^{(r)}} \right|} \quad (5)$$

where $r_i^{(0)}$ and $r_i^{(r)}$ are, respectively, the path lengths for the direct path and rth reflection, $\rho_i^{(0)} = 1$, and $\rho_i^{(r)}$ is the product of the reflection coefficients for all walls involved in the rth reflection. Note that reflection coefficients are assumed to be frequency independent. As will be shown later in this section, $g_i(\omega)$ can be estimated directly from the data, such that it need not be inferred from the room model and thus does not require a similar approximation.

Using $e^{-j\varphi_i(\omega)}$, (2) can be rewritten as

$$X_i(\omega) = g_i(\omega) e^{-j\varphi_i(\omega)} S(\omega) + H_i(\omega)S(\omega) + N_i(\omega). \quad (6)$$

Even if reflection coefficients are frequency dependent, they can always be decomposed into constant and frequency-dependent components, such that the frequency-dependent part which represents a modeling error is absorbed into the $H_i(\omega)S(\omega)$ term. In general, all approximation errors involving $\alpha_i^{(r)}(\omega)$ can be treated as unmodeled reflections, and thus absorbed into $H_i(\omega)S(\omega)$. Even if there are modeling errors, if the reflection modeling term $g_i(\omega)e^{-j\varphi_i(\omega)}S(\omega)$ is able to reduce the amount of energy carried by $H_i(\omega)S(\omega) + N_i(\omega)$, we should have an improvement over using (1).

Rewriting (6) in vector form, we obtain

$$\mathbf{X}(\omega) = S(\omega)\boldsymbol{G}(\omega) + S(\omega)\mathbf{H}(\omega) + \mathbf{N}(\omega) \quad (7)$$

where

$$\begin{aligned} \mathbf{X}(\omega) &= [X_1(\omega), \cdots, X_M(\omega)]^T \\ \boldsymbol{G}(\omega) &= \left[g_1(\omega)e^{-j\varphi_1(\omega)}, \cdots, g_M(\omega)e^{-j\varphi_M(\omega)}\right]^T \\ \mathbf{H}(\omega) &= [H_1(\omega), \cdots, H_M(\omega)]^T \\ \mathbf{N}(\omega) &= [N_1(\omega), \cdots, N_M(\omega)]^T. \end{aligned}$$

B. Noise Model

As in [15], we assume that the combined noise

$$\mathbf{N}^c(\omega) = S(\omega)\mathbf{H}(\omega) + \mathbf{N}(\omega) \quad (8)$$

follows a zero-mean, independent between frequencies, joint Gaussian distribution with a covariance matrix given by

$$\begin{aligned} \mathbf{Q}(\omega) &= \mathrm{E}\left\{\mathbf{N}^c(\omega)\left[\mathbf{N}^c(\omega)\right]^H\right\} \\ &= \mathrm{E}\left\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\right\} + |S(\omega)|^2\, \mathrm{E}\left\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\right\}. \quad (9) \end{aligned}$$

Making use of a voice activity detector, $\mathrm{E}\{\mathbf{N}(\omega)[\mathbf{N}(\omega)]^H\}$ can be directly estimated from audio frames which do not contain speech. To simplify matters, we assume that noise is uncorrelated between microphones, such that

$$\mathrm{E}\left\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\right\} \approx \mathrm{diag}\left(\mathrm{E}\left\{|N_1(\omega)|^2\right\},\cdots,\mathrm{E}\left\{|N_M(\omega)|^2\right\}\right). \quad (10)$$

We also assume that the second noise term is diagonal, such that

$$|S(\omega)|^2\,\mathrm{E}\left\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\right\} \approx \mathrm{diag}(\lambda_1,\cdots,\lambda_M) \quad (11)$$

with

$$\lambda_i = \mathrm{E}\left\{|S(\omega)|^2\,|H_i(\omega)|^2\right\} \quad (12)$$

$$\approx \gamma\left(|X_i(\omega)|^2 - \mathrm{E}\left\{|N_i(\omega)|^2\right\}\right) \quad (13)$$

where $0 < \gamma < 1$ is an empirical parameter which models the amount of reverberation residue, under the assumption that the energy of the unmodeled reverberation is a fraction of the difference between the total received energy and the energy of the background noise. This model has been used successfully [1], [14] for cases where reflections were not explicitly modeled [$R = 0$ in (5)], and good results have be achieved for a wide variety of environments with $0.1 < \gamma < 0.3$. Even though γ depends on the distance from the source to the array, previous work has shown that even a constant γ produces better results than neglecting the reverberation energy and using $\gamma = 0$. Furthermore, by modeling early reflections, the proposed method becomes even less sensitive to γ.

In reality, neither $\mathrm{E}\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\}$ nor $|S(\omega)|^2\mathrm{E}\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\}$ should be diagonal. In particular, any noise component due to reverberation should be correlated between microphones. However, estimating $\mathbf{Q}(\omega)$ would become significantly more expensive if not for these simplifications, and the algorithm's main loop would become significantly more expensive as well, since it requires computing $\mathbf{Q}^{-1}(\omega)$. In addition, the above assumptions do produce satisfactory results in practice.

Under the assumptions above

$$\mathbf{Q}(\omega) = \mathrm{diag}(\kappa_1,\cdots,\kappa_M) \quad (14)$$

$$\kappa_i = \gamma\,|X_i(\omega)|^2 + (1-\gamma)\mathrm{E}\left\{|N_i(\omega)|^2\right\} \quad (15)$$

such that $\mathbf{Q}(\omega)$ is easily invertible, and can be estimated with a voice activity detector.

C. Maximum-Likelihood Framework

The log-likelihood for receiving $\mathbf{X}(\omega)$ can be obtained as in [15], and (neglecting an additive term which does not depend on the hypothetical source location) is given by

$$J = \int_\omega \frac{1}{\sum_{i=1}^M |g_i(\omega)|^2/\kappa_i}\left|\sum_{i=1}^M \frac{g_i^*(\omega)X_i(\omega)e^{j\varphi_i(\omega)}}{\kappa_i}\right|^2 d\omega. \quad (16)$$

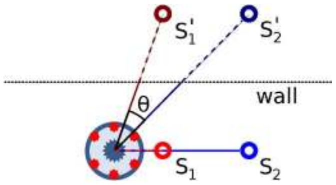


Fig. 2. Range discrimination problem with image sources. By considering image sources, range discrimination can be recast as azimuth discrimination.

The gain factor $g_i(\omega)$ can be estimated by assuming

$$|g_i(\omega)|^2\,|S(\omega)|^2 \approx |X_i(\omega)|^2 - \kappa_i \quad (17)$$

i.e., that the power received by the ith microphone due to the anechoic signal of interest and its dominant reflections can be approximated by the difference between the total received power and the combined power estimates for background noise and residual reverberation. Inserting (15) into (17) and solving for $g_i(\omega)$, we obtain

$$g_i(\omega) = \sqrt{(1-\gamma)\left(|X_i(\omega)|^2 - \mathrm{E}\left\{|N_i(\omega)|^2\right\}\right)}/\,|S(\omega)|\,. \quad (18)$$

Substituting (18) into (16)

$$J = \int_\omega \frac{\left|\sum\limits_{i=1}^M \frac{1}{\kappa_i}\sqrt{|X_i(\omega)|^2 - \mathrm{E}\left\{|N_i(\omega)|^2\right\}}X_i(\omega)e^{j\varphi_i(\omega)}\right|^2}{\sum\limits_{i=1}^M \frac{1}{\kappa_i}\left(|X_i(\omega)|^2 - \mathrm{E}\left\{|N_i(\omega)|^2\right\}\right)} d\omega. \quad (19)$$

The proposed approach for SSL consists of evaluating (19) over a grid of hypothetical source locations inside the room, and returning the location for which it attains its maximum.

To evaluate (19), one must know which reflections to use in (5), which is the only term that depends on the source location. Given the location of the walls provided by the room modeling step, we assume that the dominant reflections are the first- and second-order reflections involving only the closest walls. Using the image model [27], we analytically determine the path length of each of the first- and second-order reflections, and thereby the corresponding attenuation factors $1/r_i^{(r)}$ and time delays $\tau_i^{(r)}$ in (5). Equation (19) is then evaluated using the thus obtained value of $e^{j\varphi_i(\omega)}$.

Since $e^{j\varphi_i(\omega)}$ only depends on the room geometry and on the grid of hypothetical source locations, it can be precomputed. By assuming that γ is constant, κ_i is independent of the hypothetical source location, and has to be computed only once per frame. As we show with experiments, considering reflections from only the ceiling and one close wall is sufficient for accurate SSL.

Fig. 2 provides intuition to why the proposed method is effective. Consider two sources S_1 and S_2 which have the same azimuth and elevation angles with respect to the array. As seen in Fig. 1, it is very difficult to discriminate between both sources by using only the direct path TDOAs. However, consider image sources S_1' and S_2', which appear due to reflections off a wall. The array has good resolution in azimuth, so it can easily distinguish between S_1' and S_2'. In reality, the array always acquires the superposition of the direct path and several strong reflections, so it cannot isolate the contributions of S_1' and S_2' from those due to S_1 and S_2. Nevertheless, since signals emitted by

S_1 and S_2 have nearly identical sets of phase shifts at the microphones and because signals emitted by S_1' and S_2' have significantly different sets of phase shifts, their superposition results in measurably different sets of phase shifts for sources 1 and 2. Thus, we have transformed a detection problem for which the array had no resolution capability into another which it can solve.

An equivalent interpretation of the image model provides further insight into why this method is effective. Consider an image model which has image microphones instead of image sources. Under this model, the effective array manifold vector is written as a weighed sum of the anechoic manifold vector and its images up to a certain order. By considering images with respect to the ceiling and the walls, the resulting manifold vector no longer corresponds to that of a planar array. Since the image arrays are located outside the room, the effective manifold vector has contributions from virtual elements which are very far apart. Thus, even though the modeled array still has the same number of elements, its weaknesses due to small size and simple geometry are mitigated.

IV. Experimental Results

Since the proposed algorithm makes use of a 3-D room model, a natural question is how detailed and accurate the model needs to be. Rooms are potentially complex environments, which may contain furniture, people, partial walls, doors, windows, nonstandard corners, etc. 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, about 80 cm high. Furthermore, many objects that seem visually important are small enough to be considered acoustically transparent for most frequencies of interest. These small elements are difficult to estimate, and are sometimes moving.

It would certainly be nearly impossible to accurately model a real room. On the other hand, we need not model 100% of the reverberation. Suppose, for example, that all we can reliably estimate is the ceiling. Even if we can account for only 10% of the energy in the room added by reverberation, we would still be better off than if we had no information. Based on these observations, we adopted a simple room model: one to four walls and a ceiling. We assume the floor absorption coefficient is large enough and that sound trapping under the table will absorb most of the energy that goes below table level. To estimate the orientation and distance of these walls and ceiling, we use the method proposed in [21].

Note that this room estimation step detects only one point of reflection on each wall, indicated by the black segments in each of the four walls in Fig. 3. 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.

As we will show in this section, the proposed algorithm performs well even with one wall and the ceiling, and is quite robust to estimation errors in the room parameters.

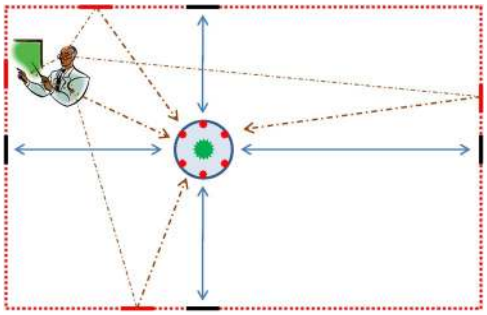


Fig. 3. Simple room model featuring reflections.

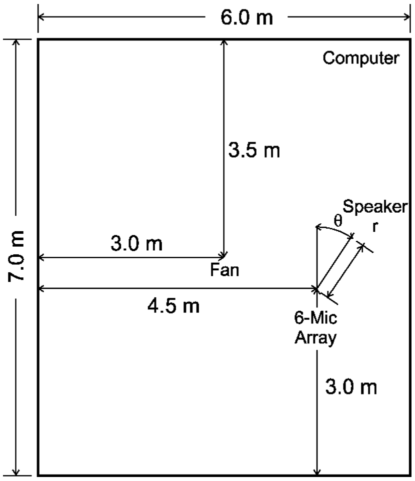


Fig. 4. Synthetic room simulated with the image model.

A. Results on Synthetic Data

Using an image model simulation [27], we generated synthetic signals to approximate what would be received by an ideal uniform circular array with a radius of 13.5 cm and six directional microphones. A three-dimensional cardioid-like gain pattern $r(\theta) = 1.1 + \cos(\theta)$ was used for each microphone. The frequency responses for each microphone were assumed to be flat, and the sampling frequency was set to 16 kHz. A virtual room with dimensions 6 × 7 × 3 m was created, with noise sources simulating a ceiling fan and a desktop computer (which were recorded from a real fan and computer), as shown in Fig. 4. The coordinates for the ceiling fan, desktop computer, and array were simulated at $3 \times 3.5 \times 3$ m, $6 \times 7 \times 0.5$ m and $4.5 \times 3 \times 1$ m, respectively. The speaker was always at a distance of 1.3 m from the array, at an elevation of 25° and at azimuth $\theta = 0°, 6°, 72°, \ldots, 324°$. Unless otherwise noted, the room was set to have a reverberation time $T_{60} = 250$ ms. The simulation does not model a conference room table, which was present in both rooms where we performed real measurements (see Section IV-C). Therefore, the dominant reflector for the synthetic scenarios is the floor (which is usually the closest surface).

The first set of synthetic data is used for modeling the room, and contains sweeps played from the loudspeaker located in the center of the device. We use this data as described in [21]. With the exception of the most distant wall (which was not detected), all walls and the ceiling were estimated within 1 cm of their

TABLE I
ERROR RATES FOR SYNTHETIC DATA, USING $T_{60} = 250$ ms

Closest	VAD	ML-SSL			R-ML-SSL		
Mic. SNR	Frames	$\lvert\Delta\theta\rvert > 10^\circ$	$\lvert\Delta\phi\rvert > 5^\circ$	$\lvert\Delta r\rvert > .15\,m$	$\lvert\Delta\theta\rvert > 10^\circ$	$\lvert\Delta\phi\rvert > 5^\circ$	$\lvert\Delta r\rvert > .15\,m$
25 dB	236	1%	3%	81%	1%	2%	2%
20 dB	222	1%	5%	83%	1%	2%	2%
15 dB	189	1%	7%	88%	1%	2%	3%
10 dB	136	1%	15%	91%	1%	4%	5%
5 dB	66	4%	28%	94%	4%	13%	16%
0 dB	18	9%	31%	94%	9%	15%	24%

TABLE II
ERROR RATES FOR SYNTHETIC DATA, USING $T_{60} = 500$ ms

Closest	VAD	ML-SSL			R-ML-SSL		
Mic. SNR	Frames	$\lvert\Delta\theta\rvert > 10^\circ$	$\lvert\Delta\phi\rvert > 5^\circ$	$\lvert\Delta r\rvert > .15\,m$	$\lvert\Delta\theta\rvert > 10^\circ$	$\lvert\Delta\phi\rvert > 5^\circ$	$\lvert\Delta r\rvert > .15\,m$
25 dB	252	6%	13%	88%	6%	8%	11%
20 dB	243	6%	15%	90%	6%	9%	11%
15 dB	220	5%	18%	91%	5%	9%	13%
10 dB	184	7%	29%	93%	7%	14%	20%
5 dB	118	12%	41%	96%	12%	26%	32%
0 dB	41	14%	51%	97%	14%	33%	45%

true position, and reflection coefficients within 0.12 of their true value, which was 0.77 for all surfaces. If the algorithm was set to only find the three closest walls, reflection coefficients were found to be exactly 0.77.

The second set of synthetic data is used to evaluate SSL performance, and simulates a male speaker standing at 1.3 m from the array. The SSL algorithm samples (19) in azimuth from 0° to 359° in 4° increments, in elevation from 0° to 60° in 1° increments, and in range from 0.5 to 5.0 m in 0.05-m increments, and only considers grid points which are inside the room. The reported results are the average for ten speaker locations distributed uniformly in azimuth around the array, all located at a distance of 1.3 m and at an elevation of 25°. At each location the speech utterance lasted 30 s, and was preceded by 2 s of background noise. The reported signal-to-noise ratio (SNR) values are for the best microphone (i.e., the one closest to the source). The MCLT [28] was used as the frequency domain transform, and the analysis frame of the SSL was set to 160 ms, overlapping by 80 ms. Only frequency taps from 200 Hz to 4 kHz were considered.

A simple energy thresholding voice activity detector (VAD) was used to estimate noise and signal powers, and to decide which frames to run the SSL algorithm on. If the VAD detected speech, first the azimuth would be estimated with the algorithm from [15], which is reasonably sensitive to elevation and completely insensitive to range. Even though the proposed algorithm produces more robust and more precise estimates for azimuth, it would require an expensive three-dimensional grid search over azimuth, range, and elevation to jointly estimate all three coordinates simultaneously. For reasonably high SNR values it would suffice to estimate azimuth by guessing a range and elevation and running a one-dimensional search, but doing so would not produce better results than completely disregarding reflections and falling back to [15]. After estimating azimuth, the proposed algorithm jointly estimated range and elevation, this time modeling first and second-order reflections.

In order to show that the proposed algorithm is robust to modeling errors, the cardioid model was not used in the SSL code, and an omnidirectional model was used instead for all microphones. Experiments show that when the microphones are known to have a nonuniform spatial pattern, it is useful to underestimate reflection coefficients. This can be justified by referring to (5), where we implicitly neglected the source and microphone directivities and assumed $\alpha_i^{(r)}(\omega) \approx \rho_i^{(r)}/r_i^{(r)}$. However, if the microphone is known to be directive, then $\alpha_i^{(r)}(\omega) \leq \rho_i^{(r)}/r_i^{(r)}$. By using an intentionally underestimated $\rho_i^{(r)}$, we can indirectly account for this attenuation. Underestimating reflection coefficients is also prudent in practical scenarios, where due to movable obstacles such as chairs and people, the reflection from the walls might not be as strong as estimated from the calibration step.

We name the proposed algorithm R-ML-SSL, and compare it to ML-SSL [15]. Table I presents simulation results for ML-SSL and R-ML-SSL in terms of frames with azimuth error $|\Delta\theta| > 10^\circ$, elevation error $|\Delta\phi| > 5^\circ$, and range error $|\Delta r| > 0.15$ m for a reverberation time $T_{60} = 250$ ms. Table II presents the corresponding simulations for $T_{60} = 500$ ms. Both algorithms use $\gamma = 0.2$ [see (13)] to model reverberation energy. It can be seen that range estimation has been dramatically improved when compared to ML-SSL. Elevation estimates have also been significantly improved. Since ML-SSL is used for azimuth estimation in both algorithms, whenever the azimuth estimate is wrong, the elevation and range joint estimation typically also produces incorrect results.

One can significantly improve the accuracy of ML-SSL and R-ML-SSL by rejecting frames without a clearly identifiable peak in the log likelihood surface. By doing so, the error rates can be made arbitrarily close to 0%, as long as the SNR values are not exceedingly small (lower than 0 dB, for example), otherwise all frames would be rejected. We describe below a version of this technique that can be applied to ML-SSL and R-ML-SSL, and delivers good results.

As mentioned previously, ML-SSL and R-ML-SSL were implemented with a simple energy thresholding VAD. In order to add a degree of noise robustness, we used a simple criterion to reject frames which had noisy log likelihood curves for

TABLE III
ERROR RATES FOR SYNTHETIC DATA, USING $T_{60} = 250$ ms AND AZIMUTH LOG LIKELIHOOD THRESHOLDING

Closest Mic. SNR	VAD Frames	ML-SSL			R-ML-SSL		
		$\|\Delta\theta\| > 10^\circ$	$\|\Delta\phi\| > 5^\circ$	$\|\Delta r\| > .15\,m$	$\|\Delta\theta\| > 10^\circ$	$\|\Delta\phi\| > 5^\circ$	$\|\Delta r\| > .15\,m$
25 dB	226	0%	1%	80%	0%	0%	0%
20 dB	208	0%	2%	82%	0%	0%	0%
15 dB	170	0%	5%	87%	0%	1%	1%
10 dB	112	0%	11%	90%	0%	1%	2%
5 dB	47	0%	21%	94%	0%	6%	7%
0 dB	10	0%	20%	91%	0%	4%	6%

TABLE IV
ERROR RATES FOR SYNTHETIC DATA, USING $T_{60} = 500$ ms AND AZIMUTH LOG LIKELIHOOD THRESHOLDING

Closest Mic. SNR	VAD Frames	ML-SSL			R-ML-SSL		
		$\|\Delta\theta\| > 10^\circ$	$\|\Delta\phi\| > 5^\circ$	$\|\Delta r\| > .15\,m$	$\|\Delta\theta\| > 10^\circ$	$\|\Delta\phi\| > 5^\circ$	$\|\Delta r\| > .15\,m$
25 dB	202	0%	5%	86%	0%	1%	1%
20 dB	187	0%	7%	88%	0%	1%	1%
15 dB	162	0%	10%	90%	0%	2%	3%
10 dB	116	0%	20%	92%	0%	4%	7%
5 dB	55	1%	27%	95%	1%	8%	12%
0 dB	15	1%	34%	96%	1%	10%	16%

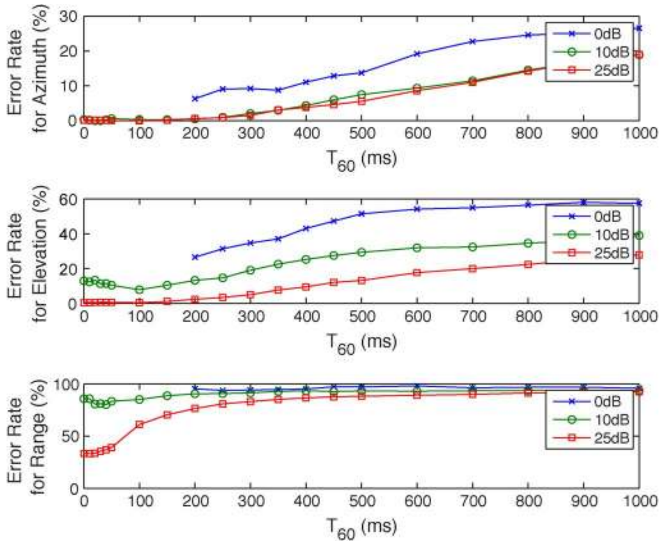


Fig. 5. ML-SSL mean error rates for varying reverberation times.

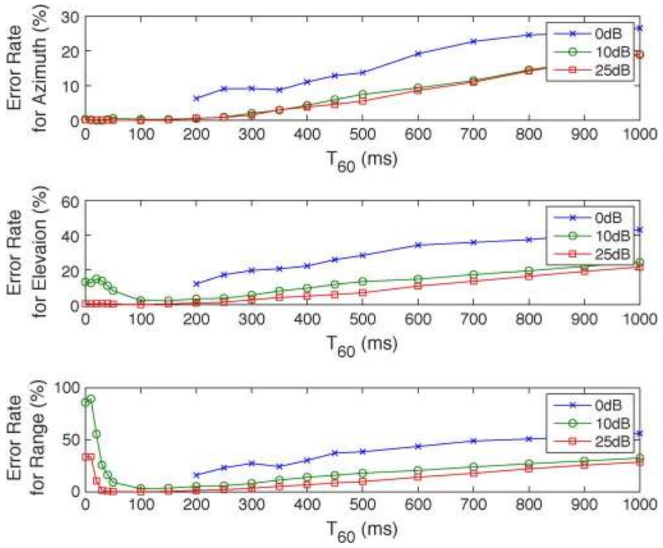


Fig. 6. R-ML-SSL mean error rates for varying reverberation times.

azimuth. If the ratio of the log likelihood's peak to its mean value was smaller than a threshold (set to 3.0 for all simulations, but which in a practical application would depend on the hardware), the frame was ignored as if the VAD had never classified it as speech. Otherwise, the algorithm would proceed as usual by computing the joint log likelihood for range and elevation. Thus, by analyzing the log likelihood for azimuth alone it is possible to reliably identify whether a frame has a sufficient amount of speech content to allow accurate three-dimensional SSL. If it does not, the frame can be immediately rejected, saving the effort of computing the joint log likelihood for range and elevation. Results are shown in Tables III and IV, and compare very favorably to the data from Tables I and II, especially for R-ML-SSL. We note that this technique was not used in any other simulation.

Figs. 5 and 6 illustrate the performance of ML-SSL and R-ML-SSL, respectively, for T_{60} varying from 0 to 1000 ms, which correspond to reflection coefficients varying from $\rho = 0.00$ to $\rho = 0.94$. Like all previous simulations, this graph considers mean SSL errors for a speaker distributed at ten locations equally spaced in azimuth, at an elevation of 25° and at a range of 1.3 m from the array. Data points are not present for 0-dB SNR and $T_{60} < 200$ ms because the voice activity detector could not identify a significant number of speech frames from at least one of the azimuth locations.

Since ML-SSL is always used for azimuth estimation, the top plot is identical for both figures. It is clear from both graphs that R-ML-SSL outperforms ML-SSL for all data points. R-ML-SSL behaves extremely well for 100 ms $\leq T_{60} \leq$ 300 ms because the walls are reflective enough to provide extra localization information, but not so reflective that the reverberation tail compromises the estimates. For range estimation, R-ML-SSL is always better in the presence of reverberation (at least for $T_{60} < 1000$ ms).

To better understand how using walls helps to estimate range and elevation, we first show on Fig. 7 the joint log likelihood for

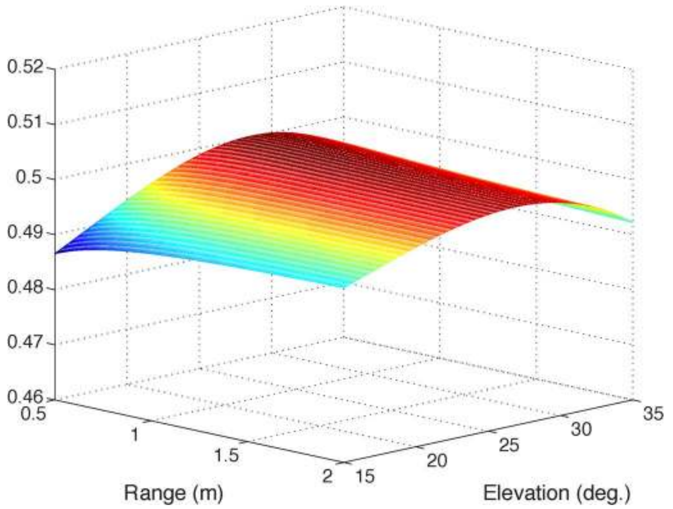


Fig. 7. ML-SSL log likelihood for range and elevation.

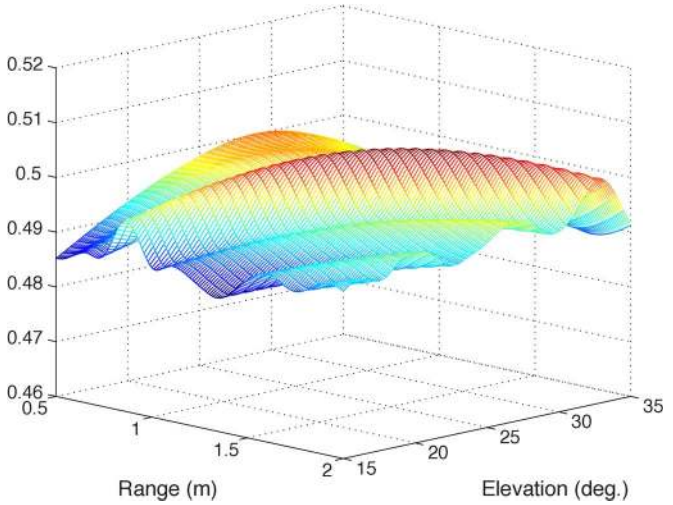


Fig. 8. R-ML-SSL log likelihood for range and elevation, considering only the ceiling.

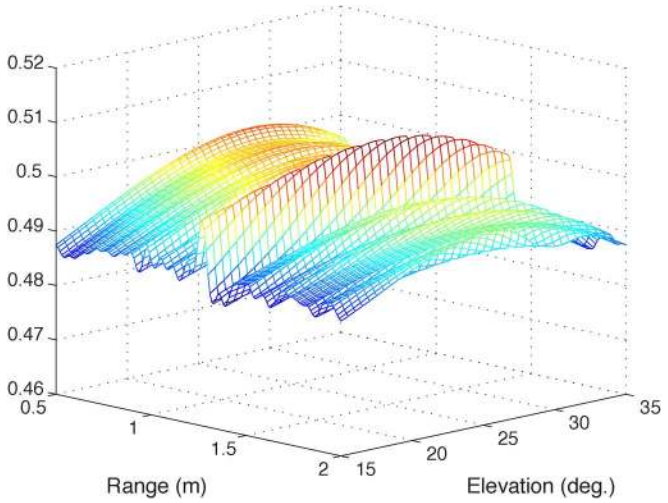


Fig. 9. R-ML-SSL log likelihood for range and elevation, considering only the closest wall.

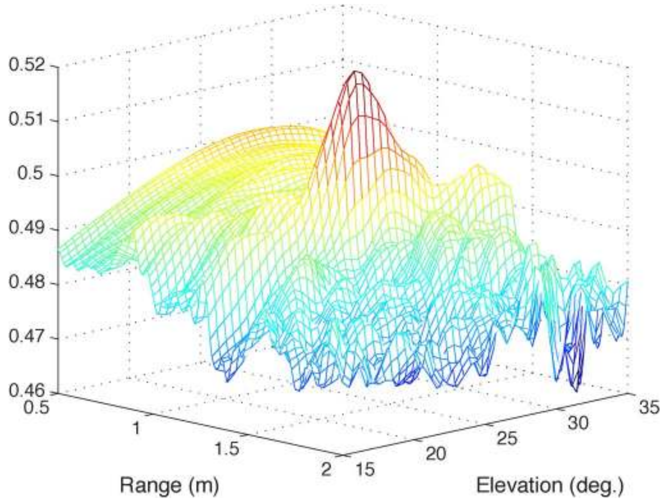


Fig. 10. R-ML-SSL log likelihood for range and elevation, considering the whole room.

range and elevation when not modeling reflections. This surface was obtained by processing a 25-dB SNR speech frame generated in the synthetic room, for a speaker located at a distance of 1.3 m and at an elevation of 25°. It has a maximum at the correct azimuth and elevation, but at an incorrect range of 0.8 m. This joint likelihood function is always very smooth, but since compact circular arrays have very poor range resolution, the maximum for range is extremely sensitive to noise and generally not a reliable estimate of the ground truth.

Now compare Fig. 7 with Fig. 8, where we introduce the modeling for the ceiling. There is now a strong ridge, which crosses the correct range-elevation value. This is introduced by the reflection of the sound source with the ceiling. Note that there is still ambiguity, as a different elevation coupled with a different range could produce similar results at the array. Compare these two figures with Fig. 9, where we introduce a single wall. Note that it also produces a ridge (similar to the one produced by introducing the ceiling), and the ridge has a different orientation. Thus, each wall, floor or ceiling produces a ridge, each with a different orientation. The correct estimate is, as one would expect, at the intersection of these ridges, as it can be seen in Fig. 10.

B. Parameter Sensitivity

We now investigate how sensitive the algorithm is to errors in the room model. First, let us look at the sensitivity to the reflection coefficient estimates. This is particularly important, since as we mentioned in Section II, many 3-D modeling tools are based on imaging or ultrasound, and may provide little or no information about reflection coefficients.

The effect of varying reflection coefficients on the error rates is shown on Fig. 11, for all values from 0.0 to 1.0 in increments of 0.1 (the ground truth being 0.77). It is clear that the proposed algorithm is relatively insensitive to the choice of reflection coefficient, as long as it is not too small (which is equivalent to disregarding reflections) or too large (which leads to a noisy log likelihood function).

As mentioned previously, the proposed algorithm does not require knowledge of all walls for good performance. As shown in the discussion associated with Figs. 7–10, for a given source location, the position of the ceiling and a dominant wall will suffice for unambiguous SSL. However, the dominant wall may not be the same for all source locations. Using a non-dominant wall will still provide SSL capability, but the method may not

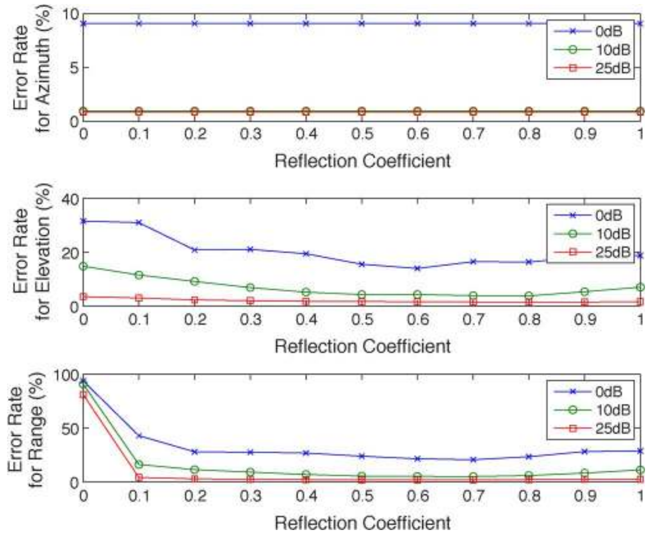


Fig. 11. Error rates for the proposed method, against increasing reflection coefficients, considering all walls, floor, and ceiling.

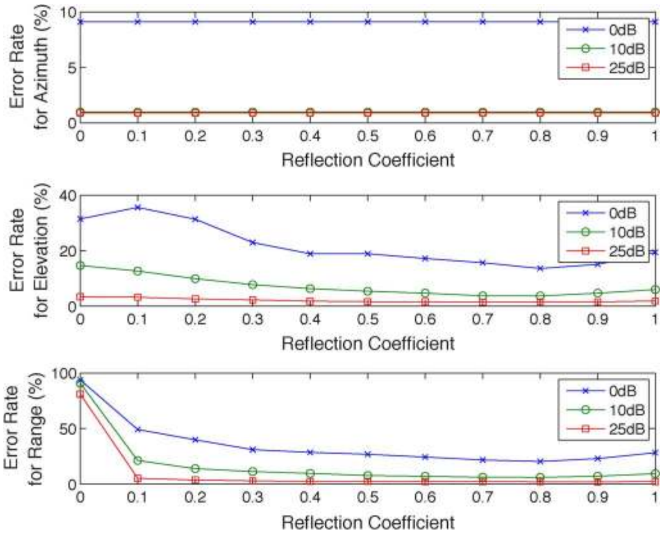


Fig. 12. Error rates for the proposed method, against increasing reflection coefficients, considering only the floor, ceiling, and closest wall.

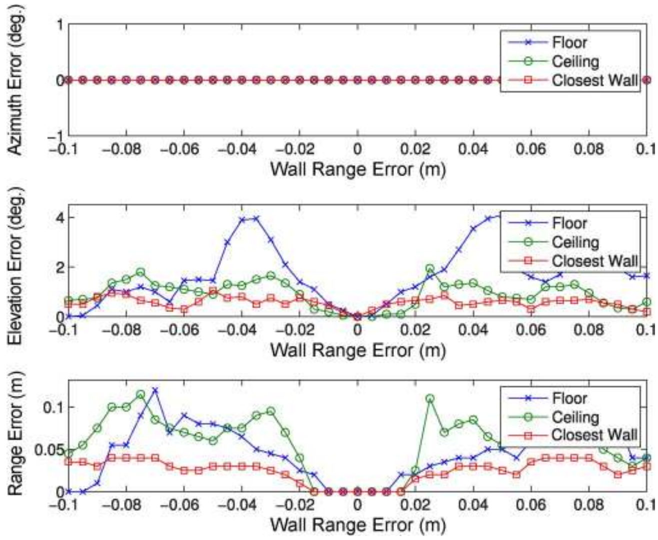


Fig. 13. Mean absolute estimation errors against wall distance perturbations, applied to one wall at a time. Top: azimuth errors, middle: elevation errors, bottom: range errors. All graphs determined using (20), with $\rho = 0.5$.

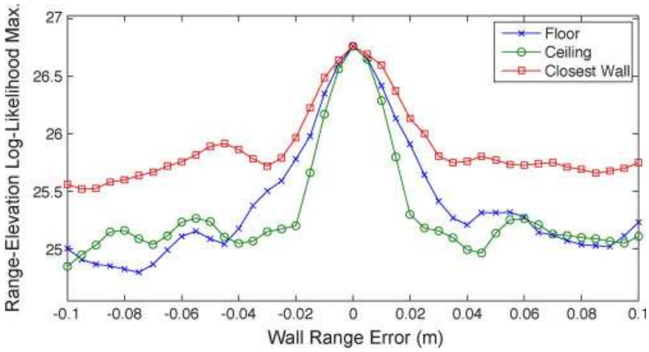


Fig. 14. Log likelihood maxima against wall distance perturbations, applied to one wall at a time, determined using (20), with $\rho = 0.5$.

perform as accurately as with a dominant wall. Fig. 12 presents error rates considering only the floor, ceiling and closest wall. One can see that SSL performance degrades very slightly in comparison with using the full model (plotted in Fig. 11). In other words, the 3-D room model may be as simplistic as three reflecting surfaces, with little reduction in performance. This can look intriguing at first, but remember that we are not interested in predicting the reverberation, but simply in capturing the extra information embedded in some of the early reflections.

Finally, since the proposed method relies on estimates of wall positions, incorrect estimates will certainly cause performance degradation. We now evaluate sensitivity to errors in wall position estimates. Performance degradation occurs in two ways: 1) the peak of the likelihood function becomes less pronounced, compromising its detection even in the absence of noise and 2) the estimates become biased. In order to perform this analysis, we consider (19) under a high SNR assumption, i.e., when $|X_i(\omega)| \gg |N_i(\omega)|$. In this case, after neglecting multiplicative constants, (19) reduces to

$$J = \int_{\omega} \left| \sum_{i=1}^{M} \frac{X_i(\omega) e^{j\varphi_i(\omega)}}{|X_i(\omega)|} \right|^2 d\omega \qquad (20)$$

which has the form of SRP-PHAT [29], [30], but with $e^{j\varphi_i(\omega)}$ in place of the direct path phase shift $e^{j\omega\tau_i}$.

Simulations show that as a wall estimate deviates from the ground truth, its corresponding log-likelihood ridge moves and decreases in height. Thus, for small wall positioning errors (on the order of a few cm), the increased error rates are mostly due to bias, since the log likelihood features remain clear, but the peak is shifted to neighboring coordinates. This effect can be observed in Fig. 13, which shows how the estimated source location shifts due to room modeling errors.

Note, additionally, that even if the initial estimates of wall positions are not perfect, the received signal can be used to refine these estimates. More specifically, the peak of the log likelihood appears due to constructive interference from the contribution of multiple walls. Thus, as shown in Fig. 14, it increases as the wall

TABLE V
ERROR RATES FOR REAL-WORLD UTTERANCES RECORDED IN ROOM 1

Speaker Position			VAD	ML-SSL			R-ML-SSL		
θ	ϕ	r	Frames	$\lvert\Delta\theta\rvert > 10^\circ$	$\lvert\Delta\phi\rvert > 5^\circ$	$\lvert\Delta r\rvert > 0.15$	$\lvert\Delta\theta\rvert > 10^\circ$	$\lvert\Delta\phi\rvert > 5^\circ$	$\lvert\Delta r\rvert > 0.15$
0°	35°	1.35 m	66	8%	8%	98%	8%	6%	21%
60°	24°	1.95 m	68	0%	0%	93%	0%	0%	1%
120°	19°	2.35 m	78	1%	29%	96%	1%	5%	13%
180°	15°	2.90 m	67	0%	30%	96%	0%	0%	1%
240°	43°	1.15 m	66	21%	20%	89%	21%	18%	21%

TABLE VI
ERROR RATES FOR REAL-WORLD UTTERANCES RECORDED IN ROOM 2

Speaker Position			VAD	ML-SSL			R-ML-SSL		
θ	ϕ	r	Frames	$\lvert\Delta\theta\rvert > 10^\circ$	$\lvert\Delta\phi\rvert > 5^\circ$	$\lvert\Delta r\rvert > 0.15$	$\lvert\Delta\theta\rvert > 10^\circ$	$\lvert\Delta\phi\rvert > 5^\circ$	$\lvert\Delta r\rvert > 0.15$
60°	26°	1.85 m	77	0%	0%	100%	0%	5%	1%
60°	18°	2.40 m	71	0%	0%	83%	0%	1%	8%
120°	32°	1.60 m	72	3%	3%	96%	3%	3%	3%
120°	22°	2.15 m	68	0%	3%	90%	0%	1%	3%
180°	15°	3.10 m	82	1%	10%	99%	1%	1%	27%
180°	12°	4.00 m	76	0%	24%	96%	0%	0%	16%
240°	36°	1.45 m	75	12%	12%	96%	12%	9%	12%
300°	40°	1.40 m	84	7%	7%	98%	7%	7%	35%

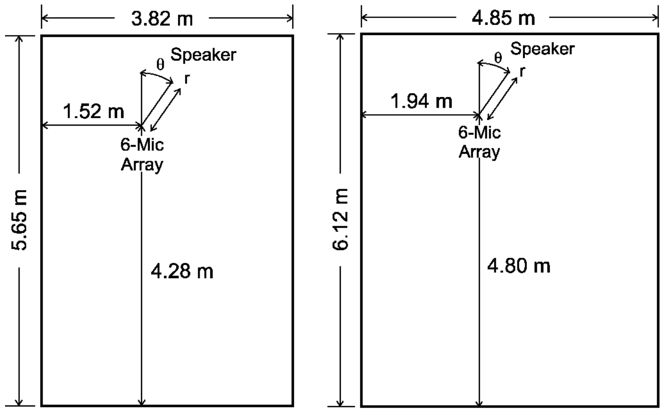


Fig. 15. Real conference rooms: room 1 (left), room 2 (right).

estimates improve. By testing wall estimates in a given neighborhood, one can select the wall coordinates which produce the largest log likelihood value.

C. Results on Real Conference Rooms

In addition to the simulated data, speech was recorded in two real, fully furnished conference rooms, which we denote Room 1 and Room 2. Room 1 measured $3.82 \times 5.65 \times 2.77$ m, and the microphone array was placed on top of a large conference table at coordinates $1.52 \times 4.28 \times 0.76$ m. Room 2 measured $4.85 \times 6.12 \times 2.73$ m, and the array was again placed on top of a large conference table, this time at coordinates $1.94 \times 4.80 \times 0.76$ m. For both rooms, $T_{60} \approx 300$ ms. Diagrams of the rooms are shown on Fig. 15. For both cases, the SSL algorithm assumed omnidirectional models for the microphones. The utterances for Room 1 have approximately a 20-dB SNR, and the utterances for Room 2 have approximately a 16-dB SNR.

To record all the experiments we used a RoundTable device. The RoundTable features a six-element uniform circular array of directional microphones, with a speaker rigidly mounted in its center and with microphones located 13.5 cm from the center. It samples audio at a rate of 16 kHz with 16-bit resolution, which allows the room modeling method detailed in [21] to estimate wall distances with better than 2-cm accuracy.

For Room 1, distances to the walls were estimated as prescribed in Section II by playing a 3-s linear sine sweep from 30 to 8 kHz through the RoundTable's internal speaker, and recorded simultaneously by all six microphones. Particularities of the device design (which was not originally designed for this purpose) produce an extremely accurate estimate of the ceiling, but less reliable estimates of walls, particularly distant walls. Fortunately, to unambiguously determine range and elevation, two strong reflectors suffice. Since the best reflector pair can change between source locations, we always used the three closest reflectors: the ceiling, the wall at -90° and the wall at 0°.

For Room 2, distances to the walls were estimated using an ultrasonic range finder with a resolution of 1 cm. Reflection coefficients were underestimated and set to 0.3 in order to account for the directivity of the microphones. For all source locations, the room model considered the ceiling, the wall at -90° and the wall at 0°.

The algorithm sampled (19) in azimuth from 0° to 359° in 4° increments, in elevation from 0° to 60° in 1° increments, and in range from 0.5 to 5.0 m in 0.05-m increments, and only considered grid points which were inside the room. All other parameters match those of the simulations (error criteria, frame size, frequency transform and γ). Tables V and VI show the error rates for Room 1 and Room 2, respectively. It is clear that R-ML-SSL shows much better range estimation than ML-SSL. It also typically outperforms ML-SSL for elevation estimation, especially for more difficult estimation problems (for example, where the source was at 4.00 m from the array, in Room 2).

Note that since the elevation and range estimates depend on a correct estimation of azimuth, the error percentages for elevation and range are in practice bounded below by the error percentages for azimuth. Furthermore, utterances with a large fraction of anomalous estimates correspond to speaker positions that are either very close to or very far from the array. Preliminary studies in an anechoic chamber showed that the microphones of the RoundTable device have a very non-smooth directivity pattern, which can be attributed to the capsule directionality, assembly, and housing. This characteristic affects the performance of azimuth estimation of close sources using ML-SSL, which in turn impacts range and elevation estimation using R-ML-SSL. Distant sources are naturally more difficult to localize due to the attenuation of the direct path and of the reflections.

V. Conclusion

We have proposed R-ML-SSL, an algorithm for sound source localization which uses strong reflections to estimate elevation and range in reverberant environments with small arrays, tasks considered very difficult with previous approaches. It uses a simple model of the room which requires only knowledge of the position and reflection coefficients for the walls closest to the array. The algorithm performs well for a large range of SNRs and reverberation times and is also robust to device modeling errors. It can be easily extended to refine previous wall estimates during the SSL step, making it more robust to room modeling errors. We have also shown with simulations that the proposed method is quite insensitive to the modeled reflection coefficients, which simplifies the room estimation step.

One of the significant contributions of this work is the incorporation of a model for reverberation that requires only knowledge of the room geometry, instead of estimates of the impulse responses from the speaker to the array. Since this room model can be obtained offline and the room geometry is assumed to be invariant, the proposed method does not require blind estimation and tracking of impulse responses, which is typically a computationally intensive and ill conditioned problem.

The use of a room model makes R-ML-SSL significantly more robust to reverberation over a large range of scenarios. Its accuracy is especially noteworthy, because the model of early reflections provides localization information which would have not been available in an anechoic environment. Since R-ML-SSL only models the strongest early reflections and does not explicitly model the reverberation tail, it is not completely immune to the effects of increasing reverberation times. Nevertheless, for many localization applications (in particular, for range discrimination under realistic reverberation times) the benefits of having strong reflections outweigh the deterioration due to the reverberation tail.

References

[1] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. WASPAA*, 1997.

[2] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: A meeting capture and broadcasting system," in *Proc. ACM Conf. Multimedia*, 2002, p. 512, ACM.

[3] E. Weinstein, K. Steele, A. Agarwal, and J. Glass, Loud: A 1020-node modular microphone array and beamformer for intelligent computing spaces Mass. Inst. of Technol., Tech. Rep. MIT-LCS-TM-642, 2004.

[4] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audio-visual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 601–616, Feb. 2007.

[5] M. Siracusa, L. Morency, K. Wilson, J. Fisher, and T. Darrell, "A multi-modal approach for determining speaker location and focus," in *Proc. ICMI*, 2003, pp. 77–80, ACM.

[6] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. Fink, and G. Sagerer, "Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot," in *Proc. ICMI*, 2003, pp. 28–35, ACM.

[7] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, no. 5, pp. 1210–1217, Oct. 1983.

[8] K. Ho and M. Sun, "An accurate algebraic closed-form solution for energy-based source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2542–2550, Nov. 2007.

[9] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[10] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.

[11] P. Georgiou, C. Kyriakakis, and P. Tsakalides, "Robust time delay estimation for sound source localization in noisy environments," in *Proc. WASPAA*, 1997.

[12] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer, 2001.

[13] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: Performance bounds and ML estimation," in *Proc. ACSSC*, 2001.

[14] Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proc. ICASSP*, 2004, no. II, pp. 133–136.

[15] C. Zhang, Z. Zhang, and D. Florencio, "Maximum likelihood sound source localization for multiple directional microphones," in *Proc. ICASSP*, 2007, vol. I, pp. 125–128.

[16] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 791–803, Nov. 2003.

[17] C. Zhang, D. Florencio, D. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 538–548, Apr. 2008.

[18] E. Weinstein, A. Oppenheim, M. Feder, and J. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Trans. Signal Process.*, vol. 42, no. 4, pp. 846–859, Apr. 1994.

[19] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP J. Appl. Signal Process.*, pp. 1110–1124, 2003.

[20] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, p. 384, 2000.

[21] D. Ba, F. Ribeiro, C. Zhang, and D. Florencio, "L1 regularized room modeling with compact microphone arrays," in *Proc. ICASSP*, 2010, pp. 157–160.

[22] D. Kimber, C. Chen, E. Rieffel, J. Shingu, and J. Vaughan, "Marking up a world: Visual markup for creating and manipulating virtual models," in *Proc. IMMERSCOM*, 2009.

[23] A. O'Donovan, R. Duraiswami, and D. Zotkin, "Imaging concert hall acoustics using visual and audio cameras," in *Proc. ICASSP*, 2008, pp. 5284–5287.

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

[25] 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. EUSIPCO*, 2009.

[26] M. Moebus and A. Zoubir, "Three-dimensional ultrasound imaging in air using a 2D array on a fixed platform," in *Proc. ICASSP*, 2007, vol. II, pp. 961–964.

[27] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

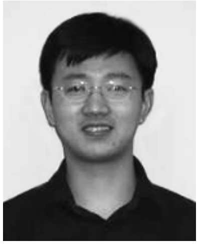
[28] H. Malvar, "A modulated complex lapped transform and its applications to audio processing," in *Proc. ICASSP*, 1999, vol. III, pp. 1421–1424.

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

[30] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. ICASSP*, 1997, vol. I, pp. 375–378.

Flavio Ribeiro (S'09) received the B.S. degree in electrical engineering from Escola Politécnica, University of São Paulo, São Paulo, Brazil, in 2005, and the B.S. degree in mathematics from the Institute of Mathematics and Statistics, University of São Paulo, in 2008. He currently pursuing the Ph.D. degree in electrical engineering, also at Escola Politécnica, University of São Paulo.

From 2007 to 2009, he was a Hardware Engineer at Licht Labs, where he developed controllers for power transformers and substations. On the summer of 2009, he was a Research Intern at Microsoft Research, Redmond, WA. His research interests include array signal processing, multimedia signal processing, active noise control, and applied linear algebra.



Cha Zhang (M'04–SM'09) received the B.S. and M.S. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1998 and 2000, respectively, and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University (CMU), Pittsburgh, PA, in 2004.

He is a Researcher in the Communication and Collaboration Systems Group at Microsoft Research, Redmond, WA. His current research focuses on applying various machine learning and computer graphics/computer vision techniques to multimedia applications, in particular, multimedia teleconferencing. During his graduate studies at CMU, he worked on various multimedia-related projects including sampling and compression of image-based rendering data, 3-D model database retrieval and active learning for database annotation, peer-to-peer networking, etc. He has published more than 40 technical papers and holds 10 U.S. patents. He coauthored a book titled *Light Field Sampling* (Morgan and Claypool, 2006). He currently serves as an Associate Editor for the *Journal of Distance Education Technologies*, *IPSJ Transactions on Computer Vision and Applications*, and *ICST Transactions on Immersive Telecommunications*. He was a guest editor for *Advances in Multimedia*, Special Issue on Multimedia Immersive Technologies and Networking.

Dr. Zhang won the Best Paper Award at ICME 2007, and the top 10% award at MMSP 2009. He was the Publicity Chair for International Packet Video Workshop in 2002, the Program Co-Chair for the first Immersive Telecommunication Conference (IMMERSCOM) in 2007, the Steering Committee Co-Chair and Publicity Chair for IMMERSCOM 2009, and the Program Co-Chair for the ACM Workshop on Media Data Integration (in conjunction with ACM Multimedia 2009). He served as TPC members for many conferences including ACM Multimedia, CVPR, ICCV, ECCV, MMSP, ICME, ICPR, ICWL, etc.

Dinei A. Florêncio (M'96–SM'05) received the B.S. and M.S. degrees from the University of Brasília, Brasília, Brazil, and the Ph.D. degree from the Georgia Institute of Technology, all in electrical engineering.

He has been a Researcher with Microsoft Research, Redmond, WA, since 1999. From 1996 to 1999, he was a member of the research staff at the David Sarnoff Research Center, Princeton, NJ. From 1994 to 1996, he was also an Associate Researcher with AT&T Human Interface Lab (now part of NCR) and a summer intern at the (now defunct) Interval Research in 1994. He has a passion for research that can have a real impact in products; his technologies have shipped in several Microsoft products and impacted the lives of millions of users. His current research interests include signal enhancement, 3-D video, signal processing for remote collaboration, and computer security. He has authored over 50 papers and holds 36 U.S. patents.

Dr. Florencio received the 1998 Sarnoff Achievement Award. He was general chair of CBSP'08 and MMSP'09, and is technical co-chair of Hot3D'2010, and WIFS'10.

Demba Elimane Ba received the B.S. degree in electrical engineering from the University of Maryland, College Park, in May 2004 and M.S. degree from the Massachusetts Institute of Technology, Cambridge, where he is currently pursuing the Doctor of Science degree in electrical engineering and computer science.

In 2006 and 2009, he worked as an summer research intern with the Communication and Collaboration Systems group at Microsoft Research, Redmond, WA. His research interests lie in the areas of mathematical and statistical signal processing, with a focus on (but not limited to) applications in multimedia and biomedical signal processing.