TURNING ENEMIES INTO FRIENDS: USING REFLECTIONS TO IMPROVE SOUND SOURCE LOCALIZATION

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

[1] Electronic Systems Eng. Dept., Universidade de São Paulo, Brazil, fr@lps.usp.br
[2] Dept. of Electrical Eng. and Computer Science, MIT, Cambridge, MA, demba@mit.edu
[3] Microsoft Research, One Microsoft Way, Redmond, WA, {chazhang,dinei}@microsoft.com

ABSTRACT

Sound Source Localization (SSL) based on microphone arrays has numerous applications, and has received significant research attention. Common to all published research is the observation that the accuracy of SSL degrades with reverberation. Indeed, early (strong) reflections can have amplitudes similar to the direct signal, and will often interfere with the estimation. In this paper, we show that reverberation is not the enemy, and can be used to improve estimation. More specifically, we are able to use early reflections to significantly improve range and elevation estimation. The process requires two steps: during setup, a loudspeaker integrated with the array emits a probing sound, which is used to obtain estimates of the ceiling height, as well as the locations of the walls. In a second step (e.g., during a meeting), the device incorporates this knowledge into a maximum likelihood SSL algorithm. Experimental results on both real and synthetic data show huge improvements in range estimation accuracy.

Keywords— Sound source localization, SSL, circular microphone array, image method, distance discrimination.

1. INTRODUCTION

Sound source localization (SSL) has been an active area of research for many years [1], and finds applications in many array processing algorithms. Several methods have been proposed with varying degrees of accuracy, noise robustness and computational complexity. Most algorithms can be classified into three categories: beamformer steering [2], subspace methods [3, 4], and methods based on time delay of arrival (TDOA) [1, 5, 6, 7]. Common to all these is the fact that performance decreases with reverberation. Another characteristic of these algorithms when used with small circular or linear arrays is their emphasis on estimating only azimuth, since in these scenarios estimating range and elevation is usually considered an almost impossible task.

The difficulty in estimating range is easily explained by considering that SSL requires implicit or explicit estimates of the relative time delay of arrival between array microphones. Relative delays for targets located at the same azimuth and elevation but at different ranges only differ due to the curvature of the wavefront. If the array is small, the curvature sampled by the microphones is negligible, and range estimation becomes essentially impossible. While elevation estimation is generally not as difficult, planar array geometries privilege azimuth estimation in detriment of elevation. Furthermore, certain array geometries are intrinsically ambiguous for elevation estimation, regardless of their size.

In this paper, we propose a novel approach to significantly improve the accuracy of range and elevation estimation: use a room model to extract the indirect source location information contained in the early reflections. We extend the method proposed in [7] to include strong reflections from walls and ceilings, accounting for reflection coefficients and attenuations due to propagation, in a method that reduces gracefully to previous model in an anechoic scenario.

We note that previous research has tried to improve robustness to reverberation by incorporating models to account for room reverberation [6, 7], or by directly trying to estimate room impulse responses (RIRs) [8]. However, both proposals have limited effect: generic reverberation models will only reduce the interference caused by reverberation, and estimating RIRs is a difficult task. Furthermore, RIRs change rapidly and significantly with the position and orientation of the source.

We solve the room estimation problem with an indirect approach: instead of trying to directly estimate RIRs, we estimate the position of main reflectors (walls and ceiling) in relation to the array. Given this information and a hypothetical source location, one can analytically compute the time delays and amplitudes for the strong reflections. As we show in the following sections, we can incorporate these reflections into the SSL algorithm to significantly improve range and elevation estimates, even with imperfect modeling.

The remainder of this paper is organized as follows: Section 2 gives a brief overview of the room estimation method. Section 3 derives a maximum likelihood SSL algorithm that incorporates the room model's early reflections. Section 4 shows experimental results on real and synthetic data, and Section 5 presents some of our conclusions and future work.

2. ROOM ESTIMATION OVERVIEW

Rooms are potentially complex environments, which may contain furniture, people, partial walls, doors, windows, non-standard corners, etc. Yet, after sampling a few conference rooms in corporate environments, several things seem so common we may take them for granted. Almost every room has four walls, a ceiling and a floor; the floor is leveled, and the ceiling

978-1-4244-7493-6/10/$26.00

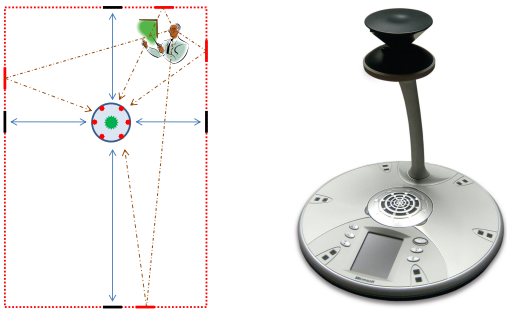

Fig. 1. Room model and RoundTable device

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. In addition, many objects that seem visually important are small enough that they 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. We assume the floor absorption coefficient is high enough and that sound trapping under the table absorbs most of the energy that goes below table level.

Even with such a simplified room model, it would be difficult to passively estimate the components of the model (wall positions and reflection coefficients) based only on unknown signals already existing in the room. Instead, we 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), as proposed in [9]. While this method is convenient since it does not require additional hardware, alternative methods for room modeling could be used [10, 11, 12]. The proposed method for SSL is quite robust and will work with underestimated reflection coefficients, such that only the knowledge of the wall geometry is strictly necessary.

Although other arrays and devices could be used in similar fashion, for the purposes of this paper we consider the RoundTable device [13], a uniform circular array with a speaker rigidly mounted in its center (see Fig. 1). The room estimation step detects 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 wall segments in Fig. 1. The underlying assumption is that the walls extend linearly and with similar acoustic characteristics.

Note that this modeling is performed during device initialization, and only has to be repeated if the array is moved or if the room geometry changes (the first being far more likely). A motion detector can be used to test if the device moves. The device could be also equipped with ultrasound emitters and microphones, which could not only aid in the wall detection phase, but could also be used to dynamically monitor and adapt the acoustic environment.

3. ML SSL FRAMEWORK

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

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

where $i \in \{1, ..., M\}$ is the microphone index; τ_i is the TDOA from the source to the i^{th} microphone; $\alpha_i(\omega)$ is a gain factor which includes the microphone frequency dependent sensitivity and directionality, the source gain and directionality, and the attenuation due to the distance to the source; $H_i(\omega) S(\omega)$ is a reverberation term corresponding to the room's response convolved with the signal of interest; $N_i(\omega)$ is the noise captured by the i^{th} microphone. This was the treatment given in [7, 14], and which resulted in an ML SSL estimator for direction of arrival.

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

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

where $\alpha_i^{(r)}(\omega)$ represents a gain factor which considers the microphone frequency dependent directionality for the reflection's direction of arrival, the source gain and directionality in the direction which results in a reflection, reflection coefficients and attenuation due to the distance to the source; $\tau_i^{(r)}$ is the time delay for the r^{th} reflection. We also define $\alpha_i^{(0)}(\omega) = \alpha_i(\omega)$, and $\tau_i^{(0)} = \tau_i$, which correspond to the direct path signal.

Let $G_i(\omega) = \sum_{r=0}^{R} \alpha_i^{(r)}(\omega) e^{-j\omega\tau_i^{(r)}}$, which can be 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 approximate the phase shift components by modeling each $\alpha_i^{(r)}(\omega)$ with only attenuations due to reflections and path lengths, such that

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

where $d_i^{(0)}$ and $d_i^{(r)}$ are respectively the path lengths for the direct path and r^{th} reflection; $\rho_i^{(0)} = 1$, and $\rho_i^{(r)}$ is the r^{th} reflection coefficient. Note that reflection coefficients are assumed to be frequency independent. We will show that $g_i(\omega)$ can be estimated from the data, such that it need not be inferred from the room model and thus does not require a similar approximation.

Considering this approximation, (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 though wall reflection coefficients have some frequency dependency, 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, approximation errors 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 $g_i(\omega)e^{-j\varphi_i(\omega)}S(\omega)$ reduces the amount of energy carried by $H_i(\omega)S(\omega) + N_i(\omega)$, one should have an improvement over (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

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

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

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

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. We assume that the noise is uncorrelated between microphones, such that

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

We also assume that term corresponding to unmodeled reverberation is diagonal, such that

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

with

$$\begin{aligned}
\lambda_i &= \mathrm{E}\{|S(\omega)|^2|H_i(\omega)|^2\} &(12)\\
&\approx \gamma(|X_i(\omega)|^2 - \mathrm{E}\{|N_i(\omega)|^2\}), &(13)
\end{aligned}$$

where $0 < \gamma < 1$ is an empirical parameter which models the amount of reverberation residue, under the assumption that the reverberation noise energy is proportional to the source signal energy. 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, reverberation terms should be correlated between microphones. However, estimating $\mathbf{Q}(\omega)$ would be intractable if not for these simplifications, and the algorithm's main loop would become significantly more expensive as well, since it requires computing $\mathbf{Q}^{-1}(\omega)$ for every frame. In addition, the above assumptions do produce satisfactory results in practice.

Under these assumptions,

$$\begin{aligned}
\mathbf{Q}(\omega) &= \mathrm{diag}(\kappa_1, \cdots, \kappa_M) &(14)\\
\kappa_i &= \gamma|X_i(\omega)|^2 + (1-\gamma)\mathrm{E}\{|N_i(\omega)|^2\}. &(15)
\end{aligned}$$

The log-likelihood for receiving $\mathbf{X}(\omega)$ can be obtained as in [7], and (neglecting an additive term which does not depend on the hypothetical source location) is given by

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

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 i^{th} microphone due to the signal of interest can be modeled by the total received power, minus the combined estimate for noise and reverberation power. Solving for $g_i(\omega)$ we obtain

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

Substituting (18) into (16),

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

The proposed method 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). Given the location of the walls determined with the room modeling step, we assume that the dominant reflections will be the first and second order reflections originating from the closest walls. We apply the image model [15] to analytically determine the contributions to gain and phase shift due to these dominant reflections, which are used in (5), and therefore, in (19). As we show with experiments, determining the position of only the ceiling and one close wall is sufficient for accurate SSL.

4. EXPERIMENTAL RESULTS

4.1. Results for Synthetic Data

Using the image model, we generated synthetic signals simulating the signals that would be received by the RoundTable

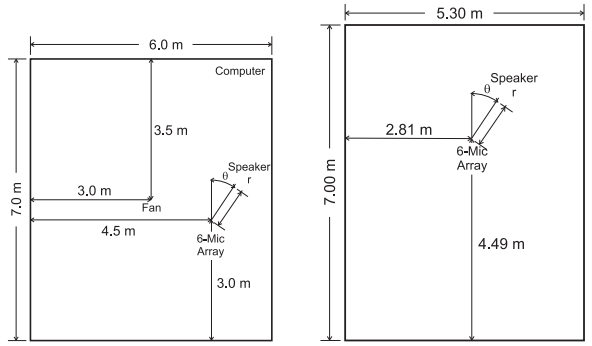

Fig. 2. Synthetic and real rooms (left and right, respectively)

device, which is a uniform circular array with six directional microphones and a radius of 13.5 cm (see Fig. 1). 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 $6 \times 7 \times 3$ 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. 2. 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 $r = 1.3$ m to the array, elevation $\phi = 0°$ and azimuth $\theta = 0°, 36°, 72°, ..., 324°$. The room was set to have a reverberation time $T_{60} = 250$ ms. Note that the simulation does not model obstacles – in particular, it does not model a conference room table. Since the RoundTable microphones are very close to the bottom of the unit (almost on the plane of the table), computing first and second order images with respect to the table would not have a significant effect.

The first set of synthetic data corresponds the the room estimation step, i.e., sweeps played from the loudspeaker, located at the center of the device. We use this data to estimate the room, as described in [9]. The wall positions were estimated within 1 cm of their true position, and reflection coefficients within 0.12 of their true value, which was 0.77 for all surfaces. If only the three dominant reflectors were needed, this method would produce their exact positions and reflection coefficients. As will be presented in this section, only two dominant reflectors are required for unambiguous three-dimensional SSL.

The second set of synthetic data simulates a male speaker, sitting at 1.3 m from the array, and is used to test the SSL step. The SSL algorithm samples (19) in azimuth over the whole circle with $4°$ resolution, in elevation from $-10°$ to $+10°$ with $0.5°$ resolution, and in range from 0.5 to 2 m with 0.05 m resolution. The reported results are the average for 10 speaker locations distributed equally in azimuth around the array, all located at a distance of 1.3 meters and a $0°$ elevation. At each location the speech utterance lasted 5 seconds, and was preceded by 2 seconds of background noise. The MCLT [16] was used as the frequency domain transform, and the analysis frame of the SSL was set to 160 ms, overlapping by 80 ms.

A simple speech activity detector (VAD) was used to estimate noise and signal powers, and to decide on which frames to run the SSL algorithm. If the VAD classified a frame as speech, the SSL algorithm consisted of estimating azimuth using the algorithm from [7], which corresponds to ignoring reflections. After estimating azimuth, the algorithm jointly estimated range and elevation. The azimuth estimation is quite insensitive to range and elevation, so the decoupled estimation does not significantly impact performance. Even though better robustness and accuracy could be achieved by jointly estimating azimuth, range and elevation, doing so would require a computationally expensive three-dimensional search. Using the proposed method for estimating only azimuth with an initial (fixed) estimate for range and elevation (for example, 1.0 m and $30°$) would also work but would deliver worse results, since it would incorrectly compute reflections.

In order to show that the method is robust to calibration and modeling errors, the SSL code assumed an omnidirectional model for the microphones. Simulations and real-world experiments show that in the presence of calibration or modeling errors, 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 $\alpha_i^{(r)}(\omega) \approx \rho_i^{(r)}/d_i^{(r)}$. However, if the microphones are known to be directive, then $\alpha_i^{(r)}(\omega) \leq \rho_i^{(r)}/d_i^{(r)}$. By using an intentionally underestimated $\rho_i^{(r)}$, we can indirectly account for the directional 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.

Table 1 presents simulation results, in terms of frames with azimuth errors larger than $5°$, elevation errors larger than $1°$ and range errors larger than 0.15 m. We name our proposal R-ML-SSL, and compare it to ML-SSL [7]. Both algorithms use $\gamma = 0.2$ to model reverberation energy. As it can be seen, estimation of elevation and range is dramatically improved.

To better understand how using walls helps to estimate range and elevation, we show on Fig. 3 the joint log likelihood for range and elevation when not accounting for reflections (i.e., with ML-SSL), obtained by processing a single 25 dB SNR speech frame obtained in the synthetic room. When compared to the likelihood surfaces obtained with reflections (Figs. 4-6), this surface is very smooth and nearly flat.

For small arrays, the ML-SSL log likelihood maximum is very sensitive to variations in the estimated delays of arrival. Indeed, varying range will have a small effect on the signals arriving at the microphones, other than the distance attenuation and a constant delay over all microphones, which cannot be detected since we do not have the original signal. Thus, range estimation becomes an ill conditioned inverse problem, and even a small amount of noise can introduce large errors in the estimation. This problem is especially severe for the simulated scenario, because a circular array has maximum noise sensitivity for sources at $\phi = 0°$, even when using omnidirectional microphones. This sensitivity increases due to the cardioid microphone models, because the source is only captured well by the 3-4 microphones facing it.

The sensitivity is such that at 25 dB SNR, even the concavity

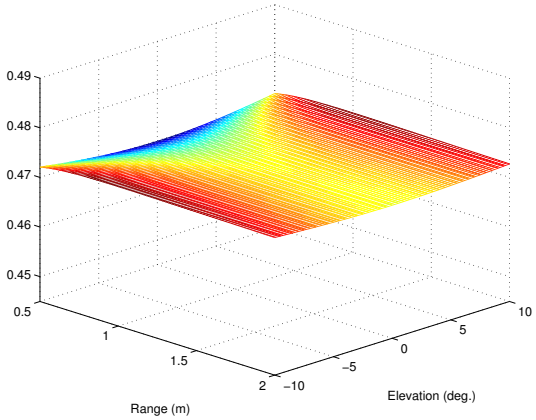

Fig. 3. Log likelihood for ϕ and r, using [7].

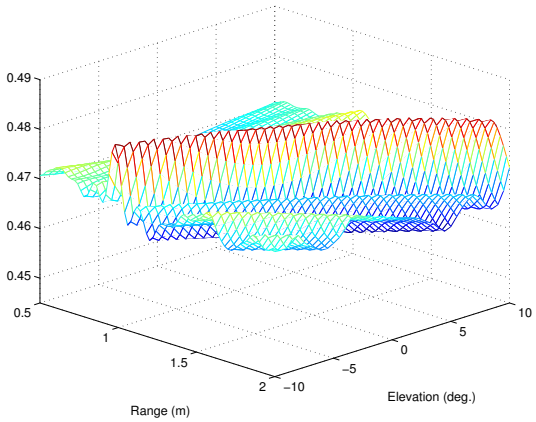

Fig. 4. Log likelihood for ϕ and r, considering ceiling only.

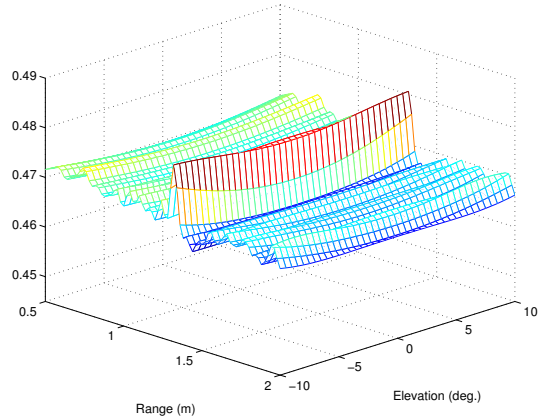

Fig. 5. Log likelihood for ϕ and r, considering one wall.

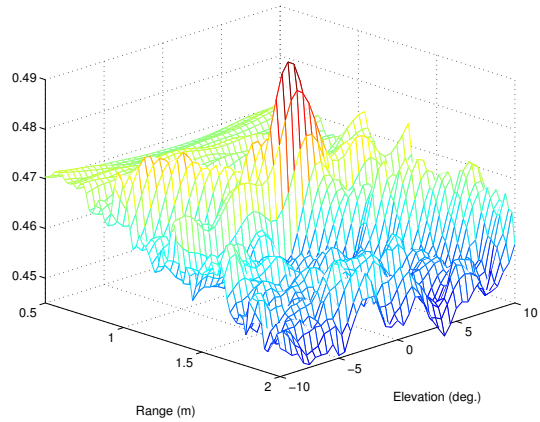

Fig. 6. Log likelihood for ϕ and r, considering ceiling, walls and floor.

of the ML-SSL surface can change from one frame to the next. In fact, close inspection of Fig. 3 shows that it features the opposite concavity that one would expect for this problem, with a minimum at $\phi = 0°$. This is why the elevation errors from Table 1 tend to decrease with increasing SNR. For low SNR, the ML-SSL range-elevation surface often changes from the saddle shown in Fig. 3 to a concave function symmetrical with respect to $\phi = 0°$, for which the estimate is coincidentally correct.

Now compare Fig. 3 with Fig. 4, where we introduce the modeling for the ceiling. There is now a strong ridge, which crosses the correct range-elevation pair. This is introduced by the ceiling reflection of the sound source. Note that there is still ambiguity, as a different elevation coupled with a different range could produce similar results at the array. Now compare those two figures with Fig. 5, 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 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 Fig. 6 shows.

4.2. Results for a Real Conference Room

In addition to the simulated data, real data was recorded in a typical conference room. The room measured $5.30 \times 7.00 \times 2.77$ m, and the RoundTable device was placed on top of a large conference table at coordinates $2.81 \times 4.49 \times 0.78$ m. A diagram of the room is presented on Fig. 2. In the first step, the room was estimated as prescribed in [9] by playing a 3-second linear sine sweep from 30 Hz to 8 kHz through the RoundTable's internal speaker, and recorded simultaneously by all 6 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, as we argue in Section 4.1, to unambiguously determine range and elevation, two reflectors are sufficient. Therefore, we decided to only use the distance to the ceiling and to the closest wall.

Table 2 shows the corresponding error percentages. Note that error thresholds were relaxed due to difficulty in measuring ground truth values. It is clear that R-ML-SSL shows much better range estimation than ML-SSL. It was interesting to notice that ML-SSL could correctly estimate $\hat{\phi} = 0°$, except for one frame. Even though the RoundTable microphones are directional, their enclosures and assembly make them significantly less directional than the cardioid model, so that the source signal is captured reasonably well by all microphones and not only by the 3-4 facing the source. Thus, even though resolution for elevation is still poor, the log likelihood surfaces are typically

Table 1. Error rates for synthetic data

Closest Mic. SNR	# Voice Frames	ML-SSL			R-ML-SSL		
		$\Delta\theta > 5^\circ$	$\Delta\phi > 1^\circ$	$\Delta r > .15\,m$	$\Delta\theta > 5^\circ$	$\Delta\phi > 1^\circ$	$\Delta r > .15\,m$
25 dB	48	0.4 %	92.0 %	84.6 %	0.4 %	0.2 %	0.2 %
20 dB	47	0.2 %	90.6 %	81.6 %	0.2 %	0.0 %	0.0 %
15 dB	44	1.1 %	88.2 %	80.9 %	1.1 %	0.9 %	0.4 %
10 dB	38	3.4 %	85.6 %	81.6 %	3.4 %	2.9 %	2.1 %
5 dB	29	15.0 %	86.0 %	90.1 %	15.0 %	14.3 %	6.9 %
0 dB	16	24.8 %	85.1 %	94.3 %	24.8 %	22.2 %	15.5 %

Table 2. Error rates for real-world utterances

Speaker Position	# Voice Frames	ML-SSL			R-ML-SSL		
		$\Delta\theta > 10^\circ$	$\Delta\phi > 5^\circ$	$\Delta r > 0.15$	$\Delta\theta > 10^\circ$	$\Delta\phi > 5^\circ$	$\Delta r > 0.15$
$\theta = 0^\circ, \phi \approx 3^\circ, r \approx 1.70\,m$	11	0%	0%	55%	0%	0%	0%
$\theta = 60^\circ, \phi \approx 3^\circ, r \approx 1.47\,m$	21	0%	5%	95%	0%	9%	9%
$\theta = 120^\circ, \phi \approx 3^\circ, r \approx 1.37\,m$	9	0%	0%	33%	0%	11%	11%

concave and symmetrical around $\phi = 0^\circ$, resulting in correct estimates. While ML-SSL was able to perform some range estimation, its performance is not reliable, as seen in the results for position #2.

5. CONCLUSIONS

We have presented an SSL algorithm which uses strong reflections to estimate source elevation and range, tasks considered very difficult with previous approaches. It performs effectively even under low SNR, and does not require an accurate model of the array characteristics. Experiments show that its azimuth localization accuracy remains comparable to that of previous methods.

Future work involves improving the method with better device models, accounting for microphone gains, directivities and phase responses, or applying similar room modeling techniques to other problems [17]. It would also be desirable to adaptively refine wall position estimates using the SSL results, given that incorrect walls or reflection coefficients typically result in a decreased peak in the likelihood function [18].

6. REFERENCES

[1] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*, Springer, 2001.

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

[3] 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, 1989.

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

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

[6] Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proc. of ICASSP*, 2004.

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

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

[9] D. Ba, F. Ribeiro, C. Zhang, and D. Florencio, "L1-regularized room modeling with compact microphone arrays," in *Proc. of ICASSP*, 2010.

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

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

[12] 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.

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

[14] 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, April 2008.

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

[16] H. Malvar, "A modulated complex lapped transform and its applications to audioprocessing," in *Proc. of ICASSP*, 1999.

[17] M. Song, C. Zhang, D. Florencio, and H. Kang, "Personal 3D Audio System with Loudspeakers," in *Proc. of Hot3D*, 2010.

[18] F. Ribeiro, C. Zhang, D. Florencio, and D. Ba, "Using reverberation to improve range and elevation discrimination for small array sound source localization," submitted to IEEE Trans. Audio, Speech, Lang. Process., 2010.