ACOUSTIC REFLECTOR LOCALIZATION AND CLASSIFICATION

Luca Remaggi\textsuperscript{1}, Hansung Kim\textsuperscript{1}, Philip J. B. Jackson\textsuperscript{1}, Filippo Maria Fazi\textsuperscript{2} and Adrian Hilton\textsuperscript{1}

\textsuperscript{1}Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK
\textsuperscript{2}Institute of Sound and Vibration Research, University of Southampton, Southampton, SO17 1BJ, UK

\section*{ABSTRACT}

The process of understanding acoustic properties of environments is important for several applications, such as spatial audio, augmented reality and source separation. In this paper, multichannel room impulse responses are recorded and transformed into their direction of arrival (DOA)-time domain, by employing a superdirective beamformer. This domain can be represented as a 2D image. Hence, a novel image processing method is proposed to analyze the DOA-time domain, and estimate the reflection times of arrival and DOAs. The main acoustically reflective objects are then localized. Recent studies in acoustic reflector localization usually assume the room to be free from furniture. Here, by analyzing the scattered reflections, an algorithm is also proposed to binary classify reflectors into room boundaries and interior furniture. Experiments were conducted in four rooms. The classification algorithm showed high quality performance, also improving the localization accuracy, for non-static listener scenarios.

\textbf{Index Terms—} Reflector Localization, Reflector Size Classification, Room Impulse Response, Beamforming

\section*{1. INTRODUCTION}

In audio processing, several areas can benefit from an accurate estimation of the environmental acoustic properties. Information about the main reflectors can aid audio forensic \cite{1} and simultaneous localization and mapping methods \cite{2}. Moreover, knowledge of the reflected signal characteristics can improve systems in source separation \cite{3}, dereverberation \cite{4}, and microphone array raking \cite{5}. Another application, that this paper targets, is the active manipulation of reverberant spatial audio objects, for next generation audio production \cite{6}. Furthermore, augmented reality (AR) systems \cite{7} can be improved by the auditory scene analysis provided here. The main contributions of this paper are: a novel method to localize acoustic reflectors given room impulse responses (RIRs); a novel classification algorithm, that analyzed the scattered reflections to classify reflectors into large (i.e. room boundaries) and small (e.g. furniture).

One of the first attempts to localize acoustic reflectors was proposed in \cite{8}. There, the time-space domain of multichannel RIRs was employed. This was done by assuming a linear microphone array to be perpendicular to the analyzed direction. Later, in \cite{9}, the direction of arrival (DOA)-time domain was analyzed, by employing the plane-wave decomposition method. Nevertheless, a large space was required to be sampled, to correctly capture the sound wave. Spherical harmonics were employed in \cite{10, 11}, to localize reflections. As in \cite{12}, they employed continuous signals, however, having issues with low SNRs. In \cite{13}, reflectors were localized by first estimating the related image source positions \cite{14}, having knowledge of the reflections’ times of arrival (TOAs). However, microphones were assumed to be sparse in space. A 2D geometric approach was presented in \cite{15}. Its full 3D extension was later proposed in \cite{16}, where reflectors were localized by constructing ellipsoids, having axes proportional to the reflection TOAs. Despite the good performance, it had a high computational cost. In this paper, we propose a novel approach, that exploits the reflection TOAs and DOAs. These are estimated by a new 2D morphological image processing method, which detects meaningful peaks in noisy spectral images, such as the DOA-time domain of multichannel RIRs.

Every state-of-the-art method that has been mentioned above aimed to localize acoustic reflectors by observing their specular reflections \cite{17}. However, a correct analysis of the scattered reflections can help achieving several goals \cite{18}. For instance, it can lead to a discrimination between walls and furniture, in a room. This kind of classification allows production of synthetic RIRs \cite{19}, and estimation of reverberant spatial audio object parameters \cite{20}. To our knowledge, in the literature, there is yet no method to classify acoustic reflectors in terms of their spatial extension. To fill this gap, in this paper, we propose a binary classification algorithm, that uses the scattered reflections’ energy as feature.

In this paper, Sec. 2 presents novel methods to localize and classify acoustic reflectors; in Sec. 3, experiments and results are discussed; Sec. 4 draws the overall conclusion.

\section*{2. THE PROPOSED METHODS}

We propose a new pipeline (see Fig. 1), to localize and classify acoustic reflectors. RIRs are employed, by defining them...
as \( r_{i,l}(n) = \sum_{e=0}^{T_{en}} h_{e,i,l}(n - n_{e,i,l}) + g_{i,l}(n) \), where \( i \) and \( l \) are the microphone and source indexes, respectively; \( n \) is the discrete time index. \( T_{en} \) is the last early reflection, and \( g_{i,l}(n) \) the late reverberation. \( n_{e,i,l} \) is the early reflection TOA, with \( e \) being the reflection index (\( e = 0 \) indicates the direct sound).

2.1. The DOA-Time Domain

Multichannel RIRs can be transformed to the joint DOA-time domain, where the acoustic energy is function of both azimuth DOA and time. This domain transformation is done by using a superdirective array (SDA) beamformer [21]. Before employing the SDA, RIRs are band-pass filtered between \( \Omega_1 = 2 \text{kHz} \) and \( \Omega_2 = 8 \text{kHz} \), to avoid the poor low frequency directivity, and the spatial aliasing at high frequencies.

The SDA is our 3D version [21] of [22], and it can be applied to any kind of microphone array. Points in the 3D Cartesian coordinates can be defined as \( \begin{bmatrix} u_x; u_y; u_z \end{bmatrix}^T = [\cos \theta \sin \phi; \sin \theta \sin \phi; \cos \phi]^T \), where \( [\cdot]^T \) is the matrix transpose operator, \( \theta \) the azimuth, and \( \phi \) the elevation. The array manifold can be then written as \( a(\omega, \theta, \phi) = \exp(j \frac{\omega}{c_0} A X) \), where \( j = \sqrt{-1} \), \( A \) contains the \( M \) microphones’ coordinates, \( \omega \) is the angular frequency, and \( c_0 \) the sound speed. The elements of the noise cross-spectral matrix \( S \) are: \( s_{i,i'} = \text{sinc}(\frac{\omega}{c_0}|A_i - A_{i'}|) \), where \( A_i \) and \( A_{i'} \) contain the \( i \)-th and \( i' \)-th microphone coordinates, respectively, and \( || \cdot || \) is the \( \ell_2 \) norm operator. The regularization matrix is \( B = |S + \eta I|^{-1} \), with \( \eta = 0.01 \) (empirically found) and \( I \) the identity matrix. The SDA weights are: \( w(\omega, \theta, \phi) = \frac{B a}{a^* B a} \), where \( [\cdot]^* \) is the complex conjugate transpose operator. \( w(\omega, \theta, \phi) \) contains the \( M \) microphones signals’ weights \( w_i(\omega, \theta, \phi) \).

Beamformed RIRs are produced by using the SDA as:

\[
\begin{align*}
    r_i^B(n, \theta, \phi) &= \mathcal{F}^{-1} \left\{ \sum_{i=1}^{M} \sum_{\omega=0}^{T_{\text{early}}} w_i(\omega, \theta, \phi) R_{i,l}(\omega) \right\},
\end{align*}
\]

where \( R_{i,l}(\omega) \) is the \( r_{i,l}(n) \)’s Fourier transform and \( \mathcal{F}^{-1} \{ \cdot \} \) the inverse Fourier transform operator. By setting \( \phi = 0 \), beamformed RIRs are produced for each \( \theta \), and placed one next to the others, to obtain the DOA-time domain (see Fig. 2).

2.2. Local Maxima Detection

In the DOA-time domain, the energy maximum coordinates are the reflection TOAs and azimuth DOAs. A peak-picking algorithm can be exploited to detect them. Automated peak detection for 2D spectral images have been actively researched for more than 30 years, in image processing and bioinformatics [23, 24, 25]. The major challenges of 2D peak detection algorithms is the distinction of artifacts from real peaks, and also detecting peaks with overlapping resonances. We propose a 2D peak detection algorithm modified from the morphology-based method in [25], for meaningful peak detection in the DOA-time domain images. Fig. 3 shows the block diagram of the proposed 2D peak detection algorithm.

In the pre-processing stage, the input image is filtered by a Gaussian kernel to reduce noise. A sound wave decreases its energy by 6 dB every double of the traveled distance [26]. To compensate this bias, the original values in the DOA-time domain are updated by \( y = y + 6 \log_2(\omega) \), where \( y \) is the energy level in dB and \( \omega \) is time in ms. The middle image of Fig. 2 shows the result of this bias compensation. This image is converted into two images by two morphological process: erosion and dilation, then subtracted from the original image. Two subtraction images are binarized and the erosion image is used as a mask to filter out the peak image from the dilation image. Finally, contours of the images are found, and their centers are defined as the detected peak locations. The right image of Fig. 2 shows the proposed peak detector output.
2.3. Elevation Estimation

From the position of the maxima in the DOA-time domain, the reflection TOAs \( n_{e,l} \) and azimuth DOAs \( \theta_{e,l} \) are obtained. However, to localize reflections in the 3D space, elevations \( \phi_{e,l} \) are also needed. By fixing \( \theta \) to the estimated values \( \theta_{e,l} \), the SDA weights \( w(\omega, \theta = \theta_{e,l}, \phi) \) are calculated considering elevation angles between 0° and 90° with a resolution of 10°. The corresponding beamformed signals \( r_{\text{B}}^{n}(n, \theta = \theta_{e,l}, \phi) \) are then obtained by following Eq. (1). By using a Hamming window of length \( D = 2.7 \text{ ms} \) (value heuristically found), centered at \( n_{e,l} \), the beamformed reflections are segmented. The elevations \( \phi_{e,l} \) are selected as the ones producing the reflection segments having the greatest energy:

\[
\phi_{e,l} = \arg \max_{\phi} \sum_{n=n_{e,l}-D}^{n_{e,l}+D} |r_{\text{B}}(n, \theta = \theta_{e,l}, \phi)|^2.
\]

2.4. The SARSE Algorithm

The novel algorithm, to discriminate between large and small reflectors, is named as scattering based reflector size estimator (SARSE). Different from the well-known scattering coefficient [18], we define here the partial reflection scattering coefficient \( K \). This measures the partial amount of sound that is scattered away from the specular reflection direction, by observing it from a single listening position. SARSE’s assumptions are: reflectors are not perfectly smooth; small objects are not closely spaced.

Reflections are segmented out from the beamformed signals, through a Hamming window of length \( D \) centered at \( n_{e,l} \). The specular component energy is calculated by using the beamformed reflection segment associated to \( \theta_{e,l} \) and \( \phi_{e,l} \):

\[
E_{e,l}^{\text{sp}} = \sum_{n=n_{e,l}-D}^{n_{e,l}+D} |r_{\text{B}}(n, \theta = \theta_{e,l}, \phi = \phi_{e,l})|^2.
\]

The energy of the partial scattered reflections is obtained by using the beamformed reflection segments \( \pm \alpha \) around the specular component:

\[
E_{e,l}^{\text{sc}} = \sum_{n=n_{e,l}-D}^{n_{e,l}+D} |r_{\text{B}}(n, \theta = \theta_{e,l}, \phi = \phi_{e,l})|^2.
\]

The energy difference between specular and scattered energy is an indicator of the reflection size. The SDA beam size, at 2 kHz, is 5°, we suggest \( \alpha = 20° \), i.e. four times it. Partial scattered energies, for a large and a small reflector, are shown Fig. 4. The partial reflection scattering coefficient is \( K_{e,l} = 1 - (E_{e,l}^{\text{sp}}/E_{e,l}^{\text{sc}}) \). By setting a threshold \( K_{\text{th}} \), a binary classification is performed, with output \( C_{e,l} \): if a reflector is large, then \( K_{e,l} > K_{\text{th}} \) (\( C_{e,l} = 1 \)), since scattered components are received at the listener from around the specular reflection direction; if a reflector is small, this does not happen, since no surface exists around the specular reflective area, thus \( K_{e,l} < K_{\text{th}} \) (\( C_{e,l} = 0 \)).

2.5. Reflector Characterization

By combining the localization and classification outputs, reflectors can be finally characterized. From the estimated \( n_{e,l} \), the distance between the reflection image sources and the microphone array are obtained as \( \rho_{e,l} = n_{e,l} c_0 \). By also having \( \theta_{e,l} \) and \( \phi_{e,l} \), the image sources are localized in spherical coordinates, as it was done in our image source direction and ranging (ISAR) [16]. The loudspeaker image bisection (LIB) [27] is then used to calculate \( M_{e,l} \), that is the midpoint between the image source position, in Cartesian coordinates, \( B_{e,l} \) and the loudspeaker \( B_{0,l} \): \( M_{e,l} = (B_{e,l} + B_{0,l})/2 \).

Midpoints that are related to reflections classified as large \( M_{e,l}, C_{e,l} = 1 \) are converted to planes by employing our ISAR-LIB method (for a detailed description, please, refer to [16]). Looking towards the six spatial directions, the closest planes to the microphone array are selected to generate the room shoebox model. If \( L \) loudspeakers are used, the shoebox estimation is refined, by applying our mean-ISAR-LIB [16], which averages each room boundary over its \( L \) estimations. Finally, the remaining midpoints \( M_{e,l}, C_{e,l} = 0 \), that are related to small reflectors, are modeled as cuboids of dimensions ten times smaller than the shoe box in which they are in. This size was chosen to approximate a typical furniture dimension.

3. EXPERIMENTS AND RESULTS

SARSE and its application to the new localization method are evaluated. Experiments aim to demonstrate the reflector classification importance, for dynamic listener applications.

3.1. The Datasets

RIRs were recorded in four rooms. A bi-circular array of microphones was employed. Each circle, of radii 85 cm and 106 cm, respectively, is composed of 24 microphones, for a total of 48. Prior knowledge about the reflector positions was used to solve the up-down ambiguity generated by planar arrays [16]. The room dimensions, reverberation times (RT60s) averaged between 500 Hz and 4 kHz, and loudspeaker positions (with respect to the microphone array) are reported on the left side of Tab. 1. All the recorded RIRs, together with relative detailed documentation, can be downloaded from [28, 29]. The first room, named as “BBC LR”, is a listening environment at the BBC MediaCity, in Salford, UK, where the interior objects were limited to the recording devices [29]. The second room is named as “BBC UL”. It is a lab at the BBC MediaCity, furnished to reproduce a living room-like environment [29]. “Vislab” was an acoustically treated lab at the University of Surrey, where the Surrey Sound Sphere was built (i.e. the main interior object) [28]. The last dataset is “VML”. It was a mock room built in a lab at the University of Surrey, with one wall and ceiling missing, like a film set [28]. The interior objects were the recording devices.

3.2. Evaluation Metrics

The first part of the experiments evaluates SARSE’s classification performance. Four metrics are used [30]: the precision,
that is calculated as $\frac{TP}{TP+FP}$, where TP stands for true positive and FP for false positive; the recall that is $\frac{TP}{TP+FN}$, where FN means false negative; the accuracy $\frac{TP+TN}{TP+FN+FP+FN}$, with TN being true negative; the $F_1$ Score, that is $\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$.

For the acoustic reflector localization evaluation, the Euclidean distances between the image source positions (related to the estimated reflectors), and their groundtruth are calculated to obtain the errors $\epsilon_e$. The first evaluated error is the average over the $E$ room boundaries, that were found without using SARSE: $\epsilon = \frac{1}{E} \sum_{e=1}^{E} \epsilon_e$. The same kind of error is calculated by introducing SARSE, and named as $\epsilon_{\text{SARSE}}$.

### 3.3. Classification and Localization Performance

Results at the center of Table 1 shows SARSE’s high performance, with an average, over the four datasets, that is always above 90%. It works perfectly in both BBC LR and Vislab. SARSE classifies as small objects the recording devices in BBC LR, and the Surrey Sound Sphere bars, in Vislab. This suggests robustness of SARSE also with non-flat objects. In BBC UL, although having a recall of 100%, the other metrics show some errors. These are caused by a coffee table, placed between one loudspeaker and the microphone array, which occludes the floor. Therefore, SARSE recognizes this table as being large. VML shows decent performance for most of the metrics. However, the accuracy is about 70%. This is caused by the floor, i.e. the furthest planar reflector. There, SARSE mistakenly classifies first order reflections as belonging to small objects, whereas higher order ones as planes.

The advantage of using SARSE is underlined by non-static listener scenarios, for applications such as AR, or next generation spatial audio production. By classifying a reflector, the image source position can be correctly dynamically readjusted with the listener movements. In fact, small reflectors are modeled as small cuboids instead of planes. To evaluate this performance gain, the errors made by localizing the reflector, with (i.e. $\epsilon_{\text{SARSE}}$) and without (i.e. $\epsilon$) SARSE, are compared on the right side of Table 1. At first glance, these errors may look greater than the ones in the literature [16, 31]. However, there, just the first arriving reflections were analyzed. Here, the error is averaged over all the estimated reflectors composing the room shoebox. Furthermore, here, gross and fine errors are not discriminated. In BBC LR, BBC UL and Vislab, the improvement is remarkable. Without SARSE, in BBC UL, three planes would be generated at the position of two sofas and a computer screen (see Fig. 5). In Vislab, several planes would surround the listener at the position of the Surrey Sound Sphere. However, SARSE classifies a second-order reflection as the VML’s floor. The related plane lies far from the groundtruth, increasing $\epsilon_{\text{SARSE}}$. In preliminary tests on a large room (RT60 = 910 ms), SARSEs classification performance was similar to that in the VML (precision = 50%, recall = 100%, accuracy = 71%, $F_1$ score = 0.67). However, the error in reflector localization, $\epsilon$, was large. Further work is needed to investigate reflector localization in large reverberant rooms for application to concert halls.

### 4. CONCLUSION

A novel pipeline, that localizes and classifies acoustic reflectors, has been proposed. Experiments were performed by employing four different rooms, showing the high performance of the novel classification algorithm SARSE and the importance of its integration within the novel reflector localization method. Future work may focus on extending the analysis to a greater room number. Moreover, alternative reflector classification methods may be studied. Investigations about combinations with vision based methods may also be carried out.
5. REFERENCES


