AUTOMATIC SOUND SOURCE LOCALIZATION FOR OBJECT-BASED AUDIO RECORDING

N. Epain and J. Daniel

b<>com, France

ABSTRACT

Together with nine European partners, b<>com is participating in the ORPHEUS research project, which aims to invent new workflows for producing, broadcasting and playing back object-oriented audio content. Among the many challenges associated with the adoption of object-oriented audio representations by the public and professionals, is the lack of dedicated sound recording techniques and tools. In particular, capturing sound objects together with their associated metadata, such as the object's position, remains difficult. In this paper, we focus on the problem of localizing a sound source by comparing the source signal, as recorded using a spot microphone, to the signals recorded by a distant microphone array. We designed an algorithm for estimating the direction and distance of the spot microphone, relative to the main microphone array. The performance of the algorithm is assessed using both simulated and recorded sound signals.

INTRODUCTION

Object-based audio

Object-based audio [1] is a relatively new paradigm for representing audio content. Traditionally, audio content such as movie soundtracks or radio programs was stored and distributed in a format that corresponded to the device used for sound reproduction. For instance, the music in a compact disc is in a stereo signal format and is assumed to be played back over two loudspeakers distributed accordingly. On the contrary, with object-based audio representations contents are decomposed into distinct objects, which typically define sound sources. These sources are characterised by one or more audio signals and some metadata defining its position, width, whether it belongs to the audio foreground or background, etc.

Object-based audio representations are revolutionary in that they allow more immersive and more interactive user experiences. More immersive, because 3D audio content represented as objects can be played back optimally for any speaker configuration or rendered binaurally over headphones. More interactive, because they offer the possibility to displace sound sources or to change their level at the user end. Rendering can also be adapted to suit the listening conditions. For instance, the dynamic range of the content can be reduced for hearing-impaired listeners, when listening in noisy environments.

The ORPHEUS project

The ORPHEUS project [2] is a research project funded by the European Commission under the Horizon 2020 programme. Its goal is to prepare the future of audio broadcasting with a highlight on object-based audio representations. Gathering broadcasters, researchers and engineers from four European countries (UK, Germany, France and the Netherlands), the project aims to define new tools, workflows and standards for producing, archiving, broadcasting and consuming object-based audio experiences.

A particular focus of ORPHEUS is the integration of the different steps along the broadcast workflow, from production to playback. As such, several object-based radio programmes are to be produced and broadcast during the project. Within the project, b<>com is primarily involved in tasks related to audio content production and quality of user experience.

Challenges in object-based audio recording

The shift from the channel-based audio paradigm to the emerging object-based and scene-based representations produces numerous challenges in audio engineering, ranging from content archiving and exchange, to rendering techniques at the user end. Among these challenges is how to produce object-based audio content. For decades, sound engineers have been perfecting recording and mixing techniques that are specifically suited to channel-based audio formats. In order to capture sound objects, new tools and algorithms must be designed. Ideally, these tools should be reliable and intuitive enough that sound professionals have no major difficulty adapting to the new audio production processes.

In this paper, we focus on the problem of acquiring object metadata. More specifically, we discuss capturing the position of sound sources in the following context: one or more sound sources (typically actors, singers or music instruments) are equipped with proximity microphones; in addition, a spherical microphone array is used to capture the entire sound scene from a given point of view. This scenario could be that of a radio drama, for instance, where the actors would typically be recorded while performing together in a studio. The signals recorded using the microphone array would then be used as a “bed” consisting of the different sound sources, reverberation and other ambient noises. In this situation, localising the position of the sound sources would serve two purposes. First, knowing the direction in which a source is located with regard to the microphone array, as well as the propagation delay between the signals recorded by the spot microphone and that recorded by the microphone array, allows the sound engineer to mix these signals in a coherent manner. Second, one could imagine exporting the source position and delay parameters as object metadata accompanying the signals.

Note that, in this study, we only consider the problem of blindly estimating the source positions and delays. In other words, we make the hypothesis that no other source of information than the audio signals is available. In the future, it can be envisioned that object-based audio capturing will be facilitated by the use of infrared or electromagnetic tracking devices. However, we believe there is merit in determining the source positions based on the audio data, because such method could be used retrospectively to process recordings that were made prior to using tracking devices.

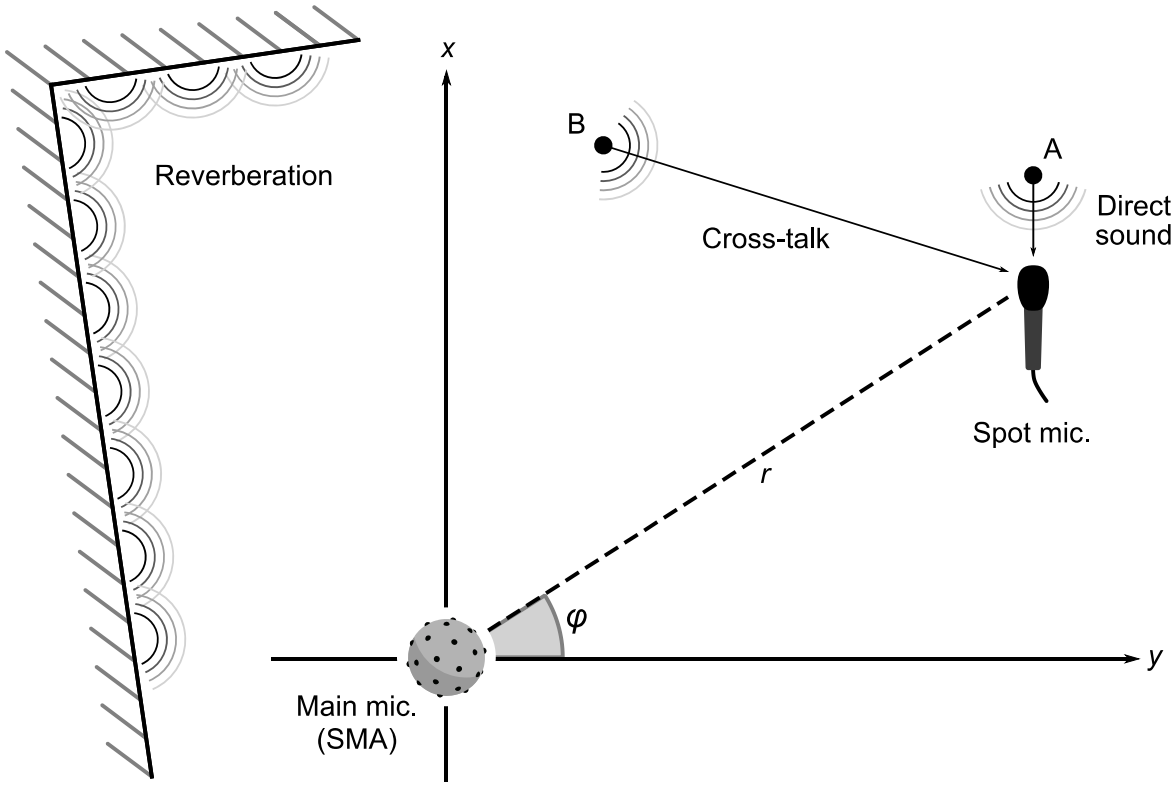

Figure 1 - Problem setup: localize sound source A using the recordings provided by the "Main" and "Spot" microphones (refer to text for more details).

METHODS

Problem formulation

Consider the scenario described by Figure 1. A spherical microphone array (SMA) is used to record a scene comprised of several sound sources (A, B), in an environment that may include walls. At least one of the sound sources is equipped with a proximity microphone. In the following we refer to the SMA as the “main” microphone and to the proximity microphone as the “spot” microphone. Our aim is to analyse the signals recorded by the main and spot microphones to determine the direction and distance of source A relative to the main microphone.

In this scenario, signals recorded by the main microphone are processed and converted to a spherical harmonic representation, i.e. in ambisonics’ B-format or as Higher-Order Ambisonics (HOA) signals up to order *L* [3]. Spherical harmonic representations constitute an elegant mathematical framework for signals recorded with SMAs.

Note that, for simplicity, we make the hypothesis that source A is close enough to the spot microphone that its position relative to the main microphone is approximately the same as that of the spot microphone. In other words, assuming source A is an actor or singer, we assume that her/his mouth is close enough to the spot microphone that the corresponding shift in position would be negligible when heard from the main microphone location. We

also assume that the signal recorded by the spot microphone is loud and clear, that is, we neglect the presence of self-noise in the spot microphone signal.

Because our aim is to determine source A's location solely based on the recorded audio signals, the success of the task primarily depends on the quality of these signals. More specifically, the following factors are expected to have a significant impact:

- The complexity of the scene, i.e. the number of sources comprising the scene and the relative strength of source A in this scene;
- The presence of ambient noise or reverberation in the main microphone signals;
- The presence of crosstalk between sources: the signal emitted by source B can be picked up by the spot microphone with an amplitude that depends on the distance separating the sources.

Source localization algorithm

In previous work [4], we presented an algorithm to determine the position of source A by analysing the signals recorded by a main microphone in B-format, i.e. $w(t)$, $x(t)$, $y(t)$ and $z(t)$, together with that recorded by a spot microphone, $s(t)$. The algorithm, which we refer to as "MainSpot" is summarised in Figure 2. First, the delay between the main and spot signals is estimated by cross-correlating the "omni" signal $w(t)$ with the spot signal $s(t)$. Next, the estimated delay is applied to the spot signal so that it is in phase with the part of the main signals that correspond to source A. The delayed spot signal is then projected on the x, y and z "main" signals to obtain a vector that points to the estimated source direction. Lastly, the direction vector's coordinates can be converted to elevation and azimuth values θ and φ.

One advantage of the MainSpot algorithm is that it requires very little computational power and therefore it can be implemented in real-time very easily. However, testing with signals recorded during music and drama performances revealed that it was not very robust in the presence of multiple sources and reverberation. We identified several possible causes for this issue and designed a new algorithm which improves the robustness of the estimation in actual recording conditions. In the following, we present results demonstrating the performance of this new algorithm.

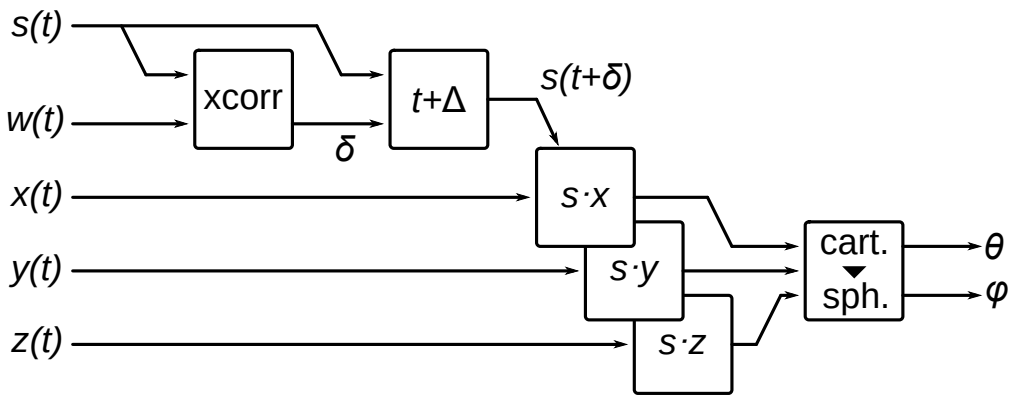

Figure 2 – Flow diagram of the MainSpot algorithm.

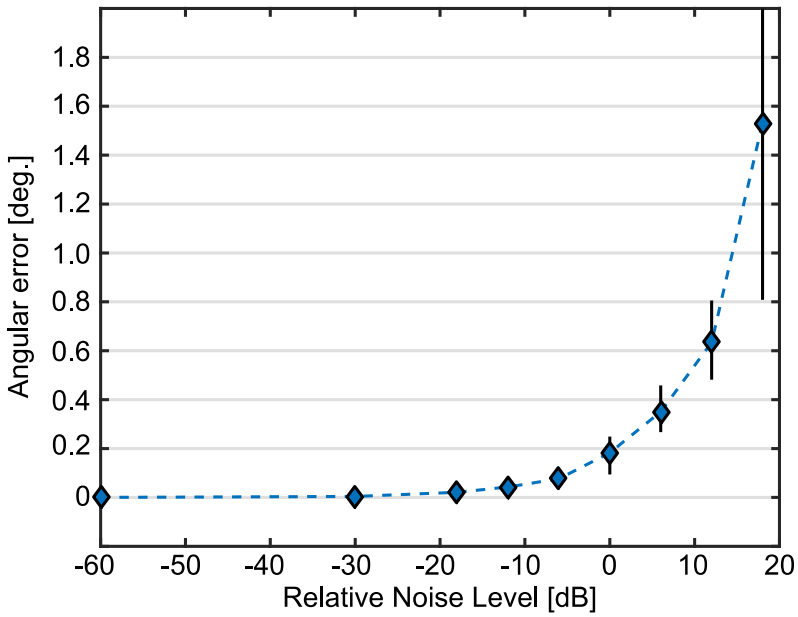

Figure 3 – Source direction estimation error in the presence of diffuse noise.

RESULTS

Numerical simulations

We assessed the performance of our source position estimation method via numerical simulations. In a first simulation, we tested the robustness of our method to the presence of noise in the main signals. In the simulation, there was only one sound source. The source signal was a clean, anechoic, speech recording. This signal modelled the spot microphone. The main microphone signals, **m**(t), were calculated as follows:

$$\mathbf{m}(t) = \mathbf{y} \cdot s(t+\delta) + \mathbf{n}(t)$$

where $s(t)$ is the spot signal, **y** is the vector of the HOA components corresponding to the source direction, δ is the source propagation delay and **n**(t) represents the noise signals. The main factor that varied in the simulation was the relative amount of noise in the main signals, which we define as the ratio of the energy of **n**(t) over the energy of $\mathbf{y} \cdot s(t)$. For each noise level under assessment, 20 trials were run. For each trial, a source position and a delay value were randomly picked, and noise signals were randomly generated with the desired noise level. Note that the noise signals consisted of perfectly uncorrelated Gaussian white noise, which corresponds to a diffuse sound field situation. Lastly, note that the main microphone was modelled as an ideal SMA providing HOA signals up to order 3.

The results of the simulation are presented in Figure 3. Diamond symbols represent the average error in the estimation of the source direction, while the vertical bars represent the interquartile range over the different trials. Up to a noise level of 20 dB, which means that the signal emitted by source A constitute less than a tenth of the main signal energy, the error in the estimation of the source direction is less than 2 degrees. Although the results

are not displayed here, the delay estimation was perfect for every trial and every noise level (0 sample error).

In a second simulation, we tested the robustness of the algorithm to the presence of an interfering source, denoted B. The spot signal was then calculated as:

$$s(t) = a(t) + \beta \cdot b(t)$$

where $a(t)$ and b(t) denote the signals corresponding to sources A and B, respectively, and β is a gain factor varying from -36 to -6 dB. Both source signals consisted of clean, anechoic male and female speech with equal RMS level. Assuming source A was located 10 cm from the spot microphone, these gain values would correspond to source B being located 10 m and 20 cm from the spot microphone, respectively. For simplicity, we assumed that sources A and B were located at the same distance from the main microphone. The main microphone signals were thus calculated as:

$$\mathbf{m}(t) = \mathbf{y}_A \cdot a(t+\delta) + \mathbf{y}_B \cdot b(t+\delta) + \mathbf{n}(t)$$

where $\mathbf{y}_A$ and $\mathbf{y}_B$ are the vectors of the HOA components for sources A and B, respectively. The relative amount of diffuse noise was set to -10 dB. Lastly, similar to the previous simulation, we ran the algorithm for 20 random trials for each value of β.

Results obtained with recorded signals

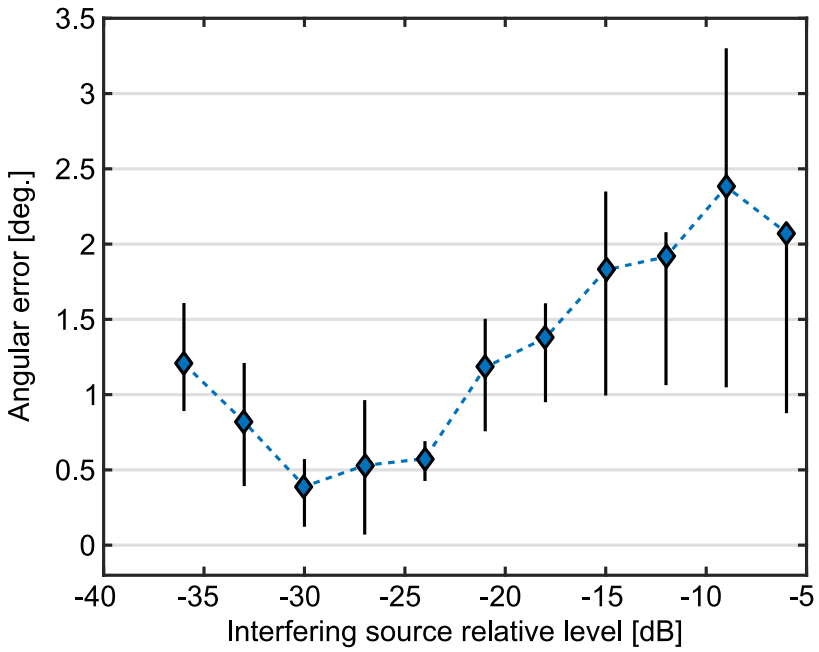

Figure 4 – Source direction estimation error in the presence of an interfering source.

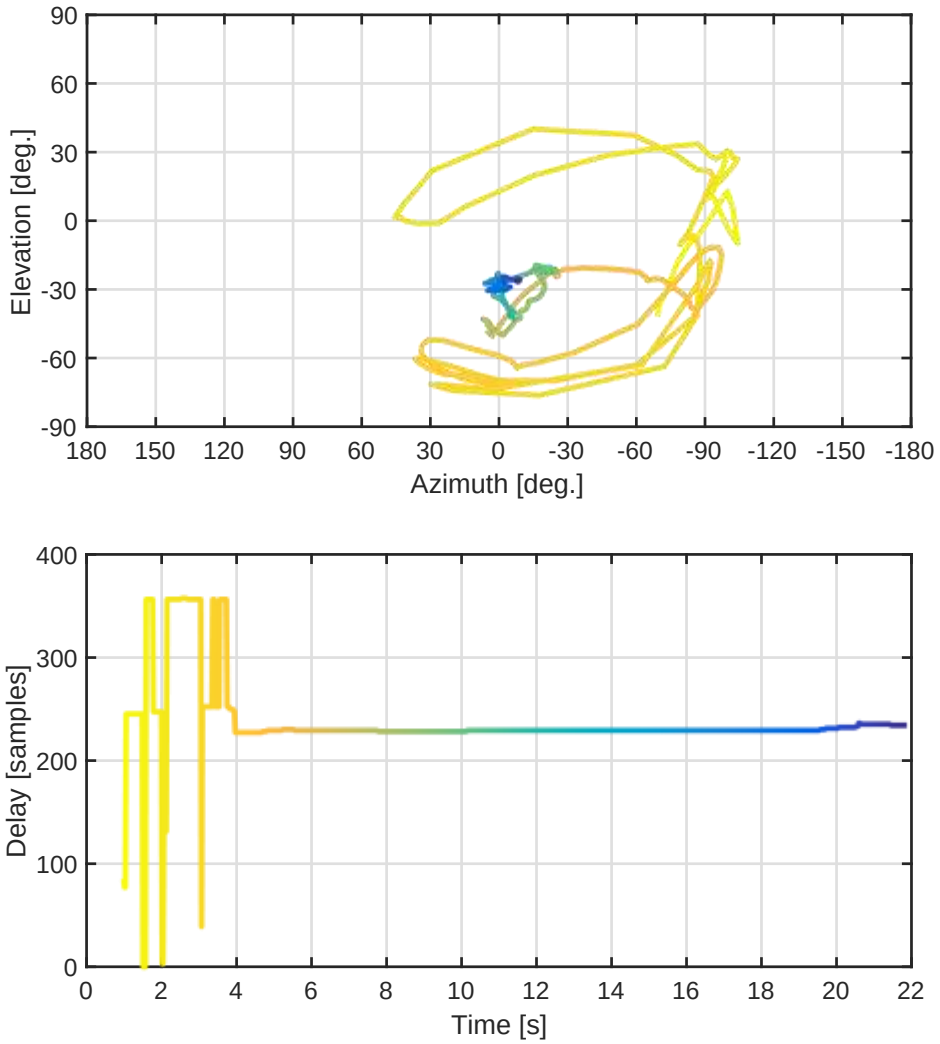

Figure 5 – Output of the source localisation algorithm in the case of recorded signals.

The simulation results are presented in Figure 4. The error in the estimation of source A's direction is of the order of a few degrees for every level of interference from source B. Again, the error in the estimation of the delay was zero samples in every trial.

In order to assess the performance of our source localisation algorithm, we used it to analyse files recorded during a musical performance. The recording setup was the following. The musicians were distributed around a circle, about two metres in radius. Each sound source was recorded with a proximity microphone. In addition, a 32-channel SMA was located at the centre of the circle, approximately 2 metres above the floor. Directly in front of the SMA was located a seated singer. On the sides were a clarinet, cello, harp, flute and percussions.

Results of the analysis are presented in Figure 5. During the first 4 seconds of the recording, both the direction and delay estimations are unstable. This is because the singer has not started singing at this time, and the spot microphone picks up signals originating from the percussions and cello. After the singer starts singing (4s onwards), the delay estimation becomes very stable at approximately 230 samples and remains so until the end of the recording. It takes a bit longer for the estimated source direction to converge to the expected value (0° azimuth and -30° elevation). Nevertheless, after convergence (12s onwards), the estimated source direction remains within 5 to 10 degrees of the expected direction.

ALGORITHM IMPLEMENTATION

We implemented the original version of the MainSpot algorithm (Figure 2) in a VST plugin prototype. In its current state the plugin displays the estimated source direction (azimuth and elevation) and delay. These values can then be used by a sound engineer to pan the signals recorded by spot microphones. A possible evolution for this plugin would be to output the panned spot signals in the HOA format.

Regarding the new and more accurate version of the algorithm, it is computationally costlier than the original MainSpot algorithm and therefore it is not very well suited for a real-time audio plugin. Our objective is now to integrate this algorithm into standalone post-production software that will import “main” and “spot” recordings and export sound object metadata, including the estimated source positions with time. These metadata could be exported as Audio Definition Model (ADM) data and added to the headers of the spot microphone recording files, for instance.

CONCLUSIONS AND PERSPECTIVES

In this paper, we discussed how to localise sound sources by analysing microphone signals in a 3D audio recording context. We designed a localisation algorithm and demonstrated that this algorithm could provide accurate estimations of the source location and corresponding propagation delay in real recording conditions.

One remaining issue with the algorithm is that it takes a relatively long time to converge. In the example illustrated in Figure 5, the estimated source position stabilises about 8 s after the singer starts singing. In future work, we will investigate how to deal with sections when the source that is recorded by the spot microphone is not playing or when the source localization has not finished converging. On envisioned improvement is to add a source activity detection module. During the “inactive” sections the source parameters would then be deduced from the results obtained in the active sections.

ACKNOWLEDGEMENTS

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 687645).

REFERENCES

1. Parmentier, M. 2015. Object-Based Audio: The Next Big Turn. J. Audio Eng. Soc. Volume 63. pp. 659 to 660.

2. The ORPHEUS Consortium. 2015. The ORPHEUS project webpage. https://orpheus-audio.eu.

3. Moreau, S., Daniel, J., and Bertet, S. 2006. 3D sound field recording with higher order Ambisonics – Objective measurements and validation of a 4th order spherical microphone. Proceedings of the AES 120th Convention, May 2006, Paris, France.

4. Fedosov, A., Pallone, G., Daniel, J. and Marchand S. 2015. Automatic HOA Mixing. Proceedings of the third International Conference on Spatial Audio (ICSA), September 2015, Graz, Austria.