3D-Audio Matting, Post-editing and Re-rendering from Field Recordings

Emmanuel Gallo[1,2], Nicolas Tsingos[1] and Guillaume Lemaitre[1]
[1]REVES/INRIA and [2]CSTB,
Sophia-Antipolis, France*

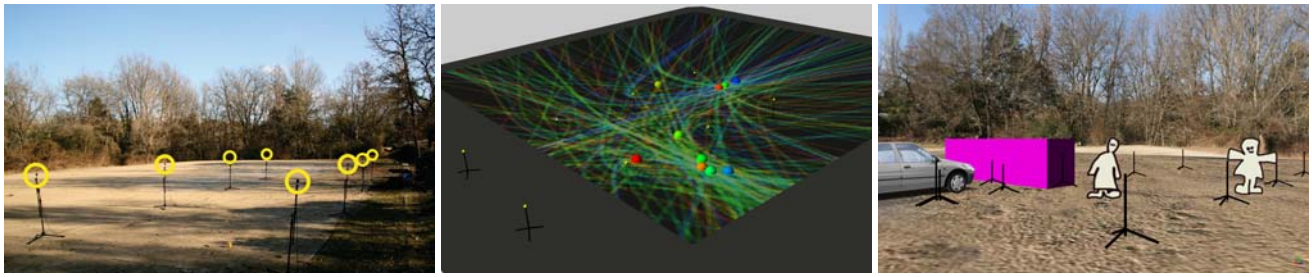


Figure 1: Left: We use multiple arbitrarily positioned microphones (circled in yellow) to simultaneously record real-world auditory environments. Middle: We analyze the recordings to extract the positions of various sound components through time. Right: This high-level representation allows for post-editing and re-rendering the acquired soundscape within generic 3D-audio rendering architectures.

Abstract

We present a novel approach to real-time spatial rendering of realistic auditory environments and sound sources recorded live, in the field. Using a set of standard microphones distributed throughout a real-world environment we record the sound-field simultaneously from several locations. After spatial calibration, we segment from this set of recordings a number of auditory components, together with their location. We compare existing time-delay of arrival estimations techniques between pairs of widely-spaced microphones and introduce a novel efficient hierarchical localization algorithm. Using the high-level representation thus obtained, we can edit and re-render the acquired auditory scene over a variety of listening setups. In particular, we can move or alter the different sound sources and arbitrarily choose the listening position. We can also composite elements of different scenes together in a spatially consistent way. Our approach provides efficient rendering of complex soundscapes which would be challenging to model using discrete point sources and traditional virtual acoustics techniques. We demonstrate a wide range of possible applications for games, virtual and augmented reality and audio-visual post-production.

Keywords: Virtual Environments, Spatialized Sound, Audio Recording Techniques, Auditory Scene Analysis, Image-based rendering, Matting and compositing

*{Emmanuel.Gallo|Nicolas.Tsingos}@sophia.inria.fr
http://www-sop.inria.fr/reves/
Guillaume Lemaitre is now with IRCAM.

1 Introduction

While hardware capabilities allow for real-time rendering of increasingly complex environments, authoring realistic virtual audio-visual worlds is still a challenging task. This is particularly true for interactive spatial auditory scenes for which few content creation tools are available.

Current models for authoring interactive 3D-audio scenes often assume that sound is emitted by a set of monophonic point sources for which a signal has to be individually generated. In the general case, source signals cannot be completely synthesized from physics-based models and must be individually recorded, which requires enormous time and resources. Although this approach gives the user the freedom to control each source and freely navigate throughout the auditory scene, the overall result remains an approximation due to the complexity of real-world sources, limitations of microphone pick-up patterns and limitations of the simulated sound propagation models.

On the opposite end of the spectrum, spatial sound recordings which encode directional components of the sound-field can be directly used to acquire live auditory environments as a whole [44, 66]. They produce lifelike results but offer little control, if any, at the playback end. In particular, they are acquired from a single location in space, which makes them insufficient for walkthrough applications or rendering of large near-field sources. In practice, their use is mostly limited to the rendering of an overall ambiance. Besides, since no explicit position information is directly available for the sound sources, it is difficult to tightly couple such spatial recordings with matching visuals.

This paper presents a novel analysis-synthesis approach which bridges the two previous strategies. Our method builds a higher-level spatial description of the auditory scene from a set of field recordings (Figure 1). By analyzing how different frequency components of the recordings reach the various microphones through time, it extracts both spatial information and audio content for the most significant sound events present in the acquired environment. This spatial mapping of the auditory scene can then be used for post-processing and re-rendering the original recordings. Re-rendering is achieved through a frequency-dependent warping of the recordings, based on the estimated positions of several frequency subbands of the signal. Our approach makes positional

information about the sound sources directly available for generic 3D-audio processing and integration with 2D or 3D visual content. It also provides a compact encoding of complex live auditory environments and captures complex propagation and reverberation effects which would be very difficult to render with the same level of realism using traditional virtual acoustics simulations.

Our work complements image-based modeling and rendering approaches in computer graphics [16, 28, 12, 5]. Moreover, similar to the *matting* and *compositing* techniques widely used in visual effects production [54], we show that the various auditory components segmented out by our approach can be pasted together to create novel and spatially consistent soundscapes. For instance, foreground sounds can be integrated in a different background ambiance.

Our technique opens many interesting possibilities for interactive 3D applications such as games, virtual/augmented reality or off-line post-production. We demonstrate its applicability to a variety of situations using different microphone setups.

2 Related work

Our approach builds upon prior work in several domains including spatial audio acquisition and restitution, structure extraction from audio recordings and blind source separation. A fundamental difference between the approaches is whether they attempt to capture the spatial structure of the wavefield through mathematical or physical models or attempt to perform a higher-level auditory scene analysis to retrieve the various, perceptually meaningful, sub-components of the scene and their 3D location. The following sections give a short overview of the background most relevant to our problem.

2.1 Spatial sound-field acquisition and restitution

Processing and compositing live multi-track recordings is of course a widely used method in motion-picture audio production [73]. For instance, recording a scene from different angles with different microphones allows the sound editor to render different audio perspectives, as required by the visual action. Thus, producing synchronized sound-effects for films requires carefully planned microphone placement so that the resulting audio track perfectly matches the visual action. This is especially true since the required audio material might be recorded at different times and places, before, during and after the actual shooting of the action on stage. Usually, simultaneous monaural or stereophonic recordings of the scene are composited by hand by the sound designer or editor to yield the desired track, limiting this approach to off-line post-production. Surround recording setups (e.g., *Surround Decca Trees*) [67, 68], which historically evolved from stereo recording, can also be used for acquiring a sound-field suitable for restitution in typical cinema-like setups (e.g., 5.1-surround). However, such recordings can only be played-back directly and do not support spatial post-editing.

Other approaches, more physically and mathematically grounded, decompose the wavefield incident on the recording location on a basis of spatial harmonic functions such as spherical/cylindrical harmonics (e.g., *Ambisonics*) [25, 44, 18, 38, 46] or generalized Fourier-Bessel functions [36]. Such representations can be further manipulated and decoded over a variety of listening setups. For instance, they can be easily rotated in 3D space to follow the listener's head orientation and have been successfully used in immersive virtual reality applications. They also allow for beamforming applications, where sounds emanating from any specified direction can be further isolated and manipulated. However, these techniques are practical mostly for low order decompositions (order 2 already requiring 9 audio channels) and, in return, suffer from limited directional accuracy [31]. Most of them also require specific microphones [2, 48, 66, 37] which are not widely available and whose bandwidth usually drops when the spatial resolution increases. Hence, higher-order microphones do not usually deliver production-grade audio quality, maybe with the exception of *Trinnov's SRP* system [37] (www.trinnov.com) which uses regular studio microphones but is dedicated to 5.1-surround restitution. Finally, a common limitation of these approaches is that they use coincident recordings which are not suited to rendering walkthroughs in larger environments.

Closely related to the previous approach is wave-field synthesis/holophony [9, 10]. Holophony uses the Fresnel-Kirchoff integral representation to sample the sound-field inside a region of space. Holophony could be used to acquire live environments but would require a large number of microphones to avoid aliasing problems, which would jeopardize proper localization of the reproduced sources. In practice, this approach can only capture a live auditory scene through small acoustic "windows". In contrast, while not providing a physically-accurate reconstruction of the sound-field, our approach can provide stable localization cues regardless of the frequency and number of microphones.

Finally, some authors, inspired from work in computer graphics and vision, proposed a dense sampling and interpolation of the *plenacoustic function* [3, 20] in the manner of *lumigraphs* [26, 39, 12, 5]. However, these approaches remain mostly theoretical due to the required spatial density of recordings. Such interpolation approaches have also been applied to measurement and rendering of reverberation filters [53, 27]. Our approach follows the idea of acquiring the plenacoustic function using only a sparse sampling and then warping between this samples interactively, e.g., during a walkthrough. In this sense, it could be seen as an "unstructured plenacoustic rendering".

2.2 High-level auditory scene analysis

A second large family of approaches aims at identifying and manipulating the components of the sound-field at a higher-level by performing auditory scene analysis [11]. This usually involves extracting spatial information about the sound sources and segmenting out their respective content.

Spatial feature extraction and restitution

Some approaches extract spatial features such as binaural cues (interaural time-difference, interaural level difference, interaural correlation) in several frequency subbands of stereo or surround recordings. A major application of these techniques is efficient multi-channel audio compression [8, 23] by applying the previously extracted binaural cues to a monophonic down-mix of the original content. However, extracting binaural cues from recordings requires an implicit knowledge of the restitution system.

Similar principles have also been applied to flexible rendering of directional reverberation effects [47] and analysis of room responses [46] by extracting direction of arrival information from coincident or near-coincident microphone arrays [55].

This paper generalizes these approaches to multi-channel field recordings using arbitrary microphone setups and no *a priori* knowledge of the restitution system. We propose a direct extraction of the 3D position of the sound sources rather than binaural cues or direction of arrival.

Blind source separation

Another large area of related research is blind source separation (BSS) which aims at separating the various sources from one or several mixtures under various mixing models [71, 52]. Most recent BSS approaches rely on a sparse signal representation in some space of basis functions which minimizes the probability that a

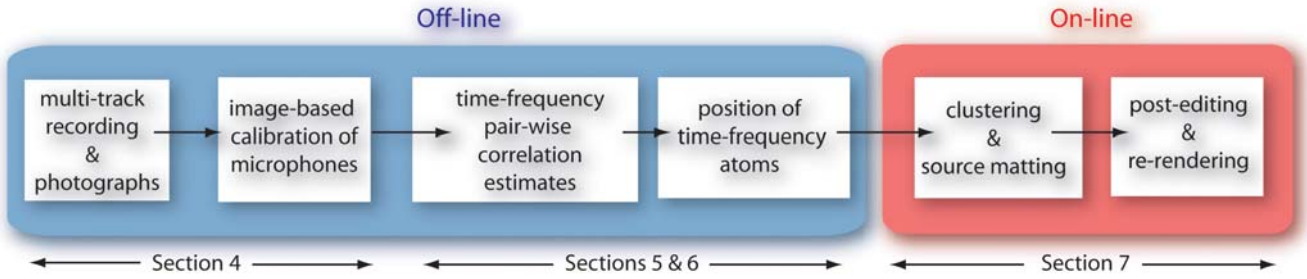


Figure 2: Overview of our pipeline. In an off-line phase, we first analyze multi-track recordings of a real-world environment to extract the location of various frequency subcomponents through time. At run-time, we aggregate these estimates into a target number of clustered sound sources for which we reconstruct a corresponding signal. These sources can then be freely post-edited and re-rendered.

high-energy coefficient at any time-instant belongs to more than one source [58]. Some work has shown that such sparse coding does exists at the cortex level for sensory coding [41]. Several techniques have been proposed such as independent component analysis (ICA) [17, 63] or the more recent *DUET* technique [32, 74] which can extract several sources from a stereophonic signal by building an inter-channel delay/amplitude histogram in Fourier frequency domain. In this aspect, it closely resembles the aforementioned binaural cue coding approach. However, most BSS approaches do not separate sources based on spatial cues, but directly solve for the different source signals assuming *a priori* mixing models which are often simple. Our context would be very challenging for such techniques which might require knowing the number of sources to extract in advance, or need more sensors than sources in order to explicitly separate the desired signals. In practice, most auditory BSS techniques are devoted to separation of speech signals for telecommunication applications but other audio applications include upmixing from stereo to 5.1 surround formats [6].

In this work, however, our primary goal is not to finely segment each source present in the recorded mixtures but rather to extract enough spatial information so that we can modify and re-render the acquired environment while preserving most of its original content. Closer in spirit, the DUET technique has also been used for audio interpolation [57]. Using a pair of closely spaced microphones, the authors apply DUET to re-render the scene at arbitrary locations along the line passing through the microphones. The present work extends this approach to arbitrary microphone arrays and rerendering at any 3D location in space.

3 Overview

We present a novel acquisition and 3D-audio rendering pipeline for modeling and processing realistic virtual auditory environments from real-world recordings.

We propose to record a real-world soundscape using arbitrarily placed omnidirectional microphones in order to get a good acoustic sampling from a variety of locations within the environment. Contrary to most related approaches, we use widely-spaced microphone arrays. Any studio microphones can be used for this purpose, which makes the approach well suited to production environments. We also propose an image-based calibration strategy making the approach practical for field applications. The obtained set of recordings is analyzed in an off-line pre-processing step in order to segment various auditory components and associate them with the position in space from which they were emitted. To compute this spatial mapping, we split the signal into short time-frames and a set of frequency subbands. We then use classical time-difference of arrival techniques between all pairs of microphones to retrieve a position for each subband at each time-frame. We evaluate the performance of existing approaches in our context and present an improved hierarchical source localization technique from the obtained time-differences.

This high-level representation allows for flexible and efficient on-line re-rendering of the acquired scene, independent of the restitution system. At run-time during an interactive simulation, we use the previously computed spatial mapping to properly warp the original recordings when the virtual listener moves throughout the environment. With an additional clustering step, we recombine frequency subbands emitted from neighboring locations and segment spatially-consistent sound events. This allows us to select and post-edit subsets of the acquired auditory environment. Finally the location of the clusters is used for spatial audio restitution within standard 3D-audio APIs.

Figure 2 shows an overview of our pipeline. Sections 4, 5 and 6 describe our acquisition and spatial analysis phase in more detail. Section 7 presents the on-line spatial audio resynthesis based on the previously obtained spatial mapping of the auditory scene. Finally, Section 8 describes several applications of our approach to realistic rendering, post-editing and compositing of real-world soundscapes.

4 Recording setup and calibration

We acquire real-world soundscapes using a number of omnidirectional microphones and a multi-channel recording interface connected to a laptop computer. In our examples, we used up to 8 identical *AudioTechnica AT3032* microphones and a *Presonus Firepod* firewire interface running on batteries. The microphones can be arbitrarily positioned in the environment. Section 8 shows various possible setups. To produce the best results, the microphones should be placed so as to provide a compromise between the signal-to-noise ratio of the significant sources and spatial coverage.

In order to extract correct spatial information from the recordings, it is necessary to first retrieve the 3D locations of the microphones. Maximum-likelihood autocalibration methods could be used based on the existence of pre-defined source signals in the scene [50], for which the time-of-arrival (TOA) to each microphone has to be determined. However, it is not always possible to introduce calibration signals at a proper level in the environment. Hence, in noisy environments obtaining the required TOAs might be difficult, if not impossible. Rather, we use an image-based technique from photographs which ensures fast and convenient acquisition on location, not requiring any physical measurements or homing device. Moreover, since it is not based on acoustic measurements, it is not subject to background noise and is likely to produce better results. We use REALVIZ *ImageModeler* (www.realviz.com) to extract the 3D locations from a small set of photographs (4 to 8 in our test examples) taken from several angles, but any standard algorithm can be applied for this step [24]. To facilitate the process we

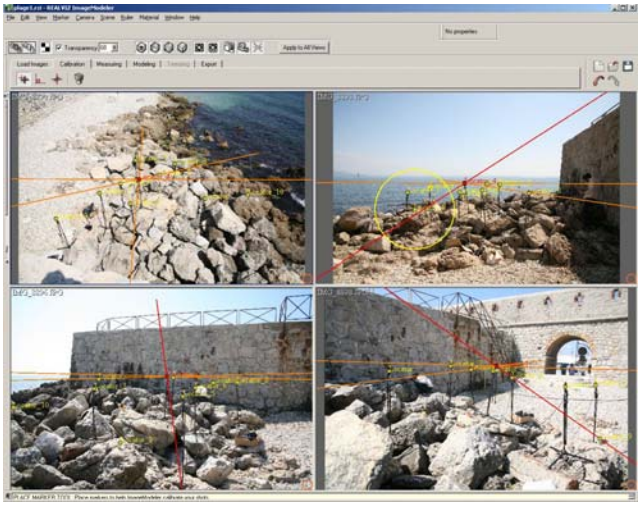


Figure 3: We retrieve the position of the microphones from several photographs of the setup using a commercial image-based modeling tool. In this picture, we show four views of a recording setup, position of the markers and the triangulation process yielding the locations of the microphone capsules.

place colored markers (tape or balls of modeling clay) on the microphones, as close as possible to the actual location of the capsule, and on the microphone stands. Additional markers can also be placed throughout the environment to obtain more input data for calibration. The only constraint is to provide a number of non-coplanar calibration points to avoid degenerate cases in the process. In our test examples, the accuracy of the obtained microphone locations was of the order of one centimeter. Image-based calibration of the recording setup is a key aspect of our approach since it allows for treating complex field recording situations such as the one depicted in Figure 3 where microphones stands are placed on large irregular rocks on a seashore.

5 Propagation model and assumptions for source matting

From the M recorded signals, our final goal is to localize and re-render a number J of *representative sources* which offer a good perceptual reconstruction of the original soundscape captured by the microphone array. Our approach is based on two main assumptions.

First, we consider that the recorded sources can be represented as point emitters and assume an ideal anechoic propagation model. In this case, the mixture $x_m(t)$ of N sources $s_1(t),..,s_n(t)$ recorded by the m^{th} microphone can be expressed as:

$$x_m(t) = \sum_{n=1}^{N} a_{mn}(t)s_n(t-\delta_{mn}(t)), \tag{1}$$

where parameters $a_{mn}(t)$ and $\delta_{mn}(t)$ are the attenuation coefficients and time-delays associated with the n^{th} source and the m^{th} microphone.

Second, since our environments contain more than one active source simultaneously, we consider K frequency subbands, $K >= J$, as the basic components we wish to position in space at each time-frame (Figure 5 (a)). We choose to use non-overlapping frequency subbands uniformly defined on a Bark scale [49] to provide a more psycho-acoustically relevant subdivision of the audible spectrum (in our examples, we experimented with 1 to 32 subbands).

In frequency domain, the signal x_m filtered in the k^{th} Bark band can be expressed at each time-frame as:

$$Y_{km}(z) = W_k(z)\sum_{t=1}^{T} x_m(t)e^{-j(2\pi zt/T)} = W_k(z)X_m(z), \tag{2}$$

where

$$W_k(f) = \begin{cases} 1 & \frac{25k}{K} < Bark(f) < \frac{25(k+1)}{K} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$Bark(f) = 13atan(\frac{0.76f}{1000}) + 3.5atan(\frac{f^2}{7500^2}), \tag{4}$$

$f = z/Zf_s$ is the frequency in Hertz, f_s is the sampling rate and $X_m(z)$ is the $2Z$-point Fourier transform of $x_m(t)$. We typically record our live signals using 24-bit quantization and $f_s = 44.1KHz$. The subband signals are computed using $Z = 512$ with a Hanning window and 50% overlap before storing them back into time-domain for later use.

At each time-frame, we construct a new representation for the captured soundfield at an arbitrary listening point as:

$$\hat{x}(t) \approx \sum_{j=1}^{J}\sum_{k=1}^{K} \hat{\alpha}^j_{km} y_{km}(t+\hat{\delta}_{km}), \forall m \tag{5}$$

where $y_{km}(t)$ is the inverse Fourier transform of $Y_{km}(z)$, $\hat{\alpha}^j_{km}$ and $\hat{\delta}_{km}$ are correction terms for attenuation and time-delay derived from the estimated positions of the different subbands. The term $\hat{\alpha}^j_{km}$ also includes a matting coefficient representing how much energy within each frequency subband should belong to to each representative source. In this sense, it shares some similarity with the *time-frequency masking* approach of [74].

The obtained representation can be made to match the acquired environment if $K >= N$ and if, following a sparse coding hypothesis, we further assume that the contents of each frequency subband belong to a single source at each time-frame. This hypothesis is usually referred to as *W-disjoint orthogonality* [74] and given N sources $S_1,..,S_N$ in Fourier domain, it can be expressed as:

$$S_i(z)S_j(z) = 0 \;\; \forall i \neq j \tag{6}$$

When the two previous conditions are not satisfied, the representative sources will correspond to a mixture of the original sources and Equ. 5 will lead to a less accurate approximation.

6 Spatial mapping of the auditory scene

In this step of our pipeline, we analyze the recordings in order to produce a high-level representation of the captured soundscape. This high-level representation is a mapping, global to the scene, between different frequency subbands of the recordings and positions in space from which they were emitted (Figure 5).

Following our previous assumptions, we consider each frequency subband as a unique point source for which a single position has to be determined. Localization of a sound source from a set of audio recordings, using a single-propagation-path model, is a well studied problem with major applications in robotics, people tracking and sensing, teleconferencing (e.g, automatic camera steering) and defense. Approaches rely either on time-difference of arrival (TDOA) estimates [1, 34, 30], high-resolution spectral estimation (e.g., MUSIC) [64, 35] or steered response power using a

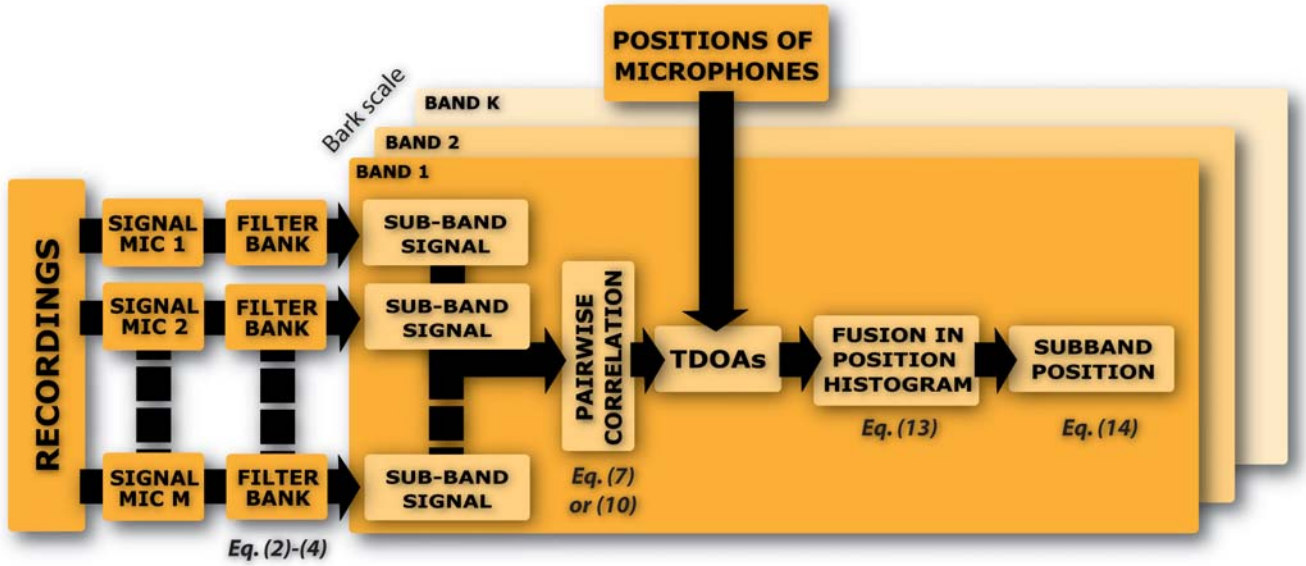


Figure 4: Overview of the analysis algorithm used to construct a spatial mapping for the acquired soundscapes.

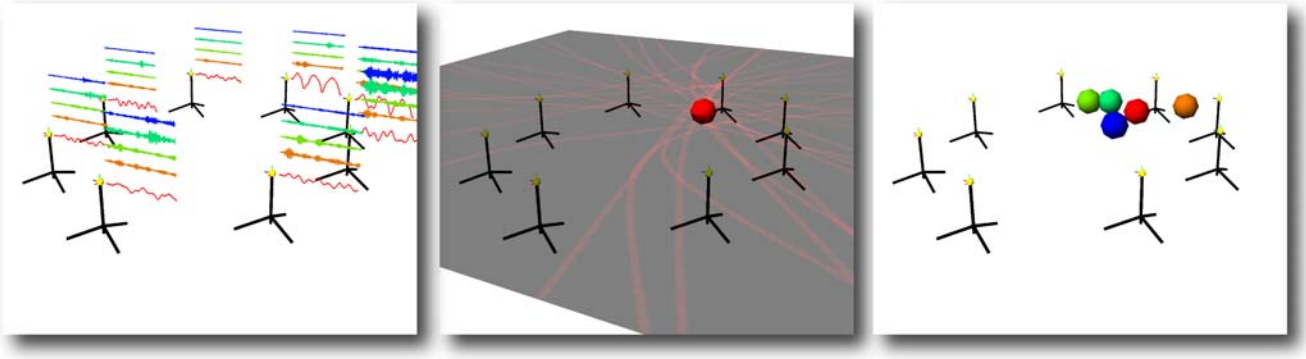


Figure 5: Illustration of the construction of the global spatial mapping for the captured sound-field. (a) At each time-frame, we split the signals recorded by each microphone into the same set of frequency subbands. (b) Based on time-difference of arrival estimation between all pairs of recordings, we sample all corresponding hyperbolic loci to obtain a position estimate for the considered subband. (c) Position estimates for all subbands at the considered time-frame (shown as colored spheres).

beamforming strategy [19, 14, 51]. In our case, the use of freely positioned microphones, which may be widely spaced, prevents from using a beamforming strategy. Besides, such an approach would only lead to direction of arrival information and not a 3D position (unless several beamforming arrays were used simultaneously). In our context, we chose to use a TDOA strategy to determine the location of the various auditory events. Since we do not know the directivity of the sound sources nor the response of the microphones, localization based on level difference cannot be applied.

Figure 4 details the various stage of our source localization pipeline.

6.1 Time-frequency correlation analysis

Analysis of the recordings is done on a frame by frame basis using short time-windows (typically 20ms long or 1024 samples at CD quality). For a given source position and a given pair of microphones, the propagation delay from the source to the microphones generates a measurable time-difference of arrival. The set of points which generate the same TDOA defines an hyperboloid surface in 3D (or an hyperbola in 2D) which foci are the locations of the two microphones (Figure 5 (b)).

In our case, we estimate the TDOAs, $\hat{\tau}_{mn}$, between pairs of microphones $< m,n >$ in each frequency subband k using standard generalized cross-correlation (GCC) techniques in frequency domain [34, 56, 15]:

$$\hat{\tau}_{mn} = \arg\max_{\tau} GCC_{mn}(\tau), \tag{7}$$

where the GCC function is defined as:

$$GCC_{nm}(\tau) = \sum_{z=1}^{Z} \psi_{nm}(z) E\left\{Y_{kn}(z) Y_{km}^{*}(z)\right\} e^{j(2\pi\tau z/Z)}. \tag{8}$$

Y_{kn} and Y_{km} are the $2Z$-point Fourier transforms of the subband signals (see Eq. 2), $E\left\{Y_{kn}(z)Y_{km}^{*}(z)\right\}$ is the cross spectrum and $.^{*}$ denotes the complex conjugate operator.

For the weighting function, ψ, we use the PHAT weighting

which was shown to give better results in reverberant environments [15]:

$$\psi_{mn}(z) = \frac{1}{|Y_n(z)Y_m^*(z)|} \tag{9}$$

Note that phase differences computed directly on the Fourier transforms, e.g. as used in the DUET technique [32, 74], cannot be applied in our framework since our microphones are widely spaced.

We also experimented with an alternative approach based on the average magnitude difference function (AMDF) [46, 13]. The TDOAs are then given as:

$$\hat{\tau}_{nm} = \arg\min_{\tau} AMDF_{nm}(\tau), \tag{10}$$

where the AMDF function is defined as:

$$AMDF_{nm}(\tau) = \frac{1}{Z}\sum_{z=1}^{Z} |y_{kn}(\tau) - y_{km}(k+\tau)| \tag{11}$$

We compute the cross-correlation using vectors of 8192 samples (185 ms at 44.1KHz). For each time-frame, we search the highest correlation peaks (or lowest AMDF values) between pairs of recordings in the time-window defined by the spacing between the corresponding couple of microphones. The corresponding time-delay is then chosen as the TDOA between the two microphones for the considered time-frame.

In terms of efficiency, the complexity of AMDF-based TDOA estimation (roughly $O(n^2)$ in the number n of time-domain samples) makes it unpractical for large time-delays. In our test-cases, running on a *Pentium4 Xeon* 3.2GHz processor, AMDF-based TDOA estimations required about 47 s. per subband for one second of input audio data (using 8 recordings, i.e., 28 possible pairs of microphones). In comparison, GCC-based TDOA estimations require only 0.83 s. per subband for each second of recording.

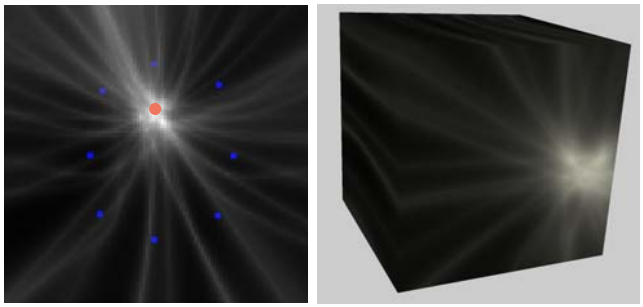
As can be seen in Figure 8, both approaches resulted in comparable subband localization performance and we found both approaches to perform reasonably well in all our test cases. In more reverberant environments, an alternative approach could be the adaptive eigenvalue decomposition [30]. From a perceptual point-of-view, listening to virtual re-renderings, we found that the AMDF-based approach lead to reduced artifacts, which seems to indicate that subband locations are more perceptually valid in this case. However, validation of this aspect would require a more thorough perceptual study.

6.2 Position estimation

From the TDOA estimates, several techniques can be used to estimate the location of the actual sound source. For instance, it can be calculated in a least-square sense by solving a system of equations [30] or by aggregating all estimates into a probability distribution function [60, 1]. Solving for possible positions in a least-square sense lead to large errors in our case, mainly due to the presence of multiple sources, several local maxima for each frequency subband resulting in an averaged localization. Rather, we choose the latter solution and compute a histogram corresponding to the probability distribution function by sampling it on a spatial grid (Figure 6) whose size is defined according to the extent of the auditory environment we want to capture (in our various examples, the grid covered areas ranging from 25 to 400 m^2). We then pick the maximum value in the histogram to obtain the position of the subband.

For each cell in the grid, we sum a weighted contribution of the distance function $D_{ij}(\mathbf{x})$ to the hyperboloid defined by the TDOA for each pair of microphones $< i, j >$:

$$D_{ij}(\mathbf{x}) = |(||M_i - \mathbf{x}|| - ||M_j - \mathbf{x}||) - DDOA_{ij}|, \tag{12}$$



(a) (b)

Figure 6: (a) A 2D probability histogram for source location obtained by sampling a weighted sum of hyperbolas corresponding to the time-difference of arrival to all microphone pairs (shown in blue). We pick the maximum value (in red) in the histogram as the location of the frequency band at each frame. (b) A cut through a 3D histogram of the same situation obtained by sampling hyperboloid surfaces on a 3D grid.

where M_i resp. M_j is the position of microphone i resp. j, $\mathbf{x}$ is the center of the cell and $DDOA_{ij} = TDOA_{ij}/c$ is the signed distance-difference obtained from the calculated TDOA (in seconds) and the speed of sound c.

The final histogram value in each cell is then obtained as :

$$H(\mathbf{x}) = \sum_{ij} \left[\frac{e^{(\gamma(1-D_{ij}(\mathbf{x})))}}{e^{\gamma}} (1 - DDOA_{ij}/||M_i - M_j||) \right. \\ \left. \text{if } D_{ij}(\mathbf{x}) < 1,\ 0 \text{ otherwise} \right]. \tag{13}$$

The exponentially decreasing function controls the "width" of the hyperboloid and provides a tradeoff between localization accuracy and robustness to noise in the TDOA estimates. In our examples, we use $\gamma = 4$. The second weighting term reduces the contribution of large TDOAs relative to the spacing between the pair of microphones. Such large TDOAs lead to "flat" ellipsoids contributing to a large number of neighboring cells in the histogram and resulting into less accurate position estimates [4].

The histogram is re-computed for each subband at each time-frame based on the corresponding TDOA estimates. The location of the k_{th} subband is finally chosen as the center point of the cell having the maximum value in the probability histogram (Figure 5 (c)):

$$B_k = \arg\max_{\mathbf{x}} H(\mathbf{x}) \tag{14}$$

In the case where most of the sound sources and microphones are located at similar height in a near planar configuration, the histogram can be computed on a 2D grid. This yields faster results at the expense of some error in localization. A naive calculation of the histogram at each time-frame (for a single frequency band and 8 microphones, i.e., 28 possible hyperboloids) on a 128×128 grid requires 20 milliseconds on a *Pentium4 Xeon* 3.2GHz processor. An identical calculation in 3D requires 2.3 s. on a $128 \times 128 \times 128$ grid. To avoid this extra computation time, we implemented a hierarchical evaluation using a quadtree or octree decomposition [61]. We recursively test only a few candidate locations (typically 16 to 64), uniformly distributed in each cell, before subdividing the cell in which the maximum of all estimates is found. Our hierarchical localization process supports real-time performance requiring only 5 ms to locate a subband in a $512 \times 512 \times 512$ 3D grid. In terms of accuracy, it was found to be comparable to the direct, non-hierarchical, evaluation at maximum resolution in our test examples.

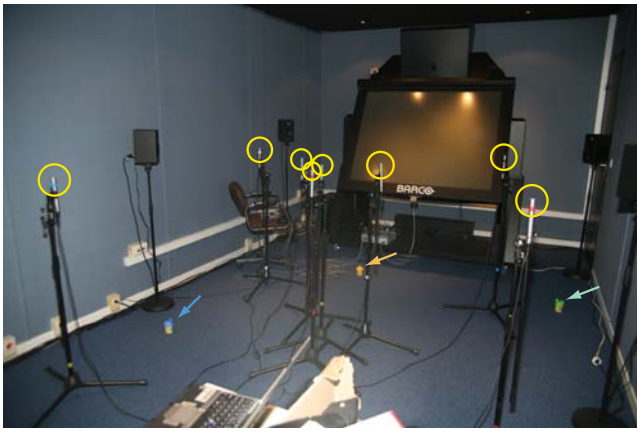


Figure 7: Indoor validation setup using 8 microphones. The 3 markers (see blue, yellow, green arrows) on the ground correspond to the location of the recorded speech signals.

6.3 Indoor validation study

To validate our approach, we conducted a test-study using 8 microphones inside a 7m×3.5m×2.5m room with limited reverberation time (about 0.3 sec. at 1KHz). We recorded three people speaking while standing at locations specified by colored markers. Figure 7 depicts the corresponding setup. We first evaluated the localization accuracy for all subbands by constructing spatial energy maps of the recordings. As can be seen in Figure 8, our approach properly localizes the corresponding sources. In this case, the energy corresponds to the signal captured by a microphone located at the center of the room.

Figure 11 shows localization error over all subbands by reference to the three possible positions for the sources. Since we do not know *a priori* which subband belongs to which source, the error is simply computed, for each subband, as the minimum distance between the reconstructed location of the subband and each possible source position. Our approach achieves a maximum accuracy of one centimeter and, on average, the localization accuracy is of the order of 10 centimeters. Maximum errors are of the order of a few meters. However, listening tests exhibit no strong artefacts showing that such errors are likely to occur for frequency subbands containing very little energy. Figure 11 also shows the energy of one of the captured signals. As can be expected, the overall localization error is also correlated with the energy of the signal.

We also performed informal comparisons between reference binaural recordings and a spatial audio rendering using the obtained locations, as described in the next section. Corresponding audio files can be found at:
`http://www-sop.inria.fr/reves/projects/audioMatting`.

They exhibit good correspondence between the original situation and our renderings showing that we properly assign the subbands to the correct source locations at each time-frame.

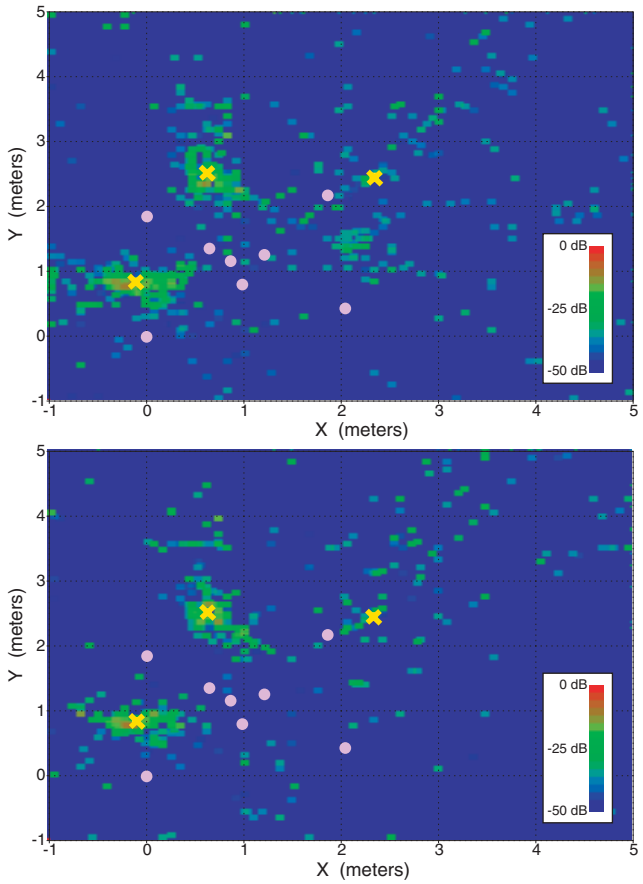


Figure 8: Energy localization map for a 28s.-long audio sequence featuring 3 speakers inside a room (indicated by the three yellow crosses). Light-purple dots show the location of the 8 microphones. The top map is computed using AMDF-based TDOA estimation while the bottom map is computed using GCC-PHAT. Both maps were computed using 8 subbands and corresponding energy is integrated over the entire duration of the sequence.

7 3D-audio resynthesis

The final stage of our approach is the spatial audio resynthesis. During a real-time simulation, the previously pre-computed subband positions can be used for re-rendering the acquired soundfield while changing the position of the sources and listener. A key aspect of our approach is to provide a spatial description of a real-world auditory scene in a manner independent of the auditory restitution system. The scene can thus be re-rendered by standard 3D-audio APIs: in some of our test examples, we used *DirectSound 3D* accelerated by a *CreativeLabs Audigy2 NX* soundcard and also implemented our own software binaural renderer, using head-related

transfer function (HRTF) data from the LISTEN HRTF database[1].

Inspired by binaural-cue coding [23], our re-rendering algorithm can be decomposed in two steps, that we detail in the following sections:

- First, as the virtual listener moves throughout the environment, we construct a *warped monophonic signal* based on the original recording of the microphone closest to the current listening position.

- Second, this warped signal is spatially enhanced using 3D-audio processing based on the location of the different frequency subbands.

These two steps are carried out over small time-frames (of the same size as in the analysis stage). To avoid artefacts we use a 10% overlap to cross-fade successive synthesis frames.

7.1 Warping the original recordings

For re-rendering, a monophonic signal best matching the current location of the virtual listener relative to the various sources must be synthesized from the original recordings.

At each time-frame, we first locate the microphone closest to the location of the virtual listener. To ensure that we remain as faithful as possible to the original recording, we use the signal captured by this microphone as our reference signal $R(t)$.

We then split this signal into the same frequency subbands used during the off-line analysis stage. Each subband is then warped to the virtual listener location according to the pre-computed spatial mapping at the considered synthesis time-frame (Figure 9).

This warping involves correcting the propagation delay and attenuation of the reference signal for the new listening position, according to our propagation model (see Eq.1). Assuming an inverse distance attenuation for point emitters, the warped signal $R'_i(t)$ in subband i is thus given as:

$$R'_i(t) = r^i_1 / r^i_2 R_i(t + (\delta^i_1 - \delta^i_2)), \tag{15}$$

where r^i_1,δ^i_1 are respectively the distance and propagation delay from the considered time-frequency atom to the reference microphone and r^i_2,δ^i_2 are the distance and propagation delay to the new listening position.

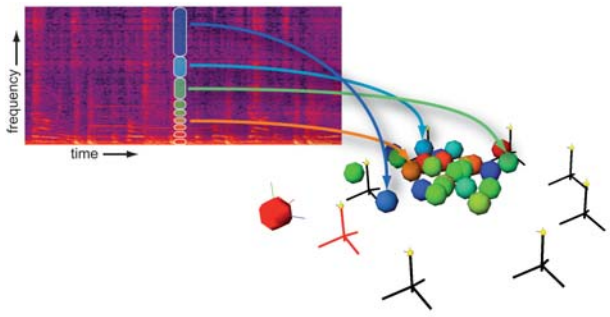


Figure 9: In the resynthesis phase, the frequency components of the signal captured by the microphone closest to the location of the virtual listener (shown in red) is warped according to the spatial mapping pre-computed in the off-line stage.

[1] http://recherche.ircam.fr/equipes/salles/listen/

7.2 Clustering for 3D-audio rendering and source matting

To spatially enhance the previously obtained warped signals, we run an additional clustering step to aggregate subbands which might be located at nearby positions using the technique of [69]. The clustering allows to build groups of subbands which can be rendered from a single representative location and might actually belong to the same physical source in the original recordings. Thus, our final rendering stage spatializes N representative point sources corresponding to the N generated clusters, which can vary between 1 and the total number of subbands. To improve the temporal coherence of the approach we use an additional Kalman filtering step on the resulting cluster locations [33].

With each cluster we associate a weighted sum of all warped signals in each subband which depends on the Euclidean distance between the location of the subband B_i and the location of the cluster representative C_k. This defines matting coefficients α_k, similar to alpha-channels in graphics [54]:

$$\alpha(C_k, B_i) = \frac{1.0/(\varepsilon + ||C_k - B_i||)}{\sum_i \alpha(C_k, B_i)}. \tag{16}$$

In our examples, we used $\varepsilon = 0.1$. Note that in order to preserve the energy distribution, these coefficients are normalized in each frequency subband.

These matting coefficients control the blending of all subbands rendered at each cluster location and help smooth the effects of localization errors. They also ensure a smoother reconstruction when sources are modified or moved around in the re-rendering phase.

The signal for each cluster $S_k(t)$ is finally constructed as a sum of all warped subband signals $R'_i(t)$, as described in the previous section, weighted by the matting coefficients $\alpha(C_k, B_i)$:

$$S_k(t) = \sum_i \alpha(C_k, B_i) R'_i(t). \tag{17}$$

The representative location of each cluster is used to apply the desired 3D-audio processing (e.g., HRTFs) without *a priori* knowledge of the restitution setup.

Figure 10 summarizes the complete re-rendering algorithm.

8 Applications and results

Our technique opens many interesting application areas for interactive 3D applications, such as games or virtual/augmented reality, and off-line audio-visual post-production. Several example renderings demonstrating our approach can be found at the following URL:

http://www-sop.inria.fr/reves/projects/audioMatting.

8.1 Modeling complex sound sources

Our approach can be used to render extended sound sources (or small soundscapes) which might be difficult to model using individual point sources because of their complex acoustic behavior. For instance, we recorded a real-world sound scene involving a car which is an extended vibrating sound radiator. Depending on the point of view around the scene, the sound changes significantly due to the relative position of the various mechanical elements (engine, exhaust, etc.) and the effects of sound propagation around the body of the car. This makes an approach using multiple recordings very interesting in order to realistically capture these effects. Unlike other techniques, such as *Ambisonics O-format* [43], our approach captures the position of the various sounding components and not only their directional aspect. In the accompanying examples, we demonstrate a re-rendering with a moving listening point of a car

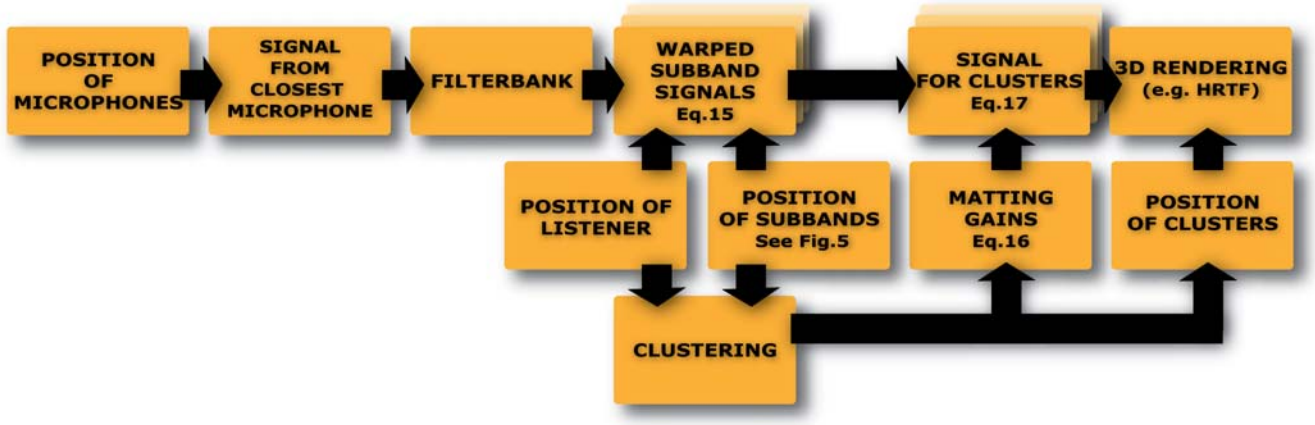
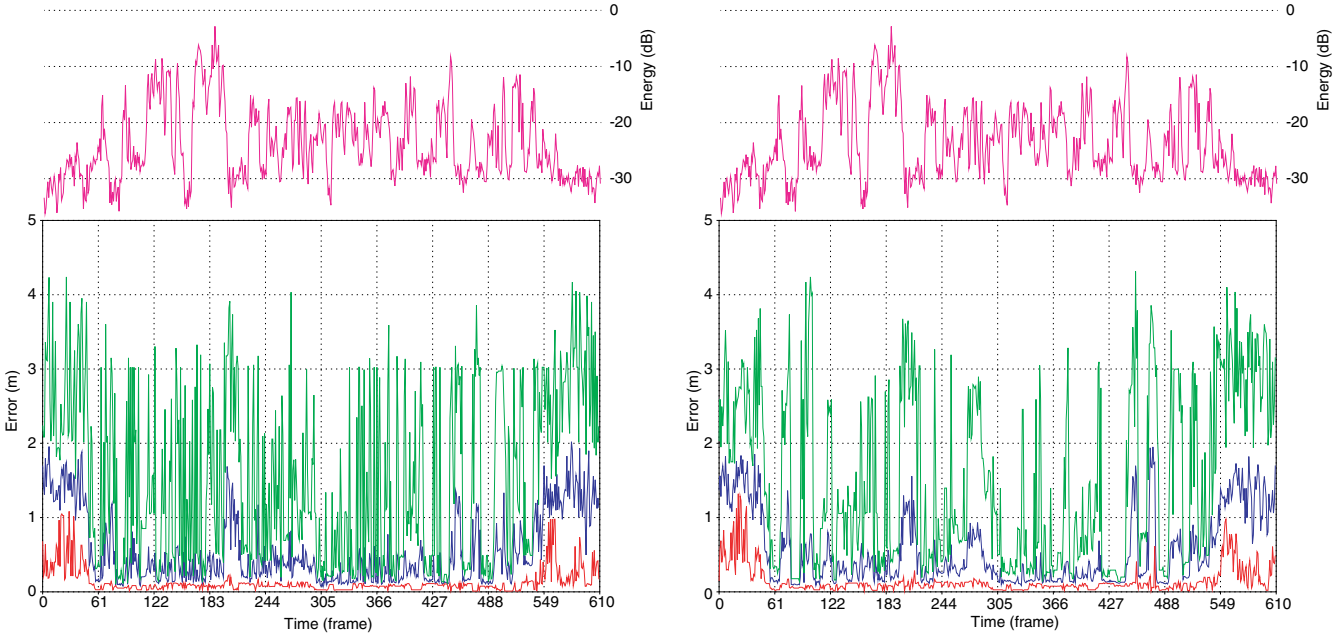


Figure 10: Overview of the synthesis algorithm used to re-render the acquired soundscape based on the previously obtained subband positions.



(a) GCC-PHAT based localization

(b) AMDF-based localization

Figure 11: Localization error for the same audio sequence as in Figure 8. computed over 8 subbands. Averaged error over all subbands is displayed in blue, maximum error in green and minimum error in red. The top (magenta) curve represents the energy for one of the input recordings and shows its correlation with the localization error (clearly larger when the energy drops out).

scenario acquired using 8 microphones surrounding the action (Figure 12). In this case, we used 4 clusters for re-rendering. Note in the accompanying video available on-line, the realistic distance and propagation effects captured by the recordings, for instance on the door slams. Figure 13 shows a corresponding energy map clearly showing the low frequency exhaust noise localized at the rear of the car and the music from the on-board stereo audible through the driver's open window. Engine noise was localized more diffusely mainly due to interference with the music.

Figure 12: We capture an auditory environment featuring a complex sound source (car engine/exhaust, passengers talking, door slams and on-board stereo system) using 8 microphones surrounding the action.

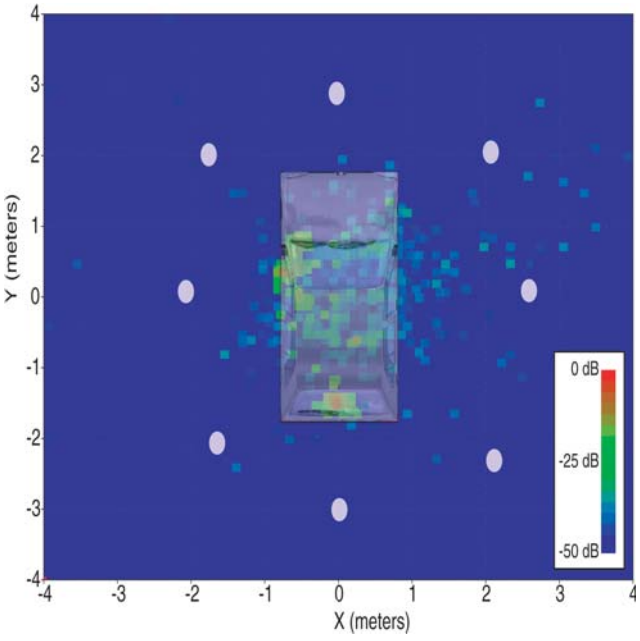


Figure 13: Energy localization map for a 15 sec.-long recording of our car scenario featuring engine/exhaust sounds and music (on the on-board stereo system and audible through the open driver-window). Positions were computed over 8 subbands using GCC-PHAT-based TDOA estimation. Energy is integrated over the entire duration of the input audio sequence.

8.2 Spatial recording and view-interpolation

Following binaural cue coding principles, our approach can be used to efficiently generate high-resolution surround recordings from monophonic signals. To illustrate this application we used 8 omni-directional microphones located in a circle-like configuration about 1.2 meters in diameter (Figure 14) to record three persons talking and the surrounding ambiance (fountain, birds, etc.). Then, our pre-processing was applied to extract the location of the sources. For re-rendering, the monophonic signal of a single microphone was used and respatialized as described in Section 7.1, using 4 clusters (Figure 16). Please, refer to the accompanying video provided on the web site to evaluate the result.

Figure 14: Microphone setup used to record the fountain example. In this case the microphones are placed at the center of the action.

Another advantage of our approach is to allow for re-rendering an acquired auditory environment from various listening points. To demonstrate this approach on a larger environment, we recorded two moving speakers in a wide area (about 15×5 meters) using the microphone configuration shown in Figure 1 (Left). The recording also features several background sounds such as traffic and road-work noises. Figure 15 shows a corresponding spatial energy map. The two intersecting trajectories of the moving speakers are clearly visible.

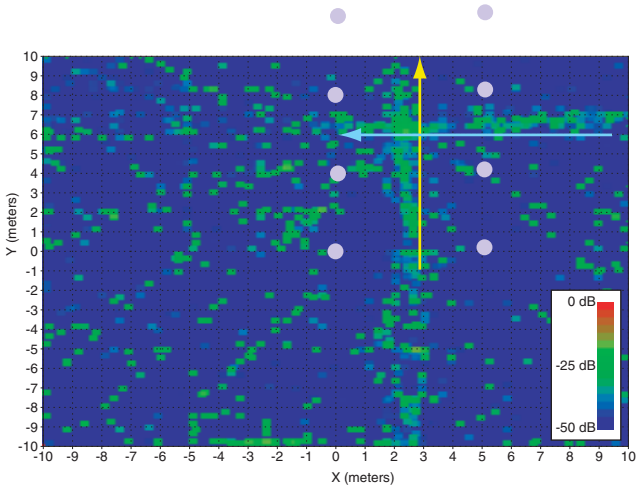


Figure 15: Energy map for a recording of our moving speaker scenario. The arrows depict the trajectory of the two speakers. Energy is integrated over the entire duration of the input audio sequence. Note how the two intersecting trajectories are clearly reconstructed.

Applying our approach, we are able to re-render this auditory scene from any arbitrary viewpoint. Although the rendering is based only on the *monophonic* signal of the microphone closest to the virtual listener at each time-frame, the extracted spatial mapping allows for convincingly reproducing the motion of the sources. Note in the example video provided on the accompanying web-site

how we properly capture front-to-back and left-to-right motion for the two moving speakers.

8.3 Spatial audio compositing and post-editing

Finally, our approach allows for post-editing the acquired auditory environments and composite several recordings.

Source re-localization and modification

Using our technique, we can selectively choose and modify various elements of the original recordings. For instance, we can select any spatial area in the scene and simply relocate all clusters included in the selected region. We demonstrate an example interactive interface for spatial modification where the user first defines a selection area then a destination location. All clusters entering the selection area are translated to the destination location using the translation vector defined by the center of the selection box and the target location. In the accompanying video, we show two instances of source re-localization where we first select a speaker on the left-hand side of the listener and move him to the right-hand side. In a second example, we select the fountain at the rear-left of the listener and move it to the front-right (Figure 16).

Compositing

Since our recording setups are spatially calibrated, we can integrate several environments into a single composite rendering which preserves the relative size and positioning of the various sound sources. For instance, it can be used to integrate a close-miked sound situation into a different background ambiance. We demonstrate an example of sound-field compositing by inserting our previous car example (Figure 12) into the scene with the two moving speakers (Figure 1). The resulting composite environment is rendered with 8 clusters and the 16 recordings of the two original soundscapes. Future work might include merging the representations in order to limit the number of composite recordings (for instance by "re-projecting" the recordings of one environment into the recording setup of the other and mixing the resulting signals).

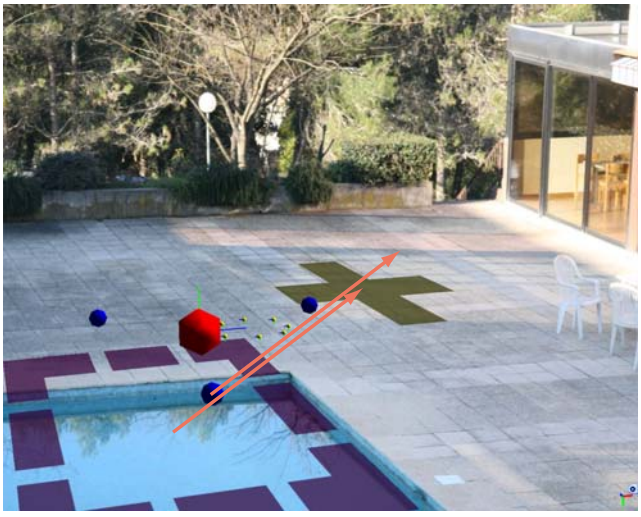


Figure 16: An example interface for source re-localization. In this example we select the area corresponding to the fountain (in purple) and translate it to a new location (shown as a yellow cross). The listener is depicted as a large red sphere, the microphone array as small yellow spheres and the blue spheres show cluster locations.

Real/Virtual integration

Our approach permits spatially consistent compositing of virtual sources within real-world recordings. We can also integrate virtual objects, such as walls, and make them interact with the original recordings. For instance, by performing real-time ray-casting between the listener and the location of the frequency subbands, we can add occlusion effects due to a virtual obstacle using a model similar to [70]. Please, refer to the accompanying examples at the previously mentioned URL for a demonstration. Of course, perfect integration would also require correcting for the reverberation effects between the different environments to composite. Currently, we experimented only in environments with limited reverberation but blind extraction of reverberation parameters [7] and blind deconvolution are complementary areas of future research in order to better composite real and virtual sound-fields.

9 Discussion

Although it is based on a simple mixing model and assumes W-disjoint orthogonality for the sources, we were able to apply our approach to real-world recording scenarios. While not production-grade yet, our results seem promising for a number of interactive and off-line applications.

While we tested it for both indoor and outdoor recordings, our approach is currently only applicable to environments with limited reverberation. Long reverberations will have a strong impact on our localization process since existing cross-correlation approaches are not very robust to interfering sound reflections. Other solutions based on blind channel identification in a reverberant context could lead to improved results [15].

Errors in localization of the frequency subbands can result in noticeable artefacts especially when moving very close to a source. These errors can come from several factors in our examples particularly low signal-to-noise ratio for the source to localize, blurring from sound reflections, correlation of two different signals in the case of widely spaced microphones or several sources being present in a single frequency subband. As a result, several overlapping sources are often fused at the location of the louder source. While the assumption of W-disjoint orthogonality has been proven to be suitable for speech signals [59], it is more questionnable for more general scenarios. It will only be acceptable if this source can perceptually mask the others. However, recent approaches for efficient audio rendering have shown that masking between sources is significant [69], which might explain why our approach can give satisfying results quite beyond the validity domain of the underlying models. Alternate decompositions [45, 40] could also lead to sparser representations and better results within the same framework.

The signal-to-noise ratio of the different sound sources is also directly linked to the quality of the result when moving very close to the source since our warping is likely to amplify the signal of the original recording in this case.

We are working on several improvements to alleviate remaining limitations of the system and improve the rendering quality:

Currently, we do not interpolate between recordings but select the signal of the microphone closest to the listener location for subsequent warping and re-rendering. This provides a correct solution for the case of omnidirectional anechoic point sources. In more general situations, discontinuities might still appear when switching from one microphone to the next. This can be caused, for instance, by the presence of a sound source with a strong directionality. A solution to this problem would be to warp the few microphones closest to the listener and blend the result at the expense of a higher computing cost. Note that naive blending between microphone signals before warping would introduce unwanted interferences, very

noticeable in the case of widely-spaced microphones. Another option would be to experiment with morphing techniques [65] as an alternative to our position-based warping. We could also use different microphones for each frequency subband, for instance choosing the microphone closer to the location of each subband rather than the one closest to the listener. This would increase the signal-to-noise ratio for each source and could be useful to approximate a close-miking situation in order to edit or modify the reverberation effects for instance.

The number of bands also influences the quality of the result. More bands are likely to increase the spatial separation but since our correlation estimates are significantly noisy, it might also make artefacts more audible. In our case, we obtained better sounding results using a limited number of subbands (typically 8 to 16). Following the work of Faller et al. [8, 23, 22], we could also keep track of the inter-correlation between recordings in order to precisely localize only the frames with high correlation. Frames with low correlation could be rendered as "diffuse", forming a background ambiance which cannot be as precisely located [47]. This could be seen as explicitly taking background noise or spatially extended sound sources into account in our mixing model instead of considering only perfect anechoic point sources. We started to experiment with an explicit separation of background noise using noise-removal techniques [21]. The obtained foreground component can then be processed using our approach while the background-noise component can be rendered separately at a lower spatial resolution. Example renderings available on the web site demonstrate improved quality in complex situations such as a seashore recording.

Sound source clustering and matting also strongly depends on the correlation and position estimates for the subbands. An alternative solution would be to first separate a number of sources using independent component analysis (ICA) techniques and then run TDOA estimation on the resulting signals [62, 29]. However, while ICA might improve separation of some sources, it might still lead to signals where sources originating from different locations are combined.

Another issue is the microphone setup used for the recordings. Any number of microphones can be used for localization starting from two (which would only give directional information). If more microphones are used, the additional TDOA estimates will increase the robustness of the localization process. From our experience, closely spaced microphones will essentially return directional information while microphone setups surrounding the scene will give good localization accuracy. Microphones uniformly spaced in the scene provide a good compromise between signal-to-noise ratio and sampling of the spatial variations of the sound-field. We also experimented with cardioid microphone recordings and obtained good results in our car example. However, for larger environments, correlation estimates are likely to become noisier due to the increase in separation between different recordings, making them difficult to correlate. Moreover, it would make interpolating between recordings more difficult in the general case. Our preferred solution was thus to use a set of identical omnidirectional microphones. However, it should be possible to use different sets of microphones for localization and re-rendering which opens other interesting possibilities for content creation, for instance by generating consistent 3D-audio flythroughs while changing the focus point on the scene using directional microphones.

Finally, our approach currently requires an off-line step which prevents it from being used for real-time analysis. Being able to compute cross-correlations in real-time for all pairs of microphones and all subbands would make the approach usable for broadcast applications.

10 Conclusions

We presented an approach to record, edit and re-render real-world auditory situations. Contrary to most related approaches, we acquire the sound-field using an unconstrained, widely-spaced, microphone array which we spatially calibrate using photographs. Our approach pre-computes a spatial mapping between different frequency subbands of the acquired live recordings and the location in space from which they were emitted. We evaluated standard TDOA-based techniques and proposed a novel hierarchical localization approach. At run-time, we can apply this mapping to the frequency subbands of the microphone closest to the virtual listener in order to resynthesize a consistent 3D sound-field, including complex propagation effects which would be difficult to simulate. An additional clustering step allows for aggregating subbands originating from nearby location in order to segment individual sound sources or small groups of sound sources which can then be edited or moved around. To our knowledge, such level of editing was impossible to achieve using previous state-of-the-art and could lead to novel authoring tools for 3D-audio scenes.

We believe our approach opens many novel perspectives for interactive spatial audio rendering or off-line post-production environments, for example to complement image based rendering techniques or free-viewpoint video. Moreover, it provides a compact encoding of the spatial sound-field, which is independent of the restitution system. In the near future, we plan to run more formal perceptual tests in order to compare our results to binaural or high-order *Ambisonics* recordings in the case of fixed-viewpoint scenarios and to evaluate its quality using various restitution systems. From a psychophysical point of view, this work suggests that real-world sound scenes can be efficiently encoded using limited spatial information.

Other promising areas of future work would be to exploit perceptual localization results to improve localization estimation [72] and apply our analysis-synthesis strategy to the real-time generation of spatialized audio textures [42]. Finaly, making the calibration and analysis step interactive would allow the approach to be used in broadcasting applications (e.g., 3D TV).

Acknowledgments

This research was made possible by a grant from the *région PACA* and was also partially funded by the RNTL project OPERA (http://www-sop.inria.fr/reves/OPERA). We acknowledge the generous donation of *Maya* as part of the *Alias* research donation program, Alexander Olivier-Mangon for the initial model of the car, and Frank Firsching for the animation.

References

[1] P. Aarabi. The fusion of distributed microphone arrays for sound localization. *EURASIP Journal on Applied Signal Processing*, 4:338–347, 2003.

[2] T.D. Abhayapala and D.B. Ward. Theory and design of high order sound field microphones using spherical microphone array. *Proceedings of Intl. Conf. on Acoustics, Speech and Signal Processing*, 2002.

[3] T. Ajdler and M. Vetterli. The plenacoustic function and its sampling. *Proc. of the 1st Benelux Workshop on Model-based processing and coding of audio (MPCA2002), Leuven, Belgium*, November 2002.

[4] Thibaut Ajdler, Igor Kozintsev, Rainer Lienhart, and Martin Vetterli. Acoustic source localization in distributed sensor networks. *Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA*, 2:1328–1332, 2004.

[5] Daniel G. Aliaga and Ingrid Carlbom. Plenoptic stitching: a scalable method for reconstructing 3d interactive walk throughs. In *SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 443–450, New York, NY, USA, 2001. ACM Press.

[6] C. Avendano. Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppresssion and re-panning applications. *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2003), New Paltz, NY, USA*, October 2003.

[7] A. Baskind and O. Warusfel. Methods for blind computational estimation of perceptual attributes of room acoustics. *proceedings of the AES 22nd Intl. Conf. on virtual, synthetic and entertainment audio, Espoo, Finland*, June 2002.

[8] Frank Baumgarte and Christof Faller. Binaural cue coding - part I: Psychoacoustic fundamentals and design principles. *IEEE Trans. on Speech and Audio Proc*, 11(6), 2003.

[9] A.J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *J. of the Acoustical Society of America*, 93(5):2764–2778, may 1993.

[10] M.M. Boone, E.N.G. Verheijen, and P.F. van Tol. Spatial sound-field reproduction by wave-field synthesis. *J. of the Audio Engineering Society*, 43:1003–1011, December 1995.

[11] A.S. Bregman. *Auditory Scene Analysis, The perceptual organization of sound*. The MIT Press, 1990.

[12] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. *Proc. of ACM SIGGRAPH*, 2001.

[13] J. Chen, J. Benesty, and Y. Huang. Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments. *EURASIP Journal on Applied Signal Processing*, 1:25–36, 2005.

[14] J.C. Chen, K. Yao, and R.E. Hudson. Acoustic source localization and beamforming: Theory and practice. *EURASIP Journal on Applied Signal Processing*, 4:359–370, 2003.

[15] Jingdong Chen, Jacob Benesty, and Yiteng (Arden) Huang. Time delay estimation in room acoustic environments: An overview. *EURASIP Journal on Applied Signal Processing*, 2006:Article ID 26503, 2006.

[16] S.E. Chen and L. Williams. View interpolation for image synthesis. *Computer Graphics*, 27(Annual Conference Series, Proc. of ACM SIGGRAPH93):279–288, 1993.

[17] P. Comon. Independent component analysis: A new concept. *Signal Processing*, 36:287–314, 1994.

[18] J. Daniel, J.-B. Rault, and J.-D. Polack. Ambisonic encoding of other audio formats for multiple listening conditions. *105th AES convention, preprint 4795*, August 1998.

[19] J.H. DiBiase, H.F. Silverman, and M.S. Branstein. *Microphone Arrays, Signal Processing Techniques and Applications, Chapter 8*. Springer Verlag, 2001.

[20] M.N. Do. Toward sound-based synthesis: the far-field case. *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada*, May 2004.

[21] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 32(6):1109–1121, December 1984.

[22] C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. of the Acoustical Society of America*, 116(5):3075–3089, November 2005.

[23] Christof Faller and Frank Baumgarte. Binaural cue coding - part II: Schemes and applications. *IEEE Trans. on Speech and Audio Proc*, 11(6), 2003.

[24] O. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. The MIT Press, Cambridge, Mass., 1993.

[25] M.A. Gerzon. Ambisonics in multichannel broadcasting and video. *J. of the Audio Engineering Society*, 33(11):859–871, 1985.

[26] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54, New York, NY, USA, 1996. ACM Press.

[27] U. Horbach, A. Karamustafaoglu, R. Pellegrini, P. Mackensen, and G. Theile. Design and applications of a data-based auralization system for surround sound. *106th Convention of the Audio Engineering Society, preprint 4976*, 1999.

[28] Youichi Horry, Ken-Ichi Anjyo, and Kiyoshi Arai. Tour into the picture: using a spidery mesh interface to make animation from a single image. In *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 225–232, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.

[29] G. Huang, L. Yang, and Z. He. Multiple acoustic sources location based on blind source separation. *Proc. of the First International Conference on Natural Computation (ICNC'05)*, 2005.

[30] Y. Huang, J. Benesty, and G.W. Elko. Microphone arrays for video camera steering. *Acoustic Signal Processing for Telecommunications*, 2000.

[31] J.-M. Jot, V. Larcher, and J.-M. Pernaux. A comparative study of 3D audio encoding and rendering techniques. *Proceedings of the AES 16th international conference, Spatial sound reproduction, Rovaniemi, Finland*, april 1999.

[32] Alexander Jourjine, Scott Rickard, and Ozgur Yilmaz. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00), Istanbul, Turkey*, June 2000.

[33] R.E. Kalman. A new approach to linear filtering and prediction problems. *Transaction of the ASME-Journal of Basic Engineering*, 82 (Series D):35–45, 1960.

[34] C.H. Knapp and G.C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 24(4):320–327, August 1976.

[35] H. Krim and M. Viberg. Two decades of array signal processing research. *IEEE Signal Processing Magazine*, pages 67–93, July 1996.

[36] A. Laborie, R. Bruno, and S. Montoya. A new comprehensive approach of surround sound recording. *Proc. 114th convention of the Audio Engineering Society, preprint 5717*, 2003.

[37] A. Laborie, R. Bruno, and S. Montoya. High spatial resolution multi-channel recording. *Proc. 116th convention of the Audio Engineering Society, preprint 6116*, 2004.

[38] Martin J. Leese. Ambisonic surround sound FAQ (version 2.8), 1998. http://members.tripod.com/martin_leese/Ambisonic/.

[39] Marc Levoy and Pat Hanrahan. Light field rendering. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, New York, NY, USA, 1996. ACM Press.

[40] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

[41] M.S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, 2002.

[42] L. Lu, L. Wenyin, and H.-J. Zhang. Audio textures: Theory and applications. *IEEE Transactions on Speech and Audio Processing*, 12(2):156–167, 2004.

[43] D.G. Malham. Spherical harmonic coding of sound objects - the ambisonic 'O' format. *Proc. of the 19th AES Conference, Surround Sound, Techniques, Technology and Perception, Schloss Elmau, Germany*, June 2001.

[44] D.G. Malham and A. Myatt. 3D sound spatialization using ambisonic techniques. *Computer Music Journal*, 19(4):58–70, 1995.

[45] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[46] J. Merimaa. Applications of a 3D microphone array. *112th AES convention, preprint 5501*, May 2002.

[47] J. Merimaa and V. Pullki. Spatial impulse response rendering. *Proc. of the 7th Intl. Conf. on Digital Audio Effects (DAFX'04), Naples, Italy*, October 2004.

[48] J. Meyer and G. Elko. Spherical microphone arrays for 3d sound recording. *chap. 2 in Audio Signal Processing for next-generation multimedia communication systems, Eds. Yiteng (Arden) Huang and Jacob Benesty, Bosten, Kluwer Academic Publisher*, 2004.

[49] Brian C.J. Moore. *An introduction to the psychology of hearing*. Academic Press, 4th edition, 1997.

[50] Randolph L. Moses, Dushyanth Krishnamurthy, and Robert Patterson. An auto-calibration method for unattended ground sensors. *Acoustics, Speech, and Signal Processing (ICASSP '02)*, 3:2941– 2944, May 2002.

[51] B. Mungamuru and P. Aarabi. Enhanced sound localization. *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*, 34(3), June 2004.

[52] P.D. O'Grady, B.A. Pearlmutter, and S.T. Rickard. Survey of sparse and non-sparse methods in source separation. *Intl. Journal on Imaging Systems and Technology (IJIST), special issue on Blind source separation and deconvolution in imaging and image processing*, 2005.

[53] R.S. Pellegrini. Comparison of data and model-based simulation algorithms for auditory virtual environments. *106th Convention of the Audio Engineering Society, preprint 4953*, 1999.

[54] T. Porter and T. Duff. Compositing digital images. *Proceedings of ACM SIGGRAPH 1984*, pages 253–259, July 1984.

[55] V. Pulkki. Directional audio coding in spatial sound reproduction and stereo upmixing. *Proc. of the AES 28th Int. Conf, Pitea, Sweden*, June 2006.

[56] D.V. Rabinkin, R.J. Renomeron, J.C. French, and J.L. Flanagan. Estimation of wavefront arrival delay using the cross-power spectrum phase technique. *132th meeting of the Acoustical Society of America, Honolulu*, December 1996.

[57] Richard Radke and Scott Rickard. Audio interpolation. In *the Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio (AES'22), Espoo, Finland*, pages 51–57, June 15-17 2002.

[58] S. Rickard. Sparse sources are separated sources. *Proceedings of the 16th Annual European Signal Processing Conference, Florence, Italy*, 2006.

[59] S. Rickard and O. Yilmaz. On the approximate w-disjoint orthogonality of speech. *Proceedings of Intl. Conf. on Acoustics, Speech and Signal Processing*, 2002.

[60] Y. Rui and D. Florencio. New direct approaches to robust sound source localization. *Intl. Conf. on Multimedia and Expo (ICME)*, July 2003.

[61] H. Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, 1990.

[62] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano. Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Applied Signal Processing*, 11:1135–1146, 2003.

[63] H. Sawada, S. Araki, R. Mukai, and S. Makino. Blind extraction of dominant target sources using ica and time-frequency masking. *IEEE Trans. Audio, Speech, and Language Processing.* accepted for future publication.

[64] R.O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, AP-34(3), March 1986.

[65] M. Slaney, M. Covell, and B. Lassiter. Automatic audio morphing. *Proceedings of Intl. Conf. on Acoustics, Speech and Signal Processing*, May 1996.

[66] Soundfield. http://www.soundfield.com.

[67] R. Streicher. The decca tree – it's not just for stereo anymore. http://www.wesdooley.com/pdf/Surround_Sound_Decca_Tree-urtext.pdf.

[68] R. Streicher and F.A. Everest, editors. *The new stereo soundbook, 2nd edition*. Audio Engineering Associate, Pasadena (CA), USA, 1998.

[69] N. Tsingos, E. Gallo, and G. Drettakis. Perceptual audio rendering of complex virtual environments. *ACM Transactions on Graphics, Proceedings of SIGGRAPH 2004*, August 2004.

[70] Nicolas Tsingos and Jean-Dominique Gascuel. Fast rendering of sound occlusion and diffraction effects for virtual acoustic environments. *Proc. 104th Audio Engineering Society Convention, preprint 4699*, May 1998.

[71] E. Vincent, X. Rodet, A. Röbel, C. Févotte, E. Le Carpentier, R. Gribonval, L. Benaroya, and F. Bimbot. A tentative typologie of audio source separation tasks. *Proc. of the 4th Intl. Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan*, April 2003.

[72] K.W. Wilson and T. Darell. Learning a precedence effect-like weighting function for the generalized cross-correlation framework. *IEEE Journal of speech and audio processing. Special issue on statistical and perceptual audio processing*, 2006.

[73] D.L. Yewdall. *Practical Art of Motion Picture Sound (2nd edition)*. Focal Press, 2003.

[74] Ozgur Yilmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.