Extracting and Re-rendering Structured Auditory Scenes from Field Recordings

Emmanuel Gallo, Nicolas Tsingos

Extracting and Re-rendering Structured Auditory Scenes from Field Recordings

Emmanuel Gallo[1,2] and Nicolas Tsingos[1]

[1]*REVES, INRIA, Sophia Antipolis, France*

[2]*CSTB, Sophia Antipolis, France*

Correspondence should be addressed to Emmanuel Gallo — Nicolas Tsingos
(emmanuel.gallo|nicolas.tsingos@sophia.inria.fr)

ABSTRACT

We present an approach to automatically extract and re-render a structured auditory scene from field recordings obtained with a small set of microphones, freely positioned in the environment. From the recordings and the calibrated position of the microphones, the 3D location of various auditory events can be estimated together with their corresponding content. This structured description is reproduction-setup independent. We propose solutions to classify foreground, well-localized sounds and more diffuse background ambiance and adapt our rendering strategy accordingly. Warping the original recordings during playback allows for simulating smooth changes in the listening point or position of sources. Comparisons to reference binaural and B-format recordings show that our approach achieves good spatial rendering while remaining independent of the reproduction setup and offering extended authoring capabilities.

1. INTRODUCTION

Current models for interactive 3D audio scene authoring often assume that sounds are emitted by a set of monophonic point sources for which a signal has to be individually generated [33, 5]. In the general case, source signals cannot be completely synthesized from physics principles 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. This is 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 recording techniques which encode directional components of the soundfield [34, 35, 23, 25] can be directly used to acquire and playback real-world auditory environments as a whole. They produce realistic results but offer little control, if any, at the playback end. In particular, they are acquired from a single location in space and only encode directional information, which makes them insufficient for free-walkthrough applications or rendering of large near-field sources. In such spatially-extended cases, correct reproduction requires sampling the soundfield at several locations and encoding the 3D position and not only the incoming direction of the sounds. In practice, the use of such single-point recordings is mostly limited to the rendering of an overall surround ambiance that can possibly be rotated around the listener.

We previously developed a novel analysis-synthesis approach which bridges the two previous strategies [12]. Inspired by spatial audio coding [10, 4, 6, 28, 13] and blind source separation [38, 37, 30], our method builds a higher-level spatial description of the auditory scene from a small set of monophonic recordings. Contrary to previous spatial audio coding work, the recordings are made from widely-spaced locations and sample both content and spatial information for the sound sources present in the scene. Our approach is also mostly dedicated to live recordings since it reconstructs estimates of the 3D locations of the sound sources from physical propagation delays. This information might not be available in studio recordings which rely on non-physical panning strategies. The obtained description can then be

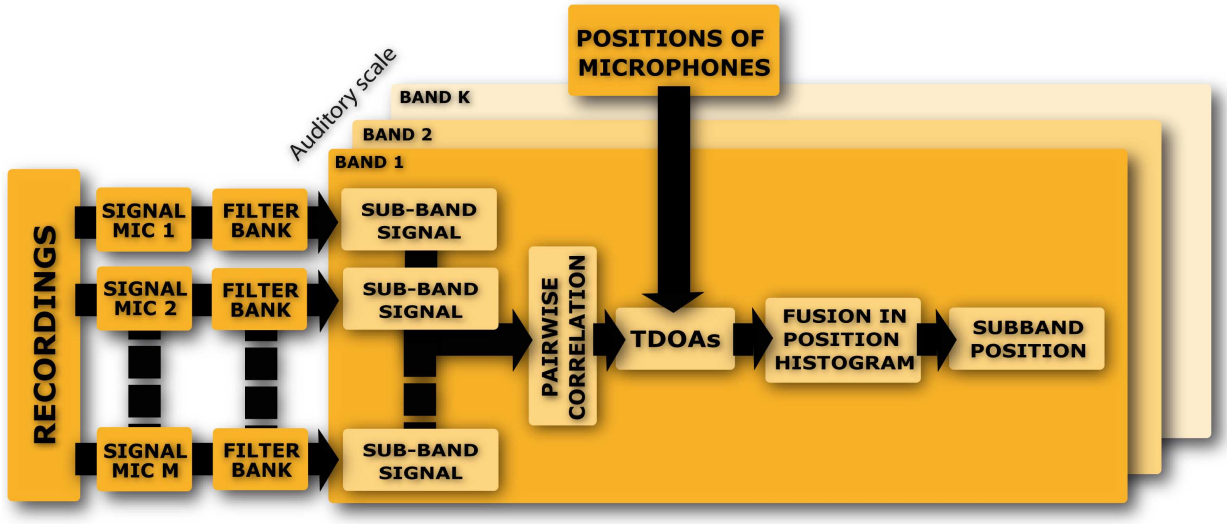


Fig. 1: Overview of our analysis pipeline.

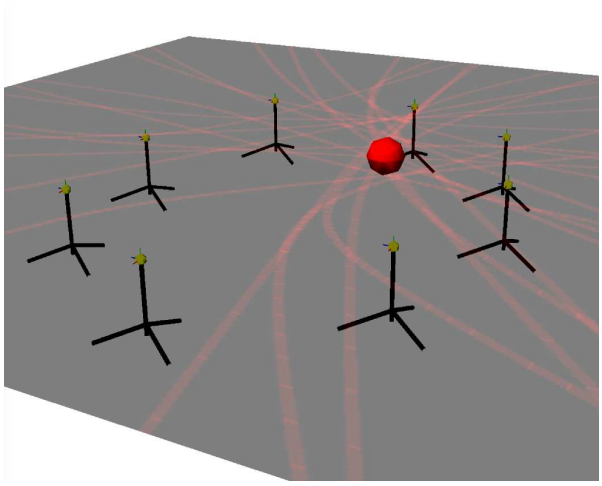


Fig. 2: Construction of a global spatial mapping for the captured sound-field. Based on calculated time-differences of arrival estimation between all pairs of recordings, we sample all corresponding hyperbolic loci to obtain a position estimate for each considered subband (shown as colored sphere).

used for real-time post-processing and re-rendering of the original recordings, for instance by smoothly varying the listening point inside the environment and editing/moving sound sources. We briefly review key aspects of this approach but refer the reader to [12] for additional details.

We first acquire real-world soundscapes using a small number (e.g., 8) of omnidirectional microphones arbitrarily positioned in the environment. In order to extract correct spatial information from the recordings, it is necessary to retrieve the 3D locations of the microphones. We use an image-based technique from photographs [11] which ensures fast and convenient acquisition on location, not requiring any physical measurements or homing device.

The obtained sparse sampling of the soundfield 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 (Figures 1 and 2). To compute this spatial mapping, we split the signal into short time-frames and decompose them onto a set of frequency subbands defined on a Bark scale [36] or, alternatively, using a gammatone filter bank [26]. Assuming that the sound sources do not overlap in time-frequency domain (*W-disjoint orthogonality* hypothesis [38]), we then use classical time-difference of arrival techniques (e.g., [18, 7]) between all pairs of microphones to retrieve a position for each subband at

each time-frame. We developed an improved hierarchical source localization technique from the obtained time-differences, using a quadtree or octree decomposition of space [32].

Real-time re-rendering is achieved through a frequency-dependent warping of the original recordings, based on the estimated positions of each frequency subband. This warping assumes an omnidirectional, anechoic, point source model. For instance, for any desired virtual listener position we first determine the closest microphone and use its signal as a reference. We then warp this reference signal by resampling and equalizing its different subbands. This warping first compensates the original propagation delay and attenuation from each calculated subband location to the location of the reference microphone. It then applies the updated propagation delay and attenuation corresponding to the new position of the virtual listener. Finally, the obtained monophonic signal is enhanced by the spatial cues computed for each subband (e.g., using head-related transfer functions), creating a spatialized rendering. Example re-renderings are available at http://www-sop.inria.fr/reves/projects/audioMatting.

However, in the case of live field recordings, this approach suffers from several limitations. First, the underlying hypothesis of time-frequency sparseness for the acquired signals is often not true in practice, especially in the presence of significant background noise [31]. This results in noisy position estimates and low quality signal reconstruction when virtually moving throughout the environment. Second, our approach uses a limited number of frequency subbands, acting as representative point sources, to model the auditory environment at each time-frame. While point sources might be appropriate to render well-localized events, background ambiance and extended sources (e.g., the sea on a seashore) cannot be convincingly reproduced using this model (Figure 3).

In this paper, we propose a solution to these shortcomings based on an *a priori* segmentation of foreground sound events and background ambiance which we describe in Section 2.1. We also present an improved re-rendering solution specifically adapted to these two components which preserves the independence from the reproduction setup. In particular, we propose to render the foreground sound events using a set of separate point sources while the background component is encoded using a smoother low-order spherical harmonics representation. Details can be found in Sections 2.2 and 2.3.

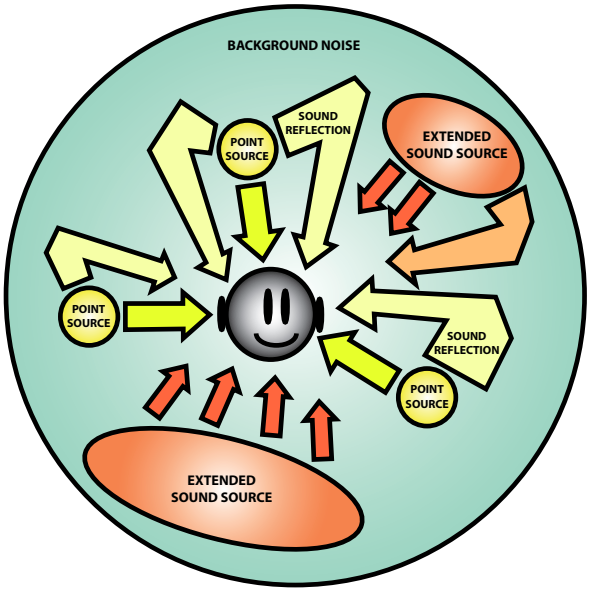


Fig. 3: Typical components of a real-world auditory scene. In this paper, we propose to explicitly separate *foreground*, non-stationary and well localized, sound events from *background* components that are more stationary and spatially diffuse.

Section 3 describes the results of a pilot perceptual evaluation study aimed at assessing the quality of our approach relative to reference binaural and B-format recordings in the case of fixed-listening-point scenarios.

Finally, our approach introduces additional authoring capabilities by allowing separate manipulation of each component, which we briefly outline in Section 4 before concluding.

2. IMPROVED ANALYSIS AND RE-SYNTHESIS

This section addresses a set of possible improvements to our previous technique. They are based on an *a priori* segmentation of background and foreground components leading to a two-layer model, similar in spirit to the pairwise/non-directional and direct/diffuse decompositions used in some spatial audio coding approaches [28, 13, 29, 6]. However, since we are warping the direct component when re-rendering from different listening points, switching between localized/diffuse models on a per-subband basis would introduce audible artefacts in our case. We chose to perform a finer-grain segmentation

of the input recordings as a pre-processing step which does not rely on position estimates. Such an approach was already reported to improve results for blind source separation problems [8]. We also propose re-rendering strategies tailored to each component.

2.1. Background/foreground segmentation

We chose to segment stationary background noise from non-stationary sound events using the technique by Ephraim and Malah [9], originally developed for denoising of speech signals. This approach assumes that the distributions of Fourier coefficients for signal and noise are statistically independent zero-mean Gaussian random variables. Under this assumption, the spectral amplitude of the denoised signal is estimated using a minimum mean-square error criterion. The background noise signal is then simply obtained by subtracting the denoised signal from the original. We found the algorithm to perform quite well. While not perfect, it leads to a foreground component with limited musical noise. In most cases, this noise is masked when re-combined with the background component at re-rendering time. The extracted foreground component, containing non-stationary sounds is also better suited to our underlying assumption of time-frequency sparseness than the original recordings (see Figure 5). However, several foreground sound sources might still overlap in time-frequency. Background and foreground segmentation is performed independently on the signals from all microphones.

2.2. Background "panorama" generation

The separated foreground and background components are both processed using the analysis pipeline described in Section 1 (see also Figure 1). However, in the case of the background component, we obtain noisier position estimates since this component will generally correspond to background noise and sources with low signal-to-noise ratios. In order to produce a smooth spatial background texture, we use the obtained positions to encode the corresponding subband signals on a 1^{st}-order spherical harmonic basis. No warping is applied to the background component in this case (Figure 4).

As our signals are real-valued, we encode them with real spherical harmonics defined as:

$$y_l^m(\theta,\phi) = \begin{cases} \sqrt{2}K_l^m cos(m\phi)P_l^m(cos\theta) & m>0 \\ \sqrt{2}K_l^m cos(-m\phi)P_l^{-m}(cos\theta) & m<0 \\ K_l^0 P_l^0(cos\theta) & m=0 \end{cases} \quad (1)$$

where l is the order, $m \in [-l;+l]$, P is the associated Legendre polynomial and K is a scaling factor defined as:

$$K_l^m = \sqrt{\frac{(2l+1)}{4\pi}\frac{(l-|m|)!}{(l+|m|)!}}. \quad (2)$$

For each subband signals, we compute the minimum and maximum elevation and azimuth of the obtained positions over the entire duration of the recording. Then, we uniformly expand the background signal in this area. We choose the background signal to encode from the monophonic recording closest to the center of the acquired scene. Accordingly, the background texture is encoded relative to a fixed reference point, for instance the central point of the scene.

This background panorama can thus be encoded in a pre-processing stage so that only the decoding is performed at run-time, e.g., when freely navigating in the recordings. Several decoding options are available depending on the desired reproduction setup [15].

2.3. Improved foreground re-synthesis

At re-rendering time, we perform a warping of the original foreground recordings in order to generate a signal as consistent as possible with the desired virtual listening position (Figure 4). Assuming an inverse distance attenuation for point emitters, the warped signal $R'_i(t)$ in subband i is given as:

$$R'_i(t) = r_1^i / r_2^i R_i(t + (\delta_1^i - \delta_2^i)), \quad (3)$$

where r_1^i,δ_1^i are respectively the distance and propagation delay from the considered time-frequency atom to the reference microphone and r_2^i,δ_2^i are the distance and propagation delay to the desired listening position.

This warping heavily relies on the fact that we consider the subband signals to be re-emitted by anechoic point sources. In real-world environments this model is challenged, due to the strong directionality of some sound sources. As a result, discontinuities can appear when the virtual listener is moving around if the signal from a single reference microphone is used (e.g., the one closest to

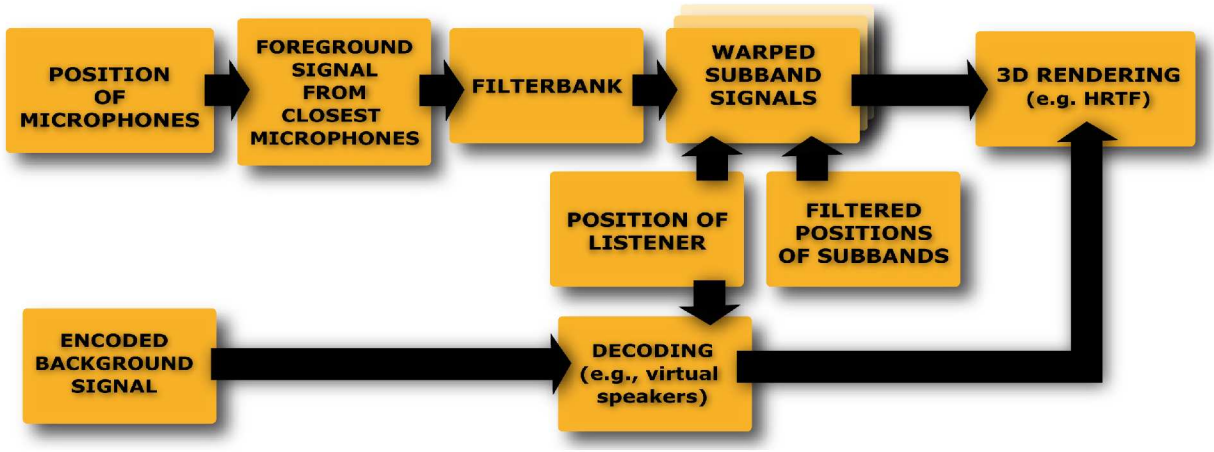
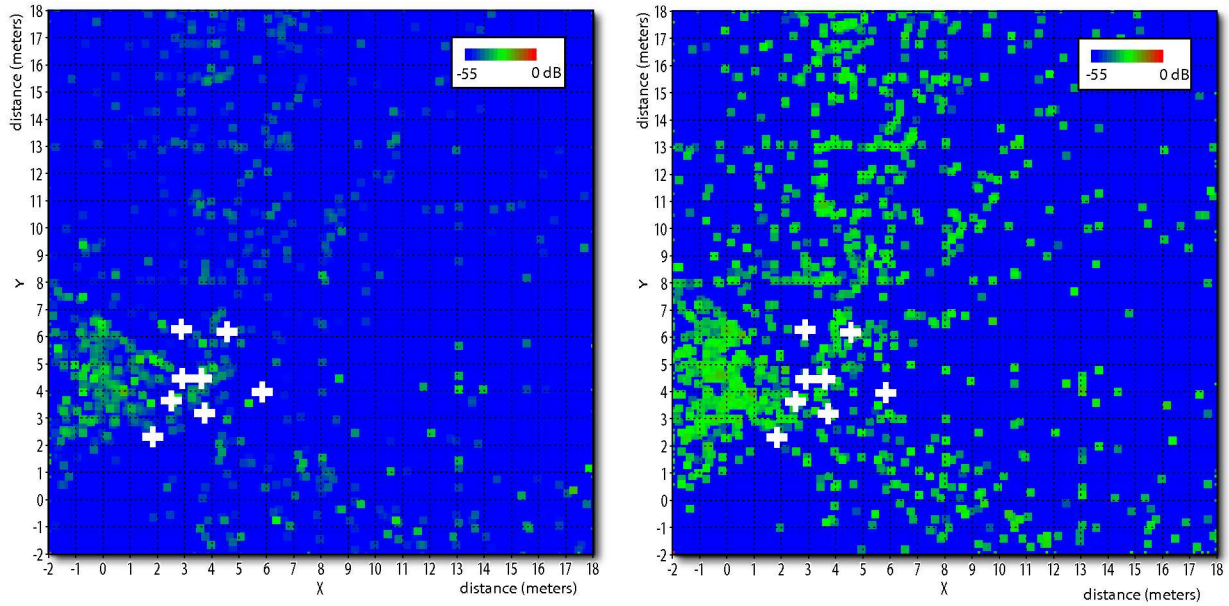


Fig. 4: Overview of our re-synthesis pipeline. Foreground sound events are rendered as point sources while background sounds are encoded using a low-order spherical harmonics decomposition.



(a) localization of the energy of the foreground component only

(b) localization of the total energy

Fig. 5: Comparison between energy localization in the seashore example of Section 4 for (a) the foreground component only and (b) the complete recording. The figure shows the reconstructed location of all subbands integrated through the entire duration of the sequence. White crosses indicate the locations of the microphones used for recording.

the desired virtual position). To avoid such problems and roughly compensate for the limitations of our anechoic point source model, we propose to continuously warp the signals of the two microphones closest to the desired virtual listening position and blend them together to generate a smoothly varying monophonic signal. Blending can be simply controlled by the relative distance of the virtual listener to these two reference microphones. Note that blending the signals prior to warping would introduce comb filtering effects that can be very noticeable

when the microphones are widely spaced. To further improve the re-rendering quality of the foreground component, we also smooth our position estimates for the subbands using Kalman filtering [16]. This prevents large and fast position changes and limits possible "wobbling" effects due to jittery subband positions.

3. PILOT SUBJECTIVE EVALUATION

In order to evaluate the quality of a spatial audio reproduction system based on our approach, we compared it to binaural and B-format recordings in the context of various scenarios with fixed listening points.

3.1. Test stimuli and procedure

We recorded test scenarios in two different environments: indoors in a moderately reverberant room (RT60 $\approx$ 0.3 sec. at 1KHz) and outdoors (see Figure 6). For each scenario, we used 8 monophonic recordings made with *AudioTechnica 3032* omnidirectional microphones to run our localization and re-rendering approach. A pair of *Sennheiser MKE-2 gold* microphones was placed inside the ears of a subject to capture reference binaural recordings and we also acquired a B-format version of the scenes using a *Soundfield ST250* microphone. Eventually, four recordings (one indoors, three outdoors), each about 50 sec. long, were chosen for quality testing.

We used 8 non-overlapping subbands uniformly distributed on a Bark scale to run our spatial analysis (Figure 1). Then, a binaural rendering from a point of view similar to the binaural and B-format recordings was generated from the monophonic input of the closest omnidirectional microphone and the time-varying locations obtained for the subbands. The signal of the same microphone was used to generate both a binaural rendering of the foreground events and the 1^{st}-order spherical harmonic background decoded over headphones using a *virtual loudspeakers* technique. In both cases, we used head related transfer functions (HRTFs) of the *LISTEN* database (http://recherche.ircam.fr/equipes/salles/listen/) for re-rendering. We also generated a re-rendering without explicit background/foreground segmentation considering the original recording to be entirely foreground. B-format recordings were also converted to binaural using a similar virtual loudspeaker approach.

We used a protocol derived from *Multiple Stimuli with Hidden Reference and Anchors* procedure (MUSHRA, ITU-R BS.1534) [1, 2, 14] to evaluate each scenario, using four tests stimuli (binaural reference, B-format, our approach with foreground only, our approach with background/foreground segmentation) and a hidden reference. We also provided one of our 8 monophonic recordings and the omnidirectional (W) component of the B-format recordings as anchors, resulting in a total of 7 signals to compare. Corresponding test stimuli are available at the following URL: http://www-sop.inria.fr/reves/projects/aes30. Test stimuli were presented over *Sennheiser HD600* headphones. Monaural anchor signals were presented at both ears.

Five subjects, aged 23 to 40 and reporting normal hearing, volunteered for this evaluation. They were asked to primarily focus on the spatial aspects of the sounds, paying particular attention to the position of the sources. Since the recordings were made with different microphones, we asked them to avoid specific judgments comparing the general timbre of the recordings. However, the subjects were instructed to keep track of any artefact compromising the quality of the reproduction. Their comments were gathered during a short post-screening interview. Subjects were instructed to rank the signals on a continuous [0,100] quality scale and give the highest possible score to the signal closest to the reference. They were also instructed to give the lowest possible score to the signal with the worst *spatial degradation* relative to the reference.

3.2. Results

Figures 7 and 8 summarize the results of this study. The subjects were able to identify the hidden reference and it received a maximal score in all test cases. In most cases, our approach was rated higher than B-format recordings in terms of quality of spatial reproduction. This is particularly true for the foreground-only approach which does not smooth the spatial cues and obtains a very high score. However, the subjects reported artefacts due to subbands whose localization varies rapidly through time, which limits the applicability of the approach in noisier environments. Our approach including background/foreground separation leads to smoother spatial cues since the low order background signal may mask the foreground signal. Hence, it was rated only slightly better than the B-format recordings. Subjects did not report specific artefacts with this approach, showing an improved signal quality. As could be expected, the monophonic anchors received the lowest scores. However, we can note that in some of our test cases, they

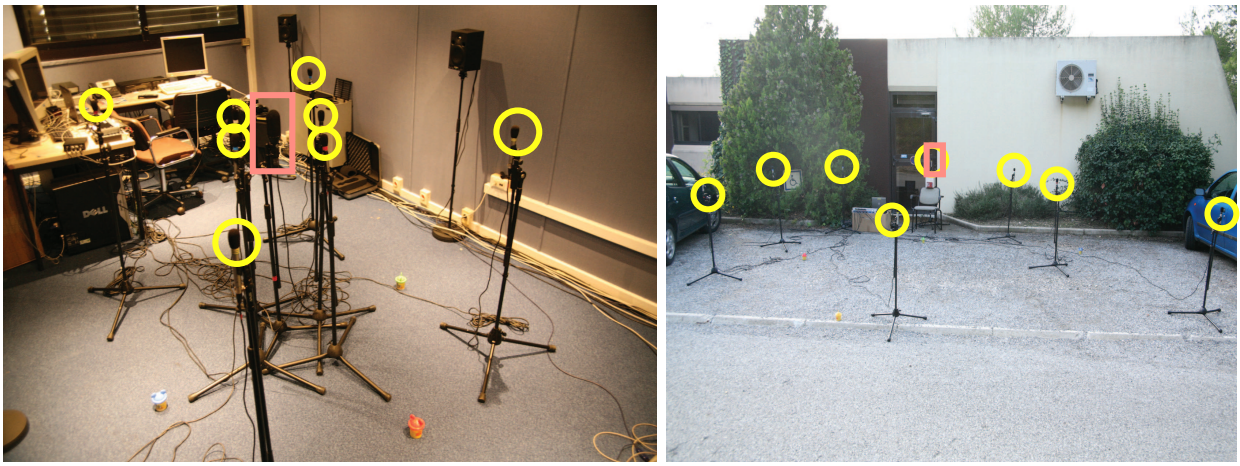


Fig. 6: Example recording setups. We used 8 omnidirectional microphones (circled in yellow) to capture the auditory scene as well as a *Soundfield* microphone (highlighted with a light red square) to simultaneously record a B-format version. A binaural recording using microphones placed in the ears of a subject provided a reference recording in each test case.

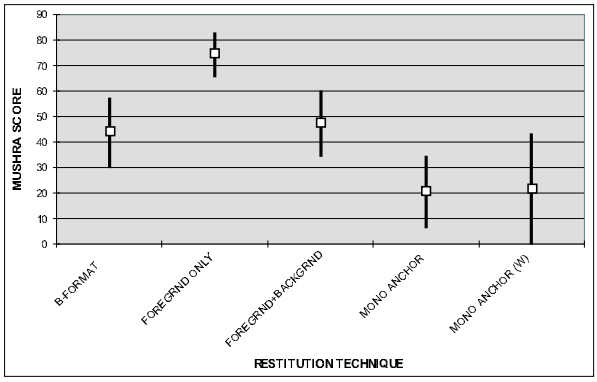


Fig. 7: Average MUSHRA scores and 95% confidence intervals for all subjects and all scenarios.

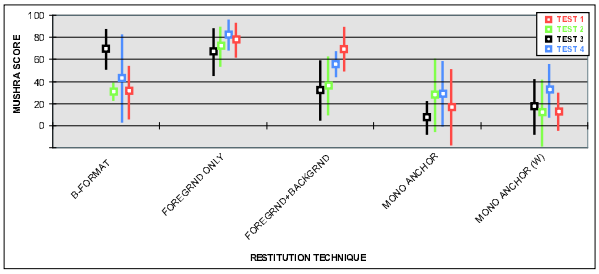


Fig. 8: Average MUSHRA scores and 95% confidence intervals for all subjects in each of our 4 test scenarios.

received scores very close to the B-format reproduction. This is probably due to the low spatial resolution of B-format but could also arise from a non-optimal HRTF-based decoding.

Looking at the various test-cases in more detail, Figure 8 highlights a significantly different behavior for the indoor scenario (TEST#3). In this case, very little background sound was present, hence our approach based on background and foreground separation did not lead to any improvement and, in fact, resulted in a degraded spatial impression. The B-format reproduction, however, obtained significantly better scores in this case, probably due to the favorable configuration of the three speakers (one in front, one to the left, and one to the right).

3.3. Discussion

In terms of audio quality, feedback from the subjects of the tests shows that our improved algorithm outperforms the previous foreground-only solution. This is of course due to the smoothly varying background and more robust foreground estimates. However, our proposed approach appears less convincing in terms of localization accuracy. Significant parts of the foreground sounds can still be present in the background component and will be spatialized using a different strategy. The resulting

Fig. 9: Recording setup used for the seashore recordings.

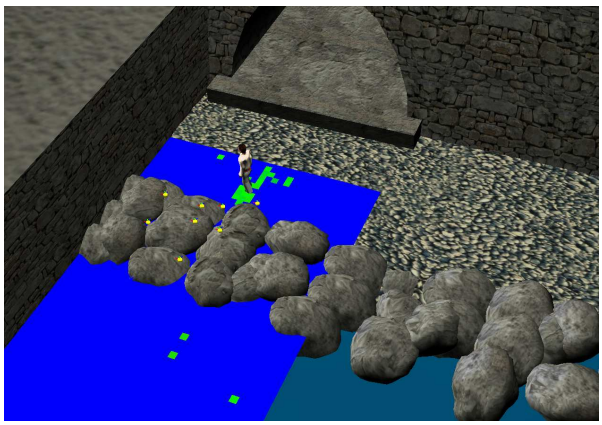


Fig. 10: Example virtual reconstruction of a seashore with walking pedestrian. Yellow spheres correspond to the locations of the microphones used for recording.

blend tends to blur out the localization cues leading to a poorer spatial impression. Improving the quality of the segmentation would probably lead to better results. Another possibility would be to use energy and not only time-differences of arrival to extract possible localization information for the background component.

We used a small number of frequency subbands in our tests which can challenge our time-frequency orthogonality assumption resulting in noisier position estimates for the foreground component. However, we obtained less convincing results with an increased number of frequency subbands due to less accurate correlation estimates for narrower subbands signals.

We do not currently model sources "at infinity", which may appear in the background but also in the foreground component. Our position estimation can return erroneous position estimates in this case due to the limited extent of our position histogram. This could also explain the perceived degradation of spatial cues compared to the reference. Explicit detection of far-field sources is a component we are planning to add in the near future. Finally, non-individualized HRTF processing could also be a major cause of spatial degradation. Running the test with head-tracking and individualized HRTFs might lead to improved results.

4. APPLICATIONS

Our approach can lead to spatial audio coding applications for live audio footage in a way similar to [28, 13, 29, 6], but it also offers novel decoding/authoring capabilities not available with previous techniques such as *free-viewpoint* walkthroughs. Figure 10 illustrates the virtual reconstruction of a seashore scene with a pedestrian walking on a pebble beach recorded with the setup shown in Figure 9. A spatial energy map is overlaid, highlighting the location of foreground time-frequency atoms. Note how the position of the footsteps sounds is well reconstructed by our approach. The sound of sea waves hitting the rocks on the shore is mostly captured by the background component (see also Figure 5). Please, visit the web pages mentioned in Sections 1 and 3 for example audio files and videos.

Spatial re-synthesis with free-moving listener

Our approach allows for a "free-viewpoint" spatial audio rendering of the acquired soundscapes. As the virtual listener moves throughout the scene, the foreground component is rendered using a collection of point sources corresponding to each time-frequency atom, as described in section 2.3. The background component is simply rotated based on the current orientation of the listener in order to provide a consistent rendering. Our representation encodes spatial cues in world space and can thus be rendered on a variety of reproduction setups (headphones, multichannel, *etc.*).

Background/foreground editing

Our two-layer model allows for independent control of the background and foreground components. Their overall level can be adjusted globally or locally, for instance to attenuate foreground sounds with local virtual occluders while preserving the background. The foreground

events can also be copied and pasted over a new background ambiance.

Re-rendering with various microphones

Finally, the microphones used for the analysis process can be different from the one used for re-rendering. For instance, it is possible to use any directional microphone to get a combined effect of spatial rendering and beamforming.

5. CONCLUSION

We presented an approach to convert field recordings into a structured representation suitable for generic 3D audio processing and integration with 2D or 3D visual content. It applies both to outdoor environments or indoor environments with limited reverberation, provides a compact encoding of the spatial auditory cues and captures propagation and reverberation effects which would be very difficult to render with the same level of realism using traditional virtual acoustics simulations.

Perceptual comparisons with reference binaural and B-format recordings showed that our approach outperforms B-format recordings and can get close to reference binaural recordings when all time-frequency atoms are rendered as foreground point sources. However, artefacts due to background noise lead to reduced signal quality. An alternative solution was proposed based on the explicit segmentation of stationary "background noise" and non-stationary "foreground events". While the signal quality is significantly improved when re-rendering, spatial cues were perceived to be degraded, probably due to non-optimal background separation.

In the future, we would like to improve on our background/foreground segmentation approach, possibly based on auditory *saliency* models [17] or taking advantage of the signals from all microphones. Alternative sparse representations of the signals [22, 21] could also be explored in order to improve our approach. Further comparisons to other sound-field acquisition techniques, for instance based on high-order spherical harmonic encoding [3, 24], Fourier-Bessel decomposition [19, 20] or directional audio coding [27, 28] would also be of primary interest to evaluate the quality vs. flexibility/applicability tradeoffs of the various approaches. 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.

6. ACKNOWLEDGMENTS

This research was made possible by a grant from the *région PACA* and was partially funded by the RNTL project OPERA (http://www-sop.inria.fr/reves/OPERA).

7. REFERENCES

[1] EBU subjective listening tests on internet audio codecs. *EBU TECHNICAL REVIEW, European Broadcast Union (EBU)*, june 2000.

[2] EBU subjective listening tests on low-bitrate audio codecs. *Technical report 3296, European Broadcast Union (EBU), Projet Group B/AIM*, june 2003.

[3] T. Abhayapala and D. 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.

[4] F. Baumgarte and C. Faller. Binaural cue coding - part I: Psychoacoustic fundamentals and design principles. *IEEE Transaction on Speech and Audio Processing*, 11(6), 2003.

[5] D. R. Begault. *3D Sound For Virtual Reality and Multimedia.* Academic Press, Inc., 1994.

[6] J. Breebaart, J. Herre, C. Faller, J. Rödén, F. Myburg, S. Disch, H. Purnhagen, G. Hotho, M. Neusinger, K. Kjörling, and W. Oomen. MPEG spatial audio coding/MPEG surround: Overview and current status. *Proc. 119th AES Convention, New York, USA. Preprint 6599*, October 2005.

[7] J. Chen, J. Benesty, and Y. A. Huang. Time delay estimation in room acoustic environments: An overview. *EURASIP Journal on Applied Signal Processing*, 2006.

[8] C. Choi. Real-time binaural blind source separation. *Proc. of the 4th Intl. Symp. on Independant Component Analysis and Blind Source Separation (ICA2003), Nara, Japan, april*, 2003.

[9] 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*, ASSP-32(6):1109–1121, December 1984.

[10] C. Faller and F. Baumgarte. Binaural cue coding - part II: Schemes and applications. *IEEE Transaction on Speech and Audio Processing*, 11(6), 2003.

[11] O. Faugeras. *Three-dimensional computer vision: a geometric viewpoint*. MIT Press, 1993.

[12] E. Gallo, N. Tsingos, and G. Lemaitre. 3D-Audio matting, post-editing and re-rendering from field recordings. *EURASIP Journal on Applied Signal Processing, special issue on Spatial Sound and Virtual Acoustics*, 2007.

[13] M. Goodwin and J.-M. Jot. Analysis and synthesis for universal spatial audio coding. *In 121th AES Convention, San Francisco, USA. Preprint 6874*, 2006.

[14] International Telecom. Union. Method for the subjective assessment of intermediate quality level of coding systems. *Recommendation ITU-R BS.1534-1*, 2001-2003.

[15] 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.

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

[17] C. Kayser, C. Petkov, M. Lippert, and N. Logothetis. Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, 15:1943–1947, Nov. 2005.

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

[19] 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.

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

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

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

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

[24] 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.

[25] 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.

[26] R. D. Patterson and B. C. J. Moore. *Auditory filters and excitation patterns as representations of auditory frequency selectivity, in Frequency Selectivity in Hearing, Academic Press, London*, pages 123–177. 1986.

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

[28] V. Pulkki and C. Faller. Directional audio coding: Filterbank and stft-based design. *In 120th AES Convention, Paris, France, Preprint 6658.*, May 20-23 2006.

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

[30] R. Radke and S. Rickard. Audio interpolation. *In the Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio (AES'22), Espoo, Finland*, pages 51–57, 2002.

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

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

[33] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen. Creating interactive virtual acoustic environments. *J. of the Audio Engineering Society*, 47(9):675–705, Sept. 1999.

[34] SOUNDFIELD. http://www.soundfield.com.

[35] R. Streicher. The decca tree - it's not just for stereo anymore, http://www.wesdooley.com/pdf/surround sound decca tree-urtext.pdf., 2003.

[36] H. Traunmüller. Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88:97–100, July 1990.

[37] E. Vincent, C. Févotte, R. Gribonval, X. Rodet, É. L. Carpentier, L. Benaroya, A. Rödel, 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*, 2003.

[38] Ö. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7), July 2004.