SPATIAL IMPULSE RESPONSE RENDERING: A TOOL FOR REPRODUCING ROOM ACOUSTICS FOR MULTI-CHANNEL LISTENING

VILLE PULKKI AND JUHA MERIMAA

Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, P.O.Box 3000, FI-02015 TKK, Finland

Spatial Impulse Response Rendering (SIRR) can be used to reproduce room acoustics with any multichannel loudspeaker system. In SIRR the time-dependent direction of arrival and diffuseness of measured room responses are analyzed within frequency bands. A multichannel response suitable for reproduction with any chosen surround loudspeaker setup is synthesized using the analysis data. The resulting loudspeaker responses are used in a multichannel convolving reverberator, and the synthesized responses create a natural perception of space corresponding to the measured room. This paper provides a technical description of the analysis-synthesis method, and discusses the differences and similarities of SIRR to convenctional microphone techniques.

1. INTRODUCTION

Extensive possibilities are available with current multichannel loudspeaker systems for reproduction of spatial sound. A standard 5.1 loudspeaker setup is able to create a surrounding perception of sound with fairly good directional accuracy especially in front of the listener. By adding more channels, the precision can be further enhanced. Nevertheless, conventional microphone
techniques have problems coping with the wide variety of existing loudspeaker setups, as well as with fully utilizing the spatial resolution available for reproduction.

In close microphone techniques, several spot microphones are placed close to sound sources to yield fairly "dry" source signals with ideally no audible
room effect. An artificial scene is then constructed by positioning these signals in desired directions using, for instance, amplitude panning. Spatial impression is created with the help of reverberators or by adding the signals of additional microphones placed further away from the source(s) in the recording room.

With convolving reverberators it has become possible to simulate recording in any performance venue using close microphone recordings and measured room responses. However, the problems of conventional microphone techniques also apply to capturing the responses. Spatial Impulse Response Rendering (SIRR) [1]–[6] has been designed to overcome some of the problems by using a perceptually motivated analysis-synthesis approach. The SIRR processing consists of analysis of direction and diffuseness of sound within frequency bands, followed by synthesis yielding multichannel impulse responses that can be tailored for an arbitrary loudspeaker system. Although applicable to general recording as well, SIRR is especially suitable for processing room responses for convolving reverberators, which is the context it will be described in in this paper.

2. PSYCHOACOUSTICAL BACKGROUND

Due to practical limitations (as will be discussed in Section 5), no established recording and reproduction technique can perfectly recreate the sound field of a recording space in a listening room. However, typically this is not even the goal of recording. Instead of physical accuracy, it is more important to relay a perception. In order to be able to recreate the perception in an existing room or a hall, it is thus important to know what kind of information human listeners uti-

lize in spatial sound perception. In the following, the current knowledge on related binaural psychoacoustics is briefly reviewed.

Human sound localization is based on five frequency-dependent cues. (1) Interaural time difference (ITD) and (2) interaural level difference (ILD) are the dominant cues for determining in which cone of confusion a sound source is located, i.e. in which cone forming a constant angle with the line connecting both ear canal entrances of a listener. Although also depending on the cone of confusion, the most important role of (3) monaural spectral cues is in resolving the direction of the source within the cone of confusion. Furthermore, (4) the effect of head rotation on the previous cues helps in determining the correct direction [7]. Additionally, human listeners are sensitive to (5) interaural coherence (IC, e.g. [8] and references therein), which has been recently proposed to be an important cue for localization in reverberant environments and multi-source scenarios [9]. All the localization cues are individual depending on the shape of the head, pinnae and torso of a listener, and they can be analyzed from measured head-related transfer functions (HRTFs).

Although the current challenges of spatial sound reproduction lie in the spatial impression, timbral perception should not be neglected. Timbre is a complex phenomenon depending on spectral, temporal, as well as spatial properties of sound. Specifically in (simulated) room environments, Brüggen [10] has shown that the perceived coloration due to the acoustical environment cannot be explained without considering binaural perception. This result suggests that incorrect spatial reproduction of sound may also create timbral artifacts.

3. ROOM RESPONSES

Room responses are usually divided into direct sound, early reflections, and late reverberation, all of which are perceived somewhat differently. The direct sound and the early reflections are typically discrete sound events arriving to the receiver from a clear single direction. As the impulse emitted by the source propagates in the room, the density of the arriving reflections grows, and the sound field turns into diffuse late reverberation. There is no specific time instant when this change happens, and the density of reflections as a function of time depends on the size of the room.

Note that when listening to a sound source in a room, the actual signal arriving at the listening position is, of course, the convolution of the source signal with the room response. The convolution modifies the localization cues in a signal dependent manner, which makes it impossible to give a general description of the final cues. For a demonstration of the resulting fluctuations in ITD cues the reader is referred to [11].

4. SPATIAL IMPULSE RESPONSE RENDERING

Based on the physical and psychoacoustical consideration in the previous section, the Spatial Impulse Response Rendering (SIRR) method can now be formulated. The description starts with an explanation of the underlying assumptions, followed by a description of the analysis and synthesis parts of SIRR and an application example.

4.1. Assumptions for spatial sound reproduction

The first assumption for the SIRR processing is that it is not necessary to perfectly reconstruct the original sound field in order to be able to faithfully reproduce the spatial impression of an existing performance venue. Instead of sound field reconstruction, SIRR aims at a time and frequency dependent recreation of features that are relevant for human perception:

- ILD and ITD cues
- monaural localization cues
- interaural coherence
- timbre

The most straightforward analysis method would, of course, be to use the signals of a binaural microphone. However, the translation from the analyzed auditory cues to multichannel reproduction cannot be solved easily. Furthermore, direct analysis and synthesis of binaural cues suffers from individual differences in HRTFs. An easier way to approach the problem is to analyze and synthesize physical properties of the sound field which transform into binaural cues. More specifically, we assume that:

1. Direction of arrival of sound will transform into ITD, ILD and monaural localization cues.
2. Diffuseness of sound will transform into interaural coherence cues.
3. Timbre depends on monaural (time-dependent) spectrum together with ITD, ILD, and IC.
4. When the direction of arrival, diffuseness, and spectrum of sound are reproduced with the temporal and spectral resolution of human hearing, the perceptual quality of the spatial reproduction is good.

5. If the perceptual quality of the spatial reproduction of a room response is good, the perceptual quality of reproduction of sound convolved with the response is also good.

Note that if the direction of arrival is correctly reproduced in a scheme like this, the localization cues will be created by interaction of the sound field with the head of the listener. Thus, the cues will be correct for each individual listener and behave correctly when the listener rotates his/her head.

The following section present the SIRR analysis-synthesis method operating according to the previous assumptions. In the analysis part, the direction of arrival and diffuseness of sound are analyzed. In the synthesis part, an omnidirectional response is then rendered to the channels of a multichannel loudspeaker system according to the analysis data. Both steps are performed with a time-frequency resolution motivated by the psychoacoustics of human hearing.

4.2. SIRR implementation

The SIRR method can be implemented in various ways. The implementation considered in this paper is based on using B-format microphone signals. B-format stands for four audio tracks recorded with four coincident microphones: one omnidirectional pressure microphone and three figure-of-eight (velocity) microphones directed towards three orthogonal Cartesian coordinate axis. B-format recording has been chosen instead of other suitable systems because of its availability, and technical simplicity.

In the analysis part of SIRR, the use of a B-format microphone suggests sound intensity and energy for estimation of the direction of arrival, and diffuseness. The sound intensity vector corresponds to direction and magnitude of net flow of sound energy. Opposite direction of the sound intensity vector is used as an estimate of the direction of arrival, and the ratio between the magnitude of intensity vector and sound energy density is used to compute an estimate for the diffuseness of sound. A short-time Fourier transform (STFT) based method is used for time-frequency analysis.

In the synthesis part of SIRR, each frequency channel and time instant of the omnidirectional signal is reproduced either point-like or diffused or cross-faded between the previous methods depending on the analyzed diffuseness. Pair- or triplet-wise amplitude panning is employed to render non-diffuse sound to the analyzed direction of arrival. The panning is formulated as vector base amplitude panning (VBAP) [12]. The gain factors are normalized using equal-power law, i.e. $\sqrt{\sum g^2} = 1$. For details of the analysis and synthesis procedure, see [4, 6] and Figs. 1 2 and 3.

In the current implementation the diffuse sound is created as a hybrid of two methods. The first method is not to use any specific diffusion technique. With diffuse sound, the analyzed directions of arrival of sound behave in a stochastic manner. When sound is applied to such directions, it will be spread all around and different frequencies are panned to different directions. This produces a fairly good perception of diffuseness, although at high frequencies some movement instability may be perceived, and the reverberant tail appears too short.

The second method is called phase randomization. The phase randomization is performed by creating continuous uncorrelated noise for each loudspeaker, and by setting the magnitude spectrum of each frequency component in each time window equal to the magnitude spectrum of the diffuse energy divided into the loudspeakers. The method corresponds thus to replacing the diffuse parts of the room response with random noise samples with a similar time-frequency envelope. The phase randomization method can create highly decorrelated signals and the time-variant nature of the algorithm avoids most timbral artifacts which might be caused by continuous coherent summation of the signals from different loudspeakers. However, the frequency domain equalization of the magnitudes in the STFT processing may result in time domain aliasing, which manifests itself as unnatural changes in magnitude spectrum at low frequencies.

The hybrid method uses phase randomization at high frequencies, namely above 1000 Hz, and no specific diffusion technique at lower frequencies. In this way, the artifacts of phase randomization at lower frequencies and artifacts of not using a specific diffusion technique at higher frequencies are avoided.

4.3. Modification of room responses using SIRR

The previous parts of the paper have described the analysis and synthesis of a room response such that the perceptual properties of the measured room are preserved as well as possible. However, the utilized parametrization provides also efficient means for modification of the responses. In addition to simple weightings of the time-frequency envelope, the sound field can be easily rotated and it is even possible to change the directions of arrival of some early reflections. Furthermore, the balance between diffuse and non-diffuse sound energy can be adjusted according to the artistic needs in a recording application.

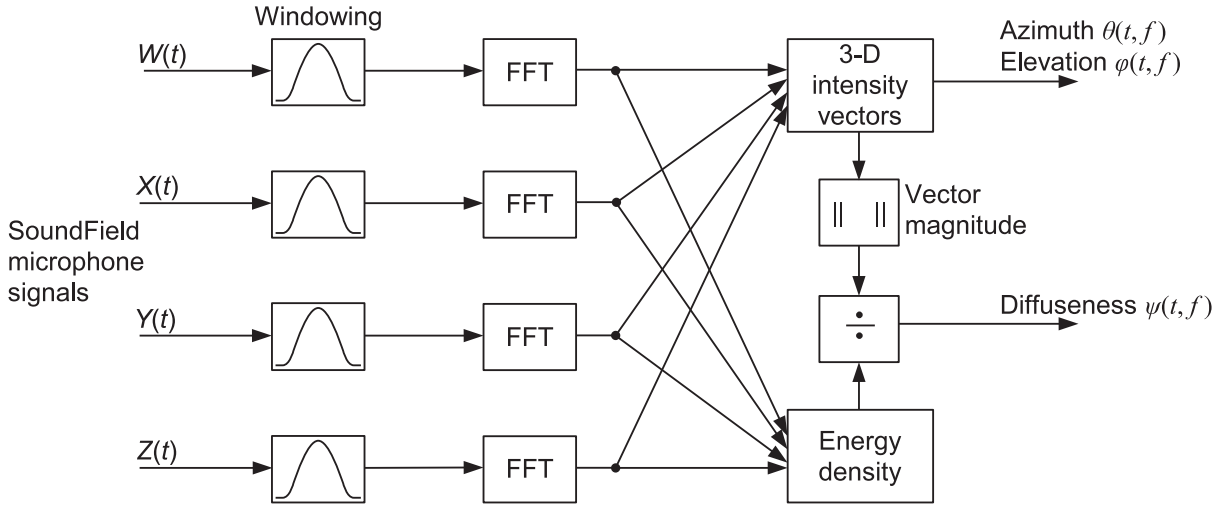

Figure 1: *Directional energy analysis of a B-format room response.*

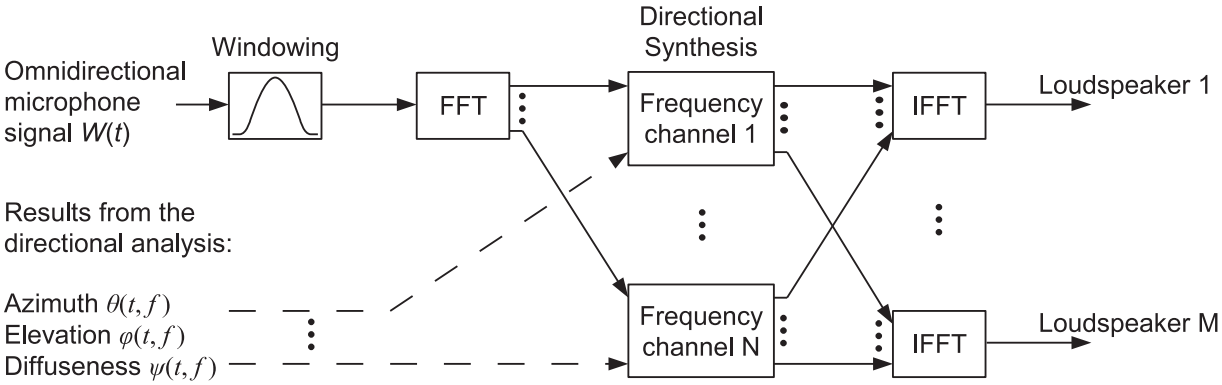

Figure 2: *Directional synthesis based on an omnidirectional room response.*

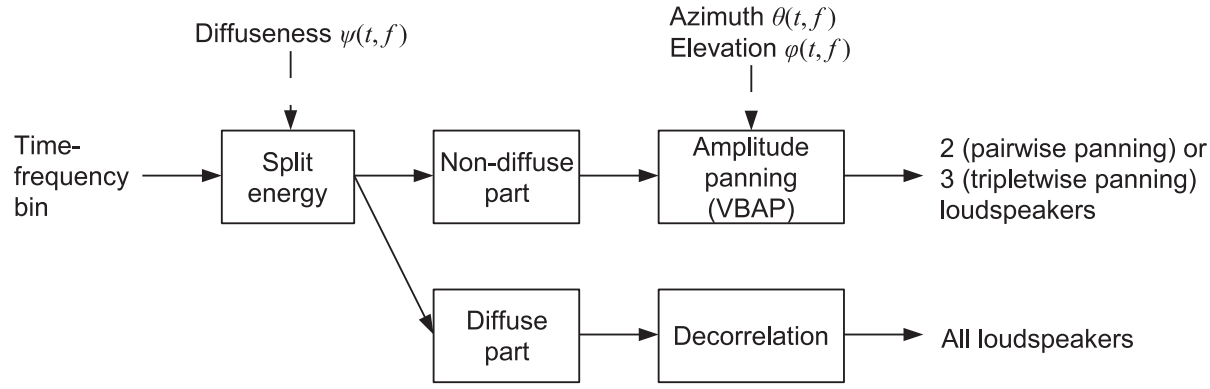

Figure 3: *Processing of a single time-frequency bin in the synthesis.*

4.4. Application example

In this section, the SIRR algorithm is illustrated with an application to a measured B-format room response. The response was taken from a published data set [13] including, among others, B-format and SIRR processed responses of the Promenadikeskus concert hall located in Pori, Finland. The hall is roughly shoebox shaped with 700 seats on a tilting floor. The sound is very diffuse due to diffusers on the walls and on the ceiling. The selected response S1-R4 was measured with an omnidirectional sound source on the stage and a SoundField MkV microphone system at the raised rear part of the floor.

4.4.1. Analysis

Fig. 4 shows the analysis data for a 100 ms sample of the response starting approximately 5 ms before the arrival of the direct sound. The topmost panel is for reference, presenting the envelope of the omnidirectional response $W(t)$. The two middle panels illustrate the time-frequency distribution of the active intensity, and the bottom panel shows the time-frequency distribution of the diffuseness estimate.

The active sound intensity is shown separately for two planes: the horizontal plane and the median plane. The directions of the vectors describe the direction of net flow of sound energy within a time window. In both planes, vectors pointing to the right represent sound propagating towards the back of the hall. In the horizontal plane a vector pointing down signifies sound arriving from the right side when looking towards the front of a hall, and in the median plane a vector pointing down represents sound arriving from above. The lengths of the vectors are proportional to the logarithmic magnitude of the sound intensity component in the plane in question. The vectors are plotted on top of a sound pressure spectrogram calculated from the omnidirectional response W.

The data are shown for a frequency range from 100 Hz to 5 kHz, which appears to give reasonable results with the utilized microphone system. The time-frequency representations have been adjusted for illustrational purposes. The spectrograms and the diffuseness have been calculated from 128 sample long Hann windowed time frames at 48 kHz sampling frequency with zero-padded FFT and largely overlapping windows in order to provide a smoother graphical representation. The intensity vectors are plotted only for positions of local maxima within each frequency band in order to reduce the amount of shown data. Furthermore, the spectrograms and the intensity vectors have been thresholded such that only a range of 25 dB from the maxima is shown.

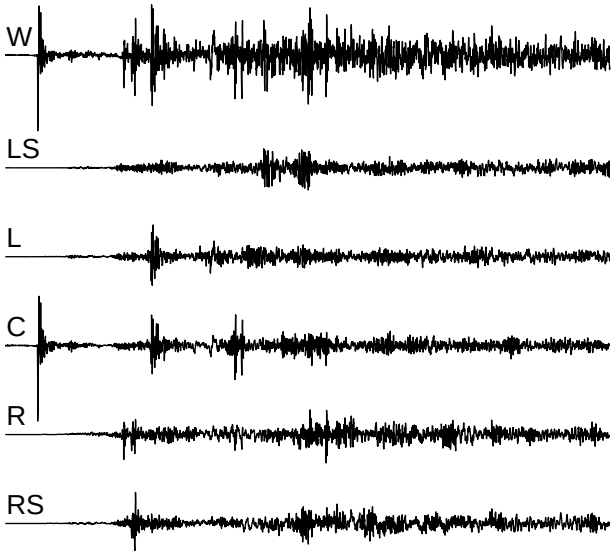

Figure 5: An omnidirectional impulse response W rendered to a 5.0 loudspeaker setup. The amplitudes of the responses are plotted as a function of time.

What can be seen from the analysis data is that the direct sound arrives at approximately 65 ms emanating from a little below the horizontal plane. It is followed by two slightly bandlimited reflections close to each other: one from slightly above on the right side at approximately 80 ms and one from the left at 85 ms. For these sound events the diffuseness estimate shows low values across the frequency. Due to the highly diffuse design of the hall, the next discrete reflections are already more constrained in frequency and in the late part it is impossible to identify any single reflections. Note that the diffuseness estimate appears fairly random apart from the first few discrete events. Investigating the diffuseness in this form is difficult because it has not been weighted with the sound energy. The diffuseness plot includes a lot of statistical fluctuations during low energy parts and even during the measurement noise before the arrival of the direct sound. However, this is the form in which the diffuseness data are used in the synthesis.

4.4.2. Synthesis

Fig. 5 illustrates the synthesis of a set of impulse responses for reproduction with a 5.0 loudspeaker system. The topmost axis shows the same omnidirectional response W as depicted in the envelope and spectrograms plots in Fig. 4. The five lower axes present the synthesized loudspeaker responses for left surround (LS), left (L), center (C), right (R), and right surround (RS) speakers, respectively.

For 2-D reproduction, only the directional analysis

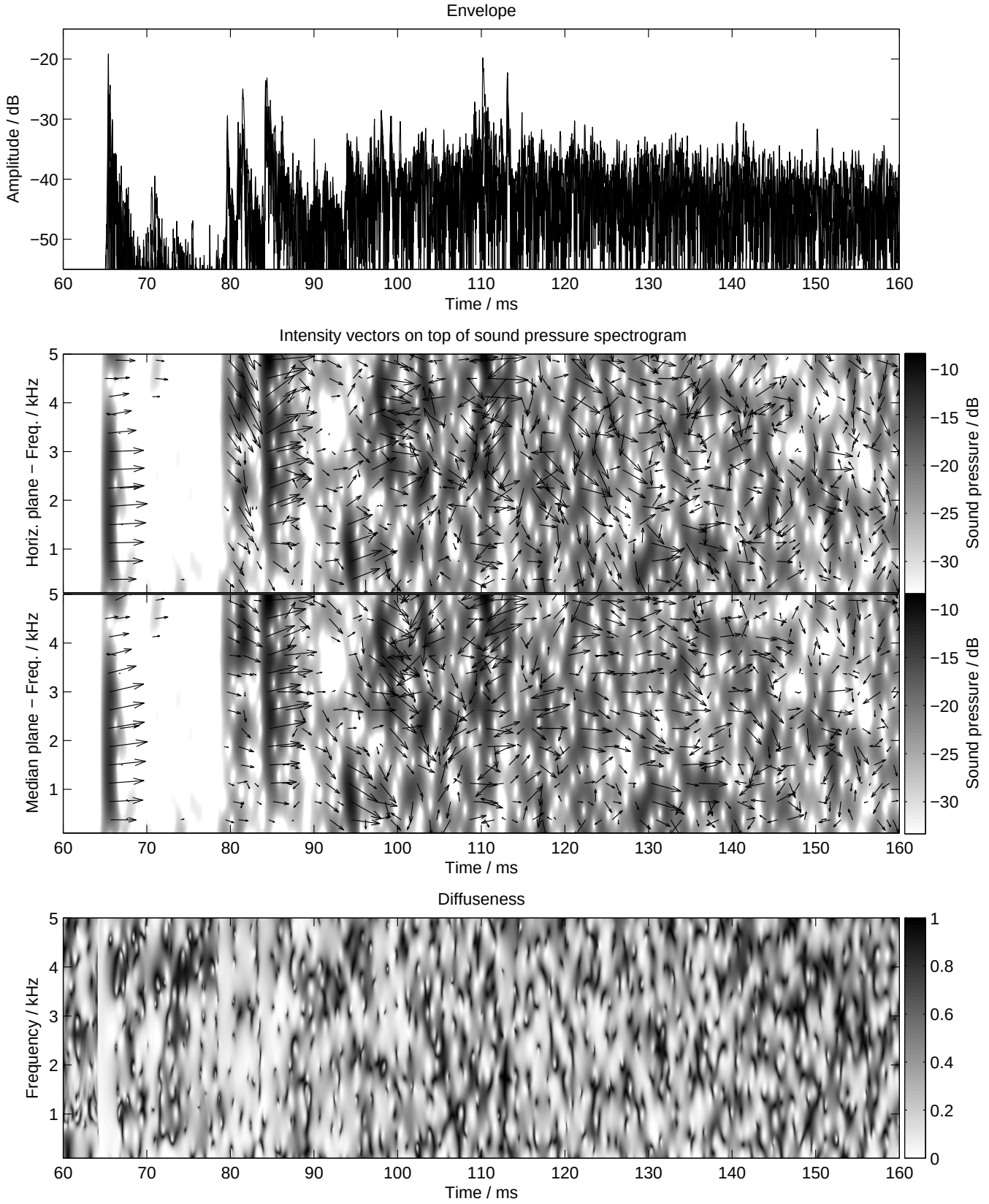

Figure 4: *Analysis results of a concert hall response. Top: Envelope of the omnidirectional response. Middle panels: time-frequency distribution of active intensity vectors plotted on top of a sound pressure spectrogram. Bottom: time-frequency representation of the diffuseness estimate.*

data in the horizontal plane has been used. However, the diffuseness estimate still includes the vertical component. Furthermore, the direct sound has been forced to the center speaker by modifying the analysis data before the synthesis. It can be seen that a major part of the first two reflections is conveyed to the rightmost loudspeakers, followed by a sound event from the front left, as expected based on the analysis data in Fig. 4. The subsequent part is still distributed unevenly to the loudspeaker channels, wheras the late more diffuse part does not show large (statistical) deviations between the channels.

5. COMPARISON OF SIRR WITH EXISTING TECHNIQUES

After presenting the SIRR algorithm, some existing methods for recording and reproduction of spatial sound are now briefly reviewed and compared to SIRR.

5.1. Conventional microphone techniques

From the psychoacoustical point of view, conventional microphone techniques are not limited in time or frequency resolution. Instead, the limitations come in the form of insufficient spatial resolution. Spatial audio or multichannel impulse responses have been typically recorded using one microphone per loudspeaker. Several different microphone configurations have been proposed in the literature, and they can be divided into coincident, quasi-coincident, and spaced setups [14].

It has been shown that coincident microphone techniques are able to produce sharpest virtual sources [14, 15]. In coincident setups, a group of directional microphones are positioned as close to each other as possible, so that the sound signal from a single sound source is ideally captured in the same phase with all the microphones. The microphones should have orientations and directivities corresponding to the loudspeaker configuration, so that sound from any specific direction would only be picked up by few microphones. Using more loudspeakers requires thus narrower directional patterns. However, with conventional microphone technology, narrow enough broad band patterns cannot be achieved. Consequently, the sound from any direction is always picked up by several microphones, which results in a blurred and colored reproduction due to the crosstalk between loudspeaker channels.

Ambisonics [16] is a special form of coincident microphone techniques, which tries to solve the directivity problem by employing a spherical harmonic decomposition of the sound field. In theory it can accurately reproduce a directional sound field in a small sweet spot by the sympathetic operation of all loudspeakers in an arbitrary surround setup. Microphone technology, however, limits the order and thus the directional resolution of Ambisonics. Furthermore, the presence of the head of the listener further disrupts the ideal operation of Ambisonics. For conventional first-order implementations, the technique reduces to using a set of virtual coincident microphones that can be adjusted in the post processing phase. The problems are also similar to those discussed in the previous paragraph.

In contrast to coincident techniques, spaced microphones are positioned at considerable distances from each other. The sound from a single source is thus captured in different phases by different microphones. In a reverberant environment the resulting microphone signals will also be to a certain degree decorrelated. The noncoincident techniques are often said to create a better feeling of “airiness” and “ambience”, and the reproduction is less sensitive to the location of the listener. However, the directional accuracy is even lower than what can be achieved with a coincident microphone setup.

Quasi-coincident microphone setups can be seen as a compromise between coincident and spaced techniques. The microphones are placed close to each other but not coincident. The resulting characteristics of reproduction also lie in between those of the coincident and spaced microphone techniques.

5.2. Characterization of SIRR

In terms of the previous discussion, room responses processed with SIRR can be characterized as follows. In a large concert hall, the direct sound and early reflections are relatively sparse in time and they can usually be individually analyzed and synthesized. As non-diffuse sound, they are synthesized as point-like virtual sources using amplitude panning. The reproduction resembles coincident microphone techniques where SIRR can be thought to adaptively narrow the microphone beams in order to get the best possible directional accuracy. On the other hand, the late reverberant part of a room response is reproduced largely as decorrelated sound emanating from all loudspeakers. This is similar to spaced microphone techniques, and the pleasent “airiness” or “ambience” of the room should be preserved. In a smaller room the reflections are more dense, which means that fewer reflections can be individually processed and some of the directional resolution is thus lost.

SIRR has been tested with subjective tests in anechoic listening [5]. In principle, the reproduction should be compared to the reference. With spatial audio this is impossible, since the listening system can not be inside the space under reproduction.

This problem was solved by generating the reference sound and reproduced sound with the same loudspeaker system in anechoic conditions. The reference was a virtual reality sample, which was created using the image-source method with 16 loudspeakers in 3-D positioning. The impulse response of the virtual reality was then reproduced with the same loudspeakers using different reproduction techniques. It was found that when the reverberation time RT_{60} was 1.5 s, the listeners could not distinguish between reference and SIRR reproduction. With shorter reverberation times, the difference was perceived larger, however it was never graded annoying.

6. SUMMARY

Spatial Impulse Response Rendering (SIRR) is a method for reproduction of measured room responses with an arbitrary multichannel loudspeaker system. Compared to conventional microphone techniques, SIRR is able to improve the directional quality of the reproduction by using a psychoacoustically motivated analysis-synthesis procedure. When loaded to a convolving reverberator, the synthesized responses create a very natural spatial impression.

7. ACKNOWLEDGMENTS

The work of Juha Merimaa has been supported by the research training network for Hearing Organisation And Recognition of Speech in Europe (HOARSE, HPRN-CP-2002-00276). Ville Pulkki has received funding from the Academy of Finland (project 105780).

The SIRR method is protected by an international patent application [17] and it has been licensed to Waves ltd.

REFERENCES

[1] J. Merimaa and V. Pulkki. Perceptually-based processing of directional room responses for multichannel loudspeaker reproduction. In *IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, pages 51–54, New Paltz, NY, USA, 2003.

[2] V. Pulkki, J. Merimaa, and T. Lokki. Multichannel reproduction of measured room responses. In *International Congress on Acoustics*, pages II 1273–1276, Kyoto, Japan, 2004.

[3] V. Pulkki, J. Merimaa, and T. Lokki. Reproduction of reverberation with spatial impulse response rendering. In *AES 116th Convention*, Berlin, Germany, 2004. Preprint 6057.

[4] J. Merimaa and V. Pulkki. Spatial Impulse Response Rendering. In *7th International Conference on Digital Audio Effects*, pages 139–144, Naples, Italy, 2004.

[5] V. Pulkki and J. Merimaa. Spatial impulse response rendering: Listening tests and applications to continuous sound. In *AES 118th Convention*, Barcelona, Spain, 2005. Accepted for publication.

[6] J. Merimaa and V. Pulkki. Spatial Impulse Response Rendering I: Analysis and synthesis. *J. Audio Eng. Soc.*, 2005. Submitted.

[7] J. Blauert. *Spatial Hearing*. The MIT Press, Cambridge, MA, USA, revised edition, 1997.

[8] S. E. Boehnke, S. E. Hall, and T. Marquadt. Detection of static and dynamic changes in interaural correlation. *J. Acoust. Soc. Am.*, 112(4):1617–1626, 2002.

[9] C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Am.*, 116(5):3075–3089, 2004.

[10] M. Brüggen. Coloration and binaural decoloration in natural environments. *Acta Acustica — Acustica*, 87:400–406, 2001.

[11] R. Mason and F. Rumsey. Interaural time difference fluctuations: Their measurement, subjective perceptual effect, and application in sound reproduction. In *AES 19th International Conference*, Schloss Elmau, Germany, 2001.

[12] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc.*, 45(6):456–466, 1997.

[13] J. Merimaa, T. Peltonen, and T. Lokki. Concert hall impulse responses — Pori, Finland. http://www.acoustics.hut.fi/projects/poririrs/, 2005.

[14] S. P. Lipshitz. Stereo microphone techniques... Are the purists wrong? *J. Audio Eng. Soc.*, 34(9):716–744, 1986.

[15] V. Pulkki. Microphone techniques and directional quality of sound reproduction. In *AES 112th Convention*, Munich, Germany, 2002. Preprint 5500.

[16] M. A. Gerzon. Periphony: With-height sound reproduction. *J. Audio Eng. Soc.*, 21(1):2–10, 1973.

[17] V. Pulkki, J. Merimaa, and T. Lokki. A method for reproducing natural or modified spatial impression in multichannel listening. Patent application FI 20030294 / PCT/FI 2004/000093, 2004.

