Multi-channel reproduction of measured room responses

Ville Pulkki[1], *Juha Merimaa*[1,2] *and Tapio Lokki*[3]

[1]Laboratory of Acoustics and Audio Signal Processing,
Helsinki University of Technology
[2]Institut für Kommunikationsakustik,
Ruhr-Universität Bochum
[3]Telecommunications Software and Multimedia Laboratory,
Helsinki University of Technology
Ville.Pulkki@hut.fi

Abstract

Spatial Impulse Response Rendering (SIRR) has been recently proposed for spatial reproduction of room acoustics. In the method, a multi-channel impulse response of a room is measured, and responses for loudspeakers in arbitrary multi-channel setup are computed. When loaded to a convolving reverberator, the responses will then produce a similar perception of space as the measured room. The method is based on measuring with a SoundField microphone or a comparable system, and on analyzing direction-of-arrival and diffuseness at frequency bands. In this paper the reproduction quality is evaluated with listening tests, and it is found that it yields a natural spatial reproduction of the acoustics of a measured room.

1. Introduction

In the recent years multichannel loudspeaker reproduction systems have become increasingly common. A standard 5.1 setup is able to produce a surrounding sound field with fair directional accuracy especially in front of the listener. By adding more channels, the precision can be further enhanced, or the reproduction can be extended to 3-D. However, due to limitations of microphone technology, no current recording systems can fully exploit such possibilities.

In a typical recording scenario 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 by adding the signals of some microphones placed further away from the sources in the recording room, or with the help of reverberators. In some recent devices it has also become possible to use actual measured room responses with real-time convolution. The problem is—as in surround sound recording—how to capture such responses so that the perceived spatial impression of the measured room or hall is accurately reproduced.

In order to overcome some of the recording problems, a method called Spatial Impulse Response Rendering (SIRR) has been recently proposed for processing directional room responses [1]. The required responses can be measured with commercial SoundField or Microflown systems, or with a suitable custom microphone array. The method yields multichannel impulse responses that can be tailored for an arbitrary surround loudspeaker system in the postprocessing phase. In this paper, the SIRR method is reviewed, and listening tests are performed to investigate the perceptual quality of SIRR reproduction compared to some other systems.

2. Spatial Impulse Response Rendering

The SIRR method is presented in [1], and it is briefly reviewed here. Using a chosen loudspeaker setup, the method aims at recreating binaural cues that would occur to a listener in the room and position where the impulse response was measured. The cues are not measured or reproduced directly. Instead, the direction of the sound is estimated as a function of frequency and time, and in the reproduction phase the sound at each frequency band is spatialized to the estimated directions in each time window. Thus, we can assume that the the produced directional cues resemble those that would occur to a listener in the measurement position.

The angle of arrival and diffuseness of a measured directional room response at each frequency band are derived from an estimate of active sound intensity. For the time-frequency processing we have adopted a Short-Time Fourier Transform based scheme common in audio coding applications. Similar processing, including the analysis of the active intensity, could also be realized using an analysis-synthesis implementation of an auditory filter bank. However, Baumgarte and Faller [2] found the computationally more efficient FFT implementation to perform equally well with auditory filter bank in their experiments with the Binaural Cue Coding (BCC) algorithm sharing some features with our processing scheme.

In the method, the responses are first divided into short overlapping time frames. Processing of each time frame consists of the following steps:

1 Calculate the FFT of the sound pressure signal.
2 Calculate the frequency distribution of the active intensity.
3 Estimate the diffuseness of the sound field in the frequency domain, based on the ratio of the magnitudes of the sound pressure and the active intensity vector.
4 Based on the diffuseness estimate, spatialize each frequency bin of an omnidirectional sound pressure signal. If the diffuseness is low, spatialization is performed point-like to the direction of the intensity vector. If it is high, sound is spatialized in a manner which produces a perception of a spreaded source.
5 Calculate the IFFTs of the frequency domain loudspeaker signals resulting from the previous step.

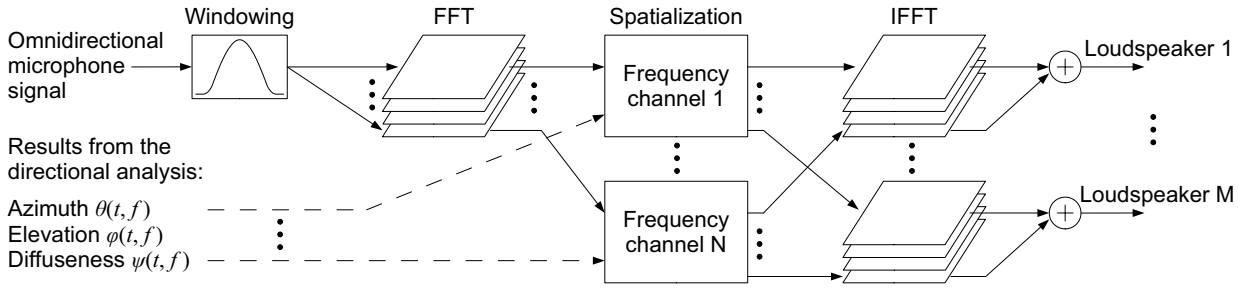

Figure 1: *Spatialization of one time window of the omnidirectional signal based on the directional analysis data.*

The synthesis part of the method is illustrated in Fig. 1. When combined, the processed time frames result in a perceptually reconstructed multichannel room response suitable for convolution for loudspeaker reproduction. For directional positioning of the frequency bins, any spatialization method can be applied. In this study, 3-D Vector Base Amplitude Panning (VBAP) [3] was used. The time frames consisted of 2.5 ms Hann windowed signal segments padded with 2.5 ms of zeros.

Diffuse spatialization was performed by applying the monophonic sound to each loudspeaker of the reproduction system, after randomizing the phase spectrum, similarly as done in [4]. The loudspeakers then emanate uncorrelated sound having the same magnitude spectrum as the monophonic sound in each time window. This technique is denoted *diffusion*.

The diffuseness of the sound field was computed as the ratio of the sound energy and magnitude of the active intensity vector in each time window. However, during the testing of SIRR a new problem was found. The diffuseness estimate yields high values whenever the net flow of sound energy is low, i.e. there are wavefronts propagating to opposite directions within an analysis window. When using very short time frames, the estimates fluctuate considerably, and even in the late diffuse part of a response high values often occur. In such a case, the spatial features of the original sound field are reproduced correctly, but the level of the late reverberation is too low. This is due to the fact that the rapidly changing spatialization direction results in loudspeaker signals that are effectively in random phase. When these signals sum up in the listening position, they partially cancel each other resulting in a lower level than intended. This problem was partially solved in this development phase by manually setting early part of response to be nondiffuse and late part to be maximally diffuse, and aligning their loudnesses separately. However, we will return to this question in near future.

3. Listening tests

Formal listening tests were conducted to evaluate the quality of SIRR and other related systems. In evaluation of spatial sound reproduction systems, there is always a problem that the spatial sound in the recording room cannot be directly compared with reproduced spatial sound in the listening room. In this study, evaluation is done by creating first as naturally-sounding virtual reality as possible with a high number of loudspeakers in an anechoic chamber, and then by reproducing this virtual reality with SIRR and other techniques. This is illustrated in Fig. 2.

3.1. Reference virtual reality

The reference virtual acoustic reality was created with the DIVA software [5], which models the direct sound and early reflections with image-source method, and late reverberation statistically. The frequency-dependent air absorption and surface reflection absorption were modeled with digital filters. To create the reference response we applied a simple room geometry, depicted in Fig. 3. With the image-source method, up to 7th order image sources were computed resulting in 285 early reflections. The walls of the virtual room were set to be concrete, and the ceiling was assumed to be mineral wool. Late reverberation was simulated using linearly rising and exponentially decaying noise, the level and decay rate of which was fitted to the early response. The reverberation time RT_{60} was 0.8 s. The summed response of all the loudspeakers is presented in Fig. 4. At frequencies above 4 kHz, the decay rate was faster. The noise simulating the reverberation was uncorrelated between loudspeakers. A loudspeaker system with 16 loudspeakers (Fig. 5) was used in the tests.

3.2. Reproduced virtual reality

Six different systems were chosen to be tested:

- Virtual reality (reference)
- SIRR for the whole IR
- SIRR for 2 first ms, rest diffused
- SIRR for 30 first ms, rest diffused
- whole IR diffused
- Ambisonics
- Ambisonics first 30 ms, rest diffused

The reproduction of the 16-channel virtual reality was simulated with a virtual microphone setup in the best listening position. This was preferred to physical measurement of the response in order to avoid any differences in the reference and reproduction due to unideal properties of the microphones and loudspeakers. The virtually measured impulse response was then reproduced using same loudspeaker layout with techniques mentioned above, resulting in an impulse response for each loudspeaker with each reproduction method. The Ambisonics system was formulated as in [6], using hypercardioid directionality in decoding stage. The loudness of each system was set to be equal in listening position.

The impulse responses were convolved with two different anechoic sound stimuli. The sounds used were a drum sample with three snare drum shots, and a male talker pronouncing the words “in language”. It was assumed that the drum shots would reveal mostly differences in spatial perception, and speech sample would reveal mostly colorations due to the systems.

4. Test procedure

The test was conducted as an A/B scale with hidden reference. The reference was the virtual reality sample, and the scale was selected according the ITU scale [7] 5.0 = imperceptible, 4.0

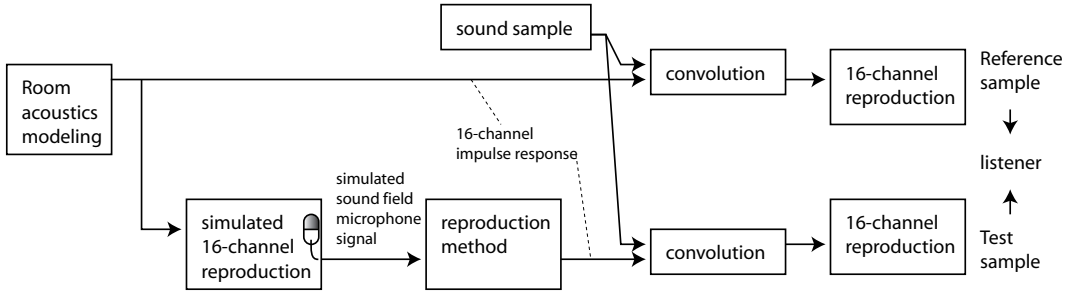

Figure 2: *Method for investigating spatial reproduction quality. A virtual reality sample is compared to its reproduction.*

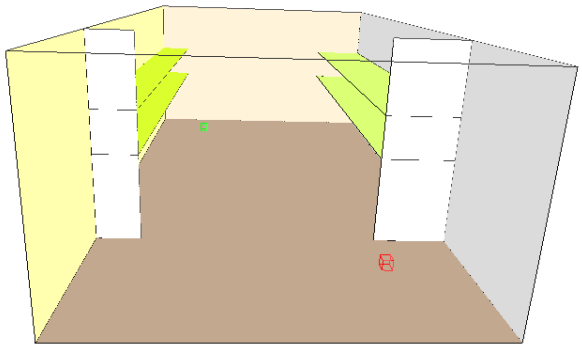

Figure 3: *The simple concert hall model (dimensions 20x12x7 meters) applied in reference response creation. The red box is the sound source and the green box is the receiver.*

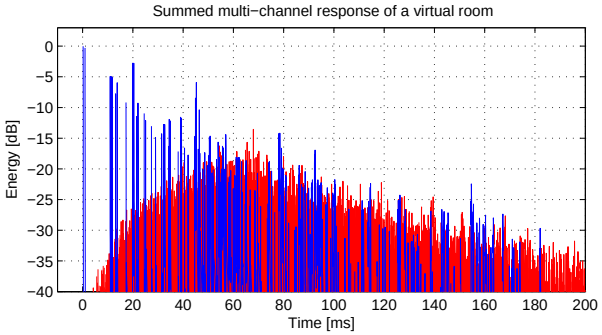

Figure 4: *The multi-channel response of a virtual room being summed to a monophonic energy-time response. The direct sound and early reflections computed with image-source method are plotted with blue, and noise simulating reverberation with red. The multi-channel response was used as reference in listening tests.*

= perceptible but not annoying, 3.0 = slightly annoying, 2.0 = annoying, and 1.0 = very annoying. Listeners could choose between 1 and 5 with increments of 0.5. They were asked to listen to three aspects in reproduction: coloration, localization, and sence of space, and to give one overall rating. The listeners could listen to each sample pair as many times as necessary.

Before the test, subjects were allowed to listen to different samples for five minutes. After this, the subjects conducted the test twice, and the results from latter run were taken to data analysis. In the test, each subject rated each sample pair four times. The order of the sample pairs was randomized. The reference was hidden, which means that it could be either of the samples in a pair. Seven listeners took part in the test, the authors of this paper did not perform it.

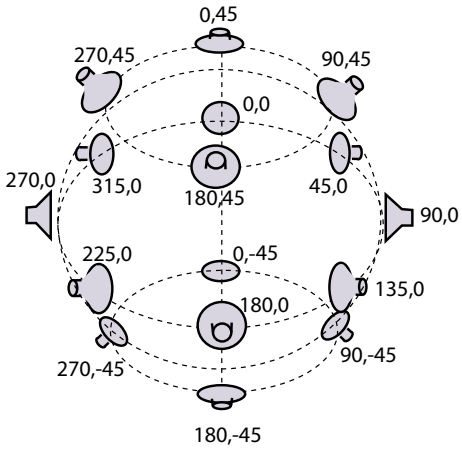

Figure 5: *The loudspeaker system employed in the listening tests.*

5. Test results

The mean values and 95% confidence intervals of the listeners' responses are shown in Fig. 6. In both cases, a similar trend can be seen. Either the SIRR method or diffusion alone did not give the best result. When direct sound was reproduced with SIRR, and all later parts with the diffusion method, quality was better. Still, Somewhat better results were obtained when most of early reflections (30 first ms) were reproduced with SIRR, and reverberant part with the diffusion method.

The reason why the SIRR method with diffusion was judged better than SIRR alone is because the reverberant part of response is reproduced with too low level in SIRR alone, as discussed in Ch. 2. When the reverberant part is reproduced with the diffusion method, the level is correct. However, when only diffusion is used, the direct sound and early reflections are not sharply localizable, which degrades the quality of spatial reproduction.

In the speech case, results for SIRR are very promising. When first 30 ms were reproduced with it and the rest with diffusion, the difference was found almost imperceptible, having a mean value of 4.7. In the drum case, the difference is larger. SIRR 30 ms was graded to a mean value of 3.7. The first thought of this is naturally that the possible flaws in reproduction of early response are more prominent with impulse-like sound signal. However, according to listeners' experiences, this is not the case. Interestingly, the most prominent difference in reproduction was that the perceived pitch of the snare drum was slightly lower. It is not known which phenomenon caused this. One listener reported that there was a perceivable bass boost at low frequencies. The pitch change could also be caused by

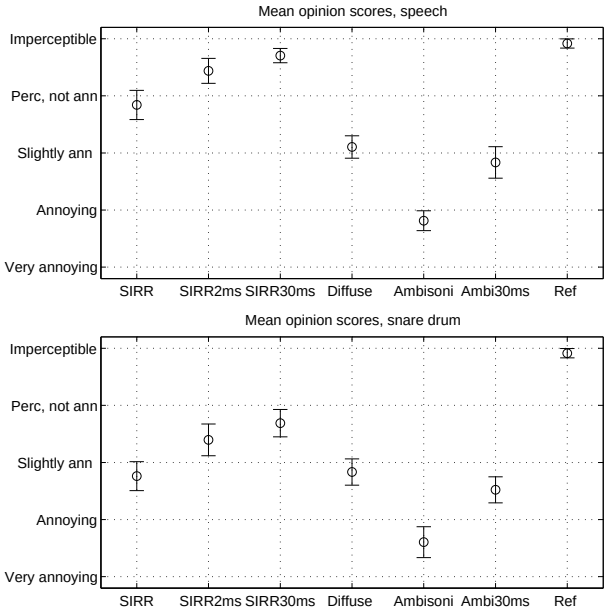

Figure 6: *Rating of different spatial room impulse response reproduction methods using a 16-channel virtual reality as a reference with speech or drum as sound material. Test procedure was A/B scale with hidden reference. Seven listeners conducted the test, and the difference between each sample pair was graded four times.*

some repetition pitch phenomena.

Ambisonics produced poor results. When the late part of the response was reproduced with the diffusion method, a better quality was achieved. However, even in that case the quality was found more annoying than with the diffusion method alone. Listeners reported Ambisonics to sound colored and to be localized intracranially. This is a consequence of the low directional resolution of first-order Ambisonics. The sound from any single direction is reproduced with virtually all loudspeakers, some of which are in opposite phase. The high amount of crosstalk and opposite phases results in severe comb filter effects, and unnatural auditory cues. Less coloring effects could have been obtained by using less loudspeakers. However, to maintain consistency, the loudspeaker setup was the same with all reproduction systems.

6. Conclusions

Spatial Impulse Response Rendering (SIRR) has been recently proposed for spatial reproduction of room acoustics. The algorithm analyzes the direction of arrival and diffuseness of the sound field at frequency bands within time frames. The resulting data is then used to spatialize an omnidirectional room response. Responses can be processed for reproduction with an arbitrary 2-D or 3-D surround loudspeaker system to be used, e.g., in a convolving reverberator. Listening tests were conducted in which the reference was a multi-channel presentation of virtual reality composed with the image-source method and artificial reverberation. This virtual reality was reproduced using SIRR, Ambisonics and diffusion. It was found that the SIRR method produced prominently better reproduction of virtual reality that Ambisonics. With a speech stimulus the difference between SIRR and the reference was almost imperceptible. With drum stimulus the difference was larger.

7. Acknowledgments

Ville Pulkki has received funding from the Academy of Finland (101339). Juha Merimaa has been supported by the research training network for Hearing Organisation And Recognition of Speech in Europe (HOARSE), and by the Graduate School in Electronics, Telecommunications and Automation (GETA),

8. References

[1] J. Merimaa and V. Pulkki, "Perceptual processing of directional room responses for multichannel loudspeaker reproduction," in *Proc. 2003 IEEE Workshop on Applications of Signal Proc. to Audio and Acoust.*, New Paltz, NJ, 2003, pp. 51–54.

[2] F. Baumgarte and C. Faller, "Binaural Cue Coding. Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. Speech Audio Proc.*, vol. 11, no. 6, 2003.

[3] V. Pulkki, "Virtual source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, June 1997.

[4] C. Faller and F. Baumgarte, "Binaural Cue Coding - Part II: Schemes and applications," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, Nov. 2003.

[5] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705, Sept 1999.

[6] G. Monro, "In-phase corrections for ambisonics," in *Proc. Int. Computer Music Conf.*, Berlin, Germany, 2000, pp. 292–295.

[7] ITU-R, "Recommendation BS.1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," International Telecommunication Union Radiocommunication Assembly, 1997.