PERCEPTUALLY-BASED PROCESSING OF DIRECTIONAL ROOM RESPONSES FOR MULTICHANNEL LOUDSPEAKER REPRODUCTION

Juha Merimaa[1,2] *and Ville Pulkki*[1]

[1] Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology
P. O. Box 3000, FIN-02015 HUT, Finland

[2] Institut für Kommunikationsakustik, Ruhr-Universität Bochum
D-44780 Bochum, Germany

juha.merimaa@hut.fi, ville.pulkki@hut.fi

ABSTRACT

A novel method for processing directional room responses is proposed. Responses measured with a SoundField microphone or a comparable system are analyzed with an auditory resolution. From this data, loudspeaker responses for an arbitrary 2-D or 3-D surround sound reproduction system are synthesized. The processed responses yield a sharp and natural directional reproduction of the acoustics of a measured room. The methodology can also be applied to low bitrate audio coding of surround sound.

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 good 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 special devices called reverberators.

The use of reverberators offers the advantage of being able to record in a studio and still create the impression of a larger performance space. Traditionally, reverberators have utilized computationally lighter algorithms to mimic the convolution of source signals with room responses. 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 this paper a novel processing method for directional room responses is proposed. The required responses can be measured with commercial SoundField or Microflown systems, or with a suitable custom microphone array. The method is able to provide multichannel impulse responses tailored for any surround loudspeaker system, resulting in a perceptually accurate reproduction of the measured room or hall within the limits of the chosen loudspeaker configuration.

2. PROBLEMS WITH CONVENTIONAL TECHNIQUES

Spatial audio or multichannel impulse responses have been typically recorded using one microphone per loudspeaker. Several different microphone configurations have been proposed. It has been shown that coincident microphone techniques are able to produce sharpest virtual sources [1, 2]. In coincident setups 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 existing 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 [3] tries to solve the directivity problem 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 practice, however, microphone technology limits the order and thus the directional resolution of Ambisonics, and the presence of the head of the listener further disrupts the ideal operation. Consequently, the technique reduces to using a set of virtual coincident microphones that can be adjusted during playback. The problems are also similar to those discussed in the previous paragraph.

In an alternative technique, noncoincident microphones are used, which are often said to create a better feeling of "airiness" and "ambience". The reproduction is also 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. In practice, recording engineers usually try to overcome the outlined problems by using their own ad hoc methods combining several different microphone techniques and the use of reverberators.

3. PSYCHOACOUSTICAL BACKGROUND

As discussed previously, the current technology has shortcomings in the recording and reproduction of spatial sound. However, in order to reproduce the perceived spatial impression of an existing room or a hall, a perfect reconstruction of the physical soundfield is not necessary. Spatial hearing is to a large extent based on two frequency-dependent binaural cues: the interaural time difference (ITD) and the interaural level difference (ILD) [4]. In a room, re-

0-7803-7850-4/03/$17.00 © 2003 IEEE

flections from several different directions affect these cues. Consequently, they no longer suggest single sharply localizable sound sources. For any nonstationary source signal, the reflections also produce time-varying fluctuations in these binaural cues, which further contributes to the subjective spatial impression [5]. We assume that in order to reproduce the spatial sound of a room, it is adequate to create perceptually similar time-varying binaural cues in the ears of the listener.

The limited resolution of human hearing has been studied extensively for monaural conditions [6]. The frequency resolution of binaural hearing appears to be equal to that of monaural hearing [7, 8], although slightly larger analysis bandwidths have been found for some test signals [9]. This suggests that the monaurally derived ERB frequency resolution [10] is also appropriate for the analysis and synthesis of binaural cues. Determining the time resolution of binaural hearing is a little more complicated. A human listener is only capable of tracking in detail the spatial movements of sound sources corresponding to fluctuations of the ITD and ILD cues up to 2.4 and 3.1 Hz, respectively [11]. However, Grantham and Wightman [12] observed that listeners were able to detect ITD fluctuations up to 500 Hz, not based on movement but on perceptual widening of the sound sources. In binaural masking studies the perception of changing ITDs has been found to resemble time averaging with a double sided exponential window. Reported time constants for such a window range between 44–243 ms at different frequencies [13] and between 40–122 ms for different binaural masking signals at a single frequency [9]. Thus, the cutoff frequency of the binaural integrator appears to be well below 10 Hz, and faster changes in the binaural cues effectively correspond to a perception of a sound source occupying a larger area.

A perceptually correct time-varying broadening seems to be difficult to analyze and reproduce with other methods than tracking the changes in the binaural cues. Grantham and Wightman [12] already confirmed this by reporting qualitative differences in the perception of noise with fast ITD modulation and noise with a similar distribution of static ITD differences. In order to fully reproduce the spatial impression of a measured room or hall, our processing algorithm uses a time resolution higher than that of binaural hearing, combined with an analysis of the diffuseness of the sound field. Using a higher time resolution also seems correct in the light of the precedence effect [4], since the directions of individually localizable events such as the direct sound and possible echos will then be sharply reproduced.

4. DIRECTIONAL ANALYSIS

Assuming that the binaural cues need to be reproduced within the resolution of human hearing, methods for analyzing and synthesizing them are needed. Using a multichannel loudspeaker system, an obvious synthesis method for sharply localizable sound events is to simply reproduce the corresponding sound as sharply as possible from the correct direction. Conversely, broadened virtual sound sources related to diffuse time-frequency components can be created by reproducing the sound simultaneously from several different directions close to each other [14].

The binaural cues are, of course, most easily analyzed from the signals measured in the ear canals of a dummy head. However, in order to synthesize them with the methods discussed, explicit knowledge of the cues is actually not needed. The necessary estimates for the time-dependent direction of arrival and diffuseness can be more easily derived using microphone array techniques. In contrast to binaural analysis, such techniques also provide an even directional resolution, which is important for a listener being able to face any direction during the reproduction.

For directional analysis we have chosen to use the concept of sound intensity [15]. Intensity is defined as the product of the particle velocity vector and the sound pressure. The time average of the intensity vector yields the so called active intensity representing the net transport of sound energy. For a plane wave the magnitude of the active intensity $I(t)$ and the sound pressure $p(t)$ have the relation

$$|I(t)| = \frac{p^2(t)}{c\rho_0} \tag{1}$$

where c is the speed of sound and ρ_0 is the mean density of air. On the other hand, in an ideally diffuse sound field the active intensity would be zero irrespective of the sound pressure. Thus, the ratio of the sound pressure and the magnitude of the active intensity vector can be used to characterize the diffuseness of the sound field, while the direction of the vector gives an estimate for the angle of arrival of sound at each instant of time.

An estimate for a component of the sound intensity in a single direction can be derived from the signals of a closely spaced pair of omnidirectional microphones. In a Fourier transform based analysis scheme the frequency distribution of the active intensity in a time frame is given by

$$I_n(\omega) \approx -\frac{j}{\rho_0 \omega d} Im\left\{G_{p1p2}(\omega)\right\} \tag{2}$$

where ω is the angular frequency, j is the imaginary unit, and G_{p1p2} is the single-sided cross-spectral density of the microphone signals

$$G_{p1p2}(\omega) = 2P_1^*(\omega)P_2(\omega) \tag{3}$$

where * denotes the complex conjugate and $P_1(\omega)$ and $P_2(\omega)$ are the Fourier transforms of the microphone signals $p_1(t)$ and $p_2(t)$, respectively [15]. The finite spacing and the nonideal transform characteristics of the microphones, however, pose limits on the usable frequency range, which may need to be taken into account in the subsequent analysis.

The most straightforward method to measure a 3-D intensity vector is to use three orthogonally aligned concentric pairs of microphones. To alleviate some of the frequency range problems we have utilized a special array of 12 small electret microphone capsules [16]. Sound intensity can also be estimated based on measurements with a SoundField microphone. Out of the B-format signals W represents sound pressure, and the directional components X, Y, and Z are related to the particle velocity in the corresponding coordinate directions. Ideally, the active intensity in the x-coordinate direction is given by

$$I_x(\omega) \approx \frac{1}{\sqrt{2}c\rho_0} Re\left\{G_{WX}(\omega)\right\} \tag{4}$$

where G_{WX} is the single-sided cross-spectral density of the W and X signals. Other directions are obtained similarly by substituting Y and Z for X in Eq. (4). In order to get the best possible results, a careful analysis of the phase and magnitude relationships of the B-format signals would be needed. The independent equalization of the magnitude characteristics of W compared to X, Y, and Z in a SoundField preamplifier [17, 18] might be a source for some systematic errors.

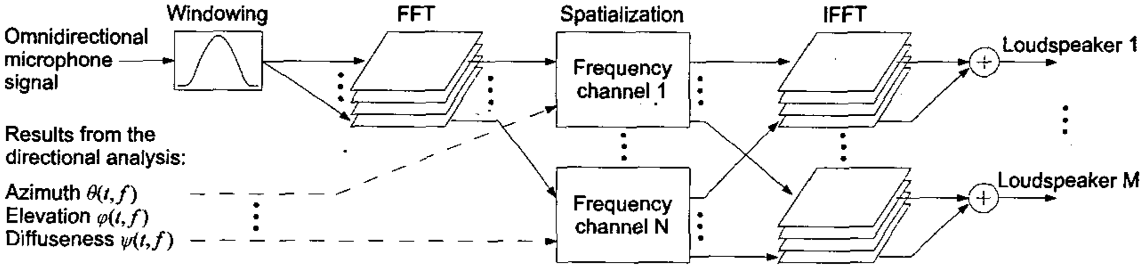

Figure 1: *Spatialization of the omnidirectional signal based on the smoothed directional analysis data.*

5. PROCESSING ALGORITHM

The proposed processing algorithm estimates the angle of arrival and diffuseness of a measured directional room response at each critical frequency band, and uses this data to spatialize an omnidirectional response. For the time-frequency processing we have adopted an FFT based scheme common in audio coding applications. At high frequencies, data from several frequency bands are averaged to form the directional estimates with a frequency resolution resembling that of the human hearing. 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 [19] found the computationally more efficient FFT implementation to perform equally well in their experiments with the Binaural Cue Coding (BCC) algorithm sharing some features with our processing scheme. Whether this holds also for processing reverberation is subject to future work.

In our algorithm, 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. If reliable intensity data for the full frequency band cannot be derived, extrapolate low and/or high frequencies.
4. Smooth the data with an ERB frequency resolution.
5. 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.
6. Based on the diffuseness estimate, spatialize each frequency bin of the unsmoothed sound pressure signal at:
   - the direction of the intensity vector, or
   - several directions around the direction of the intensity vector within a spatial angle defined by the diffuseness estimate.
7. Calculate the IFFTs of the frequency domain loudspeaker signals resulting from the previous step.

The synthesis part of the algorithm is illustrated in Fig. 1. When combined, the processed time frames result in a perceptually reconstructed multichannel room response suitable for loudspeaker reproduction. For directional positioning of the frequency bins, any spatialization method can be used. Pairwise (2-D) or triplet-wise (3-D) amplitude panning are natural candidates, since they result in a sharp reproduction of sound with a minimum number of loudspeakers. A binaural impulse response for headphone reproduction can also be created with HRTF processing.

Several related implementational issues have been discussed in [20]. The frequency domain processing results in time domain artifacts unless proper care is taken. The panning actually corresponds to deriving a filtered version of the omnidirectional signal for each loudspeaker. Time domain aliasing can be avoided by zero padding the analysis windows in the beginning and in the end before calculating the FFT. The number of zeros should be greater than the length of the impulse response of the corresponding filter.

Finding an optimal time-frequency resolution is subject to future research. Based on informal listening tests we are currently using approximately 10 ms Hann windowed time frames. The amount of overlap of subsequent frames controls the amount of spatial smoothing as a function of time and should exceed the 50 % required for perfect reconstruction. Perhaps a signal dependent application of different FFT lengths, as typically used in audio coding (see e.g. [21]), could provide even better results. Time resolution could then be momentarily increased at the expense of frequency resolution for clear discrete reflections.

6. RESULTS AND FURTHER DISCUSSION

The algorithm has not yet been formally evaluated. However, based on the comments of several experienced listeners we feel confident enough to claim that the processed responses yield a very natural spatial impression. In informal listening tests the results were compared to binaural and Ambisonics reproduction using a 5.1 and an 8-channel 3-D loudspeaker system. Spatialization was performed with the Vector Base Amplitude Panning (VBAP) algorithm [22]. The resulting spatial impression appeared to be very similar to that of the reference binaural headphone reproduction but with better externalization. Compared to Ambisonics the resulting directional sound image was considered sharper and the listening area was wider. Furthermore, no noticable coloration was perceived when the listeners moved out of the sweet spot.

A distinct advantage of the proposed method is that in a custom recording system only one studio quality omnidirectional microphone is needed for capturing high quality impulse responses. The requirements for the other microphones are less strict, although a good match between the pairs used in the intensity measurement is desirable. Furthermore, the algorithm provides means for modifying the resulting soundfield. For example, the sound arriving from below the horizontal plane can be remapped to the upper half sphere of a 3-D loudspeaker setup not having speakers below the ear level. Different directions can be weighted or rotated, and the time-frequency envelope of the reverberation can be adjusted.

The analysis and recreation of diffuseness seems to be a very important feature. The implementation of it considerably increased the naturalness of the reproduction and stabilized the sound image of diffuse reverberation. Without such processing some rotating movement could sometimes be perceived during the decay of a response of a large concert hall. It can be hypothesized that the auditory detection of movement is partly suppressed by the less clear localization cues resulting from the diffused frequency components.

The proposed method works very well for impulse responses. The problem with applying it directly to recordings of continuous sound is that the analyzed and synthesized directional spreading within critical frequency bands is currently not natural enough, resulting in a signal dependent spatial impression. However, instead of trying to process reverberant recordings, the methodology could be an attractive alternative for teleconferencing applications where complete transparency is not required. One microphone system would then be enough to track talkers in any direction in a conference room, and the sound could be transfered as a mono signal and side information to another conference room, as is done in the BCC [19, 20].

7. CONCLUSIONS

A perceptually-based processing algorithm for directional room responses has been proposed. The algorithm analyzes the direction of arrival and diffuseness of the soundfield with a critical frequency band resolution 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. The synthesis results in a natural reconstruction of the perceived spatial attributes of a measured room or a hall, and the directional accuracy of the reproduction is not limited by microphone technology but by the loudspeaker configuration.

8. ACKNOWLEDGEMENTS

The study was part of the project TAKU in the Finnish VÄRE technology program funded by TEKES (National Technology Agency). Juha Merimaa is supported by the Graduate School in Electronics, Telecommunications and Automation (GETA), and his current work in Bochum is financed by the research training network for Hearing Organisation And Recognition of Speech in Europe (HOARSE). Ville Pulkki has received funding from the Academy of Finland (101339).

9. REFERENCES

[1] S. P. Lipshitz, "Stereo microphone techniques... Are the purists wrong?" *J. Audio Eng. Soc.*, vol. 34, no. 9, pp. 716–744, 1986.

[2] V. Pulkki, "Microphone techniques and directional quality of sound reproduction," in *AES 112th Convention*, Munich, Germany, 2002, preprint 5500.

[3] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.

[4] J. Blauert, *Spatial Hearing*, revised ed. Cambridge, MA, USA: The MIT Press, 1997.

[5] 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.

[6] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. London, UK: Academic Press, 1997.

[7] A. Kohlrausch, "Auditory filter shape derived from binaural masking experiments," *J. Acoust. Soc. Am.*, vol. 84, no. 2, pp. 573–583, 1988.

[8] M. van der Hejden and C. Trahiotis, "Binaural detection as a function of interaural correlation and bandwidth of masking noise: Implications for estimates of spectral resolution," *J. Acoust. Soc. Am.*, vol. 103, no. 3, pp. 1609–1614, 1998.

[9] I. Holube, M. Kinkel, and B. Kollmeier, "Binaural and monaural auditory filter bandwidths and time constants in probe tone detection experiments," *J. Acoust. Soc. Am.*, vol. 104, no. 4, pp. 2412–2425, 1998.

[10] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," Apple Computer, Tech. Rep. 35, 1993, available at: http://www.slaney.org/malcolm/apple/tr35/PattersonsEar.pdf.

[11] J. Blauert, "On the lag of lateralization caused by interaural time and intensity differences," *Audiology*, vol. 11, pp. 265–270, 1972.

[12] D. W. Grantham and F. L. Wightman, "Detectability of varying interaural temporal differences," *J. Acoust. Soc. Am.*, vol. 63, no. 2, pp. 511–523, 1978.

[13] D. W. Grantham and F. L. Wightman, "Detectability of a pulsed tone in the presence of a masker with time-varying interaural correlation," *J. Acoust. Soc. Am.*, vol. 65, no. 6, pp. 1509–1517, 1979.

[14] V. Pulkki, "Uniform spreading of amplitude panned virtual sources," in *Proc. IEEE WASPAA*, New Paltz, NY, USA, 1999.

[15] F. J. Fahy, *Sound Intensity*. Essex, England: Elsevier Science Publishers Ltd., 1989.

[16] J. Merimaa, T. Lokki, T. Peltonen, and M. Karjalainen, "Measurement, analysis, and visualization of directional room responses," in *AES 111th Convention*, New York, NY, USA, 2001, preprint 5449.

[17] M. A. Gerzon, "The design of precisely coincident microphone arrays for stereo and surround sound," in *AES 50th Convention*, London, UK, 1975, preprint L-20.

[18] D. S. Jagger, "Recent developments and improvements in soundfield microphone technology," in *AES 75th Convention*, Paris, France, 1984, preprint 2064.

[19] F. Baumgarte and C. Faller, "Binaural Cue Coding. Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. Speech Audio Proc.*, 2003, accepted for publication.

[20] C. Faller and F. Baumgarte, "Binaural Cue Coding. Part II: Schemes and applications," *IEEE Trans. Speech Audio Proc.*, 2003, accepted for publication.

[21] K. Brandenburg and G. Stoll, "ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio," *J. Audio Eng. Soc.*, vol. 42, no. 10, pp. 780–792, 1994.

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