SEPARATION OF AUDIO SIGNALS INTO DIRECT AND DIFFUSE SOUNDFIELDS FOR SURROUND SOUND

Benjamin S. Olswang

LOUD Technologies Inc.
Wodinville, WA 98072, USA
ben.olswang@mackie.com

Zoran Cvetković

King's College London
Strand, London, WC2R 2LS, UK
zoran.cvetkovic@kcl.ac.uk

ABSTRACT

This paper presents a new method for reproduction of music in multichannel audio systems. The proposed method separates signals from individual channels into their direct and diffuse components which are then sent to different speaker elements. The direct components are sent directly to the listener, while the diffuse components are additionally scattered. The purpose of this scattering of diffuse components is twofold: first, it eliminates spurious localization cues which may be created by reproducing the sound using a small number of speakers (three to five), and second, it provides additional diffusion which improves the envelopment experience. We investigate two methods to separate direct and diffuse soundfield components, both of which assume the knowledge of the impulse response of the performance auditorium to the corresponding microphones. One method is based on techniques for multichannel equalization, while the other uses techniques for signal reconstruction after oversampled filter-bank processing. The latter method turns out to be computationally more manageable. In listening tests, subjects preferred music reproduction which separates direct and diffuse soundfields to reproduction in which both soundfields are sent to the same speaker elements.

1. INTRODUCTION

Following the advent of multichannel audio, a novel five-channel audio technology has been recently proposed that attempts to reproduce some or most of the auditory experience of an acoustic performance in its original venue [1]. The proposed audio scheme uses a specially constructed seven-channel microphone array to capture cues needed for reproduction of the original perceptual soundfield in a five-channel stereo system. The microphone array consists of five microphones in the horizontal plane, as shown in Figure 1, placed at the vertices of a pentagon, and two additional microphones laying in the vertical line in the center of the pentagon, one pointing up the other down. Each microphone receives the source sound filtered by the corresponding impulse response of the performance venue between the source and the microphone. The impulse response consists of two parts: direct, which contains the impulse which travels to the microphone directly plus several early reflections, and reverberant, which contains impulses which are reflected multiple. The soundfield component which is obtained by convolving the source sound with the direct part of the impulse response creates the so-called direct soundfield, that carries perceptual cues relevant for source localization, while the component which is the result of the convolution of the source sound with the reverberant part of the impulse response creates the diffuse soundfield, which provides the envelopment experience.

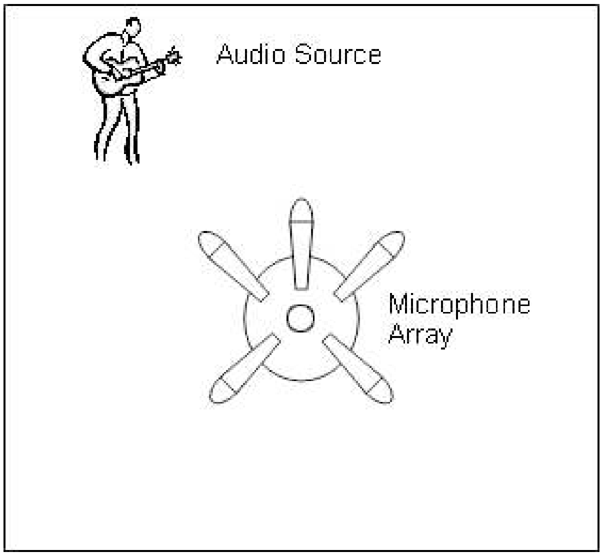


Figure 1: *Microphone array used for recording of an acoustic performance for an audio technology which aims to recreate the perceptual soundfield of the performance in its original venue.*

The seven audio signals captured by the microphone array are mixed down to five reproduction channels, front-left (FL), front-center (FC), front-right (FR), rear-left (RL), and rear-right (RR). Listening tests demonstrated significant increase of the "sweet spot" area of the new scheme compared to the standard two-channel audio in terms of sound-source localization [1]. The reproduction using only five speakers, however, cannot provide a satisfactory envelopment experience since five reproduction channels are not sufficient to produce adequate diffusion of the soundfield [2]. Besides, recreation of the diffuse soundfield using the same speaker elements which are used for recreation of the direct soundfield may produces spurious cues which affect the capability of a listener to localize the sound source. In this paper, we explore the idea of separating signals received by the microphones into their direct and diffuse components and reproducing them using different

This work was done while B. Olswang was at King's College London

1-4244-0469-X/06/$20.00 ©2006 IEEE

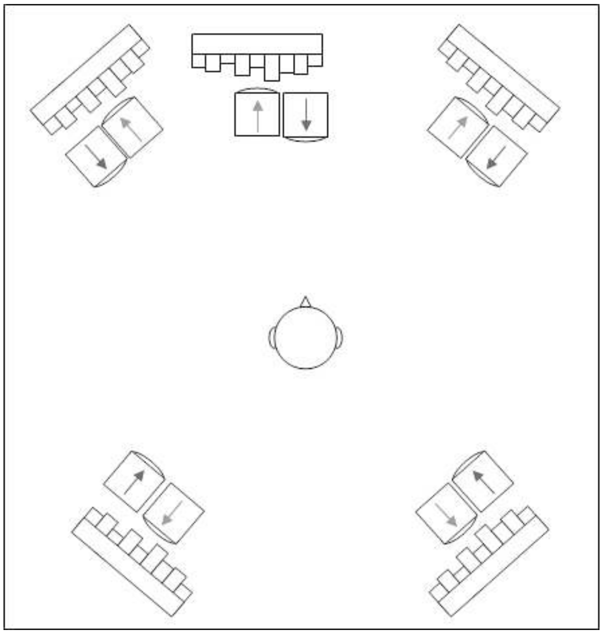


Figure 2: *The speaker set-up used in this work for improved reproduction of diffuse soundfield.*

speaker elements [2]. In particular, the direct soundfield components will be reproduced using speakers pointing toward a listener, while the diffuse soundfield components will be reproduced using speakers pointing away from the listener and toward diffuser panels which perform additional sound scattering. Such a speaker set-up is shown in Figure 2.

In Section 2, we first study techniques for separation of signals recorded by microphones into their direct and diffuse components. We restrict considerations to the case of a single sound source (solo instrument) and known impulse responses between the sound source and each of the microphones. The latter restriction is minor since the impulse responses can easily be measured. The case when there are multiple sources is currently being investigated. The results of tests which evaluate the listening experience provided by the new reproduction technology are presented in Section 3. These results demonstrate that the reproduction which separates direct and diffuse soundfields is preferable to the reproduction which does not perform this separation.

2. DIRECT/DIFFUSE SOUNDFIELD SEPARATION

Recording a musical performance using an N-channel microphone array, under the assumption of a point source, produces N signals

$$Y_i(z) = H_i(z)X(z),\ i = 1\ldots, N \tag{1}$$

where $X(z)$ is the source signal and $H_i(z)$ is the impulse response of the auditorium between the source and the i-th microphone. Each impulse response $H_i(z)$ can be represented as

$$H_i(z) = H_{i,d}(z) + H_{i,r}(z)\ , \tag{2}$$

where $H_{i,d}(z)$ and $H_{i,r}(z)$ are its direct and reverberant component, respectively. The goal is to find a method to recover direct and diffuse components $Y_{i,d}(z) = H_{i,d}(z)X(z)$ and $Y_{i,r}(z) = H_{i,r}(z)X(z)$, respectively, of all microphone signals $Y_i(z)$, given these signals and impulse responses $H_i(z)$. To this end, we shall first recover $X(z)$ from signals $Y_i(z)$ and then apply filters $H_{i,d}$ and $H_{i,r}(z)$ to obtain $Y_{i,d}(z)$ and $Y_{i,r}(z)$, respectively. Components $H_{i,d}(z)$ and $H_{i,r}(z)$ can be obtained from $H_i(z)$ in several ways, including taking $H_{i,d}(z)$ to be a given number of the first impulses of $H_i(z)$, the initial part of $H_i(z)$ in a given time interval, or extracting $H_{i,d}(z)$ from $H_i(z)$ manually. Once, $H_{i,d}(z)$ is obtained, $H_{i,r}(z)$ is the remaining component of $H_i(z)$, $H_{i,r}(z) = H_i(z) - H_{i,d}(z)$. Therefore, from now on, we focus on the problem of recovering $X(z)$ from $Y_i(z),\ i = 1, \ldots, Y_N(z)$, given $H_i(z),\ i = 1, \ldots, N$.

The problem at hand was studied in-depth in the filter bank literature. Below we review relevant results, details of which can be found in [3]. $X(z)$ can be reconstructed from $Y_i(z)$'s in a numerically stable manner if and only if impulse responses $H_i(z)$ do not have zeros in common on the unit circle. If this condition is satisfied then there exist stable filters $G_i(z),\ i = 1, \ldots, N$ such that

$$\sum_{i=1}^{N} G_i(z)H_i(z) = 1\ . \tag{3}$$

Hence, $X(z)$ can be reconstructed as

$$X(z) = \sum_{i=1}^{N} G_i(z)Y_i(z)\ . \tag{4}$$

Note that filters $G_i(z)$ are not unique, and one particular solution to (3) is given by

$$G_i(z) = \frac{H_i(z^{-1})}{\sum_{i=1}^{N} H_i(z)H_i(z^{-1})}\ . \tag{5}$$

This solution has an advantage over all other solutions in the sense that it performs maximal reduction of white additive noise which may be present in signals $Y_i(z)$. Another issue of particular interest is to be able to reconstruct $X(z)$ using FIR filters. A set of FIR filters $F_i(z)$ such that any $X(z)$ can be reconstructed from corresponding signals $Y_i(z)$ exists if and only if impulse responses $H_i(z)$ have no zeros in common. If this is satisfied, a set of FIR filters $F_i(z)$ which can be used for reconstructing $X(z)$ can be found by solving the system

$$\sum_{i=1}^{N} F_i(z)H_i(z) = 1\ . \tag{6}$$

The problem of solving (6) for a set of FIR filters was previously studied by the communications community as a multichannel equalization problem [4]. Note that both the condition for perfect reconstruction of $X(z)$ using stable filters and the condition for perfect reconstruction using FIR filters are normally satisfied since it is very unlikely that impulse responses $H_i(z)$ will have a common zero.

In this work we consider reconstructing $X(z)$ using filters FIR $F_i(z)$ obtained by solving (6), and we refer to this approach as Method 1. We also consider using FIR approximations of filters $G_i(z)$ in (5), and we refer to this approach as Method 2. In the following subsections we discuss details of these two methods, their advantages and problems.

2.1. Method 1

Finding a set of FIR filters $F_i(z)$ which satisfy (6) amounts to solving a system of linear equations for the coefficients of the unknown filters. While solving a system of linear equations may seem trivial, in the particular case which we consider here a real challenge arises from the fact that the systems in question are usually huge, since impulse responses of music auditoria are normally thousands of samples long. To illustrate an expected dimension of the linear system, consider impulse responses $H_i(z)$ and let L_h be the length of the longest one among them. Assume that we want to find filters $F_i(z)$ of length L_f. Then, the dimension of the linear system of equations which is equivalent to (6) is $L_h + L_f - 1$. The system has an exact solution if the total number of variables, which is in this case NL_f (the number of filters $F_i(z)$ times the filter length), is larger or equal to the number of equations, that is, if $NL_f \geq L_h + L_f - 1$. This implies that L_f must be greater than $L_h/(N-1)$. Hence, the dimension of the system is greater than $NL_h/(N-1)$. In the case of 44.1kHz sampling rate (CD quality), and assuming 5-channel microphone array (just the microphones in the horizontal plane), for a room which has a one second reverberation time, L_h becomes $L_h = 44100$ and the corresponding linear system has around 55000 equations. In the experiments we were conducting, MATLAB was unable to solve systems of 17000 equations.

Another problem associated with this approach is that effect of filters $F_i(z)$ obtained in this manner on possible additive noise is unclear. To ensure good noise reduction properties one needs to allow for filters longer than the minimal length required to solve the system exactly and then perform constrained optimization of an intricate function of a huge number of variables.

2.2. Method 2

Equation (5) provides a closed form solution for filters $G_i(z)$ which can be used for perfect reconstruction of $X(z)$ according to (4). Observe that filters $G_i(z)$ given by this formula are IIR filters. One way to use these filters would be to implement them directly as IIR filters, but that would require an unacceptably high number of coefficients. Another way would be to find FIR approximations. We attempted computing a given number first coefficients of filters $G_i(z)$ by filtering corresponding impulse responses $H_i(z^{-1})$ with IIR filter $1/D(z)$, $D(z) = \sum_{i=1}^{N} H_i(z)H_i(z^{-1})$ (see Equation (5)). Our attempts to execute these computations in MATLAB, however, failed. The FIR approximations to $G_i(z)$'s which are used in the experiments reported in the next section were obtained by dividing the DFT of corresponding functions $H_i(z^{-1})$ by the DFT of $D(z)$ and finding the inverse DFT of the result. The size of the DFT used for this purpose was four times larger than the length of $D(z)$. Note that it is essential that the DFT size is large since Method 2 computes coefficients of IIR filters $G_i(z)$ by finding their inverse Fourier transform using finitely many transform samples. This discretization of the Fourier transform causes time-aliasing of impulse responses of filters $G_i(z)$ and the aliasing is reduced as the size of the DFT is increased. Despite the need for the DFT of large size, Method 2 turned out to be numerically much more efficient than Method 1 and could operate on larger impulse responses. Reconstruction of $X(z)$ using this approximation also gave very accurate results, as we report in the next section.

3. EXPERIMENTAL RESULTS

In experiments reported in this section a 5-channel reduction of the 7-channel microphone array was considered which consisted of the microphones in the horizontal plane. Three test signals, $20-$ 30s long, were used: a male rock singer, jazz trumpet, and jazz guitar. All three signals were originally recorded with a close microphone technique to minimize early reflections and reverberation, and were high quality audio files. Impulse responses $H_i(z)$ were generated for hypothetical rectangular auditoria using the method of images described in [5]. The test signals were used with impulse responses of two different lengths and in that way 4 test scenarios were created, which we shall refer to as Source 1-4, with the following meaning: Source 1 - rock male singer, impulse responses 5000-samples long, 22.05kHz sampling, Source 2 - electric jazz guitar, impulse responses 5000-samples long, 22.05kHz sampling, Source 3 - jazz trumpet, impulse responses 5000-samples long, 22.05kHz sampling, Source 4 - jazz trumpet, impulse responses 17000-samples long, 44.1kHz sampling.

Table I shows the signal-to-error ratio (SER) of the reconstruction of $X(z)$ in the four test scenarios using Methods 1 and 2. The SER is measured as $SER = 10\log_{10}(\sum_n |x[n]|^2/\sum_n |x[n] - x_r[n]|^2)$, where $x[n]$ is the original signal and $x_r[n]$ is its reconstructed version. MATLAB implementation of Method 1 failed in the case of Source 4 due to the large length of impulse responses $H_i(z)$. In all cases the reconstructed signal $x_r[n]$ was audibly indistinguishable from the original despite the fact that in same cases the SER of Method 1 was low. However, the fact that Method 2 shows consistently higher SER than Method 1 should not be ignored. This lower SER of Method 1 is possibly partly due to the fact that in some cases we had to add some small amount of noise to the coefficients of the linear equations to make the system solvable in MATLAB.

Table 1: *The error of signal reconstruction using Methods 1 and 2*

	Source 1	Source 2	Source 3	Source 4
Method 1	38.5dB	117.8dB	3.5dB	
Method 2	152.6dB	208.1dB	183.6dB	553.0dB

Listening experiments were performed with 5 subjects who were all MSc student at King's College London around 25 years old. The subjects were asked to evaluate 4 different surround sound systems in terms of realism, perceived room size, and overall preference. The first system was five-channel audio where both the direct and diffuse soundfield were reproduced using the same speaker elements pointing toward the listener. The other three systems were using the 10-speaker set-up shown in Figure 2 in which direct soundfield components were sent to speakers directed toward the listener, while the diffuse components were sent to speakers directed away from the listener and toward diffuser panels. The three 10-speaker systems differed in the way in which direct and diffuse soundfield components were separated. In the system referred to as *Ideal* the direct and diffuse soundfield components were obtained by convolving $X(z)$ directly with corresponding $H_{i,d}(z)$ and $H_{i,r}(z)$ filters, respectively. In the systems denoted by *Method 1* and *Method 2*, $X(z)$ was first reconstructed from $Y_i(z)$, $i = 1, \ldots, N$ using Method 1 and Method 2, respectively, and the reconstructed versions of $X(z)$ are then convolved with filters $H_{i,d}(z)$ and $H_{i,r}(z)$ to obtain direct and diffuse components in individual channels. Signals used in the 5-speaker scheme were played at an increased level so to match the loudness level of the 10-speaker schemes. In the evaluations, the subjects were asked to rank the four systems and another system which is not described here in terms of realism, perceived room, size and overall preference. In this manned each of the systems was given a score on a scale from 1 to 5, where 5 was the best. The results of these listening tests are represented by the graphs in Figure 3. Note that the results for the 10-speaker scheme with Method 1 do not appear in these graphs for Source 4 since MATLAB was unable to solve the corresponding system of equations. The perceived differences between the three 10-channel schemes were very subtle, which along with the fact that the number of subjects was small explains the variations in the average scores of the three schemes. However, it is evident that, in all of the considered aspects, all three schemes which separated direct and diffuse soundfields scored significantly better than the scheme which did not perform this soundfield separation. We also perceived that separating direct and diffuse soundfields improved the localization of sound source, however, formal tests in this respect were not conducted.

4. CONCLUSION

This paper presented two results. First, two techniques for separating direct and diffuse soundfields in multichannel audio recordings were studied. One of the techniques has been studied previously by the communications community within the multichannel equalization framework, and was found here not to be suitable for the audio problem at hand due to excessively long channels (room impulse responses). The other technique, based on results of the theory of oversampled filter banks, is quite novel and has shown good results in this context. The second result is a new technology for multichannel surround audio systems. The proposed technology aims to improve the envelopment experience by separating direct and diffuse soundfields and introducing additional scattering of the diffuse soundfield using diffuser panels. The results of the listening experiments reported in this paper show that the new technology improves the listening experience.

Acknowledgement

We would like to thank J. D. Johnston for his guidance in this research, J. Hall for his help in setting up the experiments, S. Lambotharan for pointing out multichannel equalization techniques, and to the Music Department of King's College London for allowing us to use one of their practice rooms for our experiments.

5. REFERENCES

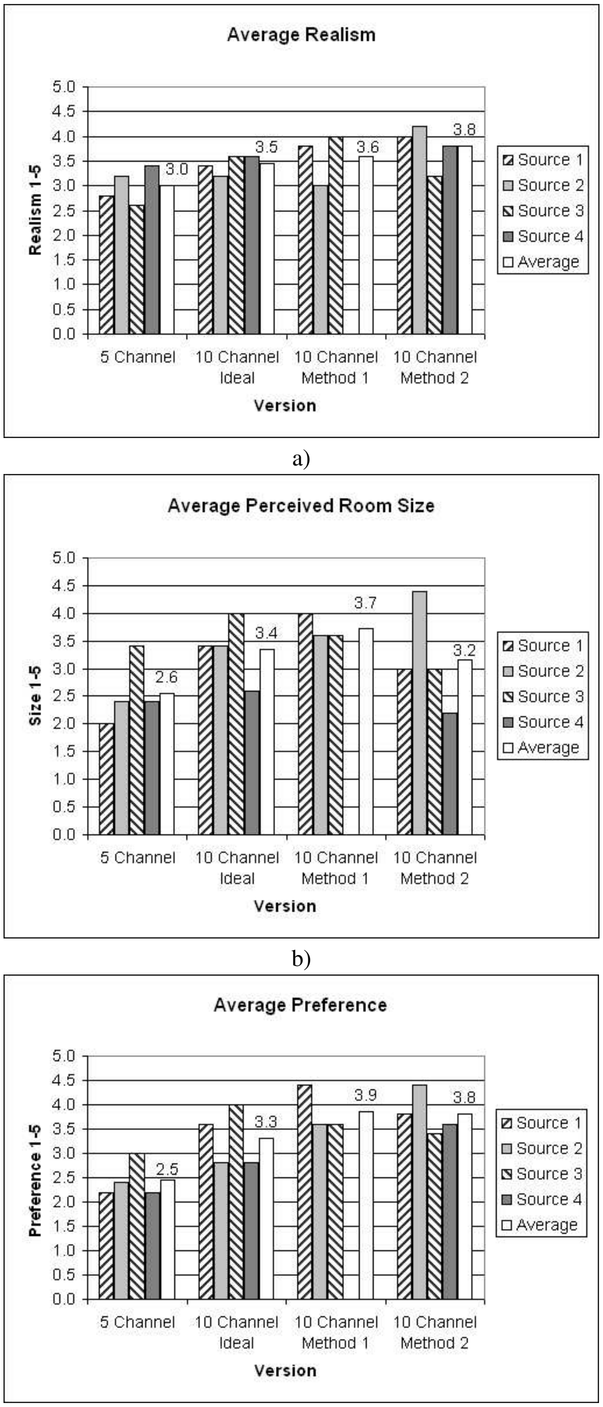
[1] J.D. Johnston and Y.H. Lam, " Perceptual soundfield reconstruction," *109th AES Convention*, paper No. 5202, September 2000.

[2] J.D. Johnston, personal communication, July, 2002.

[3] Z. Cvetković and M. Vetterli, "Oversampled filter banks," *IEEE Trans. Signal Processing*, Vol. 46, No. 5, pp. 1245–1257, May 1998.

[4] J.R. Treichler, I. Fijalkow, and C.R. Johnson, "Fractionally Spaced Equalizers," *IEEE Signal Processing Magazine*, Vol. 13, pp. 65–81, May 1996.

[5] J.B. Allen and D.A. Berkeley, "Image method for efficiently simulating small-room acoustics," *JASA*, Vol. 65, No. 4, pp. 943–950, April 1979.



a)

b)

c)

Figure 3: *The results of listening tests. a) The average perception of realism. b) The average perception of room size. c) Overall average preference.*