Modifying the Directional Responses of a Coincident Pair of Microphones by Postprocessing*

CHRISTOF FALLER, *AES Member*
(christof.faller@epfl.ch)

Illusonic LLC, CH-1006 Lausanne, Switzerland

The proposed technique processes coincident or nearly coincident microphone stereo recordings to change the angle between and the width of the directional responses of the microphones effectively. Through this modification of the responses the width and spaciousness of the stereo recording can be controlled without introducing any artificial reverberation into the recording. The algorithm is based on prediction with the aim of controlling the amount of sound corresponding to the overlapping part of the microphone responses. It is shown how this processing relates to the effective directional responses of the processed signals. The results of a headphone-based subjective test indicate that the achieved audio waveform quality is very high and B-format derived coincident cardioid stereo recordings are improved significantly.

0 INTRODUCTION

Many techniques have been proposed for stereo recording. In the context of this paper, when we discuss stereo recording we always mean recording with two microphones to get a stereo signal representing the "natural mix" of sound as it has arrived at the microphones. Various stereo recording techniques are discussed in [1], [2] from a practical point of view.

One commonly used class of techniques, originally proposed by Blumlein [3], uses a pair of coincident microphones. Blumlein proposed the use of two dipole (figure-of-eight) microphones pointing toward ±45° relative to the forward axis. Often different directional microphones and different angles are used. Commonly used are cardioid, hypercardioid, and supercardioid microphones with angles of between ±45° and ±60°.

Another class of techniques uses spaced omnidirectional microphones. Various configurations are used, such as a relatively closely spaced or a widely spaced pair of microphones. While coincident microphone recording often results in precisely localized virtual sources in the stereo signal, its weakness is the amount of ambience or spaciousness. On the other hand, spaced microphone techniques usually result in more ambience and spaciousness, but often suffer from a lack of localization precision and a "hole in the middle."

Microphone setup for stereo recording is often about finding a compromise between localization precision, avoiding a hole in the middle, and controlling ambience and spaciousness. One such compromise is the AB microphone configuration, which uses two noncoincident directional microphones. A certain degree of localization precision is obtained due to the fact that the microphones are relatively closely spaced and due to the level differences resulting from the directive nature of the microphones.

While in practice stereo recording is an art requiring skill and extensive experience, a number of works have attempted to explain the result of the various stereo recording techniques or propose theoretically "optimal" ways of stereo recording.

A detailed discussion about the various stereo recording techniques, with a clear argument for coincident stereo recording, has been given by Lipshitz in the 1980s [4]. He argues that the localization cues [5] at the ears of a listener in the sweet spot are only natural for intensity stereo. Time delays, occurring in spaced microphone setups, result in highly signal- and frequency-variant localization cues. Lipshitz also opines that the spaciousness resulting from spaced microphones is not natural as it occurs in the concert venue, but artificial, caused by the phase inconsistencies of the recording.

*Manuscript received 2007 February 28; revised 2007 August 29 and 2008 March 7.

A thorough analysis of different stereo microphone recording setups has been presented by Williams in the 1980s [6]. Among other things he relates psychoacoustic data, microphone response, and intermicrophone distance to the angular range that is mapped to ±30° in the stereo listening setup, the angular range in which the direct-to-reverberant ratio is correct, and the maximum angular error of virtual sources. This analysis was extended for the variable M/S stereophonic microphone system in [7].

Another more physical attempt to look at spatial sound recording and reproduction systematically is Ambisonics [8], [9]. One of the main ideas here is to measure the sound field in one point (sound pressure, particle velocity) and then, by loudspeaker playback, to reproduce that sound field in the sweet spot. Coincident microphones [10] are used to measure the sound field in one point, such as the Soundfield Microphone [11] capturing B-format signals. The B-format signals are related to the sound pressure and the particle velocity vector.

The perceived azimuth of virtual sources not located at the left loudspeaker, center position, or right loudspeaker is frequency dependent when time delay or amplitude panning is used [12]. This imposes limitations on the stereo technique, no matter what microphone setup is used.

The signal processing technique proposed in this paper is applicable to coincident stereo microphones and simulates a stereo microphone with directional responses different from those of the original stereo microphone. In this way a stereo microphone signal can be modified to have a wider sound stage, and the amount of spaciousness can be controlled.

Since the proposed technique is postprocessing of fixed microphone signals, it is flexible in terms of optimizing the left and right directional responses during or after recording. While the variable M/S stereophonic microphone system and the Soundfield Microphone are flexible in this sense, the result is always a first-order linear response. Similar to the M/S stereophonic microphone system, and less than the Soundfield Microphone, only two microphone capsules are needed for the proposed technique.

The conceptual idea behind the proposed technique is the following. The microphone signals are processed such that the amount of sound that is picked up within the overlapping part of the left and right microphone responses is reduced. This processing has an effect on the responses of the left and right microphones. It is shown that the responses can be made more directive and less overlapping. An example of the result of the proposed processing is shown in Fig. 1. Fig. 1(a) shows the responses of a given pair of coincident cardioid microphone signals, and Fig. 1(b) shows the responses (for direct sound) after processing has been applied to the signals.

The paper is organized as follows. Section 1 describes the proposed processing. An analysis of the effect of the proposed processing for direct and diffuse sounds is presented in Sections 2 and 3, respectively. Section 4 describes parameter estimation for the general case of a mix of direct and diffuse sound. Section 5 describes the subjective test that was carried out for evaluating the proposed technique. Various aspects of the proposed technique are discussed in Section 6. Conclusions are given in Section 7.

1 PROPOSED MICROPHONE SIGNAL PROCESSING

The proposed scheme adapts to signal statistics as a function of time and frequency. Thus the signals are processed in a time–frequency representation, as is illustrated in Fig. 2. A suitable choice for such a time–frequency representation is the use of critical bands as, for example, is described in [13], [14]. The signals are assumed to be stationary in each time–frequency tile. Given a signal $x(n)$, its time–frequency representation is denoted by $X(k, i)$, where k is the (usually downsampled) time index and i is the frequency (or subband) index.

We are assuming that the microphone signal $X_2(k, i)$ can be written as

$$X_2(k, i) = a(k, i)X_1(k, i) + N_2(k, i) \tag{1}$$

where $a(k, i)$ is a time- and frequency-dependent gain factor related to the crosstalk between both microphone signals $X_1(k, i)$ and $X_2(k, i)$. It is assumed that all signals are zero mean and that $X_1(k, i)$ and $N_2(k, i)$ are independent.

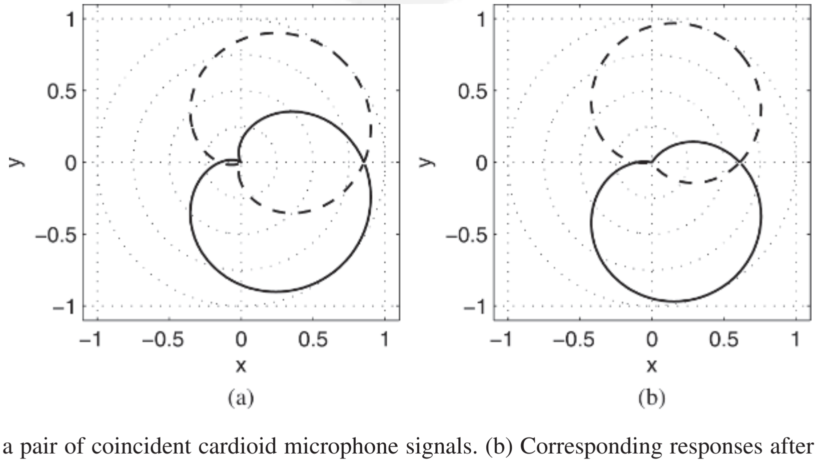

Fig. 1. (a) Responses of a pair of coincident cardioid microphone signals. (b) Corresponding responses after proposed processing has been applied.

The goal of the proposed algorithm is to modify the directional response of the microphone signal $X_2(k, i)$ [and similarly for $X_1(k, i)$] by eliminating or partially eliminating the signal components in $X_2(k, i)$, which are correlated with $X_1(k, i)$,

$$Y_2(k, i) = c(k, i)[X_2(k, i) - \tilde{a}(k, i)X_1(k, i)]. \tag{2}$$

Note that if the weights are chosen to be $c(k, i) = 1$ and $\tilde{a}(k, i) = a(k, i)$, then $N_2(k, i)$ is recovered. If the weights are chosen $\tilde{a}(k, i) < a(k, i)$, then some signal components correlated with $X_1(k, i)$ remain in $Y_2(k, i)$. As will be shown later, $\tilde{a}(k, i)$ is computed as a function of $a(k, i)$ and the desired properties of the directional response. The postscaling factor $c(k, i)$ is used to scale the signal such that the maximum response is 0 dB. For simplicity of notation, in the following we are often ignoring the time and frequency indexes k and i.

To compute a the following equation is used:

$$E\{X_1X_2\} = aE\{X_1^2\} \tag{3}$$

where $E\{.\}$ is a short time averaging operation for estimating a mean in a time–frequency tile. Eq. (3) solved for a yields

$$a = \frac{E\{X_1X_2\}}{E\{X_1^2\}}\,. \tag{4}$$

This can be written as

$$a = \sqrt{\frac{E\{X_2^2\}}{E\{X_1^2\}}}\,\Phi_{12} \tag{5}$$

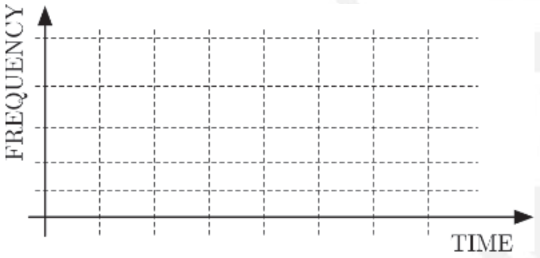

Fig. 2. Signals are analyzed and processed in a time–frequency representation.

where Φ_{12} is the normalized cross-correlation coefficient between X_1 and X_2,

$$\Phi_{12} = \frac{E\{X_1X_2\}}{\sqrt{E\{X_1^2\}E\{X_2^2\}}}\,. \tag{6}$$

If at a specific time and frequency, sound is arriving from only one direction, the two signals X_1 and X_2 are coherent. Thus N_2 [Eq. (1)] will be zero. To prevent that Y_2 [Eq. (2)] is zero, $\tilde{a}$ is computed by limiting a,

$$\tilde{a} = \min\{a, q\} \tag{7}$$

where q is the value at which a is limited. The directional response corresponding to the so computed Y_2 signal can be controlled with parameter q, as shown in the following sections.

Fig. 3 summarizes the processing carried out by the proposed scheme. The two given directional microphone signals $x_1(n)$ and $x_2(n)$ are converted to their corresponding time–frequency representations by a filterbank (FB) or time–frequency transform. Further processing is shown for one subband signal. The parameters $\tilde{a}$ and c are estimated, and the subband signal of the output signal $Y_2(k, i)$ is computed. The subbands of the output signal are converted back to the time domain using an inverse filterbank (IFB) or time–frequency transform, resulting in the time-domain output signal $y_2(n)$.

In the next two sections it is shown how q in Eq. (7) relates to the resulting directional response for direct and diffuse sound, respectively. Further, the postscaling factor c in Eq. (2) is derived for direct and diffuse sound. Then we explain in Section 4 the use of the proposed scheme for general scenarios, where direct and diffuse sound is mixed.

2 RESPONSE FOR DIRECT SOUND

For sound arriving from only one direction, the signals measured by two coincident cardioid microphones, pointing toward the directions of ϕ_0 and $-\phi_0$, can be written as

$$X_1 = \frac{1}{2}[1 + \cos(\phi - \phi_0)]S$$
$$X_2 = \frac{1}{2}[1 + \cos(\phi + \phi_0)]S \tag{8}$$

where S is the short-time spectrum of the sound and ϕ is the direction from which the sound is arriving. Fig. 1(a)

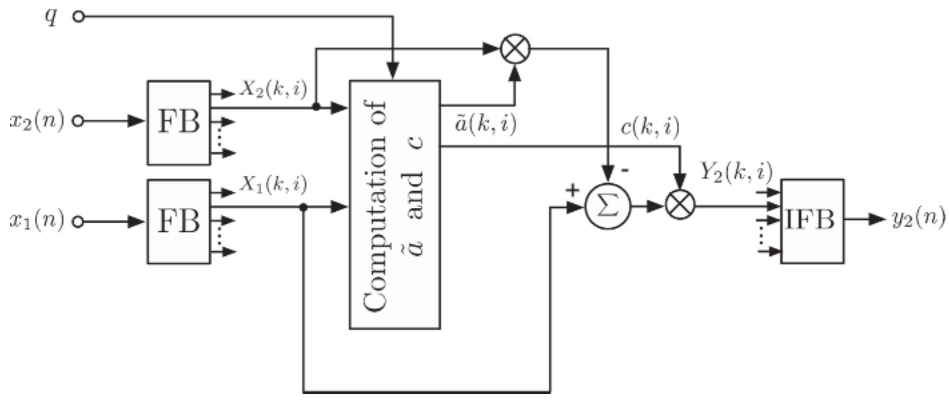

Fig. 3. Schematic diagram of processing of proposed scheme (one subband signal).

shows the directional responses of X_1 and X_2 for $\phi_0 = 45°$. Without loss of generality, the proposed scheme is derived for cardioid microphones. Note that the proposed scheme can be applied with microphones having other directional responses.

The estimated signal Y_2 [Eq. (2)] is equal to

$$Y_2 = \frac{c}{2}[1 - \tilde{a} + \cos(\phi + \phi_0) - \tilde{a}\cos(\phi - \phi_0)]S. \tag{9}$$

The corresponding directional response is

$$d(\phi) = \frac{c}{2}[1 - \tilde{a} + \cos(\phi + \phi_0) - \tilde{a}\cos(\phi - \phi_0)]. \tag{10}$$

Y$_2$ [Eq. (2)] is zero, except when the gain factors [Eq. (4)] are limited [Eq. (7)], that is, $a > q$. Thus the effective directional response is obtained by substituting $\tilde{a} = q$ in Eq. (10) and lower bounding it by zero,

$$d_{Y_2}(\phi) = \max\left\{\frac{c}{2}[1 - q + \cos(\phi + \phi_0) - q\cos(\phi - \phi_0)], 0\right\}. \tag{11}$$

This is equivalent to

$$d_{Y_2}(\phi) = \max\left\{\frac{c}{2}[1 - q + R\cos(\phi - \phi_2)], 0\right\} \tag{12}$$

where

$$R = \sqrt{(1+q)^2 \sin^2\phi_0 + (1-q)^2\cos^2\phi_0}$$
$$\phi_2 = \tan^{-1}\frac{-(1+q)\sin\phi_0}{(1-q)\cos\phi_0}. \tag{13}$$

The postscaling factor c is chosen such that the maximum gain of the resulting response is equal to 1, that is, $d_{Y_2}(\phi_2) = 1$. From Eqs. (12) and (13) it follows that this is the case for $c = c_1$, with

$$c_1 = \frac{2}{1 - q + \sqrt{(1+q)^2\sin^2\phi_0 + (1-q)^2\cos^2\phi_0}}. \tag{14}$$

The −3-dB width of the directional response as a function of q is

$$\alpha = 2\cos^{-1}\frac{R + 1 - \sqrt{2} + (\sqrt{2} - 1)q}{\sqrt{2}R}. \tag{15}$$

Later used in this paper, the range in radians where the response of Y_2 [Eq. (12)] is nonzero is

$$\beta = \pi + 2\sin^{-1}\frac{1-q}{R}. \tag{16}$$

In the following examples the angle between the original microphone responses and the forward axis is always $\phi_0 = 45°$. Fig. 4 illustrates an example for $q = 0.4$. The responses of X_1 and X_2 are shown as dotted lines. The resulting response without postscaling ($c = 1$) is indicated by the solid thin line. Note that the maximum of the response, $d_{Y_2}(\phi_2)$, is smaller than 1 in this case. The response after postscaling with $c = c_1 = 1.19$ [Eq. (14)] is shown as a bold solid line in Fig. 4. The response after postscaling, in polar coordinates, is also illustrated in Fig. 1(b) (solid, bold). The direction ϕ_2 [Eq. (13)] of the response is indicated with a dashed vertical line in Fig. 4. Further, the −3-dB width of the response α is indicated with another two dashed vertical lines. The signal Y_1 with a response like the one pointing upward in Fig. 1(b) (dashed, bold) is obtained by similar processing (by exchanging X_2 and X_1 in the equations).

Fig. 5 shows the width α [Eq. (15)], nonzero range β [Eq. (16)], and direction ϕ_2 [Eq. (13)] as a function of the gain factor limit q. These data indicate that the proposed processing allows to make the response for direct sound more narrow and pointing farther to the side than the original cardioid response of X_2.

The original cardioid microphone responses pointing toward ±45° and responses of Y_2 for different gain factor limits q are shown in Fig. 6. By exchanging X_1 and X_2 and applying the same processing, responses pointing toward $-\phi_2$ can be obtained.

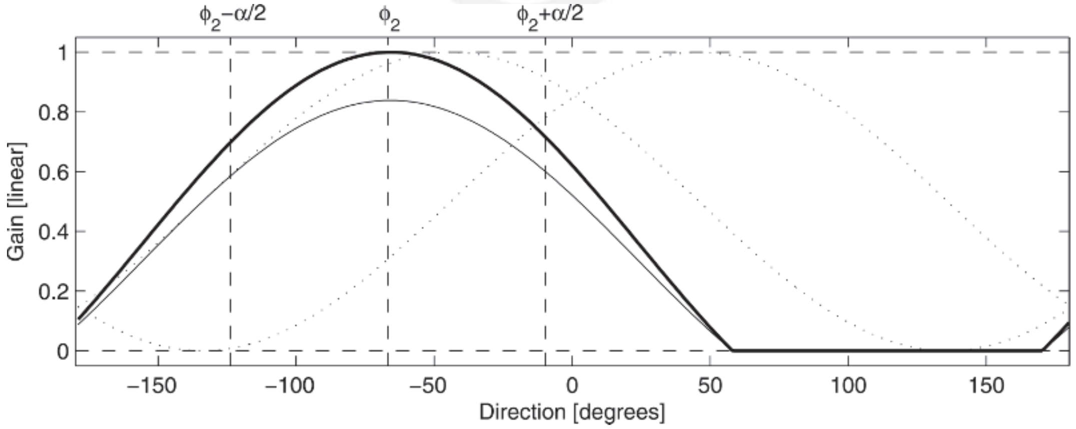

Fig. 4. Directional responses of X_1 and X_2 for $\phi_0 = 45°$ (····). Directional response of Y_2 before postscaling (——) and after postscaling (▬▬). Direction ϕ_2 and width α of response are indicated as vertical lines (– – –).

The postscaling factor c_1 [Eq. (14)] is shown in Fig. 7 as a function of the gain limit q. Note that c_1 increases as q increases. But c_1 is reasonably small, and thus no signal-to-noise ratio issues due to postscaling are expected.

3 RESPONSE FOR DIFFUSE SOUND

As opposed to the case of sound arriving only from one direction, for diffuse sound arriving from all directions N_2 [Eq. (1)] is not zero. For the analysis of this case we are first computing N_2,

$$N_2(k, i) = X_2(k, i) - a(k, i)X_1(k, i) \tag{17}$$

and then with the insights gained a postscaling factor c, most appropriate for diffuse sound, is determined.

3.1 Computation of N_2 for Diffuse Sound

It is assumed that diffuse sound can be modeled with plane waves arriving from different directions. Thus diffuse sound measured by two coincident cardioid microphones, pointing toward ϕ_0 and $-\phi_0$, can be written as

$$\begin{aligned} X_1(k, i) &= \frac{1}{2}\int_{-\pi}^{\pi} [1 + \cos(\phi - \phi_0)]D(k, i, \phi)\, \mathrm{d}\phi \\ X_2(k, i) &= \frac{1}{2}\int_{-\pi}^{\pi} [1 + \cos(\phi + \phi_0)]D(k, i, \phi)\, \mathrm{d}\phi \end{aligned} \tag{18}$$

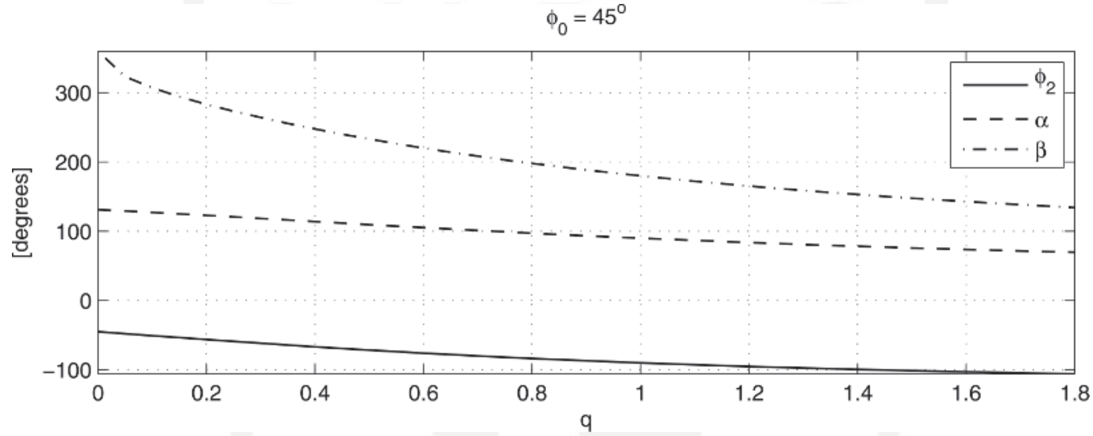

Fig. 5. Width α, nonzero range β, and direction ϕ_2 of response as a function of gain factor limit q.

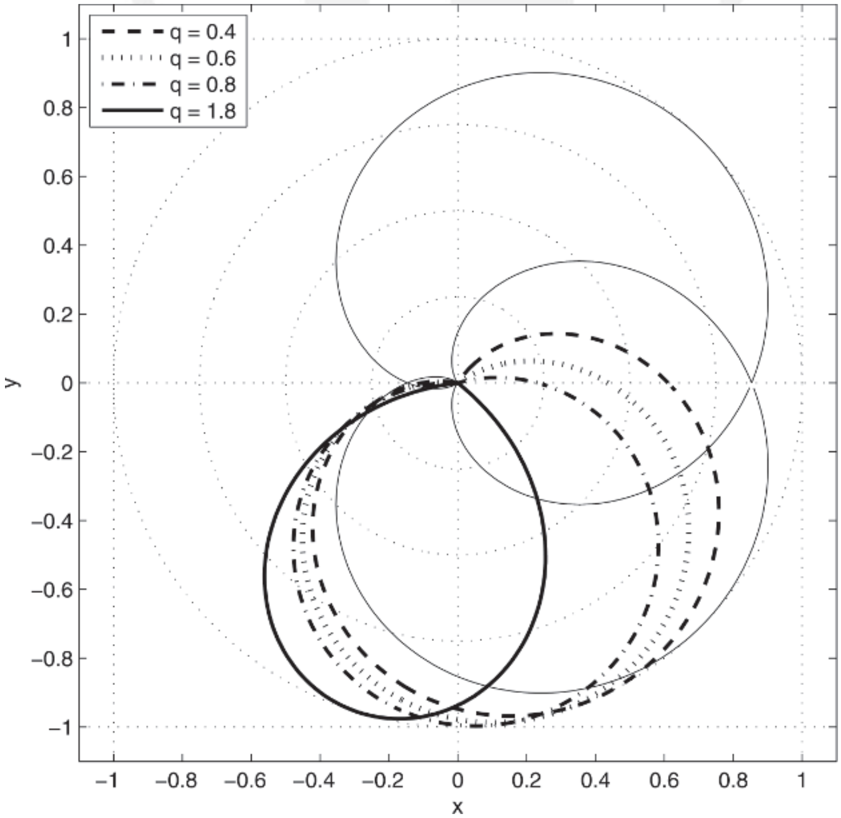

Fig. 6. Original microphone responses (——) and responses for different values of q.

where $D(k, i, \phi)$ is related to the complex amplitude of a plane wave arriving from direction ϕ [15], [16]. For the diffuse sound analysis it is assumed that the power of sound P is independent of direction and that the sound arriving from a specific direction is orthogonal to the sound arriving from all other directions, that is,

$$E\{D(k, i, \phi)D(k, i, \gamma)\} = P\delta(\phi - \gamma) \tag{19}$$

where $\delta(.)$ is the delta Dirac function.

For obtaining N_2 [Eq. (17)] in this case, a [Eq. (4)] is computed, and thus $E\{X_1^2\}$

and $E\{X_1X_2\}$ are needed. $E\{X_1^2\}$ is equal to

$$E\{X_1^2\} = \frac{1}{4}E\left\{\int_{-\pi}^{\pi}[1 + \cos(\phi - \phi_0)]D(k, i, \phi)\, d\phi \times \int_{-\pi}^{\pi}[1 + \cos(\gamma - \phi_0)]D(k, i, \gamma)\, d\gamma\right\}. \tag{20}$$

With Eq. (19) this can be simplified and solved,

$$E\{X_1^2\} = \frac{P}{4}\int_{-\pi}^{\pi}(1 + \cos^2\phi)\, d\phi = \frac{3\pi P}{4}. \tag{21}$$

In a similar fashion $E\{X_1X_2\}$ can be computed,

$$E\{X_1X_2\} = \frac{\pi[2 + \cos(2\phi_0)]P}{4}. \tag{22}$$

Substituting Eqs. (21) and (22) into Eq. (4) yields $a = r$, with

$$r = \frac{2 + \cos(2\phi_0)}{3}. \tag{23}$$

The corresponding directional response is

$$d_{N_2}(\phi) = \frac{1}{2}[1 - r + \cos(\phi - \phi_0) - r\cos(\phi + \phi_0)]. \tag{24}$$

For example, for $\phi_0 = 45°$ the weight [Eq. (23)] is $a = r = 2/3$. The corresponding directional response [Eq. (24)] is shown as a thin solid line in Fig. 8. The responses of X_1 and X_2 are shown as dotted lines.

3.2 Computing Y_2 for Diffuse Sound

The directional response for Y_2 is

$$d_{Y_2}(\phi) = \frac{c}{2}[1 - \tilde{r} + \cos(\phi + \phi_0) - \tilde{r}\cos(\phi - \phi_0)] \tag{25}$$

where $\tilde{r}$ is equal to $\tilde{a}$ [Eq. (7)], which is equal to min (r, q).

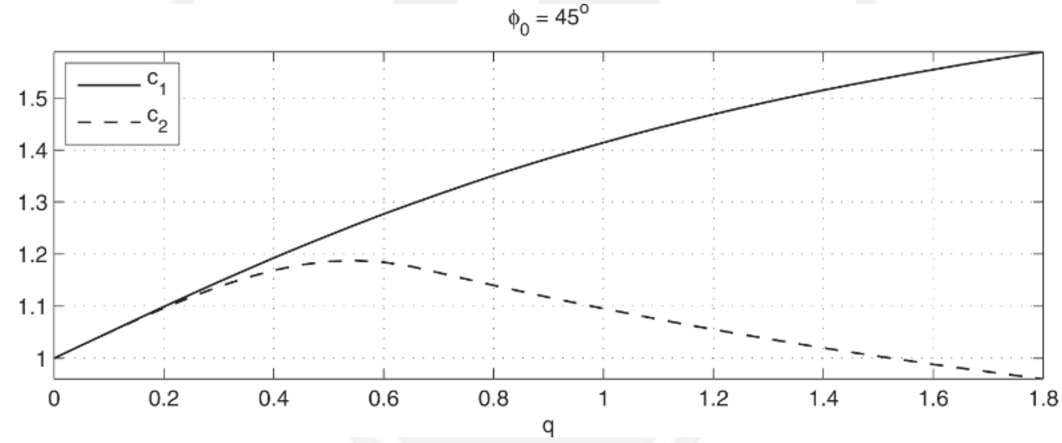

Fig. 7. Postscaling factors for direct and diffuse sound c_1 and c_2 as a function of gain factor limit q.

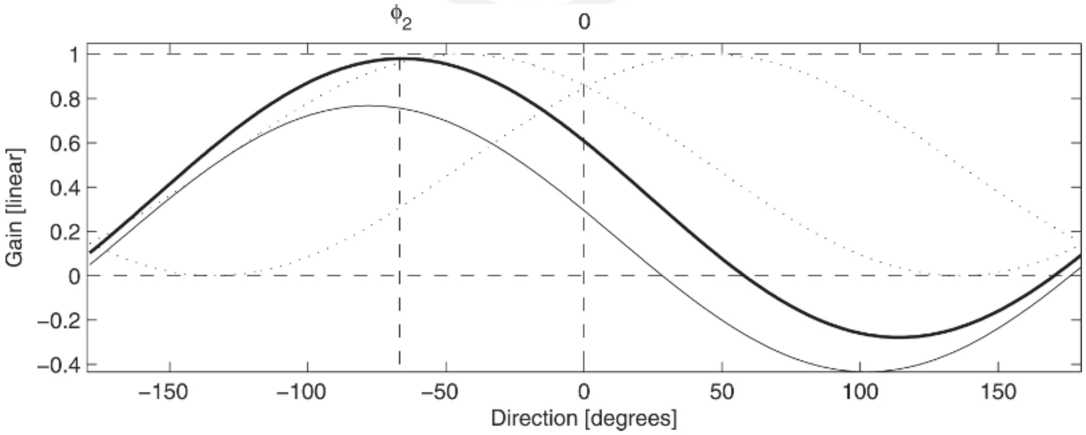

Fig. 8. Directional responses of X_1 and X_2 for $\phi_0 = 45°$ (····). Directional response of N_2 (——) and Y_2 (━━) for diffuse sound. Direction of direct sound response ϕ_2 is indicated as vertical line (–––).

The directional response obtained for the case of sound arriving from one direction [Eq. (12)] is considered to be the desired directional response. However, for diffuse sound the directional response [Eq. (25)] is different. The goal now is to choose c in [Eq. (25)] such that this response is as similar as possible to the desired response, Eq. (12).

In order to match these two different directional responses better, the postscaling factor c for the diffuse sound case is computed such that the power of the resulting Y_2 is the same as the power that would result if the true desired response [Eq. (12)] would pick up the diffuse sound. That is, the postscaling factor is computed as $c = c_2$ with

$$c_2 = \sqrt{\frac{P_{\mathrm{dir}}}{P_{\mathrm{diff}}}} \tag{26}$$

where P_{diff} is the power of Y_2 [Eq. (25)] for the diffuse sound case (with $c = 1$) and P_{dir} is the power of the Y_2 signal if the diffuse sound were picked up by the desired response [Eq. (12)]. The resulting response of Y_2 [Eq. (25)] for diffuse sound is shown in Fig. 8 as bold solid line. Note that the direction (location of maximum) of this response is slightly different than the direction ϕ_2 of the response for direct sound.

In the following, P_{diff} and P_{dir}, needed for the computation of c_2 [Eq. (26)] are computed.

Computation of P_{diff} From Eq. (25) it follows that the signal Y_2 is ($c = 1$)

$$Y_2(k, i) = \frac{1}{2}\int_{-\pi}^{\pi} [1 - \tilde{r} + \cos(\phi + \phi_0) - \tilde{r}\cos(\phi - \phi_0)] \times D(k, i, \phi)\, d\phi. \tag{27}$$

This is equivalent to

$$Y_2(k, i) = \frac{1}{2}\int_{-\pi}^{\pi} [(1 - \tilde{r})(1 + \cos\phi\cos\phi_0) - (1 + \tilde{r})\sin\phi\sin\phi_0]D(k, i, \phi)\, d\phi. \tag{28}$$

Thus the power of Y_2 for diffuse sound P_{diff} can be written as

$$\begin{aligned} P_{\mathrm{diff}} = \frac{1}{4}E\bigg\{&\int_{-\pi}^{\pi} [(1 - \tilde{r})(1 + \cos\phi\cos\phi_0) \\ &+ (1 + \tilde{r})\sin\phi\sin\phi_0]D(k, i, \phi)\, d\phi \\ &\times \int_{-\pi}^{\pi} [(1 - \tilde{r})(1 + \cos\gamma\cos\phi_0) \\ &+ (1 + \tilde{r})\sin\gamma\sin\phi_0]D(k, i, \gamma)\, d\gamma\bigg\}. \end{aligned} \tag{29}$$

Considering the assumption about diffuse sound [Eq. (19)], this can be simplified and solved,

$$\begin{aligned} P_{\mathrm{diff}} &= \frac{P}{4}\int_{-\pi}^{\pi} [(1 - \tilde{r})^2(1 + \cos^2\phi\cos^2\phi_0) + (1 + \tilde{r})^2\sin^2\phi\sin^2\phi_0]\, d\phi \\ &= \frac{\pi P}{4}[(1 - \tilde{r})^2(2 + \cos^2\phi_0) + (1 + \tilde{r})^2\sin^2\phi_0] \end{aligned} \tag{30}$$

Computation of P_{dir} Applying the desired directional response [Eq. (12)] to diffuse sound yields the signal

$$\begin{aligned} Y_2(k, i) &= \frac{c_1}{2}\int_{\phi_2 - \frac{\beta}{2}}^{\phi_2 + \frac{\beta}{2}} [1 - q + R\cos(\phi - \phi_2)]D(k, i, \phi)\, d\phi \\ &= \frac{c_1}{2}\int_{-\frac{\beta}{2}}^{\frac{\beta}{2}} (1 - q + R\cos\phi)D(k, i, \phi)\, d\phi \end{aligned} \tag{31}$$

where β [Eq. (16)] is the width for which the response is nonzero.

The power of Y_2, P_{dir}, can be written as

$$P_{\mathrm{dir}} = \frac{c_1^2}{4}E\left\{\int_{-\frac{\beta}{2}}^{\frac{\beta}{2}} (1 - q + R\cos\phi)D(k, i, \phi)\, d\phi \times \int_{-\frac{\beta}{2}}^{\frac{\beta}{2}} (1 - q + R\cos\gamma)D(k, i, \gamma)\, d\gamma\right\} \tag{32}$$

Considering the assumption about diffuse sound [Eq. (19)], this can be simplified and solved,

$$\begin{aligned} P_{\mathrm{dir}} &= \frac{c_1^2 P}{4}\int_{-\frac{\beta}{2}}^{\frac{\beta}{2}} (1 - q + R\cos\phi)^2\, d\phi \\ &= \frac{c_1^2 P\beta}{4}(1 - q)^2 + \frac{c_1^2 P}{8}R^2\left(\beta + 2\cos\frac{\beta}{2}\sin\frac{\beta}{2}\right) \\ &\quad + c_1^2 PR(1 - q)\sin\frac{\beta}{2}. \end{aligned} \tag{33}$$

Computation of the Postscaling Factor c_2 Thus for diffuse sound the postscaling factor [Eq. (26)] is $c = c_2$, where

$$c_2 = \sqrt{\frac{A + B + C}{2\pi[(1 - \tilde{r})^2(2 + \cos^2\phi_0) + (1 + \tilde{r})^2\sin^2\phi_0]}} \tag{34}$$

with

$$\begin{aligned} A &= 2c_1^2\beta(1 - q)^2 \\ B &= c_1^2 R^2\left(\beta + 2\cos\frac{\beta}{2}\sin\frac{\beta}{2}\right) \\ C &= 8c_1^2 R(1 - q)\sin\frac{\beta}{2}. \end{aligned} \tag{35}$$

The postscaling factor c_2 for diffuse sound as a function of the gain factor limit q is shown as dashed line in Fig. 7.

4 POSTSCALING IN THE GENERAL CASE WHEN THERE IS A MIX OF DIRECT AND DIFFUSE SOUND

In the previous sections it was described how q in Eq. (7) relates to properties of the resulting directional response. Further, the postscaling factors for direct and diffuse sound, c_1 and c_2, respectively, were determined. In practice usually a mix of direct and diffuse sound reaches the microphones. In the following it is described how the postscaling factor c in Eq. (2) is determined for a general scenario when direct sound and diffuse sound reach the microphones.

The first step is to determine how much the sound reaching the microphones is direct or diffuse. For direct sound X_1 and X_2 are coherent, that is, $\Phi_{12} = \Phi_{\text{dir}} = 1$ [Eq. (6)]. For diffuse sound Φ_{12} can be computed using Eqs. (6), (21), (22), and noting that $E\{X_1^2\} = E\{X_2^2\}$,

$$\Phi_{12} = \Phi_{\text{diff}} = \frac{2 + \cos(2\phi_0)}{3} . \tag{36}$$

For the previously used examples with $\phi_0 = 45°$, $\Phi_{\text{diff}} = 2/3$.

Next the postscaling factor c is determined to be a value between c_1 [Eq. (14)] and c_2 [Eq. (26)] as a function of an estimation of how direct or diffuse the sound reaching the microphones at time k and frequency i is,

$$\begin{aligned} c(k, i) &= \frac{\max\{\Phi_{12}(k, i) - \Phi_{\text{diff}}, 0\}}{\Phi_{\text{dir}} - \Phi_{\text{diff}}} (c_1 - c_2) + c_2 \\ &= \frac{\max\{\Phi_{12}(k, i) - \Phi_{\text{diff}}, 0\}}{1 - \Phi_{\text{diff}}} (c_1 - c_2) + c_2 \end{aligned} \tag{37}$$

where $\Phi_{12}(k, i)$, [Eq. (6)] is the normalized cross correlation estimated in the time–frequency tile with time index k and frequency index i.

5 SUBJECTIVE EVALUATION

A subjective test was conducted to show that the proposed processing can improve coincident stereo recordings and that high audio quality is achieved. Ideally we would have liked to have a panel of sound engineers experienced with stereo recording and mixing for the subjective test. Since that was not available, we asked experienced listeners to participate in the test and collaborated with one professional sound engineer with many years of experience with coincident stereo recording, mastering, and mixing. The sound engineer determined “optimal” stereo recording parameters and also participated in the subjective test.

5.1 Stimuli

In order to simulate different coincident microphone setups of the same recording, we used professionally recorded B-format audio material, recorded with a Soundfield Microphone. Five 10-s long audio excerpts, specified in Table 1, were used. The sampling rate of the excerpts was 48 kHz, and time–frequency processing using a 1024-point fast Fourier transform (FFT) with sine windowing and 512-sample hop size was used. FFT spectral bins were combined to form 22 perceptual partitions [13] representing the subbands of the proposed processing. The time constant for short-time averaging ($E\{.\}$) was chosen to be 50 ms. The postscaling factor c [Eq. (37)] is also smoothed with a 50-ms time constant.

Table 1. List of audio excerpts used.

	Excerpt	Category
1	Applause	Ambient
2	Birdland	Jazz
3	Philharmonic	Classical
4	Verbier	Classical
5	Voix Bulgaire	Choir

Table 2 specifies the differently processed stereo signals that were used in the test. The Cardioid case corresponds to the input signal of the proposed processing, that is, signals of two cardioids pointing toward ±45°. Dipole corresponds to a traditional Blumlein stereo recording with two dipoles pointing toward ±45°. Table 3 shows the weights given to the B-format *w, x,* and *y* signals for forming the corresponding responses. The responses for these two cases are shown in Fig. 9(a) and (b).

The Optimized case used directional responses that were chosen by the sound engineer for each excerpt according to his preference. We used a real-time software, which converted the B-format to a stereo signal, where the angle between the left and right responses and the response shape could be chosen freely. Table 4 shows the weights given to the B-format *w, x,* and *y* signals for generating the stereo signals. The corresponding responses are shown in Fig. 9(c). The response for applause is shown as a dashed line and the other responses as solid lines. Note that all responses, except the one for applause, are fairly similar.

The Proposed 1 stereo signals were also optimized by the sound engineer. All parameters were used as described in this paper, and the free parameter to optimize was q. The right-hand column in Table 4 shows the parameter q as determined by the sound engineer. Note that for applause q was chosen larger for more spaciousness.

The Proposed 2 stereo signals are the same as those described previously, except that the diffuse sound gain c_2 was chosen 4 dB larger than as computed in Eq. (26) to amplify diffuse sound.

For Proposed 1 and 2 the responses for the applause excerpt are shown in Fig. 9(d) as dashed lines and the responses for the other excerpts as solid lines. Interestingly

Table 2. Different methods compared in subjective test.

Name	Description
Cardioid	Two cardioid responses pointing toward ±45°
Dipole	Two dipole responses pointing toward ±45°
Optimized	Linear responses optimized manually for each excerpt
Proposed 1	Proposed processing with default parameters
Proposed 2	Proposed processing with 6-dB diffuse sound emphasis

Table 3. B-format decoding parameters for Cardioid and Dipole methods.

Method	*w*	*x*	*y*
Cardioid	1.00	0.50	±0.50
Dipole	0.00	1.00	±1.00

the manually optimized responses for Optimized and Proposed 1 and 2 are very similar, except that the negative backward lobe is missing in Proposed 1 and 2. Fig. 10 shows the Optimized responses averaged, ignoring the applause excerpt (dashed line) and the Proposed 1 and 2 response (solid line), indeed indicating similarity for the front angular range of about ±60°.

The excerpts processed with all methods were scaled to have the same average level (variance).

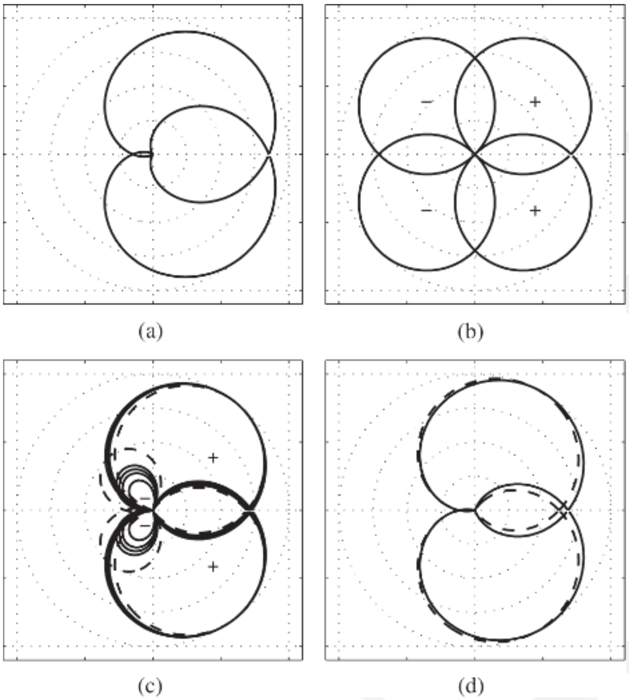

Fig. 9. Responses for different test methods. (a) Cardioid. (b) Dipole. (c) Optimized. (d) Proposed 1 and 2.

Table 4. Manually optimized parameters for Optimized and Proposed 1 and 2 methods.

	Excerpt	*w*	*x*	*y*	*q*
1	Applause	0.49	0.63	±0.86	0.39
2	Birdland	0.69	0.53	±0.76	0.3
3	Philharmonic	0.63	0.55	±0.80	0.3
4	Verbier	0.66	0.54	±0.78	0.3
5	Voix Bulgaire	0.75	0.50	±0.73	0.3

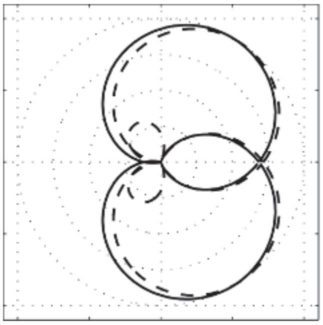

Fig. 10. Directional responses averaged over all excerpts except applause for Optimized (–––) and Proposed 1 and 2 (——) methods.

5.2 Subjects and Test Setup

We asked the sound engineer and six other experienced listeners to participate in the test. Four of the subjects were particularly experienced with spatial audio due to their research related to spatial audio and spatial hearing. Since the subjects were located all over the world, we let them carry out the test autonomously with an automated subjective test software. The subjects used high-quality digital-to-analog converters and headphones (Sennheiser HD 600, Sennheiser HD 650, Stax Lambda Classic).

For judging only the spatial aspect of the stimuli, a monitoring stereo loudspeaker system may have been better since stereo recordings are usually optimized for loudspeaker playback. Besides the distributed subjects, we carried out a headphone-based test for two more reasons. First it is often easier to detect signal distortions with headphones, and we found it important to show that the proposed processing does yield high signal waveform quality. Second we found it easier to hear differences between those methods that were very similar when using headphones.

5.3 Test Method

A MUSHRA [17] type subjective test using a relative grading scale was conducted. The subjects were asked to grade the overall preference of the methods relative to the reference (Cardioid). Each corresponding method had to be graded relative to the reference, while a hidden reference was used to test reliability of the subjects.

Fig. 11 shows the graphical user interface that was used for the test. The subject was presented with a frozen slider for the reference and sliders for the corresponding other methods. With the Play buttons the subject could listen to either the reference or any other method. The subject could switch between the stimuli at any time while the sound instantly faded from one method to the other. Informal listening indicated that such instant switching facilitates comparison of the methods.

We asked the subjects to carry out the subjective test for training until they felt familiar with the stimulus differences and confident in their judgments. The test software showed written instructions on the computer screen before the test started. The test contained the five excerpts listed in Table 1. The excerpt and method order were randomized differently for each subject, and each excerpt was tested twice.

The duration of the test session varied between the listeners due to the freedom to repeat the stimuli as often as requested. Typically the test duration was between 30 and 50 min.

5.4 Results

The results and 95% confidence intervals[1] for each subject, averaged over the five excerpts, are shown in Fig. 12.

[1] The confidence intervals for each subject and method were computed assuming a different mean for each excerpt to indicate how similar the subjects graded the same excerpts and to not indicate variations between the grading of different excerpts.

For each method the average grading for each subject is shown. The leftmost grading for each method is the result of the sound engineer who optimized the Optimized and Proposed 1 and 2 parameters.

The grading for all subjects follows a similar trend, except that the relative gradings of Dipole differ between the subjects. Most subjects judge Dipole as significantly worse than the reference and some judge it as significantly better. All subjects prefer Optimized, Proposed 1, and Proposed 2 over Cardioid and Dipole. The small confidence intervals indicate that the subjects graded the same excerpt very similar both times it occurred in the test.

The results averaged for all listeners and the 95% confidence intervals,[2] shown in Fig. 13, give more indication about the relative performance between Optimized, Proposed 1, and Proposed 2. The mean grading of Proposed 2 is best, followed by Proposed 1, which is slightly above Optimized. A *t*-test with 5% significance level indicates that the mean grading of Proposed 2 is significantly higher than for Optimized. *t*-tests also indicate that the pairs Optimized/Proposed 1 and Proposed 1/Proposed 2, do not have significantly different mean gradings.

To see how the gradings for the different methods depend on the specific excerpt, the mean gradings and 95% confidence intervals[3] averaged over the subjects are shown in Fig. 14 for each method and excerpt. The ordering of the excerpts from left to right in Fig. 14 is the same as the ordering in Table 2, that is, the left grading for each method indicates the grading for Applause, the second grading from the left indicates the grading for Birdland, and so on.

The data shown in Fig. 14 indicate that the performance of the different methods is relatively independent of the excerpt, except for the applause excerpt, which performs

[2] The confidence intervals were computed considering a single mean for each method.

[3] The confidence intervals were computed considering a single mean for each method and excerpt.

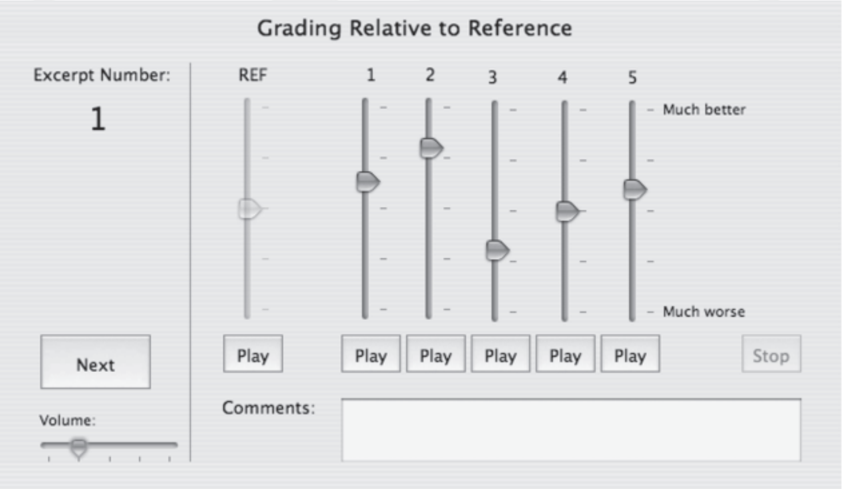

Fig. 11. Graphical user interface used for test. Left frozen slider corresponds to reference, right five sliders correspond to other methods and hidden reference.

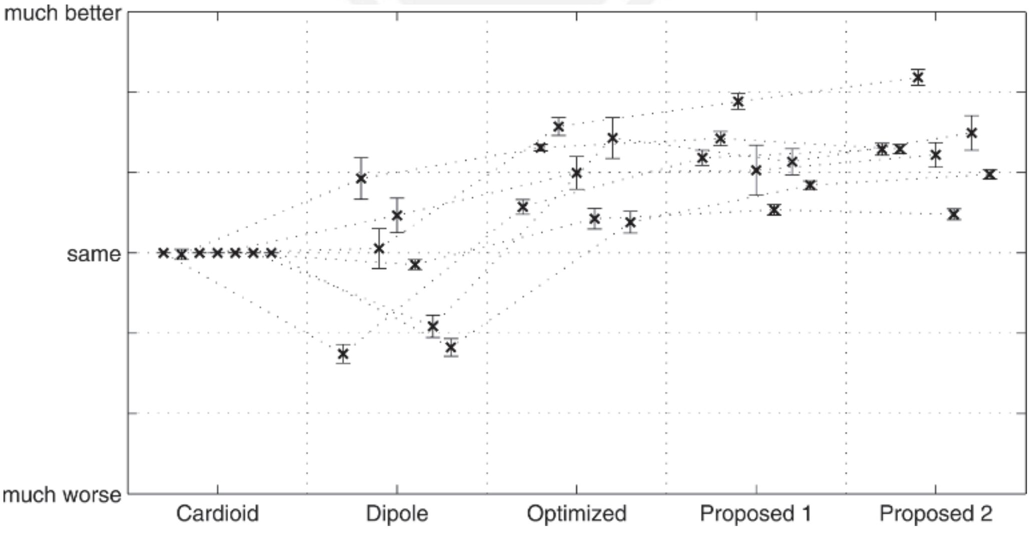

Fig. 12. Mean grading and 95% confidence intervals for each method and subject averaged over all excerpts.

significantly better than the other excerpts for the Dipole method. This may be due to the fact that applause is the only excerpt that is not front sound-stage oriented, that is, sound reaches the microphone from all directions and the Dipole method equally captures sound from all directions. This may give applause an advantage relative to the other excerpts for the Dipole method.

None of the subjects reported waveform degradations (artifacts) that were not present in all methods, indicating that the proposed technique results in a good waveform quality.

6 DISCUSSION

Coincident stereo recordings are known for good localization but often lack spaciousness due to limited directivity of microphones. The lack of spaciousness can be mitigated or partially mitigated by using microphones with directional responses with negative lobes pointing toward the rear (hypercardioids, supercardioids, dipoles). The result is more spaciousness but also lower rear sound rejection.

The proposed technique improves the directivity of two cardioid microphone responses by means of postprocessing. The resulting more narrow responses with less overlap enable stereo recording with good localization and spaciousness while rear sound is rejected and sound from the side is captured in phase.

The proposed technique results in signals very similar to the signals of microphones that would have truly the desired responses. As was shown, the desired responses are achieved perfectly for direct sound. For diffuse sound the responses are not exactly the same, but in an energetic sense similar signals are captured. Since diffuse sound is not directive, one could argue that perceptually it also does not matter if the desired response is truly imposed, as long as energetically the captured signal is correct.

Note that the postscaling factor for diffuse sound c_2 [Eq. (26)] was computed assuming horizontal diffuse

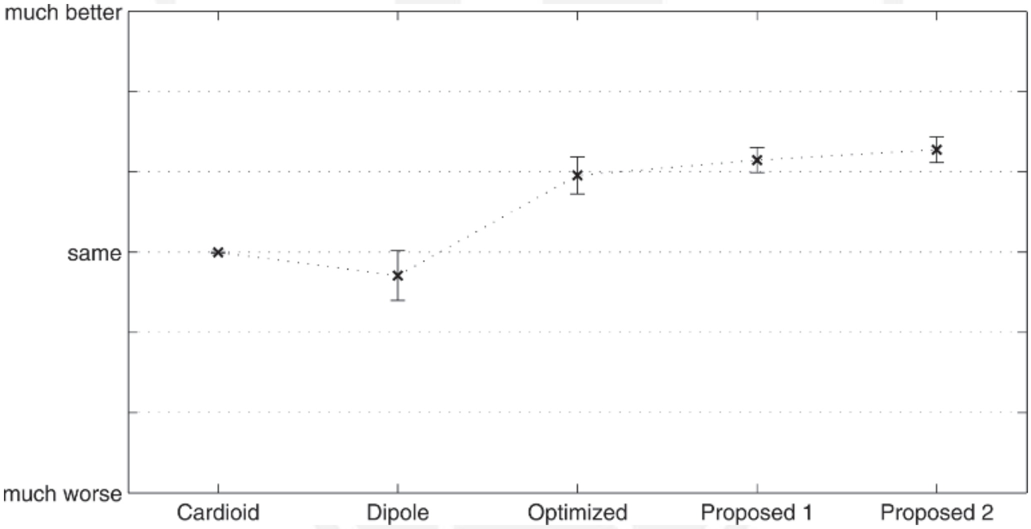

Fig. 13. Mean grading and 95% confidence intervals for each method averaged over all excerpts and subjects.

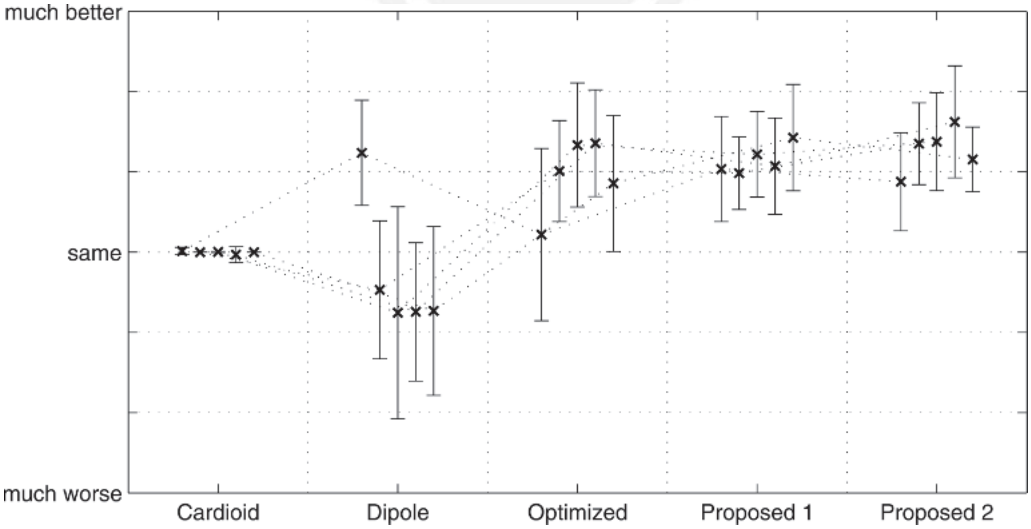

Fig. 14. Mean grading and 95% confidence intervals for each method and excerpt averaged over all subjects.

sound [Eq. (18)]. Using similar derivations, c_2 could also be computed using a diffuse sound model considering elevated sound as well, which would result in a smaller c_2.

The merit of the proposed technique depends on whether or not high signal fidelity is maintained, which is the case as indicated by the subjective test. The headphone-based subjective test also indicates that the subjects preferred the stereo signal generated with the proposed processing over cardioid stereo, dipole stereo, and manually optimized B-format stereo. The proposed technique offers flexibility as M/S and B-format microphone configurations also do while merely being based on two cardioid microphone capsules. The subjects participating in the subjective test favored the proposed processing over all other methods tested, indicating that using merely two cardioids with the proposed processing, one can achieve good stereo recording performance.

The proposed technique is not directly related to beamforming. But there is one adaptive beamforming technique which in concept has some similarities. A sidelobe canceler [18] also uses signals corresponding to various directional responses (the output signals of the "blocking matrix") to achieve a certain goal. A sidelobe canceler is based on optimizing for signal-to-noise ratio by effectively placing nulls in its response at noise source locations, and thus its directional response depends on the locations of the noise sources and is not determined independently of the signal, making conventional adaptive beamforming techniques unsuitable for spatial sound recording.

7 CONCLUSIONS

A technique for modifying the effective directional responses of the left and right signals of coincident stereo recordings was proposed. By means of time–frequency-based signal processing, the widths and directions of the original microphone responses are modified, resulting in less overlap between the left and right responses. This enables coincident stereo recording with good localization and spaciousness, while rear sound is rejected and all sound is in phase.

A subjective test using headphones was carried out, comparing cardioids, dipoles, and manually optimized B-format stereo with the proposed technique applied to cardioid signals. The subjects preferred the proposed technique for headphone playback over all other methods tested.

8 REFERENCES

[1] F. Rumsey, *Spatial Audio,* Music Technology ser. (Focal Press, Oxford, UK, 2001).

[2] J. Eargle, *The Microphone Book* (Focal Press, Oxford, UK, 2004).

[3] A. Blumlein, "Improvements in and Relating to Sound Transmission, Sound Recording and Sound Reproduction Systems," British Patent 394325 (1931); reprinted in *Stereophonic Techniques* (Audio Engineering Society, New York, 1986).

[4] S. P. Lipshitz, "Stereo Microphone Techniques: Are the Purists Wrong?," *J. Audio Eng. Soc.* (*Features*), vol. 34, pp. 716–744 (1986 Sept.).

[5] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization,* rev. ed. (MIT Press, Cambridge, MA, 1997).

[6] M. Williams, "Unified Theory of Microphone Systems for Stereophonic Sound Recording," presented at the 82nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 35, p. 390 (1987 May), preprint 2466.

[7] M. Williams, "Operational Limits of the Variable M/S Stereophonic Microphone System," presented at the 88th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 38, p. 384 (1990 May), preprint 2931.

[8] M. A. Gerzon, "Periphony: Width-Height Sound Reproduction," *J. Audio Eng. Soc.,* vol. 21, pp. 2–10 (1973 Jan./Feb.).

[9] M. A. Gerzon, "Practical Periphony: The Reproduction of Full-Sphere Sound," presented at the 65th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 28, p. 364 (1980 May), preprint 1571.

[10] M. A. Gerzon, "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound," presented at the 50th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 23, pp. 402, 404 (1975 June).

[11] K. Farrar, "Soundfield Microphone," *Wireless World,* pp. 48–50 (1979 Oct.).

[12] D. Griesinger, "Stereo and Surround Panning in Practice," presented at the 112th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 50, p. 513 (2002 June), convention paper 5564.

[13] C. Faller and F. Baumgarte, "Binaural Cue Coding—Part II: Schemes and Applications," *IEEE Trans. Speech Audio Process.,* vol. 11, pp. 520–531 (2003 Nov.).

[14] C. Faller, "Parametric Coding of Spatial Audio," Ph.D. thesis, 3062, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland (2004 July), http://library.epfl.ch/theses/?nr = 3062.

[15] M. A. Poletti, "A Unified Theory of Horizontal Holographic Sound Systems," *J. Audio Eng. Soc.,* vol. 48, pp. 1155–1182 (2000 Dec.).

[16] C. Faller, "Signal Processing for Speech, Audio, and Acoustics," Course Notes, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland (2006).

[17] ITU-R BS.1116.1, "Methods for Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Surround Systems," International Telecommunications Union, Geneva, Switzerland (1997), http://www.itu.org.

[18] L. J. Griffiths and C. W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beamforming," *IEEE Trans. Antennas Propag.,* vol. 30, pp. 27–34 (1982 Jan.)

THE AUTHOR

Christof Faller received an M.S. (Ing.) degree in electrical engineering from ETH Zurich, Switzerland, in 2000, and a Ph.D. degree for his work on parametric multichannel audio coding from EPFL Lausanne, Switzerland, in 2004.

From 2000 to 2004 he worked in the Speech and Acoustics Research Department at Bell Laboratories, Lucent Technologies and Agere Systems (a Lucent company), where he worked on audio coding for digital satellite radio, including parametric multichannel audio coding. He is currently a part-time postdoctoral employee at EPFL Lausanne. In 2006 he founded Illusonic LLC, an audio and acoustics research company.

Dr. Faller has won a number of awards for his contributions to spatial audio coding, MP3 Surround, and MPEG Surround. His main current research interests are spatial hearing and spatial sound capture, processing, and reproduction.