Multiple-Loudspeaker Playback of Stereo Signals*

CHRISTOF FALLER, *AES Member*

Audiovisual Communications Laboratory, EPFL, CH-1015 Lausanne, Switzerland

A perceptually motivated spatial decomposition for two-channel stereo audio signals, capturing the information about the virtual sound stage, is proposed. The spatial decomposition allows resynthesizing audio signals for playback over sound systems other than two-channel stereo. With the use of more front loudspeakers the width of the virtual sound stage can be increased beyond ±30° and the sweet-spot region is extended. Optionally, lateral independent sound components can be played back separately over loudspeakers on the sides of a listener to increase listener envelopment. It is also explained how the spatial decomposition can be used with surround sound and wavefield synthesis–based audio systems.

0 INTRODUCTION

Many innovations beyond two-channel stereo have failed because of cost, impracticability (such as number of loudspeakers), and last but not least a requirement for backward compatibility. While 5.1 surround multichannel audio systems [1], [2] are being adopted widely by consumers, this system is also compromised in terms of the number of loudspeakers and with a backward compatibility restriction. (The front left and right loudspeakers are located at the same angles as in two-channel stereo, namely, ±30°, resulting in a narrow frontal virtual sound stage.)

It is a fact that by far most audio content is available in the two-channel stereo[1] format. For audio systems enhancing the sound experience beyond stereo, it is thus crucial that stereo audio content can be played back, desirably with an improved experience compared to the legacy systems.

It has long been realized that the use of more front loudspeakers improves the virtual sound stage also for listeners located not exactly in the sweet spot [3], [4]. The aim has been to play back stereo signals over more than two loudspeakers for an improved listener experience. Especially a lot of attention has been given to playing back stereo signals with an additional center loudspeaker [5], [6],[2] [8], [9]. In [8] the general case of converting n_1 channels to $n_2 > n_1$ channels is also treated. However, the improvement of these techniques over conventional stereo playback has not been clear enough that they would have been used widely. The main limitations of these techniques are that they consider only localization and not explicitly other aspects such as ambience and listener envelopment. Further, the localization theory behind these techniques is based on a one-virtual-source scenario, limiting their performance when a number of sources are present simultaneously in different directions.

These weaknesses are overcome by the techniques proposed in this paper by using a perceptually motivated spatial decomposition of stereo audio signals. Given this decomposition, audio signals can be rendered for an increased number of loudspeakers, loudspeaker line arrays, and wavefield synthesis systems [10], [11].

Recently another system has been proposed, which theoretically could up-mix stereo signals to any number of audio channels [12]. The method simulates stereo playback and considers the sound field in the sweet spot by computing the B-format [13] signal. The B-format signal is then decoded into the loudspeaker signals. Our approach has several advantages. We believe that B-format computation as an intermediate step is unnecessary and reduces information by introducing more correlation between the audio channels. Further, we are separating direct and diffuse sound with least-squares estimation (multichannel Wiener filtering) and thus can obtain diffuse sound signals that are truly statistically independent, whereas the method proposed in [12] cannot obtain independent diffuse sound signals (unless artificial reverberators are used), resulting in a limited ability to generate independent loudspeaker signals.

Another class of techniques has been proposed to convert two-channel stereo signals into multichannel surround

*Manuscript received 2006 June 23; revised 2006 September 17.

[1]In the following the term "stereo" is used for two-channel stereo.

[2]Reprinted in [7].

signals, that is, most often 5.1-channel surround signals. Usually these algorithms generate three front channels (left, right, and center) for reproducing the sound stage and two surround channels for ambience reproduction. These algorithms can be compared to the proposed algorithm in terms of two-to-three front-channel conversion, but are more limited since they specifically address only this scenario.

One way of converting stereo signals to surround signals is to apply a conventional matrixing decoder to (non-matrixed) stereo signals. For example, Dolby Prologic II [14] decoders can be used for this purpose. Matrix decoders suffer from the same limitations as mentioned for the previously described algorithms. Another technique has been proposed in [15], which operates in the short-time Fourier transform (STFT) domain. This technique detects ambient signal portions in time and frequency and plays part of these back over the rear channels. The proposed technique is more general and more advanced in terms of considering psychoacoustics and signal decomposition.

The paper is organized as follows. Section 1 reviews aspects of spatial hearing that are considered in the decomposition of stereo signals. The proposed decomposition of stereo signals is presented in Section 2. Playback of the decomposed stereo signals over multiple loudspeakers, loudspeaker arrays, and wavefield synthesis systems is described in Section 3. The subjective test that was carried out to evaluate the playback quality is described in Section 4. Finally, Section 5 presents the conclusions.

1 SPATIAL HEARING AND STEREO LOUDSPEAKER PLAYBACK

The most commonly used consumer playback system for spatial audio is the stereo loudspeaker setup shown in Fig. 1. Two loudspeakers are placed in front on the left and right sides of the listener. Usually these loudspeakers are placed on a circle at angles of −30° and 30°. The width of the auditory spatial image that is perceived when listening to such a stereo playback system is limited approximately to the area between and behind the two loudspeakers.

The perceived auditory spatial image, in natural listening and when listening to reproduced sound, largely depends on the binaural localization cues, that is, the interaural time difference (ITD), interaural level difference (ILD), and interaural coherence (IC) [16]. Furthermore it has been shown that the perception of elevation is related to monaural cues [16].

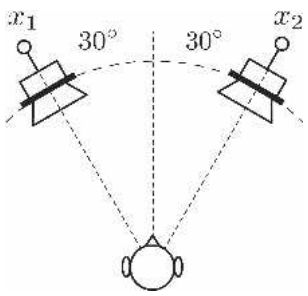

Fig. 1. Standard stereo loudspeaker setup.

The ability to produce an auditory spatial image mimicking a sound stage with stereo loudspeaker playback is made possible by the perceptual phenomenon of summing localization [16], that is, an auditory event can be made to appear at any angle between a loudspeaker pair in front of a listener by controlling the level and/or time difference between the signals given to the loudspeakers. It was Blumlein who, in the 1930s, recognized the power of this principle and filed his now famous patent on stereophony [17]. Summing localization is based on the fact that ITD and ILD cues evoked at the ears approximate crudely the dominating cues that would appear if a physical source were located in the direction of the auditory event, which appears between the loudspeakers.

As illustrated in Fig. 2, summing localization can be used to mimic a scenario where different instruments are located in different directions on a virtual sound stage [18], that is, in the region between the two loudspeakers. In the following it is described how other attributes than localization can be controlled.

Early refection widen

Important in concert hall acoustics is the consideration of reflections arriving at the listener from the sides, that is, lateral reflections. It has been shown that early lateral reflections have the effect of widening the auditory event [19]. The effect of early reflections with delays smaller than about 80 ms is approximately constant, and thus a physical measure, called lateral fraction, has been defined considering early reflections in this range [19]. The lateral fraction is the ratio of the lateral sound energy to the total sound energy that arrived within the first 80 ms after the arrival of the direct sound and measures the width of the auditory event.

An experimental setup for emulating early lateral reflections is illustrated in Fig. 3(a). The direct sound is emitted from the center loudspeaker, whereas independent early reflections are emitted from the left and right loudspeakers. The width of the auditory event increases as the relative strength of the early lateral reflections is increased. Auditory events are indicated as gray filled circles in the figure.

More than 80 ms after the arrival of the direct sound, lateral reflections tend to contribute more to the perception of the environment than to the auditory event itself. This is manifested in a sense of "envelopment" or "spaciousness of the environment," frequently called listener envelopment [20]. A similar measure as the lateral fraction for early reflections is also applicable to late reflections for measuring the degree of listener envelopment. This measure is called late lateral energy fraction [21].

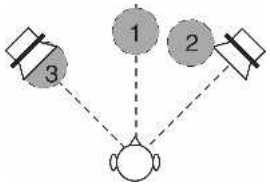

Fig. 2. Level and time difference between a pair of coherent loudspeaker signals determine the location of the auditory event that appears between the two loudspeakers.

Late lateral reflections can be emulated with a setup as shown in Fig. 3(b). The direct sound is emitted from the center loudspeaker, whereas independent late reflections are emitted from the left and right loudspeakers. The sense of listener envelopment increases as the relative strength of the late lateral reflections is increased, while the width of the auditory event is expected to be hardly affected. The perceived auditory events are indicated as gray filled objects in the figure.

Why each source goes coherently

Stereo signals are recorded or mixed such that for each source the signal goes coherently into the left and right signal channels with specific directional cues (level difference, time difference), and reflected/reverberated independent signals go into the channels determining auditory event width and listener envelopment cues. It is beyond the scope of this paper to further discuss mixing and recording techniques. The interested reader is referred to [2].

This is a model x1 is left and x2 is right.

2 SPATIAL DECOMPOSITION OF STEREO SIGNALS

As opposed to using a direct sound from a real source, as was illustrated in Fig. 3, one can use direct sound corresponding to a virtual source generated with summing localization. That is, experiments as are shown in Fig. 3 can be carried out with only two loudspeakers. This is illustrated in Fig. 4, where the signal s mimics the direct sound from a direction determined by the factor a. The independent signals n_1 and n_2 correspond to the lateral reflections. The described scenario is a perceptually motivated decomposition for stereo signals with one auditory event,

$$x_1(n) = s(n) + n_1(n)$$
$$x_2(n) = as(n) + n_2(n) \quad (1)$$

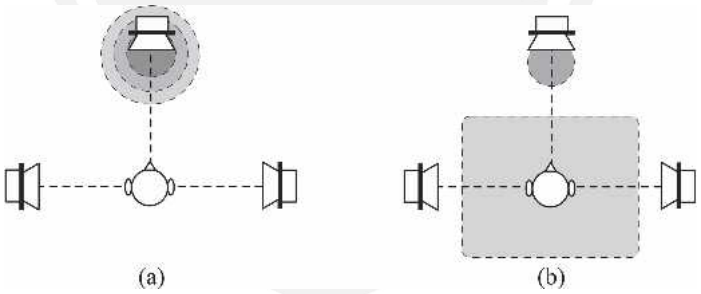

Fig. 3. (a) Early reflections emitted from side loudspeakers have the effect of widening the auditory event. (b) Late reflections emitted from side loudspeakers relate more to the environment as listener envelopment. Shaded areas indicate perceived auditory events.

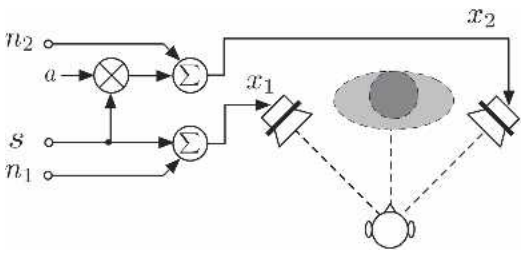

Fig. 4. Mixing a stereo signal mimicking direct sound s and lateral reflections n_1 and n_2. Factor a determines the direction in which auditory event appears.

Auditory Event?

capturing the localization and width of the auditory event and listener envelopment.

In order to obtain a decomposition that is not only effective in one auditory event scenario, but in nonstationary scenarios with multiple concurrently active sources, the described decomposition is carried out independently in a number of frequency bands and adaptively in time,

$$X_1(i, k) = S(i, k) + N_1(i, k)$$
$$X_2(i, k) = A(i, k)S(i, k) + N_2(i, k) \quad (2)$$

where i is the subband index and k is the subband time index. This is illustrated in Fig. 5. Here in each time–frequency tile with indices i and k, the signals S, N_1, N_2 and the factor A are estimated independently. For brevity of notation, the subband and time indices are often ignored in the following. We are using a subband decomposition with perceptually motivated subband bandwidths, that is, the bandwidth of a subband is chosen to be equal to one critical band [22], [23]. S, N_1, N_2, and A are estimated approximately every 20 ms in each subband. For low computational complexity, we are using a short-time Fourier transform (STFT) implemented using a fast Fourier transform (FFT). The spectral coefficients are grouped such that each group corresponds to one critical band. This type of processing, using the STFT to carry out critical-band-based processing, is described in detail in [24] or [18].

Perceptual

Given the stereo subband signals X_1 and X_2, the goal is to compute estimates of S, A, N_1, and N_2. A short-time estimate of the power of X_1 is denoted by $P_{X_1}(i, k) = E\{X_1^2(i, k)\}$, where $E\{.\}$ is a short-time averaging operation. For the other signals the same convention is used, namely, P_{X_2}, P_S, and $P_N = P_{N_1} = P_{N_2}$ are the corresponding short-time power estimates. The power of N_1 and N_2 is

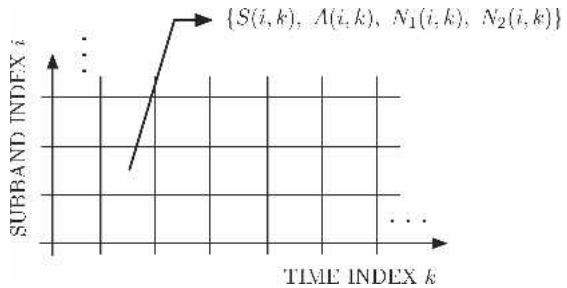

Fig. 5. Each left and right time–frequency tile of stereo signal X_1 and X_2 is decomposed into three signals S, N_1, and N_2 and a factor A.

assumed to be the same, that is, it is assumed that the amount of lateral independent sound is the same for left and right.

2.1 Estimating P_S, A, and P_N

Given the subband representation of the stereo signal, the power (P_{X_1}, P_{X_2}) and the normalized cross correlation are computed. The normalized cross correlation between left and right is

$$\Phi(i,k)=\frac{E\{X_1(i,k)X_2(i,k)\}}{\sqrt{E\{X_1^2(i,k)\}E\{X_2^2(i,k)\}}}. \tag{3}$$

A, P_S, and P_N are computed as a function of the extimated P_{X_1}, P_{X_2}, and Φ. Three equations relating the known and unknown variables are

$$\begin{aligned} P_{X_1}&=P_S+P_N\\ P_{X_2}&=A^2P_S+P_N\\ \Phi&=\frac{aS}{\sqrt{P_{X_1}P_{X_2}}}. \end{aligned} \tag{4}$$

AS

These equations solved for A, P_S, and P_N yield

$$\begin{aligned} A&=\frac{B}{2C}\\ P_S&=\frac{2C^2}{B}\\ P_N&=X_1-\frac{2C^2}{B} \end{aligned} \tag{5}$$

with

$$\begin{aligned} B&=P_{X_2}-P_{X_1}+\sqrt{(P_{X_1}-P_{X_2})^2+4P_{X_1}P_{X_2}\Phi^2}\\ C&=\Phi\sqrt{P_{X_1}P_{X_2}}. \end{aligned} \tag{6}$$

2.2 Least-Squares Estimation of S, N_1, and N_2

Next the least-squares estimates of S, N_1, and N_2 are computed as a function of A, P_S, and P_N. For each i and k the signal S is estimated as

$$\begin{aligned} \hat{S}&=w_1X_1+w_2X_2\\ &=w_1(S+N_1)+w_2(AS+N_2) \end{aligned} \tag{7}$$

where w_1 and w_2 are real-valued weights. The estimation error is

$$E=(1-w_1-w_2A)S-w_1N_1-w_2N_2. \tag{8}$$

The weights $w_1$and w_2 are optimal in a least-mean-square sense when the error E is orthogonal to X_1 and X_2 [25], that is,

Orthogonality Principle

$$\begin{aligned} E\{EX_1\}&=0\\ E\{EX_2\}&=0 \end{aligned} \tag{9}$$

yielding two equations,

$$\begin{aligned} (1-w_1-w_2A)P_S-w_1P_N&=0\\ A(1-w_1-w_2A)P_S-w_2P_N&=0 \end{aligned} \tag{10}$$

from which the weights are computed,

$$\begin{aligned} w_1&=\frac{P_SP_N}{(A^2+1)P_SP_N+P_N^2}\\ w_2&=\frac{AP_SP_N}{(A^2+1)P_SP_N+P_N^2}. \end{aligned} \tag{11}$$

Similarly, N_1 and N_2 are estimated. The estimate of N_1 is

$$\begin{aligned} \hat{N}_1&=w_3X_1+w_4X_2\\ &=w_3(S+N_1)+w_4(AS+N_2). \end{aligned} \tag{12}$$

The estimation error is

$$E=(-w_3-w_4A)S-(1-w_3)N_1-w_2N_2. \tag{13}$$

Again, the weights are computed such that the estimation error is orthogonal to X_1 and X_2, resulting in

$$\begin{aligned} w_3&=\frac{A^2P_SP_N+P_N^2}{(A^2+1)P_SP_N+P_N^2}\\ w_4&=\frac{-AP_SP_N}{(A^2+1)P_SP_N+P_N^2}. \end{aligned} \tag{14}$$

The weights for computing the least-squares estimate of N_2,

$$\begin{aligned} \hat{N}_2&=w_5X_1+w_6X_2\\ &=w_5(S+N_1)+w_6(AS+N_2) \end{aligned} \tag{15}$$

are

$$\begin{aligned} w_5&=\frac{-AP_SP_N}{(A^2+1)P_SP_N+P_N^2}\\ w_6&=\frac{P_SP_N+P_N^2}{(A^2+1)P_SP_N+P_N^2}. \end{aligned} \tag{16}$$

2.3 Postscaling

Given the least-squares estimates, these are postscaled such that the power of the estimates $\hat{S}$, $\hat{N}_1$, and $\hat{N}_2$ equals P_S and $P_N=P_{N_1}=P_{N_2}$. The power of $\hat{S}$ is

$$P_{\hat{S}}=(w_1+aw_2)^2P_S+(w_1^2+w_2^2)P_N. \tag{17}$$

Thus to obtain an estimate of S with power P_S, $\hat{S}$ is scaled

$$\hat{S}'=\frac{\sqrt{P_S}}{\sqrt{(w_1+aw_2)^2P_S+(w_1^2+w_2^2)P_N}}\hat{S}. \tag{18}$$

With similar reasoning, $\hat{N}_1$ and $\hat{N}_2$ are scaled,

$$\begin{aligned} \hat{N}_1'&=\frac{\sqrt{P_N}}{\sqrt{(w_3+aw_4)^2P_S+(w_3^2+w_4^2)P_N}}\hat{N}_1\\ \hat{N}_2'&=\frac{\sqrt{P_N}}{\sqrt{(w_5+aw_6)^2P_S+(w_5^2+w_6^2)P_N}}\hat{N}_2. \end{aligned} \tag{19}$$

2.4 Numerical Examples

The factor A and the normalized power of S and AS are shown in Fig. 6 as functions of the stereo signal level difference and Φ.

The weights w_1 and w_2 for computing the least-squares estimate of S are shown in the top two panels of Fig. 7 as functions of the stereo signal level difference and Φ. The postscaling factor for $\hat{S}$ [Eq. (18)] is shown in the bottom panel.

The weights w_3 and w_4 for computing the least-squares estimate of N_1 and the corresponding postscaling factor, Eq. (19), are shown in Fig. 8 as functions of the stereo signal level difference and Φ.

The weights w_5 and w_6 for computing the least-squares estimate of N_2 and the corresponding postscaling factor, Eq. (19), are shown in Fig. 9 as functions of the stereo signal level difference and Φ.

An example for the spatial decomposition of a stereo rock music clip with a singer in the center is shown in Fig. 10. The estimates of s, A, n_1, and n_2 are shown. The signals are shown in the time domain and A is shown for every time–frequency tile. The estimated direct sound s is relatively strong compared to the independent lat-

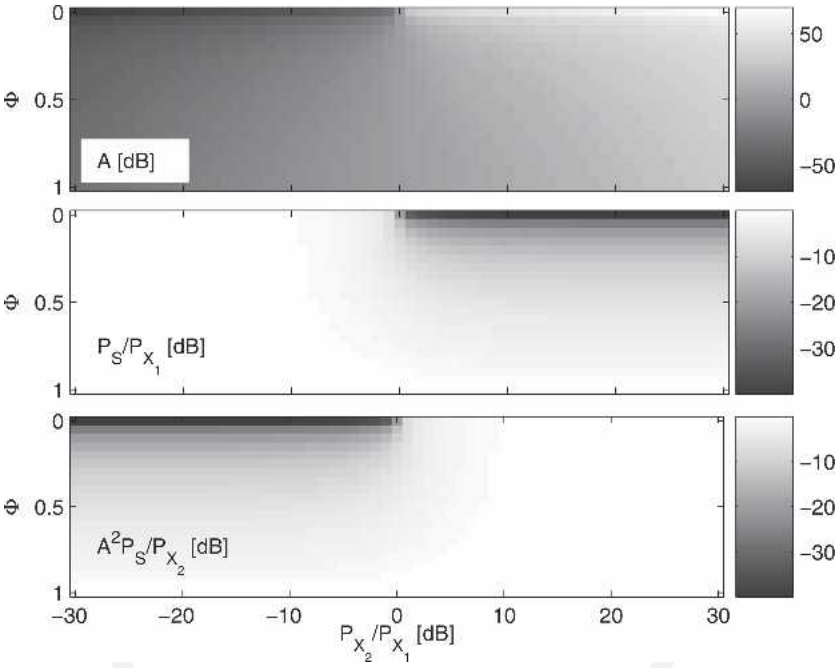

Fig. 6. Factor A and normalized power of S and AS.

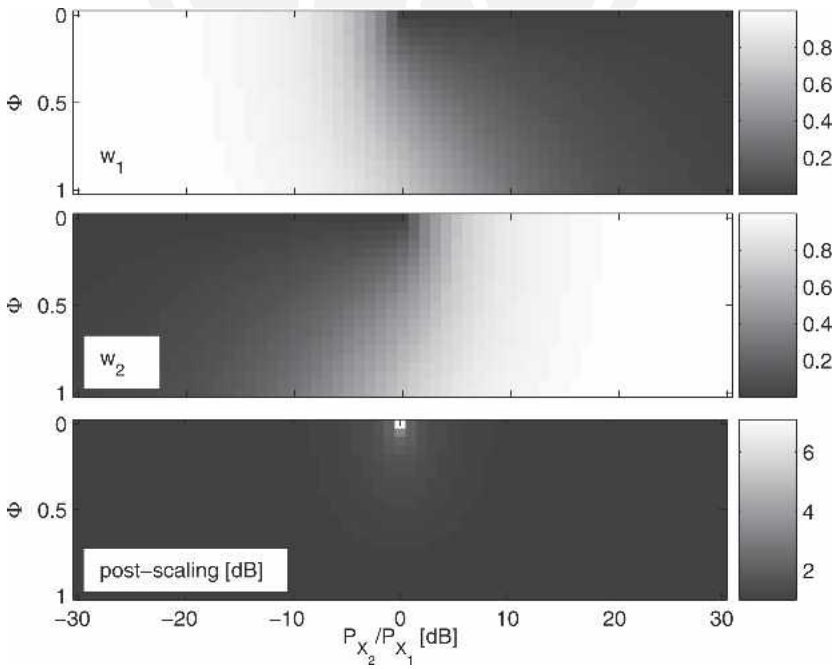

Fig. 7. Least-squares estimate weights w_1 and w_2 and postscaling factor for computation of estimate of s.

eral sounds n_1 and n_2 since the singer in the center is dominant.

3 PLAYING BACK THE DECOMPOSED STEREO SIGNALS OVER DIFFERENT PLAYBACK SETUPS

Given the spatial decomposition of the stereo signal, namely, the subband signals for the estimated localized direct sound $\hat{S}'$, the factor A, and the lateral independent sounds $\hat{N}_1'$ and $\hat{N}_2'$, one can define rules on how to emit the signal components corresponding to $\hat{S}'$, $\hat{N}_1'$ and $\hat{N}_2'$ from different playback setups.

3.1 Multiple Loudspeakers in Front of the Listener

Fig. 11 illustrates the scenario that is addressed. The virtual sound stage of width $\phi_0 = 30°$, shown in Fig. 11(a), is scaled to a virtual sound stage of width ϕ_0', which is reproduced with multiple loudspeakers, shown in Fig. 11(b).

The estimated independent lateral sounds $\hat{N}_1'$ and $\hat{N}_2'$ are emitted from the loudspeakers on the sides, namely, loudspeakers 1 and 6 in Fig. 11(b). That is, because the more the lateral sound is emitted from the side, the more it is

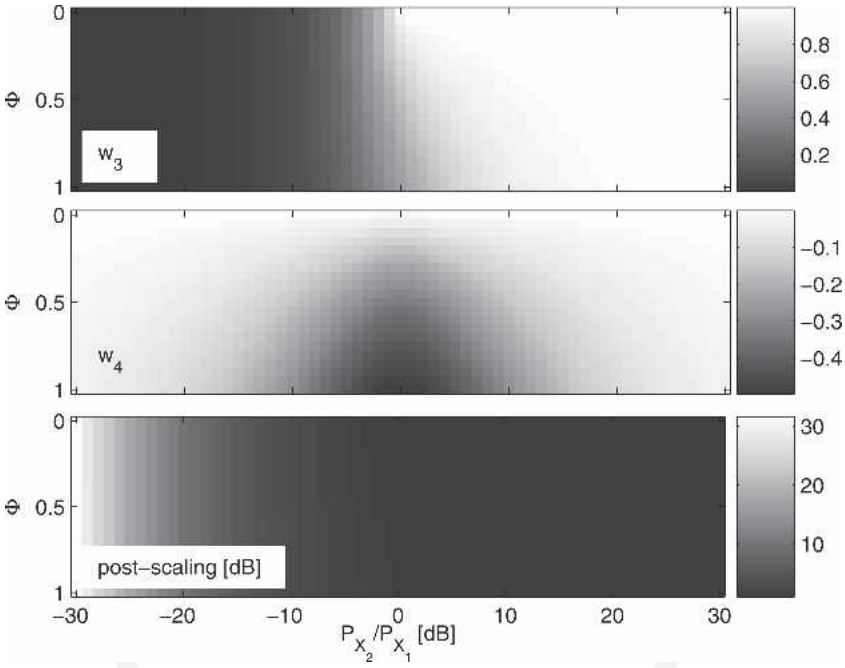

Fig. 8. Least-squares estimate weights w_3 and w_4 and postscaling factor for computation of estimate of N_1.

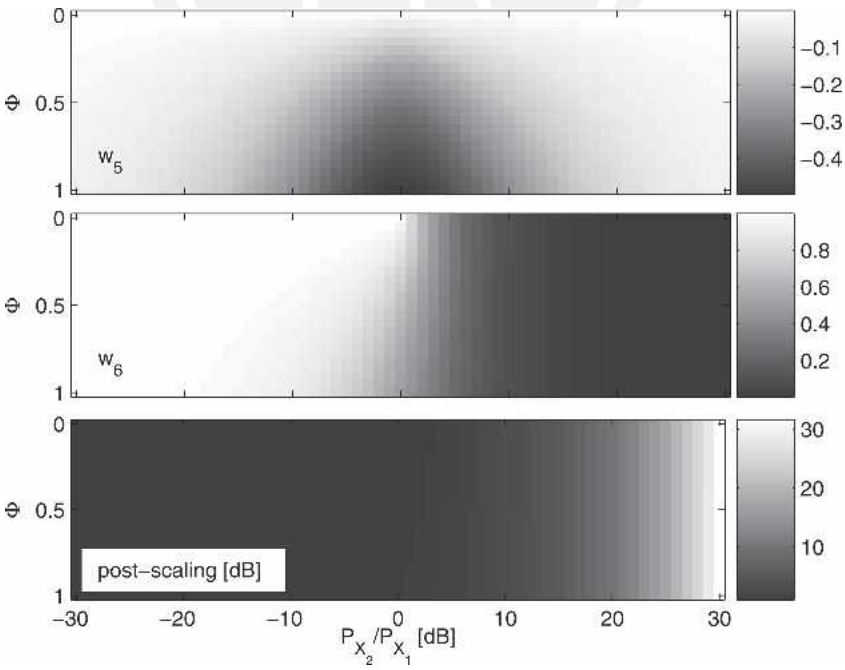

Fig. 9. Least-squares estimate weights w_5 and w_6 and postscaling factor for computation of estimate N_2.

effective in terms enveloping the listener into the sound [20]. Given the estimated factor A, the angle ϕ of the auditory event relative to the $\pm\phi_0$ virtual sound stage is estimated, using the stereophonic law of sines [26],

$$\phi = \sin^{-1}\left(\frac{A-1}{A+1}\sin\phi_0\right). \tag{20}$$

Alternatively, other panning laws, such as the stereophonic law of tangents, may be used. The angle obtained is scaled linearly to compute the angle relative to the widened sound stage,

$$\phi' = \frac{\phi_0'}{\phi_0}\phi. \tag{21}$$

The loudspeaker pair enclosing ϕ' is selected. In the example illustrated in Fig. 11(b) this pair has indices 4 and 5. The angles relevant for amplitude panning between this loudspeaker pair, γ_0 and γ, are defined as shown in the figure. If the selected loudspeaker pair has indices l and $l+1$, then the signals given to these loudspeakers are

$$a_1\sqrt{1+A^2}S, \qquad a_2\sqrt{1+A^2}S \tag{22}$$

where the amplitude panning factors a_1 and a_2 are computed using the stereophonic law of sines and normalized such that $a_1^2 + a_2^2 = 1$,

$$a_1 = \frac{1}{\sqrt{1+C^2}}, \qquad a_2 = \frac{C}{\sqrt{1+C^2}} \tag{23}$$

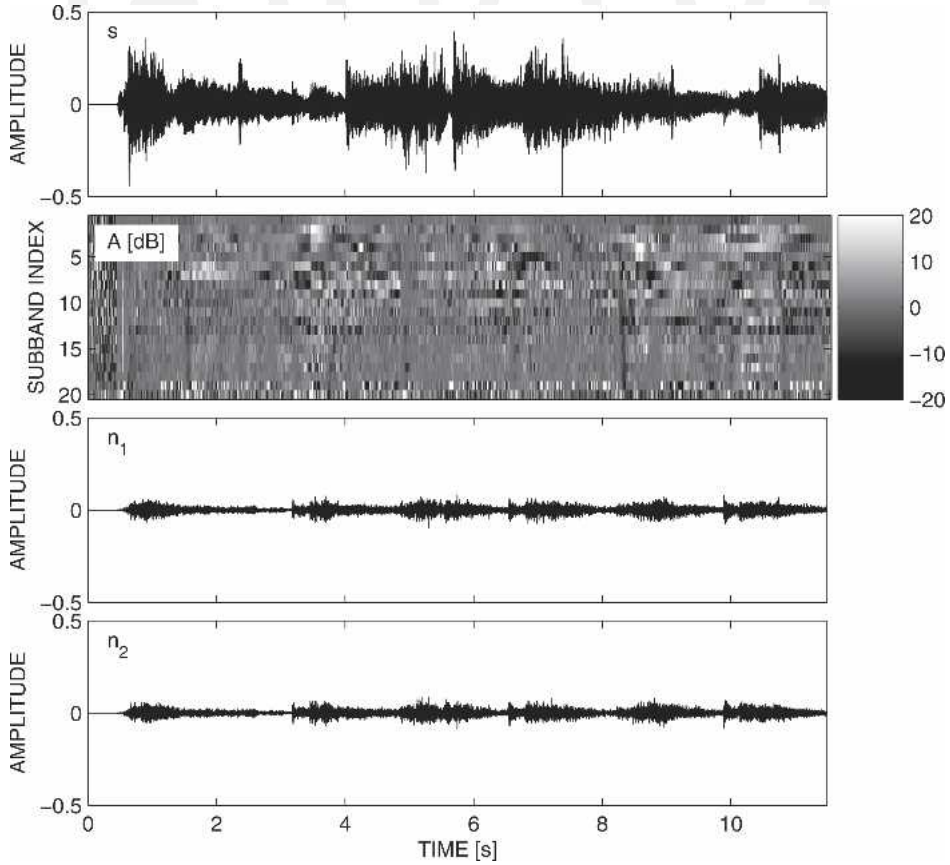

Fig. 10. Estimated s, A, n_1, and n_2.

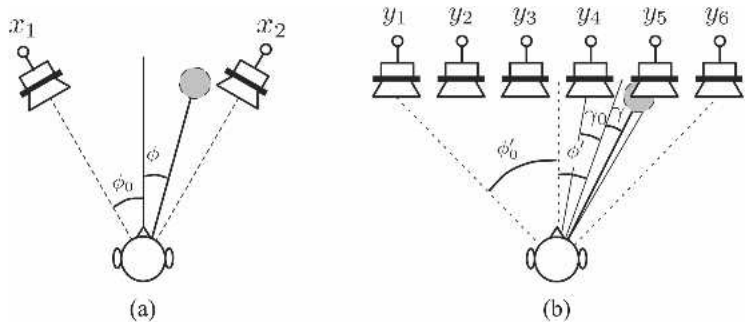

Fig. 11. (a) ±30° virtual sound stage. (b) Virtual sound stage with aperture width of a loudspeaker array.

with

$$C = \frac{\sin(\gamma_0 + \gamma)}{\sin(\gamma_0 - \gamma)}. \quad (24)$$

The factors $\sqrt{1 + A^2}$ in Eq. (22) are such that the total power of these signals is equal to the total power of the coherent components S and AS in the stereo signal.

Fig. 12 shows an example for the selection of loudspeaker pairs l and $l + 1$ and the amplitude panning factors a_1 and a_2 for $\phi_0' = \phi_0 = 30°$ for $M = 8$ loudspeakers at the angles {−30°, −20°, −12°, −4°, 4°, 12°, 20°, 30°}.

Given the foregoing reasoning, each time–frequency tile of the output signal channels i and k is computed as

$$Y_m = \delta(m-1)\hat{N}_1' + \delta(m-M)\hat{N}_2' + [\delta(m-l)a_1 + \delta(m-l-1)a_2]\sqrt{1+A^2}\hat{S}' \quad (25)$$

where

$$\delta(m) = \begin{cases} 1, & m = 0 \\ 0, & \text{otherwise} \end{cases} \quad (26)$$

and m is the output channel index, $1 \le m \le M$. The subband signals of the output channels are converted back to the time domain and form the output channels y_1 to y_M. In the following, this last step is not always explicitly mentioned again.

A limitation of the scheme described is that when the listener is on one side, such as close to loudspeaker 1, the lateral independent sound will reach him with much more intensity than the lateral sound from the other side. This problem can be circumvented by emitting the lateral independent sound from all loudspeakers with the aim of generating two lateral plane waves. This is illustrated in Fig. 13. The lateral independent sound is given to all loudspeakers with delays mimicking a plane wave with a certain direction,

$$Y_m(i,k) = \frac{\hat{N}_1'[i, k-(m-1)d]}{\sqrt{M}} + \frac{\hat{N}_2'[i, k-(M-m)d]}{\sqrt{M}} + [\delta(m-l)a_1 + \delta(m-l-1)a_2]\sqrt{1+A^2}\hat{S}' \quad (27)$$

where d is the delay,

$$d = \frac{sf_s \sin\alpha}{v} \quad (28)$$

s is the distance between the equally spaced loudspeakers, v is the speed of sound, f_s is the subband sampling frequency, and $\pm\alpha$ are the directions of propagation of the two plane waves. In our system the subband sampling frequency is not high enough such that d can be expressed as an integer. Thus we are first converting N_1' and N_2' to the

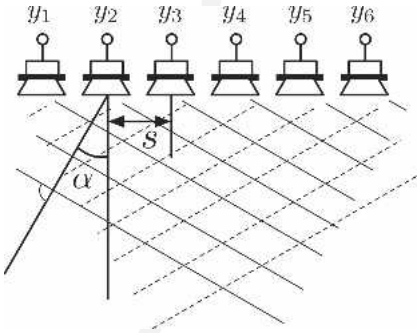

Fig. 13. $\hat{H}_1'$ and $\hat{H}_2'$ are emitted as two plane waves emitted with angles $\pm\alpha$.

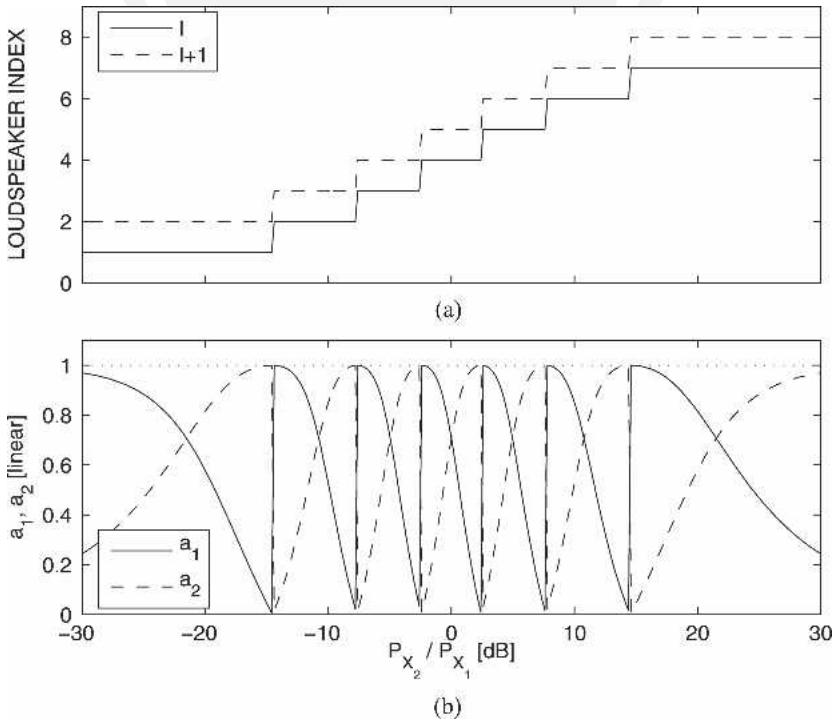

Fig. 12. (a) Loudspeaker pair selection l. (b) Factors a_1 and a_2. Both as a function of stereo signal level difference P_{X_2}/P_{x_1}.

time domain. Then we add various delayed versions of these time domain signals to the output channels.

3.2 Multiple Front Loudspeakers plus Side Loudspeakers

The previously described playback scenario aims at widening the virtual sound stage and at making the perceived sound stage independent of the location of the listener.

Optionally one can play back the independent lateral sounds N_1' and N_2' with two separate loudspeakers located more to the sides of the listener, as illustrated in Fig. 14. It is expected that this results in a stronger impression of listener envelopment. In this case the output signals are also computed by Eq. (25), where the signals with indices 1 and M are the loudspeakers on the side. The loudspeaker pair selection l and $l + 1$ is in this case such that $\hat{S}'$ is never given to the signals with indices 1 and M since the whole width of the virtual stage is projected to only the front loudspeakers, $2 \leq m \leq M - 1$.

Fig. 15 shows an example of the eight signals generated for the setup shown in Fig. 14 for the same music clip for which the spatial decomposition was shown in Fig. 10. Note that the dominant singer in the center is amplitude panned between the center two loudspeaker signals y_4 and y_5.

3.3 Conventional 5.1-Channel Surround Loudspeaker Setup

One possibility to convert a stereo signal to a 5.1-channel surround compatible multichannel audio signal is to use a setup as shown in Fig. 14(b) with three front loudspeakers and two rear loudspeakers arranged as specified in the 5.1 standard. In this case the rear loudspeakers

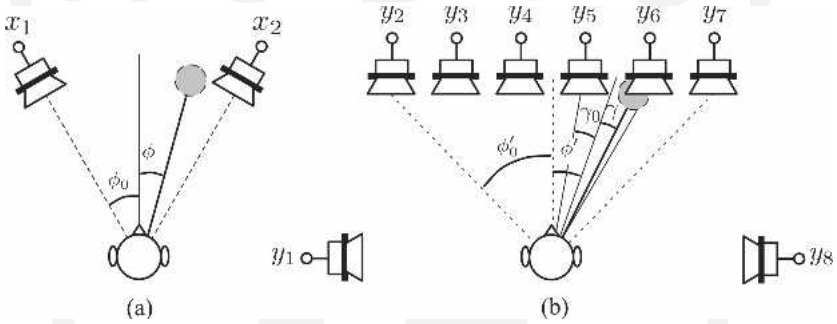

Fig. 14. (a) ±30° virtual sound stage. (b) Virtual sound stage with aperture width of a loudspeaker array. In addition, lateral independent sound is played from sides with separate loudspeakers for a stronger listener envelopment.

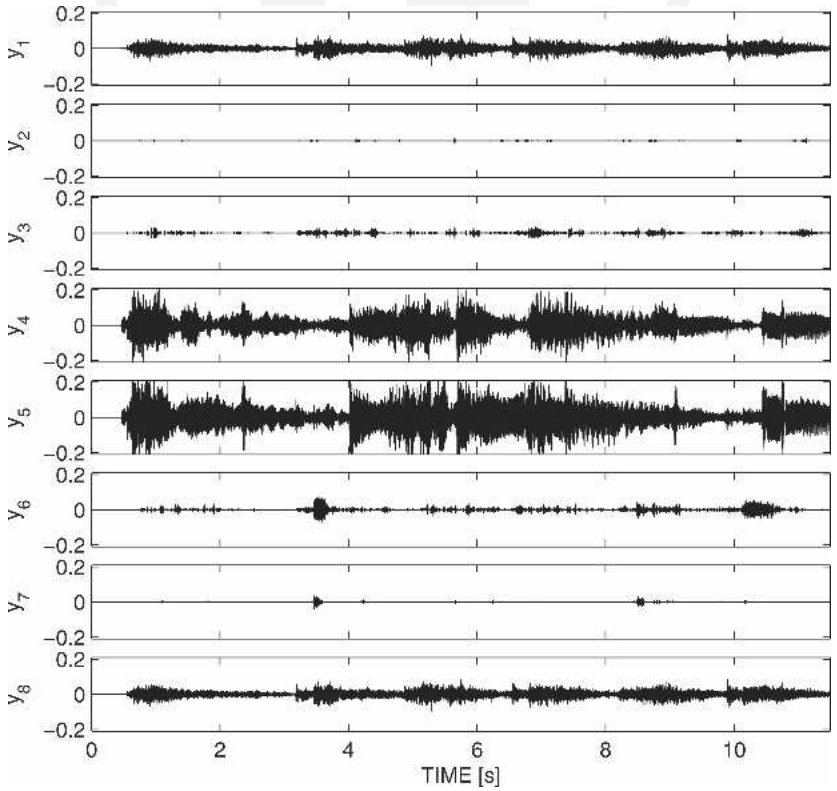

Fig. 15. Eight signals generated for a setup as in Fig. 14(b).

emit the independent lateral sound, whereas the front loudspeakers are used to reproduce the virtual sound stage. Informal listening indicates that when playing back audio signals as described, listener envelopment is more pronounced compared to stereo playback.

Another possibility to convert a stereo signal to a 5.1-channel surround compatible signal is to use a setup as shown in Fig. 11, where the loudspeakers are rearranged to match a 5.1-channel configuration. In this case the ±30° virtual stage is extended to a ±110° virtual stage surrounding the listener.

3.4 Wavefield Synthesis Playback System

First, signals are generated similarly to the setup illustrated in Fig. 14(b). Then M virtual sources are defined in the wavefield synthesis system. The lateral independent sounds y_1 and y_M are emitted as plane waves or sources in the far field, as illustrated in Fig. 16 for $M = 8$. For each of the other signals a virtual source is defined with a location as desired. In the example shown in Fig. 16 the distance is varied for the different sources, and some of the sources are defined to be in the front of the sound-emitting array, that is, the virtual sound stage can be defined with an individual distance for each defined direction.

3.5 Modifying the Decomposed Audio Signals

3.5.1 Controlling the Width of the Sound Stage

By modifying the estimated scale factors such as $A(i, k)$ one can control the width of the virtual sound stage. By linear scaling with a factor larger than 1 the instruments being part of the sound stage are moved more to the side. The opposite can be achieved by scaling with a factor smaller than 1. Alternatively one can modify the amplitude panning law, Eq. (20), for computing the angle of the direct localized sound.

3.5.2 Modifying the Ratio between Direct Localized Sound and Independent Sound

To control the amount of ambience one can scale the independent lateral sound signals $\hat{H}'_1$ and $\hat{H}'_2$ for getting more or less ambience. Similarly, the localized direct sound can be modified in strength by means of scaling the $\hat{S}'$ signals.

4 SUBJECTIVE EVALUATION

Informal listening indicated that the proposed scheme offers a benefit over conventional stereo playback, especially for off-the-sweet-spot listening. The goal for the subjective test was not to gain specific psychophysically interesting data, but to get some evidence that the proposed scheme is preferred by listeners compared to conventional stereo playback.

4.1 Subjects and Playback Setup

Eight subjects participated in the tests. Six of these subjects had already participated in the past in subjective tests for audio quality evaluation. The subjects' ages ranged between 26 and 37 years and they reported normal hearing. The test was carried out in a sound-insulated room mimicking a typical living room. For audio playback a laptop computer (Apple PowerBook G4) was used with an external digital-to-analog converter (MOTU 896) connected directly to eight active loudspeakers (Genelec 1029A).

The loudspeakers were arranged in front of the subject, as illustrated in Fig. 17. All loudspeakers were always switched on and the subjects had no explicit knowledge from which loudspeakers the sound was emitted.

4.2 Stimuli

Eleven different stereo music clips were selected. The clips were obtained from CDs and ranged in length between 10 and 15 seconds. In order to show also that the

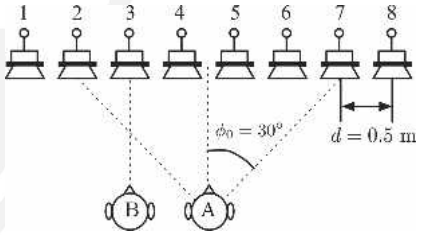

Fig. 17. Subjective test carried out with loudspeaker setup as shown.

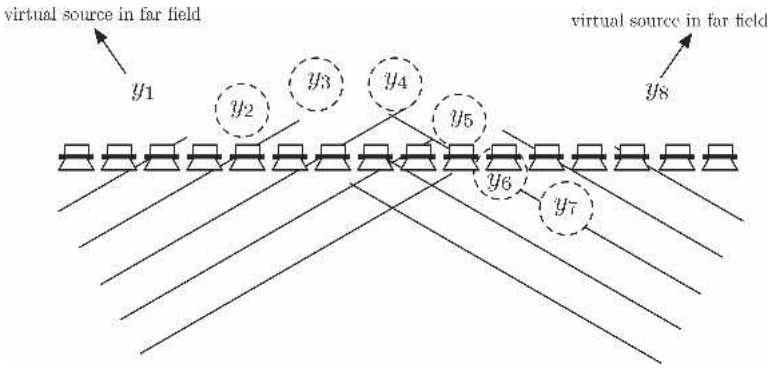

Fig. 16. Each signal corresponding to the front sound stage is defined as a virtual source. Independent lateral sound is emitted as plane waves (virtual sources in far field).

proposed scheme performs well for audio material that is encoded with a typical audio coder, we coded the clips with MP3 [27] at 192 kb/s.[3]

Three of the clips were used as training items and the other eight clips were used in the test. The clips contained classical, jazz, rock/pop, and latin music.

Each of the clips was processed to generate two types of eight-channel stimuli. One stimulus type, called standard stereo (SS), emits the stereo signal out of loudspeakers 2 and 7, mimicking a standard stereo configuration. The second type of stimulus, called front array (FA), is processed according to Eq. (27) such that the virtual sound stage is reproduced with loudspeakers 2 to 7 and plane waves with angles of ±40° are reproduced with loudspeakers 1 to 8.

4.3 Test Method

Each subject conducted the test twice, one directly after the other, but with a different listening position. The two listening positions are indicated in Fig. 17 as A and B. In position A the subject was located centered such that loudspeakers 2 and 7 formed a standard stereo listening setup with $\phi_0 = 30°$. In position B the subject was more to the side, that is, at the lateral position of loudspeaker 3. It was indicated to the listeners that the virtual stage ranges from loudspeakers 2 to 7.

The subjects were asked to grade different specific properties and the overall audio quality of the processed clips. For each corresponding stimulus pair SS and FA, one stimulus had to be graded relative to the other (reference), where either SS or FA had a 50% chance of being declared the reference. Randomization in terms of declaring SS or FA as the reference and the ordering of the clips were carried out for each subject individually. The three different grading tasks of the test are summarized in Table 1. Task 1 assesses the quality of the virtual sound stage. Task 2 evaluates distortions introduced by the processing that are not related to the spatial aspect of sound. Task 3 assesses the overall audio quality. Note that for all three tasks the ITU-R 7-grade comparison scale [28], shown in Table 2, was used.

Before the test the subject was given written instructions. Then a short training session with three clips was carried out, followed by the two tests (listener in positions A and B) containing the eight clips listed in Table 3.

Fig. 18 illustrates the graphical user interface that was used for the test. The subject was presented with (frozen) sliders for the reference and for the corresponding other stimulus. With the "Play" buttons the subjects could listen to either the reference or the corresponding other stimulus.

Table 1. Tasks and scales of the subjective test.

Task	Scale
1 Image quality	ITU-R 7-grade comparison
2 Audio quality (ignoring image quality)	ITU-R 7-grade comparison
3 Overall quality	ITU-R 7-grade comparison

Table 2. ITU-R 7-grade comparison scale for comparing item A to reference item R.

3	A much better than R
2	A better than R
1	A slightly better than R
0	A same as R
−1	A slightly worse than R
−2	A worse than R
−3	A much worse than R

Table 3. Eight music clips used for test.

	Title	Genre
a	I will survive	Pop/rock
b	Blue eyes	Pop/rock
c	Bovio	Classical
d	Slavonic dance	Classical
e	Piensa	Latin/jazz
f	He perdido contigo	Latin/jazz
g	Scoppin	Jazz
h	Y tal vez	Latin/jazz

[3]The MP3 encoder integrated with Apple QuickTime 6 was used.

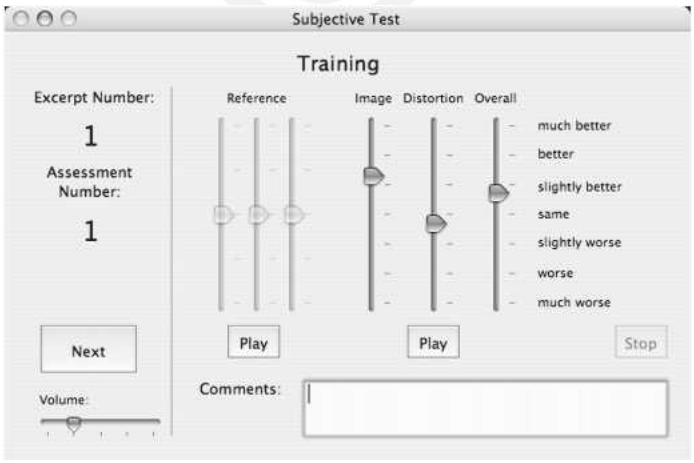

Fig. 18. Graphical user interface used for test. Left three (frozen) sliders correspond to reference, right three sliders to other stimulus.

The subject could switch between the stimuli at any time while the sound instantly faded from one type of stimulus to the other. Informal listening indicated that such instant switching greatly facilitates the comparison of the spatial attributes of the stimuli.

The duration of the test sessions (positions A and B) varied between listeners because of the freedom to repeat the stimuli as often as needed. Typically the test duration was between 30 and 50 minutes.

4.4 Results

Fig. 19 shows the results of the tests with the subjects located at listening position A (sweet spot). The letters indicated on the *x* axis correspond to the specific clip labels given in Table 3. The grading scale on the *y* axis corresponds to the comparison scale given in Table 1, where positive gradings indicate that FA (proposed scheme) is better than SS (standard stereo). Fig. 18(a) shows the gradings and 95% confidence intervals for each clip, averaged for all subjects. Fig. 19(b) shows the results averaged for all clips and subjects. The gradings are shown for the attributes image quality, distortion, and overall quality. The image quality indicates that the subjects preferred the virtual sound stage of the proposed scheme. The distortion, in most cases close to zero, indicates that the proposed scheme does only, if at all, introduce relatively few distortions. The subjects preferred the proposed scheme, as is implied by the positive overall quality gradings.

The results for the test with the subjects off the sweet spot, position B, are shown in Fig. 20. The conclusions here are similar, only that the degree of improvement compared to stereo is significantly larger, as expected, since the virtual stage for stereo with a listener not in the sweet spot is degraded.

4.5 Discussion

The goal of the subjective test with the subject in position A (sweet spot) was to assess whether the proposed playback of stereo signals over multiple loudspeakers matches the quality of stereo playback when the listener is located in the sweet spot. We were positively surprised that on average the subjects preferred the proposed playback scheme over stereo in the sweet spot.

To show the benefit of listening when the listener is not located in the sweet spot, the subjective test with the listening position B was carried out. As expected, the relative performance of the proposed scheme is better for off-the-sweet-spot listening since it maintains the extent of the virtual sound stage.

Only informally tested, the proposed playback scheme results in a virtual sound stage that does hardly depend on the listener's position. The listener can move and the stage and instruments remain at their (absolute) spatial positions. This is in contrast to wavefield synthesis systems when they are used for stereo playback. Usually the left and right stereo signals are emitted as plane waves in such

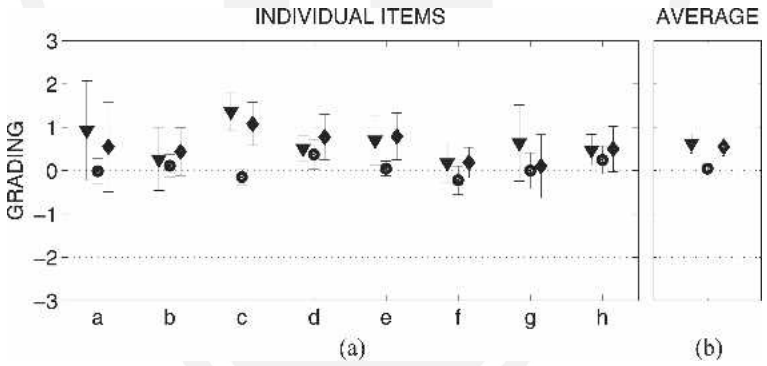

Fig. 19. Subjective test results for subjects in position A (sweet spot). (a) Grading and 95% confidence intervals for each clip averaged over all subjects. (b) Overall average gradings. ▲—image quality; ●—distortion; ♦—overall quality. Positive gradings indicate that FA is better than SS.

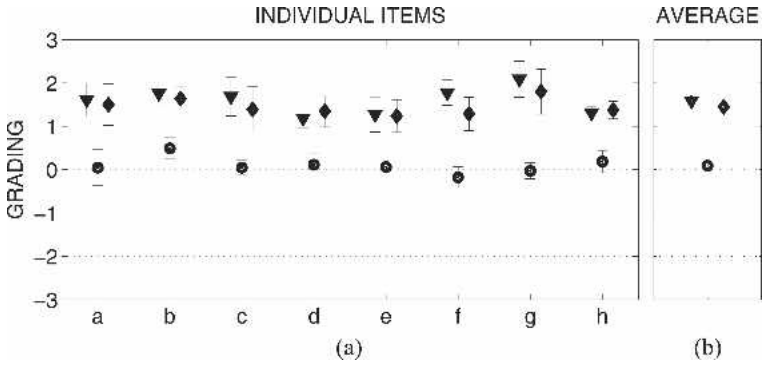

Fig. 20. Subjective test results for subjects in position B (off the sweet spot). (a) Grading and 95% confidence intervals for each clip averaged over all subjects. (b) Overall average gradings. ▲—image quality; ●—distortion; ♦—overall quality. Positive gradings indicate that FA is better than SS.

systems, resulting in virtual stage moving as the listener moves. The proposed scheme gives the flexibility of playing back stereo signals over wavefield synthesis systems where the distance of each virtual source (signal) can be freely determined. Previously a large sweet spot could only be obtained by emitting plane waves, that is, by mimicking a sound stage infinitely far away.

Informal listening revealed that the proposed algorithm, when applied for generating 5.1-channel surround signals (as described in Section 3.3), generates ambience as normally expected from discrete surround. Localization is very sharp, that is, channel separation between center and left/right is strong. On the other hand, due to the fact that localized sound is played from the front loudspeakers and lateral independent sound from the rear loudspeakers, some distortions are from time to time noticeable in the front audio channels. These distortions are masked for the scenario tested in the paper, since direct sound and independent lateral sound are more mixed (all emitted from front loudspeakers). To circumvent this problem, lateral independent sound could also be given to the front channels.

5 CONCLUSIONS

A perceptually motivated spatial decomposition for two-channel stereo signals was proposed. In a number of subbands and as a function of time, lateral independent sound and localized sound and its specific angle (or level difference) are estimated. Given an assumed signal model, the least-squares estimates of these signals are computed.

Furthermore it was described how the decomposed stereo signals can be played back over multiple loudspeakers, loudspeaker arrays, and wavefield synthesis systems. A subjective test indicates that when the decomposed stereo signals are played back over a loudspeaker array, higher quality (spatial aspect, overall quality) is achieved, even when compared to standard stereo with a listener in the sweet spot. When the listener is not in the sweet spot, the proposed technique results in more improvement, as expected.

6 REFERENCES

[1] ITU-R BS.775, "Multi-Channel Stereophonic Sound System with or without Accompanying Picture," International Telecommunications Union, Geneva, Switzerland (1993); http://www.itu.org.

[2] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.,* vol. 50, pp. 651–666 (2002. Sept.).

[3] S. Komiyama, "Subjective Evaluation of Angular Displacement between Picture and Sound Directions for HDTV Sound Systems," *J. Audio Eng. Soc.,* vol. 37, pp. 210–214 (1989 Apr.).

[4] G. Theile, "On the Performance of Two-Channel and Multi-Channel Stereophony," presented at the 88th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 38, p. 379 (1990 May), preprint 2887.

[5] J. Steinberg and W. Snow, "Auditory Perspectives–Physical Factors," *Elect. Eng.,* vol. 53, pp. 12–15 (1934 Jan.).

[6] P. W. Klipsch, "Stereophonic Sound with Two Tracks, Three Channels by Means of a Phantom Circuit (2PH3)," *J. Audio Eng. Soc.,* vol. 6, pp. 118–123 (1958 Apr.).

[7] J. Eargle, Ed., *Stereophonic Techniques* (Audio Engineering Society, New York, 1986).

[8] M. A. Gerzon, "Optimal Reproduction Matrices for Multispeaker Stereo," presented at the 91st Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 39, p. 1006 (1991 Dec.), preprint 3180.

[9] J. Bauck, "Conversion of Two-Channel Stereo for Presentation by Three Frontal Loudspeakers," presented at the 109th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 48, p. 1110 (2000 Nov.), preprint 5239.

[10] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic Control by Wave Field Synthesis," *J. Acoust. Soc. Am.,* vol. 93, pp. 2764–2778 (1993 May).

[11] E. N. G. Verheijen, "Sound Reproduction by Wave Field Synthesis," Ph.D. thesis, Delft University of Technology Delft, The Netherlands (1997).

[12] V. Pulkki, "Directional Audio Coding in Spatial Sound Reproduction and Stereo Upmixing," in *Proc. AES 28th Int. Conf.* (2006 June).

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

[14] R. Dressler, "Dolby Surround Prologic II Decoder—Principles of Operation," Tech. Rep., Dolby Laboratories (2000); www.dolby.com/tech/.

[15] C. Avendano and J. M. Jot, "Ambience Extraction and Synthesis from Stereo Signals for Multi-Channel Audio Up-Mix," in *Proc. ICASSP,* vol. 2 (Orlando, FL, 2002 May), pp. 1957–1960.

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

[17] 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).

[18] C. Faller, "Parametric Coding of Spatial Audio," Ph.D. thesis no. 3062, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland (2004 July); http://library.epfl.ch/theses/?nr=3062.

[19] M. Barron and A. H. Marshall, "Spatial Impression Due to Early Lateral Reflections in Concert Halls: The Derivation of a Physical Measure," *J. Sound Vib.,* vol. 77, pp. 211–232 (1981).

[20] M. Morimoto and Z. Maekawa, "Auditory Spaciousness and Envelopment," in *Proc. 13th Int. Congr. on Acoustics,* vol. 2 (Belgrade, Yugoslavia, 1989), pp. 215–218.

[21] J. S. Bradley and G. A. Soulodre, "Objective Measures of Listener Envelopment," *J. Acoust. Soc. Am.,* vol. 98, pp. 2590–2597 (1995).

[22] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models* (Springer, New York, 1999).

[23] B. R. Glasberg and B. C. J. Moore, "Derivation of Auditory Filter Shapes from Notched-Noise Data," *Hear. Res.,* vol. 47, pp. 103–138 (1990).

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

[25] S. Haykin, *Adaptive Filter Theory,* 3rd ed. (Prentice-Hall, Englewood Cliffs, NJ, 1996).

[26] B. B. Bauer, "Phasor Analysis of Some Stereophonic Phenomena," *J. Acoust. Soc. Am.,* vol. 33, pp. 1536–1539 (1961 Nov.).

[27] 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, pp. 780–792 (1994 Oct.).

[28] ITU-R BS.562.3, "Subjective Assessment of Sound Quality," International Telecommunications Union, Geneva, Switzerland (1990); http://www.itu.org.

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.

CORRECTIONS

CORRECTIONS TO "MULTIPLE-LOUDSPEAKER PLAYBACK OF STEREO SIGNALS"

In the above paper[1], on p. 1054, Eq. (4) and Eq. (5) should have appeared as follows:

$$\begin{aligned} P_{X_1} &= P_S + P_N \\ P_{X_2} &= A^2 P_S + P_N \\ \Phi &= \frac{aP_S}{\sqrt{P_{X_1} P_{X_2}}.} \end{aligned} \qquad (4)$$

These equations solved for A, P_S, and P_N yield

$$\begin{aligned} A &= \frac{B}{2C} \\ P_S &= \frac{2C^2}{B} \\ P_N &= P_{X_1} - \frac{2C^2}{B} \end{aligned} \qquad (5)$$

*Manuscript received 2011 June 15.

[1]C. Faller, *J. Audio Eng. Soc.,* vol. 54, pp. 1051–1064 (2006 Nov.).