Two-to-Five Channel Sound Processing*

R. IRWAN AND RONALD M. AARTS, *AES Fellow*

Philips Research Laboratories, 5656 AA Eindhoven, The Netherlands

While stereo music reproduction was a dramatic advance over mono, recently a transition to multichannel audio has created a more involving experience for listeners. An algorithm to convert stereo to five-channel sound reproduction is presented. An effective sound distribution to the surround channels is achieved by using a cross-correlation technique, and a robust stereo image is obtained using principal component analysis. Informal listening tests comparing this scheme with other methods revealed that the proposed algorithm is preferred for both on and off the reference listening position (sweet spot).

0 INTRODUCTION

Since the introduction of the digital versatile disk (DVD) and the super audio CD (SACD), a revival of multichannel audio has appeared in sound systems for consumer use today. It is, however, desirable to maintain compatibility with the existing two-channel stereo recordings and/or broadcasting. Therefore the conversion of two-channel stereo to the multichannel format has been studied extensively over the decades, and a considerable number of publications exist [1]–[8]. Among these, Gerzon and Barton's is particularly notable, in which many schemes have been proposed (see [6] and references therein).

Although many authors have introduced multichannel sound systems with a large number of channels, we restrict ourselves to a home cinema setup for which it has been shown that five channels is sufficient for creating ambience effects [9]. Hence in this paper we focus on signal format conversion from two-channel stereo to the five-channel (two-to-five) sound processing algorithm.

The desired setup is shown in Fig. 1, in which the channels are labeled L (left), C (center), R (right), S_L (left surround), and S_R (right surround) according to convention. This setting is adopted from the ITU multichannel configuration [10], with three loudspeakers placed in front of the listener, and the other two at the back.

The front channels are used to provide a high degree of directional accuracy over a wide listening area for front-stage sounds, particularly dialogues, and the rear channels produce diffuse surround sounds, providing ambience and environment affects. An additional loudspeaker (subwoofer) may be used to augment bass reproduction, which is often called 5.1 system, with .1 referring to the low-frequency enhancement (LFE) channel. In this paper, however, we do not use a subwoofer, since the system can easily be extended when necessary without affecting the algorithm.

The algorithm presented in this paper offers two improvements above the existing two-to-five channel sound systems. First a problem associated with channel crosstalk is reduced, and therefore sound localization is better. Listening tests have confirmed that good sound localization without the need to listen at the sweet spot gives more space to the listener to enjoy the program offered rather than restricting the listener to the sweet spot.

Second a better sound distribution to the surround channels is achieved by using a cross-correlation technique. Surround channels are crucial in creating the ambience effects, which is one of the main goals of multichannel audio. At the same time, the energy preservation criterion is an important constraint that has been used to design multichannel matrices [7]. The main reason for this is to maintain backward and forward stereo compatibility. Furthermore, the preservation criterion ensures that all signals present in the two-channel transmitted signals are produced at a correct power level, so that the balance between the different signal sounds in the recording is not disturbed.

This paper is organized as follows. In Section 1 a technique for deriving a robust center channel is outlined. A three-dimensional mapping to derive the surround channels is discussed in Section 2. The rest of the paper will discuss subjective assessments of some listening tests that have been performed in order to com-

*Partly presented at the AES 19th International Conference, Germany, 2001 June 21–24; revised 2001 October 5 and 2002 May 31.

pare the present method with other existing two-to-five channel sound systems. Concluding remarks are presented in Section 4.

1 CENTER LOUDSPEAKER

We consider the three-channel approach first. It is known that the sound quality of stereo sound reproduction can be improved by adding an additional loudspeaker between each adjacent pair of loudspeakers. For example, as proposed by Klipsch [1], an additional center loudspeaker C can be fed with the sum signal $\sqrt{2}(x_L + x_R)$, where x_L and x_R represent signals from left and right, respectively. The $\sqrt{2}$ factor was introduced to preserve the total energy from the three loudspeakers, assuming incoherent additive for left, center, and right sounds recorded by two widely spaced microphones. A major drawback of this approach is that crosstalk with the left and right channels is inevitable, and therefore it will narrow the stereo image considerably.

We propose an algorithm to derive the center channel without these drawbacks, using principal component analysis (PCA) [11], which produces two vectors indicating the direction of both the dominant signal y and the remaining signal q, as shown in Fig. 2 by dashed lines. Note that these two directions are perpendicular to each other, creating a new coordinate system. These two signals are then used as basis signals in the matrix decoding, a point that is different from other existing two-to-five channel sound systems.

To derive the center channel's gain using the direction of a stereo image, we process the audio signal coming from a CD (sampling frequency $F_s = 44.1$ kHz) on a sample basis. Each sample of a stereo pair at a time index k

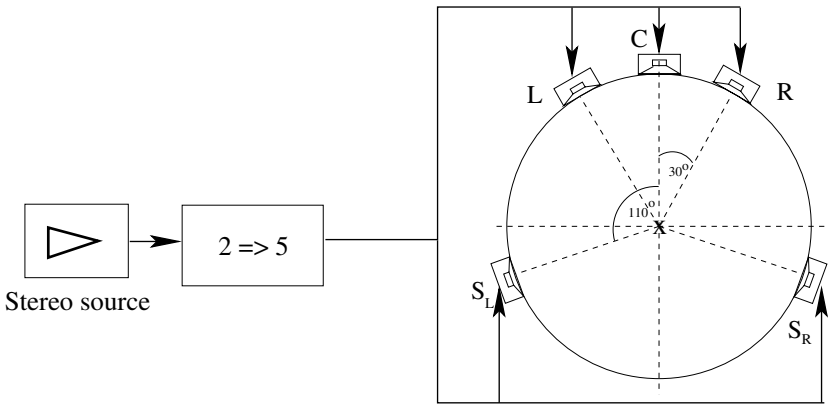


Fig. 1. ITU reference configuration [9]. ×—reference listening position (sweet spot). Left and right channels are placed at angles ±30° from C; two surround channels are placed at angles ±110° from C.

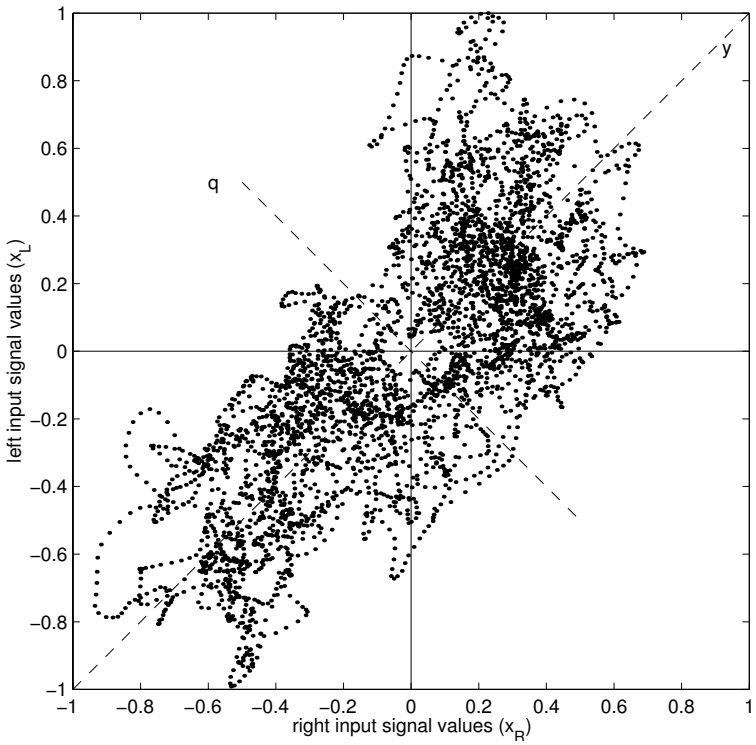


Fig. 2. Lissajous plot of stereo signal recorded from the fragment "The Great Pretender": by Freddy Mercury. Dashed lines represent new coordinate system based on both dominant signal y and remaining signal q, forming the direction of a stereo image α.

can be expressed as

$$\boldsymbol{x}(k) = \begin{bmatrix} x_{\mathrm{L}}(k) & x_{\mathrm{R}}(k) \end{bmatrix}^{\mathrm{T}} \tag{1}$$

where k is an integer.

Let us now define $y(k)$ to be a linear combination of the input signals,

$$y(k) = \boldsymbol{w}^{\mathrm{T}}(k)\,\boldsymbol{x}(k) \tag{2}$$

where

$$\boldsymbol{w}(k) = \begin{bmatrix} w_{\mathrm{L}}(k) & w_{\mathrm{R}}(k) \end{bmatrix}^{\mathrm{T}} \tag{3}$$

is a weight vector corresponding to the left and the right channels, respectively.

In order to find the optimum weighting vectors, we maximize the energy of Eq. (2) with respect to $\boldsymbol{w}$, that is,

$$\frac{\partial E\left[y(k)^2\right]}{\partial \boldsymbol{w}} = 0 \tag{4}$$

where E denotes the expected value. Using a method presented by Haykin [12], we obtain by means of the steepest descent method

$$\begin{aligned} \boldsymbol{w}(k) &= \boldsymbol{w}(k-1) + \frac{1}{2}\,\mu\,\frac{\partial E\left[y(k-1)^2\right]}{\partial \mathrm{w}} \\ &= \boldsymbol{w}(k-1) + \mu \mathrm{E}\left[y(k-1)\,\boldsymbol{x}(k-1)\right] \end{aligned} \tag{5}$$

where μ is a step size. Since $E[y(k-1)]$ and $E[\boldsymbol{x}(k-1)]$ are both scalars, Eq. (5) can be estimated as

$$\boldsymbol{w}(k) = \boldsymbol{w}(k-1) + \mu y(k-1)\,\boldsymbol{x}(k-1)\,. \tag{6}$$

Normalizing Eq. (6) such that $\|\boldsymbol{w}(k)\|_2 = 1$, gives the desired sample estimate of $\boldsymbol{w}$,

$$\boldsymbol{w}(k) = \frac{\boldsymbol{w}(k-1) + \mu y(k-1)\,\boldsymbol{x}(k-1)}{\sqrt{\sum_{\mathrm{L,R}}\left[\boldsymbol{w}(k-1) + \mu y(k-1)\,\boldsymbol{x}(k-1)\right]^2}}\,. \tag{7}$$

Assuming that the step size μ is small, Eq. (7) can be expanded as a power series in μ, yielding

$$\begin{aligned} \boldsymbol{w}(k) &= \boldsymbol{w}(k-1) + \mu y(k-1) \\ &\quad \times \left[\boldsymbol{x}(k-1) - \boldsymbol{w}(k-1)\,y(k-1)\right] \end{aligned} \tag{8}$$

which is a least-mean-square (LMS) algorithm with $y(k-1)$ as input. Writing out Eq. (8) for left and right channels, respectively, produces

$$\begin{aligned} w_{\mathrm{L}}(k) &= w_{\mathrm{L}}(k-1) + \mu y(k-1) \\ &\quad \times \left[x_{\mathrm{L}}(k-1) - w_{\mathrm{L}}(k-1)\,y(k-1)\right] \\ w_{\mathrm{R}}(k) &= w_{\mathrm{R}}(k-1) + \mu y(k-1) \\ &\quad \times \left[x_{\mathrm{R}}(k-1) - w_{\mathrm{R}}(k-1)\,y(k-1)\right]. \end{aligned} \tag{9}$$

Karhunen [13] has shown that the algorithm is stable if and only if

$$0 < \mu \boldsymbol{x}^{\mathrm{T}}(k)\,\boldsymbol{x}(k) < 2 \tag{10}$$

or the step size must satisfy the following constraint:

$$0 < \mu < \frac{2}{\boldsymbol{x}^{\mathrm{T}}(k)\,\boldsymbol{x}(k)} \tag{11}$$

and therefore it is input signal dependent.

The direction of a stereo image in terms of an angle, in radians, can easily be computed as

$$\alpha(k) = \arctan\left[\frac{\boldsymbol{w}_{\mathrm{L}}(k)}{\boldsymbol{w}_{\mathrm{R}}(k)}\right]. \tag{12}$$

Fig. 3 shows the values of α when it is calculated for a CD stereo music recording. Recalling Fig. 2 with the left channel corresponding to $\alpha = \pi/2$ and the right channel to $\alpha = 0$, we can see that α fluctuates around $\pi/4$, creating a phantom source almost equidistant between the left and right channels.

Fig. 4 shows the same response of the angle α, but now measured from a DVD movie fragment, where abrupt changes from one channel to the other are present. We intentionally take a shorter fragment in order to demonstrate that the algorithm is still able to detect abrupt changes in localizations within a short period of time.

Now we can represent a pair of stereo signals using a vector given by Eq. (3). This is a vector of unit length having the right channel gain in the horizontal axis, and the left channel gain in the vertical axis, as shown in Fig. 5(a). To map this stereo vector onto a three-channel vector, we double the angle α and produce a new mapping, as depicted in Fig. 5(b). We can then find the projections of the vector onto the LR axis and the C axis using sine and cosine rules,

$$\begin{aligned} c_{\mathrm{LR}} &= w_{\mathrm{R}}^2 - w_{\mathrm{L}}^2 \\ c_{\mathrm{C}} &= 2 w_{\mathrm{L}}\, w_{\mathrm{R}}\,. \end{aligned} \tag{13}$$

It should be pointed out that the transformation illustrated in Fig. 5 works only for nonnegative α. This is because for negative α, multiplication by a factor of 2 results in the vector being in a lower quadrant, and therefore no gain can be derived for the center channel. To overcome this problem, extra information should be used, which is described in the next section.

2 SURROUND LOUDSPEAKERS

The surround channels are generally used to create ambience effects for music. For applications in the film industry the surround channels are used for sound effects. A common technique for ambience reconstruction is the use of delayed front channel information for the surround channels. Dolby Pro Logic, for instance, has delayed the surround sounds so as to arrive at the listeners' ears at least 10 ms later than the front sounds [7].

Environmental and ambience effects can be computed by considering left and right channel variations ($x_{\mathrm{L}} - x_{\mathrm{R}}$)

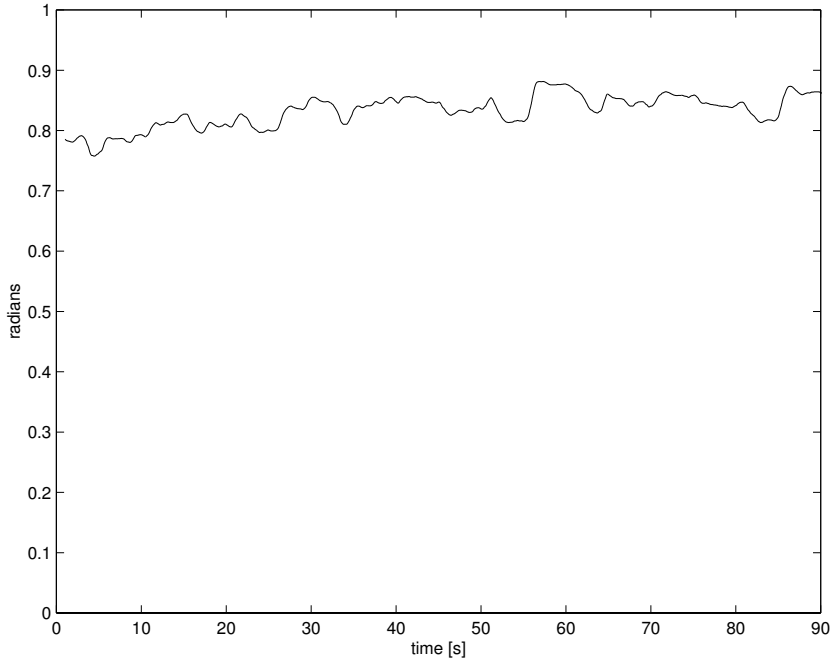


Fig. 3. Typical example of fluctuation of α, computed from a CD stereo music fragment with a stable phantom source.

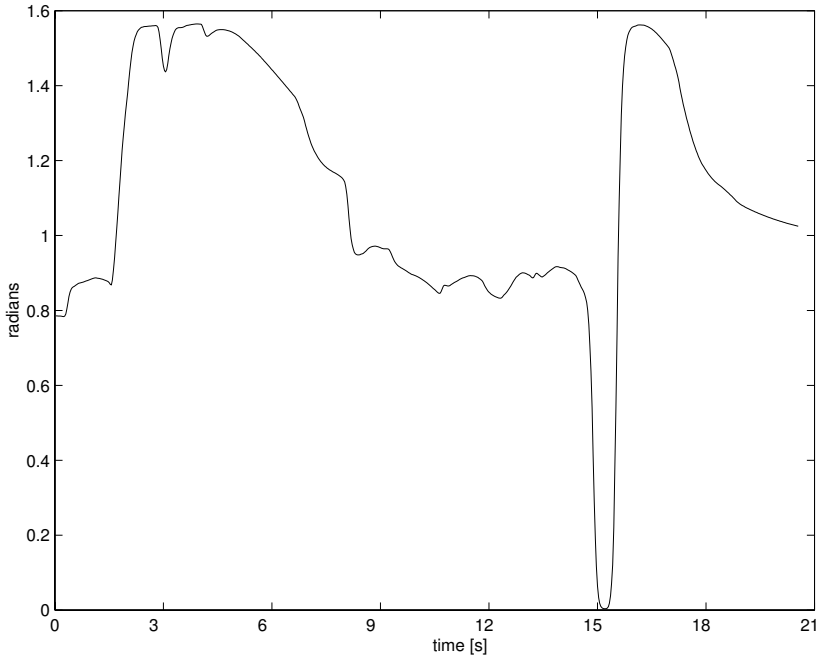


Fig. 4. Fluctuation of the direction α computed from a DVD fragment containing sounds of a car passing with high speed from one channel to the other. Total duration of fragment about 20 seconds.

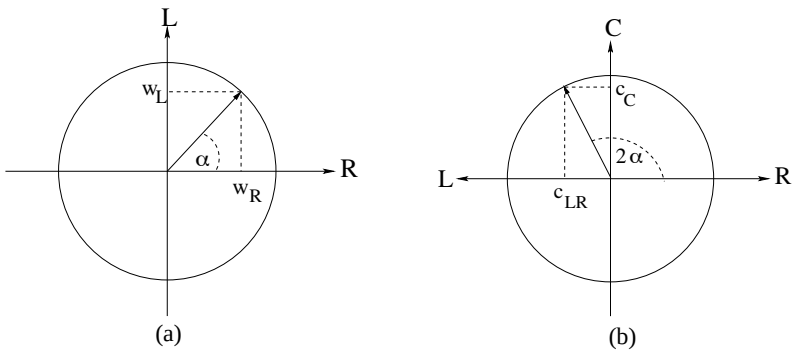


Fig. 5. (a) Direction vector plots of stereo signals. (b) Corresponding three-channel representation by doubling the angle α.

in the original signals. This variation is usually referred to as the antiphase components, the amount of which can be represented by the remaining signal q (see Fig. 2). However, it can be expected that when the amount of the dominant signal equals or almost equals that of the remaining signal, an ambiguity appears since there is no way of determining the direction vector uniquely. In this situation the distribution in Fig. 2 is no longer an ellipse but has a circlelike form ($|y| \approx |q|$), as illustrated in Fig. 6, causing α to be not well defined.

Obviously extra information is necessary when dealing with this sort of ambiguity. In this paper we propose to use a known technique to measure the amount of antiphase components, namely, the correlation coefficient, which is given in any text book on statistics as

$$\rho = \frac{\sum(x_L - \bar{x}_L)(x_R - \bar{x}_R)}{\sqrt{\sum(x_L - \bar{x}_L)^2 \sum(x_R - \bar{x}_R)^2}} \tag{14}$$

where $\bar{x}_L$ and $\bar{x}_R$ are the mean values of x_L and x_R, respectively.

Aarts et al. [14] have shown that Eq. (14) can be computed recursively by using only a few arithmetic operations,

$$\hat{\rho}(k) = \hat{\rho}(k-1) + \gamma\left\{2x_L(k)x_R(k) - \left[x_L(k)^2 + x_R(k)^2\right]\hat{\rho}(k-1)\right\} \tag{15}$$

where γ is the step size determining the time constant, and the caret (^) is used to denote that it is an estimate of the true ρ. A summary of the mathematical derivations of Eq. (15) can be found in Appendix 3.

To give some ideas how this tracking algorithm works, we present two examples of measurements using Eq. (15), which are shown in Figs. 7 and 8. The measurements are performed within a time frame of 50 seconds, with the step size γ set to 10^{-3} at $F_s = 44.1$ kHz. Fig. 7 is a typical example of a modest stereo for which the correlation varies around 0.70, and it is thus neither too strong (mono sound) nor too weak (diffuse sound). On the other hand, Fig. 8 shows an example of an uncorrelated stereo signal with many antiphase components, for which α is difficult to detect (see Fig. 6).

It is worth mentioning here that there are three other variants of Eq. (15) which are evenly robust. For further information the reader is referred to [14].

Since

$$-1 \le \rho \le 1 \tag{16}$$

it is possible that the antiphase components exceed the dominant signal($|y| < |q|$). In this case we treat the input signals as uncorrelated, and therefore

$$\rho_0 = \begin{cases} \rho, & 0 \le \rho \le 1 \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

It can be shown (see Appendix 1) that a relationship exists between this cross-correlation method and PCA described in the previous section.

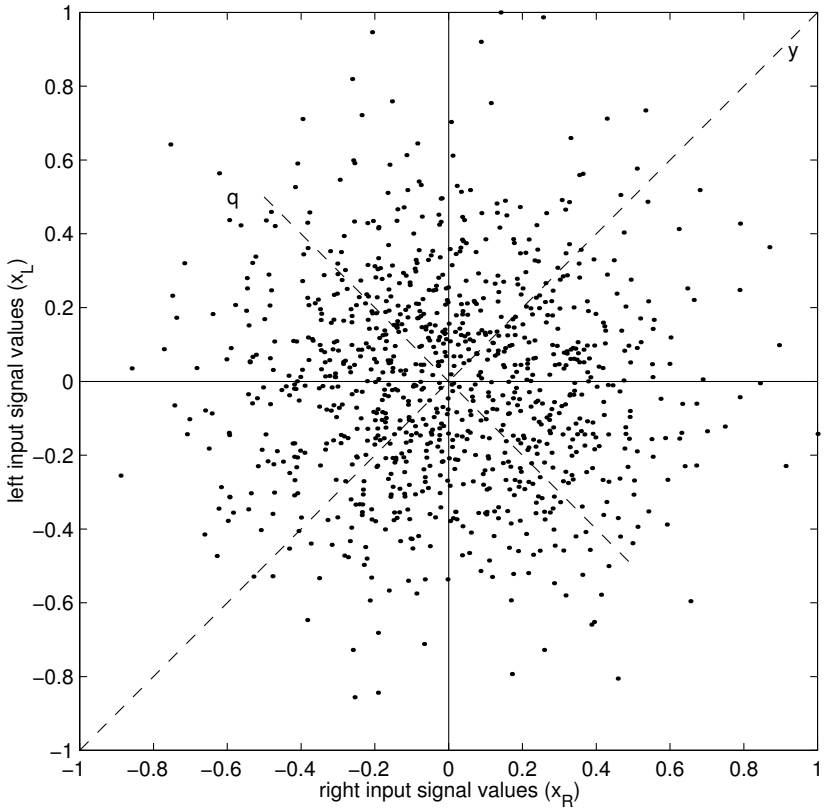


Fig. 6. Lissajous plot of the first 23-second stereo signal recorded from the fragment “Holiday” by Madonna, where the amount of a dominant signal is almost equal to that of a remaining signal, forming a circle-like distribution.

2.1 Three-Dimensional Mapping

To avoid ambiguity when the amount of the dominant signal approaches that of the remaining signals, the use of both the direction of the stereo image and the correlation coefficient is necessary. The latter is included in the mapping (see Fig. 5) by, for example, placing the surround channels in the vertical plane, as shown in Fig. 9.

The angle β can be defined to represent the actual surround information by means of the adaptive correlation coefficient, for example, by using the expression

$$\beta(k) = \arcsin\left[1 - \rho_0(k)\right] \tag{18}$$

and hence,

$$0 \le \beta(k) \le \frac{\pi}{2}\,. \tag{19}$$

Thus as the amount of the remaining signal increases (input signals become weakly correlated), the angle β also increases, which reduces the total distribution to the front channels. On the other hand, when the input signals are strongly correlated (quasi mono), β approaches zero, producing a larger contribution to the front channels. This principle satisfies the energy preservation criterion, which is discussed in more detail in Section 2.2.

Since the direction vector on the horizontal plane is now lifted by an angle β, recalculation of the projections is necessary. Using straightforward trigonometry and keeping in mind that the vector is of unit length, we obtain

$$\begin{aligned} c'_{\mathrm{LR}} &= c_{\mathrm{LR}} \cos\beta \\ c'_{\mathrm{C}} &= c_{\mathrm{C}} \cos\beta \\ c_{\mathrm{S}} &= \sin\beta\,. \end{aligned} \tag{20}$$

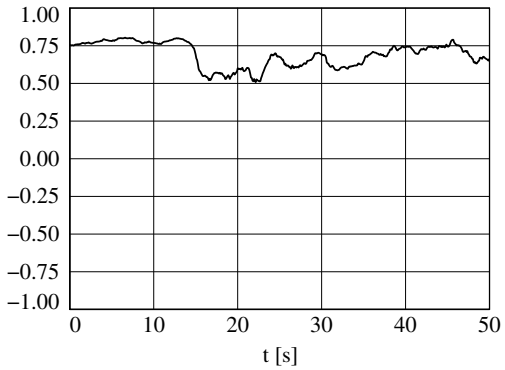


Fig. 7. Tracked cross-correlation coefficients obtained from the fragment "The Great Pretender" by Freddy Mercury. See Fig. 2 for the stereo image, which is a typical example of a modest stereo recording.

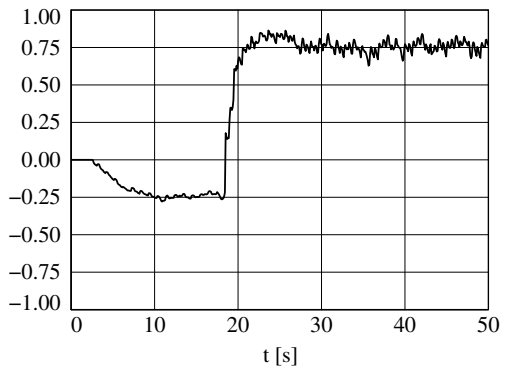


Fig. 8. Tracked cross-correlation coefficients obtained from the fragment "Holiday" by Madonna. See Fig. 6 for the corresponding Lissajous plot.

2.2 Matrixing

The system described so far reproduces four channel signals as L, C, R, and S from two input signals. Therefore we have a 4 × 2 reproduction matrix.

We now discuss the objective requirement on the energy preservation as emphasized in Section 2.1. A matrix preserves energy if and only if its columns are of unit length, and the columns are pairwise orthogonal. Since the product of any two orthogonal matrices is also orthogonal, back and forward compatibility between stereo and multichannel can also be achieved.

Following this energy criterion, we design the matrix as follows:

$$\begin{bmatrix} u_{\mathrm{L}}(k) \\ u_{\mathrm{R}}(k) \\ u_{\mathrm{C}}(k) \\ u_{\mathrm{S}}(k) \end{bmatrix} = \begin{bmatrix} c_{\mathrm{L}}(k) & g w_{\mathrm{L}}(k) \\ c_{\mathrm{R}}(k) & g w_{\mathrm{R}}(k) \\ c_{\mathrm{C}}(k) & 0 \\ 0 & c_{\mathrm{S}}(k) \end{bmatrix} \begin{bmatrix} y(k) \\ q(k) \end{bmatrix}. \tag{21}$$

The components of the left-hand side of Eq. (21) denote the signals for the left, right, and center loudspeakers, and u_{S} denotes the mono surround signal. The basis signals are obtained by rotating the coordinate system of x_{L} and x_{R},

$$\begin{aligned} y(k) &= w_{\mathrm{L}}(k)\, x_{\mathrm{L}}(k) + w_{\mathrm{R}}(k)\, x_{\mathrm{R}}(k) \\ q(k) &= w_{\mathrm{R}}(k)\, x_{\mathrm{L}}(k) - w_{\mathrm{L}}(k)\, x_{\mathrm{R}}(k) \end{aligned} \tag{22}$$

and

$$\begin{aligned} c_{\mathrm{L}} &= \begin{cases} -c_{\mathrm{LR}}\,, & c_{\mathrm{LR}} < 0 \\ 0\,, & \text{otherwise} \end{cases} \\ c_{\mathrm{R}} &= \begin{cases} c_{\mathrm{LR}}\,, & c_{\mathrm{LR}} \ge 0 \\ 0\,, & \text{otherwise} \end{cases} \end{aligned} \tag{23}$$

and g is a gain to control the energy preservation.

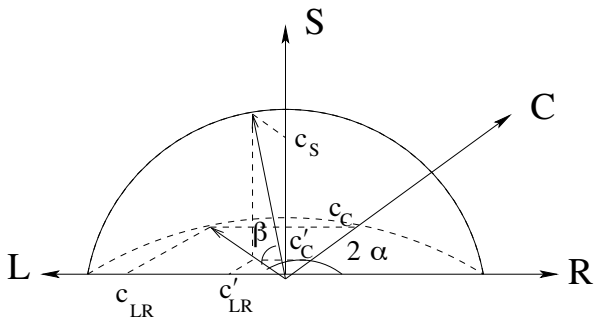


Fig. 9. Three-dimensional mapping showing front (horizontal plane) and surround channels (vertical plane). Parameter β determines the level of surround information with respect to the front channel sounds.

Since c_L and c_R can only produce one value depending on the condition in Eq. (23), the length of the first column of the matrix given in Eq. (21) is equal to $c_{LR}^2 + c_C^2$, which is unity. The second column of Eq. (21) contains mainly a projection of the vector onto the horizontal plane (see Fig. 9). The length of this column is equal to $g(w_L^2 + w_R^2) + c_S^2 = 1$. The two columns are thus of unit length and pairwise orthogonal if $g = \cos^2 \beta$, and therefore the matrix preserves the total energy.

Finally, the Lauridsen [15] decorrelator is used to obtain stereo surround because of its simplicity. This decorrelator can be viewed as two FIR comb filters (h_L and h_R) with two taps each for surround left and surround right. The impulse responses of these filters are illustrated in Fig. 10. A time delay of $\delta \approx 10$ ms (440 samples) is used between the taps, which is determined experimentally.

The choice of the time delay δ is a subtle compromise between the amount of widening and the sound diffuseness. The greater δ is, the more diffuse the sounds will be, and at some point it will lead to confusion.

Note that there are other decorrelator filters available, such as complementary comb filters, in which the "teeth" are distributed equally on a logarithmic frequency scale. Informal listening tests, however, revealed that the Lauridsen decorrelator is better appreciated when it is applied to the surround channels. Furthermore, its efficiency in the implementation also plays an important role in our application.

2.3 Analysis of Each Discrete Channel

The behavior of the proposed method can be analyzed in each channel and gives useful information for validating an implementation. In addition, such an analysis can be used to demonstrate the channel separation of our method. Such analyses are listed in Table 1.

First we feed the system with a sine wave in the left channel only and set the right channel to zero,

$$\begin{aligned} x_L(k) &= \sin(\omega k) \\ x_R(k) &= 0 \, . \end{aligned} \tag{24}$$

In this situation we have $w_L = 1$ and $w_R = 0$. Therefore,

$$\begin{aligned} y(k) &= \sin(\omega k) \\ q(k) &= 0 \, . \end{aligned} \tag{25}$$

Substituting Eqs. (24) and (25) into Eq. (21), and keeping in mind that $c_R = c_C = 0$, we obtain

$$\begin{aligned} u_L(k) &= c_L(k)\sin(\omega k) \\ u_R(k) &= 0 \\ u_C(k) &= 0 \\ u_S(k) &= 0 \, . \end{aligned} \tag{26}$$

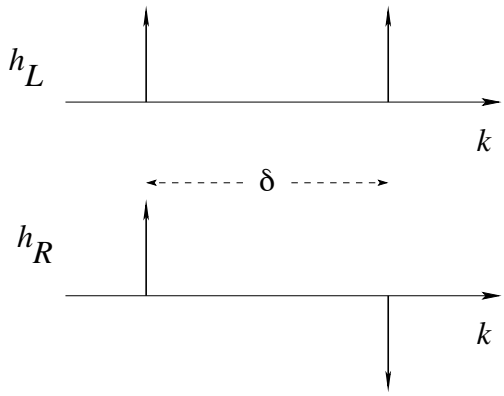


Fig. 10. Impulse response of left and right Lauridsen decorrelation filters. Time delay $\delta \approx 10$ ms (440 samples) is experimentally chosen to produce the most pleasant stereo sounds for the application concerned.

Table 1. Summary of extreme cases decribed in text.*

Input	u_L	u_R	u_C	u_S	w_L	w_R	ρ_0	β
$x_L = f, x_R = 0$	f	0	0	0	1	0	0	$\pi/2$
$x_L = 0, x_R = f$	0	f	0	0	0	1	0	$\pi/2$
$x_L = x_R = f$	0	0	κf	0	$\frac{1}{2}\sqrt{2}$	$\frac{1}{2}\sqrt{2}$	1	0
$x_L = f, x_R = -f$	κf	$-\kappa f$	0	κf	$\frac{1}{2}\sqrt{2}$	$-\frac{1}{2}\sqrt{2}$	0	$\pi/2$
$x_L, x_R \neq 0$, uncorrelated	0	0	0	κf	—	—	0	$\pi/2$

* A time signal f is used to represent any input signals fed into left, right, or a combination of left and right channels. κ represents a scalar due to mapping (Fig. 9). Note that when uncorrelated signals are fed into the system, w_L and w_R become undefined, meaning they can assume any value.

This is to be expected as any signal fed into one particular channel should be kept the same in the output.

Second, feeding the input signal into the right channel, some other combinations can be analyzed similarly. The outputs of these combinations are summarized in Table 1.

From the table it can be seen that fully correlated input signals will be reproduced in the center channel while all other channels are zero. This explains the strong sound localization that is achieved during the listening test, which is discussed in the next section.

Furthermore, when we feed the left and right channels with antiphase signals, no sound will be reproduced in the center channel, left and right outputs are in antiphase, and some sounds are going to the surround channels. This extreme case demonstrates how ambience effects are created when many antiphase signals are present in the original signals.

3 LISTENING TEST

In order to investigate the appreciation of the discussed conversion method, the system was tested together with a few other systems.

3.1 Method

The method of paired comparisons [16] was used to gather personal preference data. During each trial, subjects heard a music excerpt encoded by a certain method M_i, then the same excerpt encoded by M_j. The subject could repeat the pairs as often as desired, and finally had to indicate whether M_i was preferred above M_j or vice versa. There were four systems under test, so six pairs per repertoire per listening position were offered to the subjects. If M_i was preferred above M_j a 1 was placed in a preference matrix X at element x_{ij}, or otherwise at x_{ji}. This matrix was scaled with Thurstone's decision model [17] (see Appendix 2 for more details).

3.2 Technical Equipment and Repertoire

The subjects were either at the position advised by the ITU [10], referred to as the sweet spot, or 1 m aside of that spot, referred to as "off the sweet spot." The loudspeakers used were Philips DSS940 (digital) loudspeakers. The listening room was a rather dry listening room, which enabled a critical judgment for localization and crosstalk between the channels.

We compared four different systems: the system proposed in this paper (system 1), its variant, which puts more low-frequency to the surrounds (system 2), and two other commercially available systems (system 3 and 4, respectively).

Four different music excerpts were chosen. Three music fragments (Fish: "The Company," The Corrs: "What Can I Do," and Melanie C: "Never Be the Same Again") and a sound track from the movie picture "The Titanic" (Track #23 from the DVD) were used. In total a subject had to give 6×2 positions $\times$ 4 fragments $= 48$ assessments.

The number of different tracks was somewhat limited. To derive more general conclusions more tracks would be necessary, such as used in [19]. However, our primary aim was to focus on the theory, while more elaborate listening tests can still be done in the future.

3.3 Subjects

There were 17 subjects. Most were experienced listeners and all had no reported hearing loss.

3.4 Results

For each type of repertoire and each subject the scaled results were plotted. An example is given in Fig. 11. A high scale value means high appreciation. The value itself is of no importance. It is the value with respect to the others. The sum of the scale values equals zero.

The number on top of Figs. 11–13 denotes the coefficient of consistence ξ [16]. A value of $\xi = 1$ means fully consistent. In such a case there is never a violation of the triangular inequalities ($X_i > X_j > X_k > X_i$). $\xi = 0$ means fully inconsistent. This coefficient is important for various reasons. First it reveals how consistently a subject judges the stimuli. In this case subject SP appeared to be a very consistent judge. Second, if the subjects have different preferences with respect to each other, or for different repertoires, then summing their preference matrices will lower the ξ value.

For both positions, on the sweet spot and off the sweet spot, the results for all subjects and repertoires are scaled and plotted in Figs. 12 and 13, respectively.

It is clear that system 1—the system discussed in this paper—performs very well, both on the sweet spot and off the sweet spot. In Figs. 12 and 13 we see values of $\xi = 0.4$ and 0.6, respectively, revealing that not all the subjects act as one single fully consistent subject. However, it appeared that the individual subjects act rather consistently. Furthermore, the figures show that system 1 is rather robust in the face of a displacement from the

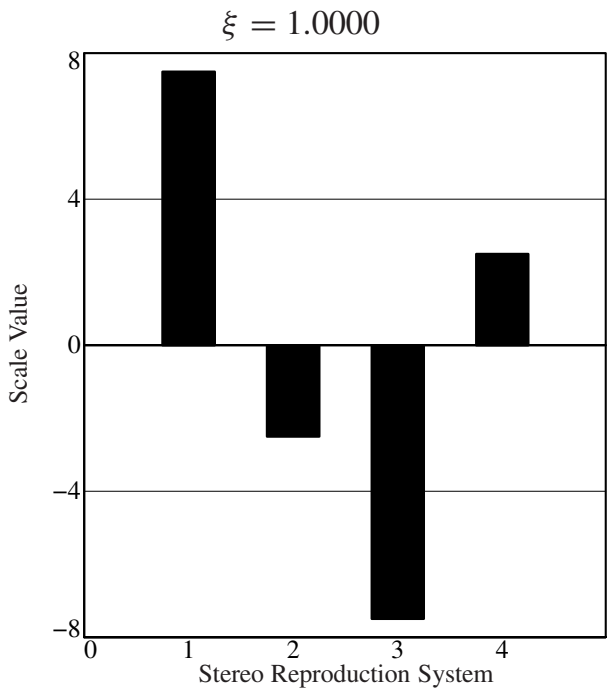


Fig. 11. Results of subject SP, four music excerpts, off the sweet spot. Scale values on vertical axis are arbitrary.

sweet spot, which is a desirable property.

Another analysis—using the Bradley–Terry model [16]—of the results of the listening test was performed [18]. This model was fitted by the maximum-likelihood method, and the goodness-of-fit was tested. It revealed that the four processing methods differed significantly from one another.

4 CONCLUSION

A new method to convert two-channel stereo to multichannel sound has been presented. A three-dimensional representation has been used to produce each channel's gain, which is time varying. PCA proved to be a powerful tool to detect the direction of a stereo image, which is then used to derive the center channel's gain. Furthermore, a robust tracking algorithm for computing the cross correlation between left and right channels has been used to improve the sound quality of the surround channels.

A listening test comparing four different systems has been carried out using both music recordings and DVD movie tracks, and the results have been analyzed using the Thurstone scaling technique as well as the Bradley–Terry model. The preliminary listening test has shown that the proposed method is very good, both on and off the sweet spot. Moreover, it has been shown that the four processing methods differed significantly from one another.

5 ACKNOWLEDGMENT

The authors thank D. Roovers for his work in the initial phase of the project, and the statisticians T. J. J. Denteneer and J. Engel for interpreting the results of the listening tests.

6 REFERENCES

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

[2] P. Scheiber, "Four Channels and Compatibility," *J. Audio Eng. Soc*., vol. 19, pp. 267–279 (1971 Apr.).

[3] J. Eargle, "4–2–4 Matrix Systems: Standards, Practice, and Interchangeability," *J. Audio Eng. Soc*., vol. 20, pp. 809–815 (1972 Dec.).

[4] M. E. G. Willcocks, "Transformations of the Energy Sphere," *J. Audio Eng. Soc*., vol. 31, pp. 29–36 (1983 Jan./Feb.).

[5] S. Julstrum, "A High-Performance Surround Sound Process for Home Video," *J. Audio Eng. Soc*., vol. 35, pp. 536–549 (1987 July/Aug.).

[6] M. A. Gerzon and G. J. Barton, "Ambisonic Decoders for HDTV," presented at the 92nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc*., vol. 40, p. 438 (1992 May), preprint 3345.

[7] R. Dressler, "Pro Logic Surround Decoder, Principles of Operation," Dolby Laboratories, http://www.dolby.com (1997).

[8] R. Irwan and R. M. Aarts, "A Method to Convert Stereo to Multichannel Sound," presented at the AES 19th International Conference, Germany (2001).

[9] G. Theile, "HDTV Sound System: How Many Channels?" in *Proc. 10th AES Conf. on Images and Audio* (1991), pp. 147–162.

[10] ITU-R BS.1116, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunications Union, Geneva, Switzerland (1994).

[11] T. W. Lee, *Independent Component Analysis:*

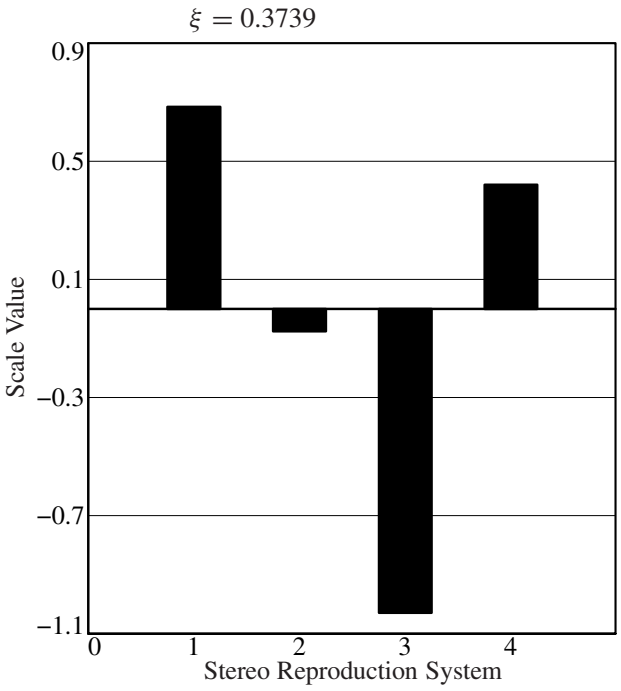


Fig. 12. Total scaling results of all 17 subjects, four music excerpts, on the sweet spot. Scale values on vertical axis are arbitrary.

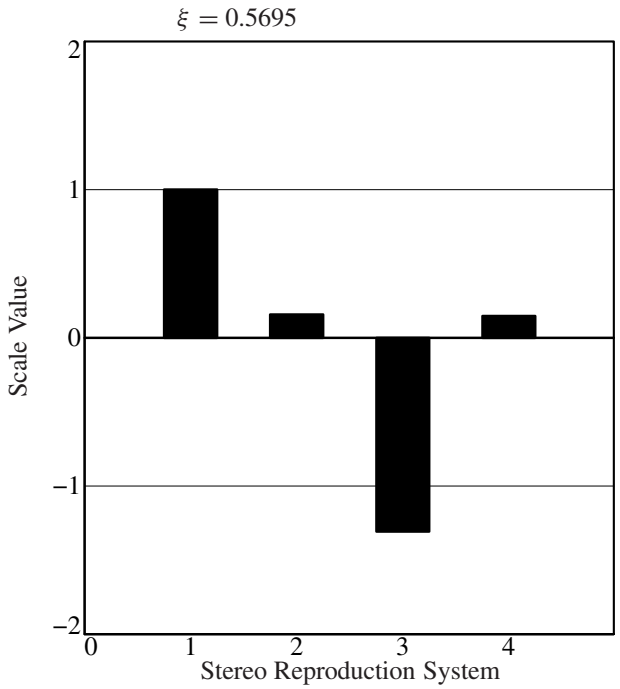


Fig. 13. Total scaling results of all 17 subjects, four music excerpts, off the sweet spot. Scale values on vertical axis are arbitrary.

Theory and Applications (Kluwer, Boston, MA, 1998).

[12] S. Haykin, *Neural Networks*, 2nd ed. (Prentice-Hall, Englewood Cliffs, NJ, 1999).

[13] J. Karhunen, "Stability of Oja's PCA Subspace Rule," *Neural Comput.*, vol. 6, pp. 739–747 (1994).

[14] R. M. Aarts, R. Irwan, and A. J. E. M. Janssen, "Efficient Tracking of the Cross-Correlation Coefficient," *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 391–402 (2002 Sept.).

[15] H. Lauridsen, "Experiments Concerning Different Kinds of Room-Acoustics Recording" (in Danish), *Ingenioren*, vol. 47 (1954).

[16] H. A. David, *The Method of Paired Comparisons*, 2nd ed. (Griffin, London, 1988).

[17] W. S. Torgerson, *Theory and Methods of Scaling* (Wiley, New York, 1958).

[18] J. Engel, "Paired Comparisons on Audio: Modelling the System Differences," Intern. Rep. E1624-20, CQM (Centre for Quantitative Methods), Eindhoven, The Netherlands (2001 July).

[19] G. A. Soulodre, T. Grusec, M. Lavoie, and L. Thibault, "Subjective Evaluation of State-of-the-Art Two-Channel Audio Codecs," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 46, pp. 164–177 (1998 Mar.).

[20] R. M. Aarts, "A New Method for the Design of Crossover Filters," *J. Audio Eng. Soc.*, vol. 37, pp. 445–454 (1989 June).

[21] L. L. Thurstone, "A Law of Comparative Judgement," *Psychol. Rev.*, vol. 34, pp. 273–286 (1927).

[22] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1972).

APPENDIX 1
RELATION BETWEEN CORRELATION COEFFICIENT ρ AND PCA

As was presented in Section 2, the cross-correlation technique is very useful in determining the surround channel distribution. In this appendix we discuss the relationship between the cross-correlation technique and the principal component analysis (PCA), also known as the Karhunen–Loéve transformation. In general, PCA maximizes the rate of decrease in variance for each of its components. The solution lies in the eigenstructure of the covariance matrix $\boldsymbol{C}$.

We may decompose $\boldsymbol{C}$ with the singular-value decomposition (SVD) [12] as

$$\boldsymbol{C} = \boldsymbol{V}\Lambda\boldsymbol{V}^{-1} \tag{27}$$

where $\boldsymbol{V}$ is an orthogonal (unitary) matrix of eigenvectors $\boldsymbol{v}_j$ and Λ is a diagonal matrix with the eigenvalues

$$\Lambda = \mathrm{diag}\left[\lambda_1, \lambda_2, \ldots, \lambda_j, \ldots, \lambda_m\right] \tag{28}$$

arranged in decreasing order,

$$\lambda_1 > \lambda_2 > \cdots > \lambda_j > \cdots > \lambda_m \tag{29}$$

so that $\lambda_1 = \lambda_{\max}$.

It appears that the PCA and the SVD of $\boldsymbol{C}$ are basically the same, just viewing the problem in different ways. From the theory of SVD it follows that the eigenvectors of the covariance matrix $\boldsymbol{C}$ define the unit vectors $\boldsymbol{v}_j$ representing the principal directions along which the variance is maximized for each component. The associated eigenvalues define the total variance of the m elements,

$$\sum_{j=1}^{m} \sigma_j^2 = \sum_{j=1}^{m} \lambda_j \, . \tag{30}$$

In the present case $\boldsymbol{C}$ is constructed in the following way. Consider a segment of left and right audio samples $\boldsymbol{x}_\mathrm{L} = [x_\mathrm{L}(1) \cdots x_\mathrm{L}(B)]$ and $\boldsymbol{x}_\mathrm{R} = [x_\mathrm{R}(1) \cdots x_\mathrm{R}(B)]$; hence $m = 2$. With their correlation coefficients ρ the covariance matrix $\boldsymbol{C}$ can be written as

$$\boldsymbol{C} = \begin{pmatrix} \sigma_{\boldsymbol{x}_\mathrm{L}}^2 & \rho\sigma_{\boldsymbol{x}_\mathrm{L}}\sigma_{\boldsymbol{x}_\mathrm{R}} \\ \rho\sigma_{\boldsymbol{x}_\mathrm{L}}\sigma_{\boldsymbol{x}_\mathrm{R}} & \sigma_{\boldsymbol{x}_\mathrm{R}}^2 \end{pmatrix} . \tag{31}$$

Now we see the relation between the correlation coefficient ρ on the left-hand side of Eq. (27) and the principal components on the right-hand side of Eq. (27). The latter we will develop more explicitly in the following.

Since $\boldsymbol{C}$ is positive definite, the eigenvalues of $\boldsymbol{C}$ are both real and positive and can be calculated as

$$\lambda_{1,2} = \frac{1}{2}\left(\sigma_{\boldsymbol{x}_\mathrm{L}}^2 + \sigma_{\boldsymbol{x}_\mathrm{R}}^2 \pm s\right) \tag{32}$$

$$s = \sqrt{\left(\sigma_{\boldsymbol{x}_\mathrm{L}}^2 - \sigma_{\boldsymbol{x}_\mathrm{R}}^2\right)^2 + \left(2\rho\sigma_{\boldsymbol{x}_\mathrm{L}}\sigma_{\boldsymbol{x}_\mathrm{R}}\right)^2} \, . \tag{33}$$

The eigenvectors of C corresponding to the eigenvalues $\lambda_{1,2}$ are

$$\boldsymbol{v}_{1,2} = \gamma\left(\frac{\sigma_{\boldsymbol{x}_\mathrm{L}}^2 - \sigma_{\boldsymbol{x}_\mathrm{R}}^2 \pm s}{2\rho\sigma_{\boldsymbol{x}_\mathrm{L}}\sigma_{\boldsymbol{x}_\mathrm{R}}}, 1\right) \tag{34}$$

where γ is such that $|\boldsymbol{v}_{1,2}| = 1$.

If we consider $\boldsymbol{V}$ as a rotation matrix over the angle α, as used in Eq. (12), then we can derive

$$\rho = \frac{\left(\sigma_{\boldsymbol{x}_\mathrm{L}}^2 - \sigma_{\boldsymbol{x}_\mathrm{R}}^2\right)\tan(2\alpha)}{2\sigma_{\boldsymbol{x}_\mathrm{L}}\sigma_{\boldsymbol{x}_\mathrm{R}}} \, . \tag{35}$$

As special cases we consider $\rho = 0$. Then the left and right channels are uncorrelated and the eigenvectors are just (1, 0), (0, 1), which are coincident with the original left and right axes. As corresponding eigenvalues we have $\lambda_1 = \sigma_{\boldsymbol{x}_\mathrm{L}}^2$ and $\lambda_2 = \sigma_{\boldsymbol{x}_\mathrm{R}}^2$, which are the powers of the left and right channels, respectively. A similar case occurs if $\sigma_{\boldsymbol{x}_\mathrm{L}}^2 = \sigma_{\boldsymbol{x}_\mathrm{R}}^2 = \sigma^2$. Then $\lambda_{1,2} = \sigma^2$ and $\alpha = \pi/4$ and/or $\rho = 0$.

Since PCA can be seen as finding a decomposition of the covariance matrix $\boldsymbol{C}$, the transformation into dominant and remaining signals using PCA described in Section 1 can also be carried out by computing $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, respectively. It has, however, a limited practical usefulness since it requires more computation effort as opposed to an efficient way of using the LMS algorithm given in Eq. (6).

APPENDIX 2
SCALING

A2.1 Introduction

A problem encountered in many disciplines is how to measure and interpret the relationships between objects. A second problem is the lack, in general, of a mathematical relationship between the perceived response and the actual physical measure. With regard to this paper, how does the appreciation of our 2-to-5-channel system differ from others? How do we measure and what scale do we need? In the following we discuss some scales and techniques and give two examples.

A2.2 Scaling

The purpose of scaling is to quantify the qualitative relationships between objects by scaling data. Scaling procedures attempt to do this by using rules that assign numbers to qualities of things or events. There are two types of scaling, univariate scaling, which is explained hereafter, and multidimensional scaling (MDS), which is an extension of univariate scaling (see, for example, [20]). Univariate scaling is usually based on the law of comparative judgment [17], [21]. It is a set of equations relating the proportion of times any stimulus *i* is judged greater or is more highly appreciated relative to a given attribute (in our case the appreciation) than any other stimulus *j*. The set of equations is derived from the postulates presented in [17]. In brief, these postulates are as follows.

1) Each stimulus when presented to an observer gives rise to a discriminal process which has some value on the psychological continuum of interest (in our case the appreciation).

2) Because of momentary fluctuations in the organism, a given stimulus does not always excite the same discriminal process. This can be considered as noise in the process. It is postulated that the values of the discriminal process are such that the frequency distribution is normal on the psychological continuum.

3) The mean and the standard deviation of the distribution associated with a stimulus are taken as its scale value and discriminal dispersion, respectively.

Consider the theoretical distributions S_j and S_k of the discriminal process for any two stimuli j and k, respectively, as shown in Fig. 14(a). Let S_j and S_k correspond to the scale values of the two stimuli and σ_j and σ_k to their discriminal dispersion caused by noise.

Now we assume that the standard deviations of the distributions are all equal and constant (as in Fig. 14), and that the correlation between the pairs of discriminal processes is constant. This is called "condition C," in Torgerson [17]. Since the distribution of the difference of the normal distributions is normal, we get

$$\bar{S}_k - \bar{S}_j = cx_{jk} \tag{36}$$

where c is a constant and x_{jk} is the transformed [see Eq. (39)] proportion of the number of times stimulus k is more highly appreciated than stimulus j. Eq. (36) is also known as Thurstone's case V. The distribution of the discriminal differences is plotted in Fig. 14(b). Eq. (36) is a set of $n(n-1)$ equations with $n+1$ unknowns, n scale values, and c. This can be solved with the least-square method. Setting $c = 1$ and the origin of the scale to the mean of the estimated scale values, that is,

$$1/n \sum_{j=1}^{n} s_j = 0 \tag{37}$$

we get

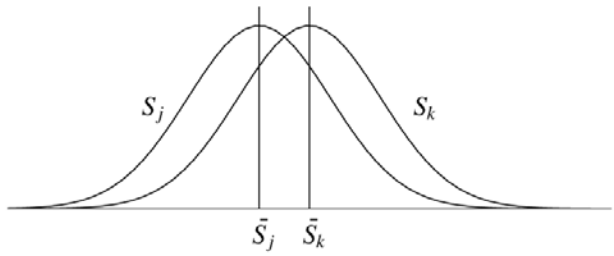
$$s_k = 1/n \sum_{j=1}^{n} x_{jk} \; . \tag{38}$$

Thus the least-square solution of the scale values can be obtained simply by averaging the columns of matrix $\boldsymbol{X}$. However, the elements x_{jk} of $\boldsymbol{X}$ are not directly available. With paired comparisons we measure the proportion p_{kj} that stimulus k was judged greater than stimulus j. This proportion can be considered a probability that stimulus k was judged greater than stimulus j. This probability is equal to the shaded are in Fig. 14(b), or

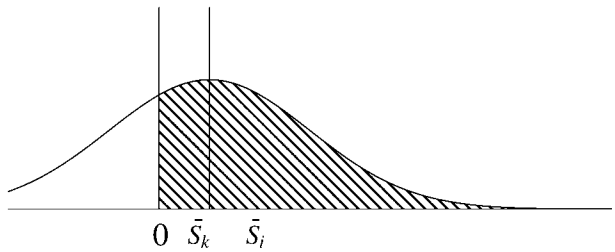
$$x_{jk} = \mathrm{erf}\left(p_{jk}\right) \tag{39}$$

where erf is the error function [22, §7, 26.2], which can easily be approximated (see, for example, [22, §26.2.23]). A problem may arise if $p_{jk} \approx \pm 1$ since $|x_{jk}|$ can be very large. In this case one can then replace x_{jk} by a large value.

It may be noted that this type of transformation is also known as Gaussian transform, where instead of the symbol x, z is used, known as the z score. Instead of using Eq. (39), other models are used, such as the Bradley–Terry model (see [16]). All forms of the law of comparative



(a)



(b)

Fig. 14. (a) Probability distributions S_j and S_k of stimuli j and k on psychological continuum, with mean values S_j and S_k. (b) Probability distributions of difference of random variables. Shaded portion gives the proportion of times stimulus k was judged greater than stimulus j. $S_k - S_i$ is proportional to the difference in scale value for both stimuli.

judgment assume that each stimulus has been compared with the other stimuli a large number of times. The direct method of obtaining the values of p_{jk} is known as the method of paired comparisons (see, for example, [16]). As an example, the measured probabilities p_{jk} for a subject are listed in Table 2. The upper triangular is calculated as $p_{jk} = 1 - p_{kj}$. Using Eq. (39) the x_{jk} values are obtained. Using Eq. (38) the final scale values are determined and plotted in Fig. 11.

APPENDIX 3
EFFICIENT COMPUTATION OF CROSS-CORRELATION COEFFICIENT

In this appendix we summarize the mathematical derivations of the tracking cross-correlation coefficients as was fully reported in [14]. For the generalization we use x and y for two signals being correlated instead of x_L and x_R, which represent specifically stereo audio signals.

We show that ρ satisfies to a good approximation (when η is small) the recursion in Eq. (15) with γ given by

$$\gamma = \frac{c\ e^{\eta}}{2x_{rms}y_{rms}} \tag{40}$$

where $c = 1 - e^{-\eta}$, and the subscripts rms refer to the root mean-square values of x and y.

Using an exponential window we can redefine the correlation of x and y at time instant k as

$$\rho(k) = \frac{S_{xy}(k)}{\left[S_{xx}(k)S_{yy}(k)\right]^{1/2}} \tag{41}$$

for k an integer and where

$$\begin{aligned} S_{xy}(k) &= \sum_{t=0}^{\infty} c\ e^{-\eta l}\, x_{k-l}\, y_{k-l} \\ &= e^{-\eta} S_{xy}(k-1) + c x_k y_k \end{aligned} \tag{42}$$

and S_{xx} and S_{yy} are defined similarly. Hence,

$$\rho(k) = \frac{S_{xy}(k-1) + c\ e^{\eta} x_k y_k}{\left\{\left[S_{xx}(k-1) + c\ e^{\eta} x_k^2\right]\left[S_{yy}(k-1) + c\ e^{\eta} y_k^2\right]\right\}^{1/2}} \tag{43}$$

Since we consider small values of η, we have that $c = 1 - e^{-\eta}$ is small as well. Expanding the right-hand side of Eq. (43) in powers of c and retaining only the constant and the linear term, we get, after some calculations,

$$\rho(k) = \rho(k-1) + \frac{c\ e^{\eta}}{2\left[S_{xx}(k-1)S_{yy}(k-1)\right]^{1/2}} \times \left\{2x_k y_k - \left\{\left[\frac{S_{yy}(k-1)}{S_{xx}(k-1)}\right]^{1/2} x_k^2 + \left[\frac{S_{xx}(k-1)}{S_{yy}(k-1)}\right]^{1/2} y_k^2\right\}\rho(k-1)\right\} + O\left(c^2\right). \tag{44}$$

Then, deleting the $O(c^2)$ term, we obtain the recursion in Eq. (15), with γ given by Eq. (40), when we identify

$$\begin{aligned} x_{rms}^2 &= S_{xx}(k) \\ y_{rms}^2 &= S_{yy}(k) \end{aligned} \tag{45}$$

for a sufficiently large k, and assuming that $x_{rms}^2 = y_{rms}^2$.

One may ask how to handle signals x and y that have nonzero, and actually time-varying, mean values. In those cases we still define $\rho(k)$ as in Eq. (42), however, with S_{xy} replaced by

$$S_{xy}(k) = \sum_{l=0}^{\infty} c\ e^{-\eta l}\left[x_{k-l} - \bar{x}(k)\right]\left[y_{k-l} - \bar{y}(k)\right] \tag{46}$$

where

$$\begin{aligned} \bar{x}(k) &= \sum_{l=0}^{\infty} c\ e^{-\eta l}\ x_{k-l} \\ \bar{y}(k) &= \sum_{l=0}^{\infty} c\ e^{-\eta l}\ y_{k-l} \end{aligned} \tag{47}$$

and S_{xx} and S_{yy} changed accordingly. It can then be shown that

$$\begin{aligned} \bar{x}(k) &= e^{-\eta}\,\bar{x}(k-1) + c x_k \\ \bar{y}(k) &= e^{-\eta}\,\bar{y}(k-1) + c y_k \end{aligned} \tag{48}$$

and

$$S_{xy}(k) = e^{-\eta} S_{xy}(k-1) + c\ p_k q_k \tag{49}$$

where $p_k = x_k - \bar{x}(k-1)$, $q_k = y_k - \bar{y}(k-1)$, while sim-

Table 2. Proportion p_{jk} of times that stimulus k was judged higher than stimulus j, obtained via paired comparisons.*

	—			
	0.0	—		
$\downarrow j$	0.0	0.0	—	
	0.5	1.0	1.0	—
	$\rightarrow k$			

*For subject SP, averaged over the four fragments off the sweep spot.

ilar recursions hold for S_{xx} and S_{yy}. This then yields

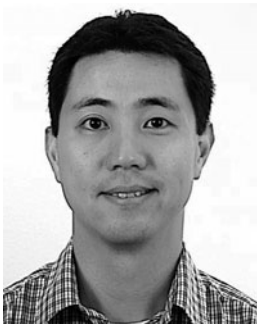
$$\rho(k) = \frac{S_{xy}(k-1) + cp_k q_k}{\left\{\left[S_{xx}(k-1) + cp_k^2\right]\left[S_{yy}(k-1) + cq_k^2\right]\right\}^{1/2}}$$

$$= \rho(k-1) + \frac{c}{2\left[S_{xx}(k-1)S_{yy}(k-1)\right]^{1/2}}$$

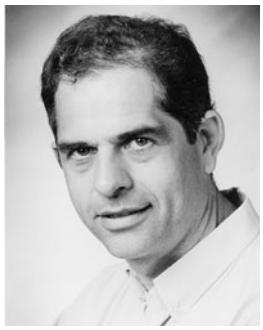
$$\times \left\{2p_k q_k - \left\{\left[\frac{S_{yy}(k-1)}{S_{xx}(k-1)}\right]^{1/2} p_k^2 + \left[\frac{S_{xx}(k-1)}{S_{yy}(k-1)}\right]^{1/2} q_k^2\right\}\rho(k-1)\right\} + O\left(c^2\right). \quad (50)$$

From this point onward, comparing with Eq. (44), one can proceed to apply many, if not all, of the developments presented in this appendix to this more general situation.

THE AUTHORS



R. Irwan



R. M. Aarts

Roy Irwan received an M.Sc. degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands in 1992, and a Ph.D. degree in electrical engineering from the University of Canterbury, Christchurch, New Zealand in 1999.

From 1993 to 1995 he was employed as a system engineer at NKF B.V. After obtaining his Ph.D. degree in 1999, he joined the Digital Signal Processing group at Philips Research Laboratories, Eindhoven, The Netherlands. Since 2002 he has been working with the State University Groningen, faculty of Medical Sciences.

Dr. Irwan has published a number of refereed papers in international journals and has more than 14 patent applications. His research interests include (medical) image and signal processing.

●

Ronald Aarts was born in 1956, in Amsterdam, The Netherlands. He received a degree in electrical engineering in 1977, and a Ph.D. degree from the Delft University of Technology in 1994.

In 1977 he joined the Optics group of Philips Research Laboratories, Eindhoven, The Netherlands, where he was involved in research into servos and signal processing for use in both Video Long Play players and Compact Disc players. In 1984 he joined the Acoustics group of the Philips Research Laboratories and was engaged in the development of CAD tools and signal processing for loudspeaker systems. In 1994 he became a member of the Digital Signal Processing group of the Philips Research Laboratories. There he has been engaged in the improvement of sound reproduction by exploiting DSP and psychoacoustical phenomena.

Dr. Aarts has published over one hundred technical papers and reports and is the holder of more than a dozen U.S. patents in his fields. He was a member of the organizing committee and chair for various conventions. He is a senior member of the IEEE, a fellow of the AES, and a member of the Dutch Acoustical Society and the Acoustical Society of America. He is past chair of the Dutch section of the AES.