Optimized Covariance Domain Framework for Time-Frequency Processing of Spatial Audio

JUHA VILKAMO[1,2], TOM BÄCKSTRÖM[2], AND ACHIM KUNTZ[2]

[1]*Aalto University, Espoo, Finland*
[2]*Fraunhofer IIS, Erlangen, Germany*

The covariance matrix of a multichannel audio signal is a measure that contains the channel energies and the inter-channel dependencies. This measure is perceptually relevant in frequency bands, since with the effect of the acoustic transfer path, it forms the inter-aural cues based on which the human spatial hearing system decodes the spatial sound. This is the foundation to several state of the art perceptual spatial processing techniques. In this paper we propose a generalized framework for spatial sound processing that operates directly in the covariance matrix domain. The technique optimizes the sound quality by a least mean square solution with respect to a defined prototype signal. Furthermore, the technique uses all available independent signal components in the input channels in order to minimize the undesired effects due to the decorrelators that are often employed in the concerned use cases. A wide variety of applications of the proposed method has been identified.

1 INTRODUCTION

Our human auditory system is able to localize sound sources, and in some degree estimate the size and characteristics of the surrounding space, based on the two audio signals entering the ear canals. The acoustic effect of the head, torso, and pinnae causes the sound to attain characteristics that are specific to the angle of the arrival. These are the inter-aural level-difference (ILD) and the inter-aural phase difference (IPD) as function of the frequency [1]. In addition, the inter-aural coherence (IC) becomes relevant in the presence of several sources or in diffuse sound environments such as reverberation.

In perceptual time-frequency processing of spatial audio [2,3,4,5,6,7,8], the processing focuses on those characteristics of the stereo or multichannel sound that contribute to the above binaural cues. These are in frequency bands the inter-channel level-difference (ICLD), inter-channel phase difference (ICPD), and the inter-channel coherence (ICC) [9]. In binaural reproduction over headphones, the inter-channel cues are the same as the inter-aural cues. In loudspeaker reproduction, the inter-channel cues map to the binaural cues with the effect of the acoustic transfer path from the loudspeakers to the listeners' ears.

Several techniques based on the inter-channel cues have emerged. Efficient multichannel transmission systems [2,3] transmit only a reduced number of channels and a low-bitrate parametric side information with which the original inter-channel cues and thus the original spatial perception are recovered. Spatial Audio Object Coding [4] is a similar technique that transmits several audio objects in downmix channels and allows an option to render the objects flexibly in the receiver end. Directional Audio Coding [5] is a DSP-driven microphone technique that analyzes the sound field direction and diffuseness in frequency bands, and uses amplitude panning and decorrelation to synthesize such a multi-channel loudspeaker signal that produces a sound field perceptually close to the original. Panning and decorrelation are tools for generating a set of ICLDs and ICCs that contribute in generating the desired spatial perception. Often such applications apply the processing in a non-uniform frequency resolution, such as the equivalent rectangular bandwidth (ERB) [10] scale.

In this paper we propose a generalized and optimized framework for spatial audio processing in terms of the signal covariance matrix in frequency bands, which is a measure that contains the described inter-channel characteristics. Applications that apply the framework determine adaptively the target covariance matrix in the time-frequency domain, and the framework performs the transform while aiming to best preserve the overall sound quality.

In addition to the covariance matrix, the signal content in the channels also matter. For example, in a scenario of teleconferencing with spectrally overlapping talkers, it is

Correspondence should be addressed to juha.vilkamo@aalto.fi

relevant to reproduce the harmonics of each talker approximately at their correct directions. Similarly, in reproduction of sound recorded in echoic spaces, due to the precedence effect [1,11] it is important to reproduce the first arriving wavefront at its correct direction. For this reason the spatial synthesis in the proposed framework is controlled also by a prototype signal. The proposed framework produces an output signal that is as close as possible to the prototype signal, with the boundary condition that the target covariance matrix, i.e., the target spatial characteristic, is attained.

The proposed framework also optimizes the usage of the decorrelators, which are processes that provide new incoherent signal components based on the input signals. By definition, they affect the time- and phase structure of the signal, which may be detrimental to the perceived signal quality. The proposed framework ensures that the decorrelated energy is injected only to the minimum necessary extent, i.e., only to complement for the missing independent signal components in the input channels.

The main expected application of the proposed framework is in the field of spatial sound recording [5,7], in which the problems related to the signal covariance are particularly apparent due to the physical limitations of building high-quality microphones that are directional enough for spatial sound reproduction. Further expected use cases include stereo and multichannel enhancement, ambience extraction, upmixing, and downmixing. The proposed framework can operate with real or complex definitions of the covariance matrices. The real-valued definitions can be sufficient in applications where the phase differences are less relevant, e.g., in spatial sound rendering using a set of coincident microphones. Complex valued processing is necessary in applications that rely also on phase differences, such as binaural rendering and spatial sound rendering from a spaced microphone array.

The paper is organized as follows. We start by the definitions, followed by the derivation of the proposed method. The adaptive mixing solution is derived first using a set of simplifying temporary assumptions. Then the generalization is provided and the means to apply a decorrelated residual signal to compensate when there is not enough independent signal energy available otherwise. Then an example implementation in context of stereo upmixing is given, followed by a set of further numerical examples, and finally the conclusion. As an appendix, an example Matlab implementation providing the optimized adaptive mixing solution as described in this paper is provided.

2 DEFINITIONS

Matrices and vectors are denoted with bold faced symbols, where uppercase denotes a matrix. Matrix and vector elements are marked with row and column subindices i and j, such that g_{ij} is an element of matrix $\mathbf{G}$, and x_i is an element of vector $\mathbf{x}$. The mixing solution is first derived assuming the same number of the input and output channels N. The method is generalized to arbitrary channel numbers in Section 3.3.

Let us define the input channels $x_i(t)$, $i = 1, \ldots, N$ and the output channels $y_j(t)$, $j = 1, \ldots, N$. Their short-time Fourier transformed (STFT) counterparts are $X_i(k, l)$ and $Y_j(k, l)$, where k is the downsampled time index and l is the frequency index. The time-frequency signal vectors are

$$\mathbf{x}(k,l) = \begin{bmatrix} X_1(k,l) \\ X_2(k,l) \\ \vdots \\ X_N(k,l) \end{bmatrix} \quad \mathbf{y}(k,l) = \begin{bmatrix} Y_1(k,l) \\ Y_2(k,l) \\ \vdots \\ Y_N(k,l) \end{bmatrix}. \tag{1}$$

Note that the time and frequency indices (k, l) are now omitted for brevity of notation. The signal covariance matrices are

$$\begin{aligned} \mathbf{C}_\mathbf{x} &= \mathrm{E}\left[\mathbf{x}\mathbf{x}^H\right] \\ \mathbf{C}_\mathbf{y} &= \mathrm{E}\left[\mathbf{y}\mathbf{y}^H\right], \end{aligned} \tag{2}$$

where E[ ] is the expectation operator and $\mathbf{x}^H$ and $\mathbf{y}^H$ are the conjugate transposes of $\mathbf{x}$ and $\mathbf{y}$. The target covariance matrix $\mathbf{C}_\mathbf{y}$ depends on the specific application, of which an example is given in Section 4.2. Based on $\mathbf{C}_\mathbf{y}$ the proposed algorithm applies a mixing procedure to the input signals $\mathbf{x}$ that generates the optimal output signals $\mathbf{y}$.

In an actual implementation the expectation operator in Eq. (2) is replaced with a mean operator over a number of time and/or frequency indices in a resolution that approximates that of the human spatial hearing. Such a time-frequency area is denoted as the time-frequency tile. The adaptive mixing solution in this paper is formulated once per each tile.

Since the covariance matrices are Hermitian and positive-semi-definite they can be decomposed as

$$\begin{aligned} \mathbf{C}_\mathbf{x} &= \mathbf{K}_\mathbf{x}\mathbf{K}_\mathbf{x}^H \\ \mathbf{C}_\mathbf{y} &= \mathbf{K}_\mathbf{y}\mathbf{K}_\mathbf{y}^H. \end{aligned} \tag{3}$$

Such decompositions can be obtained, e.g., by using the Cholesky decomposition or eigendecomposition [12]. The decomposition is not unique. For any unitary matrices $\mathbf{P}_\mathbf{x}$ and $\mathbf{P}_\mathbf{y}$, matrices $\mathbf{K}_\mathbf{x}\mathbf{P}_\mathbf{x}$ and $\mathbf{K}_\mathbf{y}\mathbf{P}_\mathbf{y}$ also fulfill the decomposition condition since

$$\begin{aligned} \mathbf{K}_\mathbf{x}\mathbf{P}_\mathbf{x}\mathbf{P}_\mathbf{x}^H\mathbf{K}_\mathbf{x}^H &= \mathbf{K}_\mathbf{x}\mathbf{K}_\mathbf{x}^H = \mathbf{C}_\mathbf{x} \\ \mathbf{K}_\mathbf{y}\mathbf{P}_\mathbf{y}\mathbf{P}_\mathbf{y}^H\mathbf{K}_\mathbf{y}^H &= \mathbf{K}_\mathbf{y}\mathbf{K}_\mathbf{y}^H = \mathbf{C}_\mathbf{y}. \end{aligned} \tag{4}$$

Finally, let us define the prototype signal $\hat{\mathbf{y}}$, which is applied in the derivation of the least mean square mixing solution. The prototype signal is constructed by applying a prototype matrix $\mathbf{Q}$ to the input signal

$$\hat{\mathbf{y}} = \mathbf{Q}\mathbf{x}. \tag{5}$$

The matrix $\mathbf{Q}$ can be defined static or time or frequency variant, depending on the application and is discussed further in the end of this section. The covariance matrix of the prototype signal is $\mathbf{C}_{\hat{\mathbf{y}}} = \mathrm{E}\left[\hat{\mathbf{y}}\hat{\mathbf{y}}^H\right]$, and typically $\mathbf{C}_{\hat{\mathbf{y}}} \neq \mathbf{C}_\mathbf{y}$. The proposed adaptive mixing solution produces an output

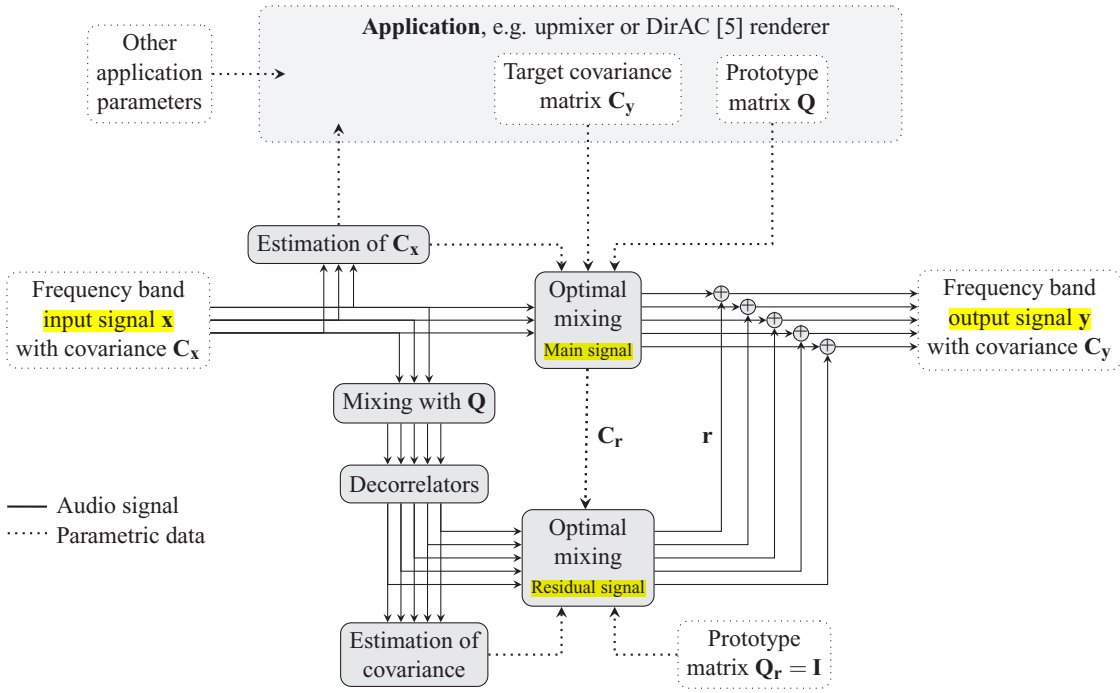

Fig. 1. The proposed framework for perceptual spatial audio processing in terms of the signal covariance matrix in frequency bands. If there are enough independent signal components in the input, the main optimal mixing block is sufficient for generating the output signal with covariance matrix $\mathbf{C_y}$. The optimal mixing block ensures that the output has the minimum distance to a prototype signal $\hat{\mathbf{y}} = \mathbf{Qx}$ after its energy normalization. In case there is not enough independent signal energy, a residual signal $\mathbf{r}$ with covariance matrix $\mathbf{C_r}$ that fully compensates for the missing signal components is generated from the decorrelated signals.

signal $\mathbf{y}$ that has $\mathbf{C_y}$, and minimizes the error

$$e = \mathrm{E}\left[\|\mathbf{y} - \mathbf{G}_{\hat{\mathbf{y}}}\hat{\mathbf{y}}\|^2\right], \tag{6}$$

where $\mathbf{G}_{\hat{\mathbf{y}}}$ is a diagonal gain matrix that normalizes the per-channel energies of $\hat{\mathbf{y}}$ to those of $\mathbf{y}$. The normalization step ensures that the error measure is weighted with the channel energies of $\mathbf{y}$, which is not guaranteed if the normalization is omitted. The diagonal elements of $\mathbf{G}_{\hat{\mathbf{y}}}$ are

$$g_{\hat{\mathbf{y}}ii} = \sqrt{\frac{c_{\mathbf{y}ii}}{c_{\hat{\mathbf{y}}ii}}}, \; i = 1, \ldots, N. \tag{7}$$

The general guideline for designing the application specific prototype matrix $\mathbf{Q}$ is to consider which kind of signal content is appropriate for each output channel. In spatial sound recording, for example, it is desirable that the sound emitted by each of the loudspeakers is most similar to the sound that in the recorded sound scene arrived from the corresponding angle. Thus, such $\mathbf{Q}$ is selected that mixes the microphone signals in the way that a spatial amplification pattern is directed toward the desired angle, while other angles are attenuated. As a further example, in stereo upmixing a reasonable $\mathbf{Q}$ is such that defines that the loudspeaker channels at one side correspond to the original stereo channel at the same side.

3 ADAPTIVE MIXING SOLUTION

The objective is to generate from the input signal $\mathbf{x}$ an output signal $\mathbf{y}$ that has the covariance matrix $\mathbf{C_y}$, while the error measure in Eq. (6) is minimized. Two processing methods are utilized in the following. The preferred method is adaptive mixing of the input channels since it best preserves the signal quality. The complementary method is the injection of a decorrelated residual signal when there are not enough independent signal components available otherwise. The input-output relation can be written as

$$\mathbf{y} = \mathbf{Mx} + \mathbf{r}, \tag{8}$$

where $\mathbf{M}$ is a mixing matrix and $\mathbf{r}$ is the residual signal processed from the decorrelated signals. The overall block diagram of the processing is shown in Fig. 1.

3.1 Adaptive Mixing Solution in Idealized Condition

In addition to the temporary assumption of the same number of the input and output channels, let us also assume that the input signal is full rank. Consequently, as will be shown, the residual signal is not necessary, i.e., $\mathbf{r} = \mathbf{0}$. Let us use a notation $\tilde{\mathbf{M}}$ in place of $\mathbf{M}$ to denote the adaptive mixing solution achieved with the temporary assumptions. From the simplified signal relation $\mathbf{y} = \tilde{\mathbf{M}}\mathbf{x}$ and the definition of the covariance matrices in Eq. (2), it follows that

$$\mathbf{C_y} = \mathrm{E}\left[\mathbf{y}\mathbf{y}^H\right] = \mathrm{E}\left[\tilde{\mathbf{M}}\mathbf{x}\mathbf{x}^H\tilde{\mathbf{M}}^H\right] = \tilde{\mathbf{M}}\mathbf{C_x}\tilde{\mathbf{M}}^H. \tag{9}$$

When the covariance matrices in Eq. (9) are decomposed according to Eq. (4) it follows that

$$\mathbf{K_y P_y P_y}^H \mathbf{K_y}^H = \tilde{\mathbf{M}} \mathbf{K_x P_x P_x}^H \mathbf{K_x}^H \tilde{\mathbf{M}}^H, \tag{10}$$

from which the set of solutions for $\tilde{\mathbf{M}}$ that fulfill Eq. (9) follows

$$\boxed{\tilde{\mathbf{M}} = \mathbf{K_y P_y P_x}^H \mathbf{K_x}^{-1} = \mathbf{K_y P K_x}^{-1}}. \tag{11}$$

The condition for these solutions is that $\mathbf{K_x}^{-1}$ exists. The unitary matrix $\mathbf{P} = \mathbf{P_y P_x}^H$ is the remaining free parameter. In Appendix A it is shown that the error measure in Eq. (6) is minimized and thus an optimal adaptive mixing solution is achieved when

$$\boxed{\mathbf{P} = \mathbf{V U}^H}, \tag{12}$$

where the unitary matrices $\mathbf{V}$ and $\mathbf{U}$ are from the singular value decomposition $\mathbf{USV}^H = \mathbf{K_x}^H \mathbf{Q}^H \mathbf{G_{\hat{y}}}^H \mathbf{K_y}$, and $\mathbf{S}$ is a non-negative diagonal matrix.

3.2 Adaptive Mixing Solution Generalized

In practice it is possible, and in some applications common, that $\mathbf{K_x}^{-1}$ in Eq. (11) does not exist, or that this inverse entails very large multipliers if some of the principal components in $\mathbf{x}$ are very small. An effective way to regularize the inverse is by the singular value decomposition $\mathbf{K_x} = \mathbf{U_x S_x V_x}^H$. The unregularized inverse is

$$\mathbf{K_x}^{-1} = \mathbf{V_x S_x}^{-1} \mathbf{U_x}^H. \tag{13}$$

The problems arise when some of the diagonal values of the non-negative diagonal matrix $\mathbf{S_x}$ are zero or very small. It is straightforward to manually tune a regularization function in this domain. In the example implementation, we first define $s_{\mathbf{x}_{\max}}$ as the largest diagonal element of $\mathbf{S_x}$ and then form the regularized diagonal matrix $\mathbf{S}'_x$ with entries

$$s'_{\mathbf{x}_{ii}} = \max\left[s_{\mathbf{x}_{ii}}, \alpha s_{\mathbf{x}_{\max}}\right], i = 1, \ldots, N. \tag{14}$$

where α is a factor between 0 and 1 that determines the amount of regularization. In the informal tests the value $\alpha = \frac{1}{5}$ was found a good compromise that avoids amplifying small energy signal components too much, but still keeping a good ability for the framework to take benefit of the available independent signal energy in the input channels. The regularized inverse is $\mathbf{K'_x}^{-1} = \mathbf{V_x S'_x}^{-1} \mathbf{U_x}^H$, and the corresponding regularized mixing matrix is $\mathbf{M} = \mathbf{K_y P K'_x}^{-1}$. The regularization effectively means that within the mixing process the amplification of some of the small principal components in $\mathbf{x}$ is reduced, and consequently their impact to the output signal $\mathbf{y}$ is also reduced, and the target covariance $\mathbf{C_y}$ is in general not reached.

The motivation for injecting the residual signal, which is discussed in the following, is to achieve $\mathbf{C_y}$ also in the regularized conditions. The residual signal $\mathbf{r}$ is incoherent with respect to $\mathbf{x}$, and its covariance matrix $\mathbf{C_r} = \mathrm{E}\left[\mathbf{rr}^H\right]$ can be derived by setting the target covariance condition

$$\begin{aligned} \mathbf{C_y} &= \mathrm{E}\left[(\mathbf{Mx} + \mathbf{r})(\mathbf{Mx} + \mathbf{r})^H\right] \\ &= \mathrm{E}\left[(\mathbf{Mxx}^H \mathbf{M}^H\right] + \mathrm{E}\left[\mathbf{rr}^H\right] \\ &= \mathbf{MC_x M}^H + \mathbf{C_r} \\ \Rightarrow & \boxed{\mathbf{C_r} = \mathbf{C_y} - \mathbf{MC_x M}^H}. \end{aligned} \tag{15}$$

Any signal that is independent with respect to $\mathbf{x}$ that is processed to have the covariance $\mathbf{C_r}$, serves as a residual signal that enables reconstructing the target covariance matrix $\mathbf{C_y}$ in situations where the regularization as in Eq. (14) was applied. As illustrated in Fig. 1, such a residual signal $\mathbf{r}$ can be generated by applying decorrelators to the prototype signal $\hat{\mathbf{y}}$ and processing the result using the proposed adaptive mixing solution to attain $\mathbf{C_r}$. In the residual mixing process, the matrix $\mathbf{Q_r}$ is defined as an identity matrix, since the prototype signal is the decorrelated signal itself. Note that although the residual signal is generated from the decorrelated signals, the residual covariance matrix $\mathbf{C_r}$ is typically not diagonal.

If the usage of decorrelators is to be completely avoided, e.g., in the higher frequency range where the human hearing is phase insensitive, at least the target channel energies can be achieved by multiplying the rows of $\mathbf{M}$ so that

$$\mathbf{M}' = \mathbf{GM}, \tag{16}$$

where $\mathbf{G}$ is a diagonal gain matrix with elements

$$\mathrm{g}_{ii} = \sqrt{\frac{\mathrm{c}_{\mathbf{y}_{ii}}}{\mathrm{c}_{\tilde{\mathbf{y}}_{ii}}}}, i = 1, \ldots, N, \tag{17}$$

where $\mathrm{c}_{\tilde{\mathbf{y}}_{ii}}$ is the ith diagonal element of $\mathbf{C_{\tilde{y}}} = \mathbf{MC_x M}^H$. Note that although the matrix $\mathbf{M}'$ is no more optimal in a mathematical sense, it is well-warranted in a perceptual sense.

3.3 Different Number of Input and Output Channels

In many applications the number of input and output channels is different. The regularization process presented in Section 3.2 provides robust mixing solutions also when some of the channels are zero-valued. The solution to the different channel numbers is thus directly available by assuming that the signal with fewer channels is extended to the higher dimension with additional zero-valued channels.

Such an approach however implies computational overhead since some rows or columns of the mixing matrix $\mathbf{M}$ correspond to the zero-valued channels. The same mixing result is obtained more efficiently by introducing a matrix Λ that is an identity matrix appended with zeros to dimension $N_y \times N_x$, where N_y is the number of the output channels, and N_x is the number of the input channels, e.g.,

$$\Lambda_{3\times 2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{18}$$

When $\mathbf{P}$ is redefined as

$$\mathbf{P} = \mathbf{V}\Lambda\mathbf{U}^H \tag{19}$$

the resulting $\mathbf{M}$ is a $N_y \times N_x$ mixing matrix that is the same as the relevant part of the $\mathbf{M}$ of the zero-padding case. Consequently, $\mathbf{C_x}$, $\mathbf{C_y}$, $\mathbf{K_x}$, and $\mathbf{K_y}$ can be of their natural dimension, and the prototype matrix $\mathbf{Q}$ is of dimension $N_y \times N_x$.

3.4 Decomposability of the Covariance Matrices

The input signal covariance matrix is always decomposable to $\mathbf{C_x} = \mathbf{K_x}\mathbf{K_x}^H$ because it is a positive semi-definite measure from an actual signal. It is, however, possible to define such target covariance matrices that are not decomposable for the reason that they represent impossible channel dependencies. There are methods to ensure decomposability, such as adjusting the negative eigenvalues to zeros and normalizing the energy [13]. However, the most meaningful usage of the proposed method is to request only valid covariance matrices.

4 EXAMPLE IMPLEMENTATION

In this section it is shown by an example how the proposed framework can be implemented in a practical spatial sound processing application. We design a processing rule for automatic stereo to surround upmixing, in which the direct and ambience components of the stereo signal are estimated and then distributed to the larger set of loudspeakers.

4.1 Prototype Matrix Q in Upmixing

Assuming 5.0 channel order L, R, C, Ls, Rs, a reasonable prototype matrix in stereo upmixing is

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \sqrt{0.5} & \sqrt{0.5} \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{20}$$

This definition means that the prototype signal for the two left channels is the original left channel, and similarly for the right, and that the prototype signal for the center channel is the mixture.

4.2 Target Covariance Matrix in Upmixing

Following the principles in [6], the direct-ambience model parameters describing each stereo time-frequency tile can be obtained: the energy of the panned direct component E_D, the energy of the incoherent ambience component E_A, and the panning angle of the direct component α_D. The target covariance matrix for each time-frequency tile is built as function of these parameters. A reasonable rule is to distribute the ambience energy to incoherently surround the listener and to amplitude pan the direct component using a pair from the three frontal channels.

A practical choice for distributing the ambience energy is incoherently using the channels L, R, Ls, and Rs. The center channel is omitted since it is assumed not to contribute much to the perceived spatial quality of the ambience elements. This procedure may improve the overall sound quality with the proposed method since having fewer independent output channels entails lesser relative amount of the decorrelated signal energy. According to this design principle, a covariance matrix describing the diffuse sound energy is

$$\mathbf{C}_\mathrm{A} = E_A \begin{bmatrix} 0.25 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & 0.25 \end{bmatrix}. \tag{21}$$

For the direct part, let us assume a column vector $\mathbf{v}$ that contains the vector base amplitude panning (VBAP) [14] gains corresponding to the panning angle α_D. Vector $\mathbf{v}$ has only two non-zero values for the pair of the front loudspeakers that is active in the panning. The direct part covariance matrix is

$$\mathbf{C}_\mathrm{D} = E_D \mathbf{v}\mathbf{v}^T. \tag{22}$$

Finally, since the direct and ambient parts were defined incoherent with respect to each other, the overall target covariance matrix is

$$\mathbf{C_y} = \mathbf{C}_\mathrm{A} + \mathbf{C}_\mathrm{D}. \tag{23}$$

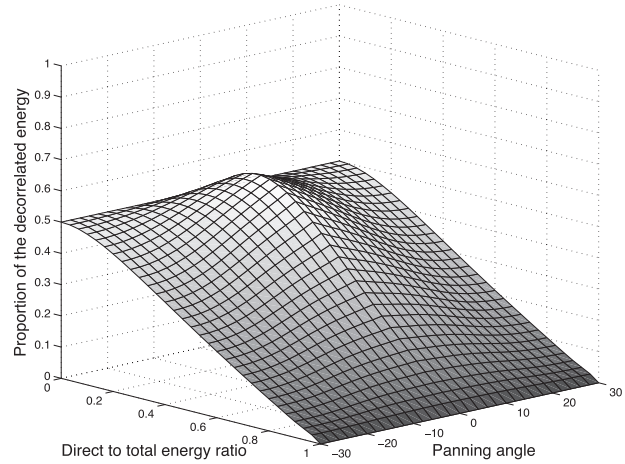

Fig. 2. The proportion of the output energy that is from the decorrelators with the example upmixing implementation.

4.3 Simulations—Applied Amount of the Decorrelated Sound Energy

A set of test scenarios was generated in Matlab by varying the panning angle and the direct-to-total energy ratio $E_D/(E_D + E_A)$ of a simulated stereo mix. In all conditions, the target covariance matrix was formulated using the example implementation in Appendix B, with the control matrices as in Eqs. (20)–(23). Fig. 2 shows the resulting relative amount of the decorrelated energy in the output signal in all simulated conditions. The maximum value 3/5 is reached at 0° semi-ambient condition, since in this condition the applied rule for the target covariance matrix generates a diagonal equal energy covariance matrix, which is the extreme in terms of the requested independent signal energy. In other conditions the target covariance matrix becomes

Table 1. Numerical examples of the proposed algorithm. When a signal with covariance is processed with a mixing matrix, and completed with a possible residual signal with, the output signal has the covariance. Although these numerical examples are static, the typical use case of the proposed method is dynamic. The channel order is assumed *L*, *R*, *C*, *Ls*, *Rs*(, *Lr*, *Rr*).

	Input parameter matrices			Formulated processing matrices	
Context	$\mathbf{C_x}$	$\mathbf{Q}$	$\mathbf{C_y}$	$\mathbf{M}$	$\mathbf{C_r}$
Decorrelation: High input ICC	$\begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1.5 & -0.75 \\ -0.75 & 1.5 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$
Decorrelation: Very high input ICC	$\begin{bmatrix} 1 & 0.97 \\ 0.97 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 2.1 & -1.4 \\ -1.4 & 2.1 \end{bmatrix}$	$\begin{bmatrix} 0.31 & -0.31 \\ -0.31 & 0.31 \end{bmatrix}$
Stereo upmixing	$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0.71 & 0.71 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.31 & -0.19 \\ -0.19 & 0.31 \\ 0.53 & 0.53 \\ 0.31 & -0.19 \\ -0.19 & 0.31 \end{bmatrix}$	$\begin{bmatrix} 0.36 & 0.11 & -0.18 & -0.14 & 0.11 \\ 0.11 & 0.36 & -0.18 & 0.11 & -0.14 \\ -0.18 & -0.18 & 0.32 & -0.18 & -0.18 \\ -0.14 & 0.11 & -0.18 & 0.36 & 0.11 \\ 0.11 & -0.14 & -0.18 & 0.11 & 0.36 \end{bmatrix}$
5.0 to 7.0 upmixing	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0 \\ 0 & 0 & 0 & 0 & 0.71 \\ 0 & 0 & 0 & 0.71 & 0 \\ 0 & 0 & 0 & 0 & 0.71 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.25 & 0 & -0.25 & 0 \\ 0 & 0 & 0 & 0 & 0.25 & 0 & -0.25 \\ 0 & 0 & 0 & -0.25 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & -0.25 & 0 & 0.25 \end{bmatrix}$
Downmixing: with some non-zero coherences	$\begin{bmatrix} 1 & 0 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 & 0 \\ 0.5 & 0 & 0 & 1 & 0 \\ 0 & 0.5 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0.71 & 1 & 0 \\ 0 & 1 & 0.71 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 2.5 & 0.5 \\ 0.5 & 2.5 \end{bmatrix}$	$\begin{bmatrix} 0.84 & 0.02 & 0.61 & 0.84 & 0.02 \\ 0.02 & 0.84 & 0.61 & 0.02 & 0.84 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$
Decorrelation: Standard 5.0 layout coincident hypercardioids in a diffuse field	$\begin{bmatrix} 1 & 0.65 & 0.91 & 0.43 & -0.22 \\ 0.65 & 1 & 0.91 & -0.22 & 0.43 \\ 0.91 & 0.91 & 1 & 0.07 & 0.07 \\ 0.43 & -0.22 & 0.07 & 1 & -0.22 \\ -0.22 & 0.43 & 0.07 & -0.22 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 2 & -0.51 & -0.83 & -0.53 & 0.41 \\ -0.51 & 2 & -0.83 & 0.41 & -0.53 \\ -0.83 & -0.83 & 2.1 & 0.04 & 0.04 \\ -0.53 & 0.41 & 0.04 & 1.2 & -0.07 \\ 0.41 & -0.53 & 0.04 & -0.07 & 1.2 \end{bmatrix}$	$\begin{bmatrix} 0.58 & -0.2 & -0.34 & -0.23 & 0.19 \\ -0.2 & 0.58 & -0.34 & 0.19 & -0.23 \\ -0.34 & -0.34 & 0.62 & 0.03 & 0.03 \\ -0.23 & 0.19 & 0.03 & 0.11 & -0.11 \\ 0.19 & -0.23 & 0.03 & -0.11 & 0.11 \end{bmatrix}$

sparser and thus there is less necessity for injecting decorrelated energy. In all conditions, the covariance rendering method ensures that the output signal is the closest to the prototype signal.

4.4 Informal Listening

The described upmixer was implemented for automatic processing of stereo music and movie recordings. The applied decorrelators were the band-wise pseudo-random delays as described in [15]. In informal listening the effect was the expected: The center amplitude panned sources were reproduced over the center loudspeaker, the ambience sounds such as the reverberation were expanded to surround the listener, and the overall sound quality was similar to that of the original stereo mix.

5 FURTHER NUMERICAL EXAMPLES

Table 1 shows a set of numerical examples to illustrate the behavior of the proposed method in different use cases. Note that for simplicity, only examples with real-valued covariance matrices were selected. The mixing matrices $\mathbf{M}$ and the residual matrices $\mathbf{C_r}$ were formulated with the Matlab code provided in the appendix based on the parameter matrices $\mathbf{C_x}$, $\mathbf{C_y}$, and $\mathbf{Q}$. Although the matrices are illustrated static, in typical applications they vary in time and frequency.

The first and the second row of the table illustrate a use case of stereo enhancement by means of adjusting the inter-channel coherence to zero. In the first row there is a small but reasonable incoherent component between the two channels, and thus a fully incoherent output is achieved with only channel mixing. In the second row, the input coherence is very high, i.e., the smaller principal component is very small. Amplifying this component in extreme degree is not desirable, and thus the built-in limiter starts to require the injection of the decorrelated energy instead, i.e., $\mathbf{C_r}$ is now non-zero.

The third row shows a case of stereo to 5.0 upmixing as described in Section 4. The residual signal is again non-zero since the dimension of the signal is increased.

The fourth row shows a case of simple 5.0 to 7.0 upmixing, where the original two rear channels are upmixed to the four new rear channels, incoherently. This example illustrates that the processing focuses on those channels where adjustments are requested and leaves the other channels unaffected.

The fifth row depicts a case of downmixing a 5.0 signal to stereo. Passive downmixing, such as applying a static downmixing matrix $\mathbf{Q}$, would amplify the coherent components over the incoherent components. Here the target covariance matrix was defined to preserve the energy, which is fulfilled by the resulting $\mathbf{M}$.

The sixth row illustrates the use case of coincident microphone spatial sound recording. The input covariance matrix $\mathbf{C_x}$ is the result of placing ideal hypercardioid microphones to an ideal diffuse field, facing toward the standard angles of the 5.0 setup. The large off-diagonal values in $\mathbf{C_x}$ illustrate an inherent disadvantage of passive first order coincident microphone techniques. In the ideal case, the covariance matrix best representing a diffuse field is diagonal, which was therefore set as the target. The relative amount of the resulting decorrelated energy in the output signal is exactly 2/5. This is because there are three independent signal components available in the first order horizontal coincident microphone signals, and two are to be added in order to reach the five-channel diagonal target covariance matrix.

6 CONCLUSION

In this paper we presented a generalized and optimized framework for time-frequency processing of spatial audio in the covariance matrix domain. The method takes into use the available independent signal components in the input channels and thus ensures that the decorrelation is applied only to the minimum necessary extent. The adaptive mixing solution was formulated to maximize the similarity of the output signal and the prototype signal describing the preferred signal content for each of the output channels.

The framework was applied to perform stereo upmixing, and simulations were provided to illustrate the optimized usage of the decorrelated energy in the different input signal conditions. A wide variety of further applications for the proposed framework have been identified.

7 ACKNOWLEDGMENTS

The authors would like to thank Fabian Küch, Giovanni Del Galdo, Emanuël Habets, and Ville Pulkki for their valuable feedback and support.

8 REFERENCES

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

[2] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K. S. Chong, "MPEG Surround—The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding," *J. Audio Eng. Soc*, vol. 56, pp. 932–955 (2008 Nov.).

[3] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric Coding of Stereo Audio," *EURASIP J. Applied Signal Processing*, vol. 2005, pp. 1305–1322 (2005).

[4] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilpert, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt, and H.-O. Oh, "MPEG Spatial Audio Object Coding—The ISO/MPEG Standard for Efficient Coding of Interactive Audio Scenes," *J. Audio Eng. Soc*, vol. 60, pp. 655–673 (2012 Sept.).

[5] V. Pulkki, "Spatial Sound Reproduction with Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 55, pp. 503–516(2007 June).

[6] C. Faller, "Multiple-Loudspeaker Playback of Stereo Signals," *J. Audio Eng. Soc.*, vol. 54, pp. 1051–1064 (2006 Nov.).

[7] C. Tournery, C. Faller, F. Küch, and J. Herre, "Converting Stereo Microphone Signals Directly to MPEG-Surround," presented at the *128th Convention of the Audio Engineering Society* (2010 May) , convention paper 7982.

[8] J. Merimaa and V. Pulkki, "Spatial Impulse Response Rendering I: Analysis and Synthesis," *J. Audio Eng. Soc*, vol. 53, pp. 1115–1127 (2005 Dec.).

[9] F. Baumgarte and C. Faller, "Binaural Cue Coding—Part I: Psychoacoustic Fundamentals and Design Principles," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 509–519 (2003).

[10] B. R. Glasberg and B. C. Moore, "Derivation of Auditory Filter Shapes from Notched-Noise Data," *Hearing Research*, vol. 47, no. 1, pp. 103–138 (1990).

[11] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The Precedence Effect," *J Acous. Soc. Am.*, vol. 106, p. 1633 (1999).

[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, vol. 3 (JHUP, 2012).

[13] R. Rebonato and P. Jäckel, "The Most General Methodology for Creating a Valid Correlation Matrix for Risk Management and Option Pricing Purposes," *Journal of Risk*, vol. 2, no. 2, pp. 17–28 (1999).

[14] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, pp. 456–466 (1997 June).

[15] J. Vilkamo, "Spatial Sound Reproduction with Frequency Band Processing of B-Format Audio Signals," Master's thesis, Helsinki University of Technology (2008).

APPENDIX A. MINIMIZING $\|\mathbf{G}_{\hat{\mathbf{y}}}\hat{\mathbf{y}} - \mathbf{y}\|^2$ WITH RESPECT TO UNITARY P

From the set of mixing solutions in Eq. (11), we now derive the optimal solution that minimizes the error measure in Eq. (6). Let us consider a signal $\mathbf{w} = \mathbf{K_x}^{-1}\mathbf{x}$, which has the identity covariance matrix

$$\begin{aligned}\mathrm{E}[\mathbf{w}\mathbf{w}^H] &= \mathrm{E}[\mathbf{K_x}^{-1}\mathbf{x}\mathbf{x}^H(\mathbf{K_x}^{-1})^H] \\ &= \mathbf{K_x}^{-1}\mathrm{E}[\mathbf{x}\mathbf{x}^H](\mathbf{K_x}^{-1})^H \\ &= \mathbf{K_x}^{-1}\mathbf{C_x}(\mathbf{K_x}^{-1})^H \\ &= \mathbf{I}. \end{aligned} \tag{24}$$

Replacing $\mathbf{w}$ in the input-output relation $\mathbf{y} = \tilde{\mathbf{M}}\mathbf{x}$ yields

$$\mathbf{y} = \tilde{\mathbf{M}}\mathbf{x} = \tilde{\mathbf{M}}\mathbf{K_x}\mathbf{w} = \mathbf{K_y}\mathbf{P}\mathbf{w}. \tag{25}$$

The error measure in Eq. (6) can then be written

$$\begin{aligned} e &= \mathrm{E}\left[\|\mathbf{G}_{\hat{\mathbf{y}}}\hat{\mathbf{y}} - \mathbf{y}\|^2\right] \\ &= \mathrm{E}\left[\|\mathbf{G}_{\hat{\mathbf{y}}}\mathbf{Q}\mathbf{x} - \tilde{\mathbf{M}}\mathbf{x}\|^2\right] \\ &= \mathrm{E}\left[\|\mathbf{G}_{\hat{\mathbf{y}}}\mathbf{Q}\mathbf{K_x}\mathbf{w} - \mathbf{K_y}\mathbf{P}\mathbf{w}\|^2\right] \\ &= \mathrm{E}\left[\|(\mathbf{G}_{\hat{\mathbf{y}}}\mathbf{Q}\mathbf{K_x} - \mathbf{K_y}\mathbf{P})\mathbf{w}\|^2\right] \\ &= \mathrm{E}\left[\mathbf{w}^H(\mathbf{G}_{\hat{\mathbf{y}}}\mathbf{Q}\mathbf{K_x} - \mathbf{K_y}\mathbf{P})^H(\mathbf{G}_{\hat{\mathbf{y}}}\mathbf{Q}\mathbf{K_x} - \mathbf{K_y}\mathbf{P})\mathbf{w}\right]. \end{aligned} \tag{26}$$

From $\mathrm{E}[\mathbf{w}\mathbf{w}^H] = \mathbf{I}$, we can readily show for a matrix $\mathbf{A}$ that $\mathrm{E}\left[\mathbf{w}^H\mathbf{A}\mathbf{w}\right] = \mathrm{tr}(\mathbf{A})$, which is the matrix trace. It follows that Eq. (26) takes the form

$$e = \mathrm{tr}\left[(\mathbf{G}_{\hat{\mathbf{y}}}\mathbf{Q}\mathbf{K_x} - \mathbf{K_y}\mathbf{P})^H(\mathbf{G}_{\hat{\mathbf{y}}}\mathbf{Q}\mathbf{K_x} - \mathbf{K_y}\mathbf{P})\right]. \tag{27}$$

For matrix traces, we can readily confirm that

$$\begin{aligned}\mathrm{tr}(\mathbf{A}+\mathbf{B}) &= \mathrm{tr}(\mathbf{A})+\mathrm{tr}(\mathbf{B})\\ \mathrm{tr}(\mathbf{A}) &= \mathrm{tr}(\mathbf{A}^H)\\ \mathrm{tr}(\mathbf{P}^H\mathbf{A}\mathbf{P}) &= \mathrm{tr}(\mathbf{A}).\end{aligned} \tag{28}$$

Using these properties, Eq. (27) takes the form

$$\begin{aligned}e = &\mathrm{tr}(\mathbf{K_x}^H\mathbf{Q}^H\mathbf{G}_{\hat{\mathbf{y}}}{}^H\mathbf{G}_{\hat{\mathbf{y}}}\mathbf{Q}\mathbf{K_x})+\mathrm{tr}(\mathbf{K_y}^H\mathbf{K_y})\\ &-2\mathrm{tr}(\mathbf{K_x}^H\mathbf{Q}^H\mathbf{G}_{\hat{\mathbf{y}}}{}^H\mathbf{K_y}\mathbf{P}).\end{aligned} \tag{29}$$

Only the last term depends on $\mathbf{P}$. The optimization problem is thus

$$\mathbf{P} = \arg\min_{\mathbf{P}} e = \arg\max_{\mathbf{P}}[\mathrm{tr}(\mathbf{K_x}^H\mathbf{Q}^H\mathbf{G}_{\hat{\mathbf{y}}}{}^H\mathbf{K_y}\mathbf{P})]. \tag{30}$$

We can readily show for a non-negative diagonal matrix $\mathbf{S}$ and any unitary matrix $\mathbf{P}_s$ that

$$\mathrm{tr}(\mathbf{S}) \geq \mathrm{tr}(\mathbf{S}\mathbf{P}_s). \tag{31}$$

Thereby, by defining the singular value decomposition $\mathbf{U}\mathbf{S}\mathbf{V}^H = \mathbf{K_x}^H\mathbf{Q}^H\mathbf{G}_{\hat{\mathbf{y}}}{}^H\mathbf{K_y}$, where $\mathbf{S}$ is non-negative and diagonal and $\mathbf{U}$ and $\mathbf{V}$ are unitary, it follows that

$$\begin{aligned}\mathrm{tr}(\mathbf{S}) &\geq \mathrm{tr}(\mathbf{S}\mathbf{V}^H\mathbf{P}\mathbf{U}) = \mathrm{tr}(\mathbf{U}\mathbf{S}\mathbf{V}^H\mathbf{P}\mathbf{U}\mathbf{U}^H)\\ &= \mathrm{tr}(\mathbf{K_x}^H\mathbf{Q}^H\mathbf{G}_{\hat{\mathbf{y}}}{}^H\mathbf{K_y}\mathbf{P}),\end{aligned} \tag{32}$$

for any unitary $\mathbf{P}$. The equality holds for

$$\boxed{\mathbf{P} = \mathbf{V}\mathbf{U}^H}, \tag{33}$$

whereby this choice for $\mathbf{P}$ yields the maximum of $\mathrm{tr}(\mathbf{K_x}^H\mathbf{Q}^H\mathbf{G}_{\hat{\mathbf{y}}}{}^H\mathbf{K_y}\mathbf{P})$ and the minimum of the error measure in Eq. (6).

APPENDIX B. EXAMPLE MATLAB CODE

The following Matlab code was used in the numerical examples and provides the general functionality of the proposed method.

Listing 1: Matlab implementation of the proposed method.

```
function [M,Cr]=formulate_M_and_Cr(Cx,Cy,Q,flag)
% flag = 0: Expect usage of residuals
% flag = 1: Fix energies instead
lambda=eye(length(Cy),length(Cx));

% Decomposition of Cy
[U_Cy,S_Cy]=svd(Cy);
Ky=U_Cy*sqrt(S_Cy);

% Decomposition of Cx
[U_Cx,S_Cx]=svd(Cx);
Kx=U_Cx*sqrt(S_Cx);

%SVD of Kx
Ux=U_Cx;
Sx=sqrt(S_Cx);
% Vx = identity matrix

% Regularization Sx
Sx_diag=diag(Sx);
limit=max(Sx_diag)*0.2+1e-20;
Sx_reg_diag=max(Sx_diag,limit);

% Formulate regularized Kx^-1
Kx_reg_inverse=diag(1./Sx_reg_diag)*Ux';

% Formulate normalization matrix G_hat
Cy_hat_diag=diag(Q*Cx*Q');
limit=max(Cy_hat_diag)*0.001+1e-20;
Cy_hat_diag=max(Cy_hat_diag,limit);
G_hat=diag(sqrt(diag(Cy)./Cy_hat_diag));

% Formulate optimal P
[U,S,V]=svd(Kx'*Q'*G_hat'*Ky);
P=V*lambda*U';

% Formulate M
M=Ky*P*Kx_reg_inverse;

% Formulate residual covariance matrix
Cy_tilde = M*Cx*M';
Cr=Cy-Cy_tilde;

% Use energy compensation instead of residuals
if flag==1
    adjustment=diag(Cy)./diag(Cy_tilde + 1e-20);
    G=diag(sqrt(adjustment));
    M=G*M;
    Cr='unnecessary';
end
```

THE AUTHORS


Juha Vilkamo


Tom Bäckström


Achim Kuntz

Juha Vilkamo studied electrical engineering at former Helsinki University of Technology, now Aalto University, Finland. He received his M.Sc. degree in acoustics and audio signal processing in 2008. He worked between 2008 and 2011 as a researcher for Fraunhofer IIS, Germany, in the fields of binaural technologies and spatial sound reproduction. Currently he is pursuing his D.Sc. in Aalto University, in a research collaboration project with Fraunhofer IIS. His enthusiasm at work is in developing optimized and perceptually motivated audio signal processing solutions. His delight in life is his lovely wife and two daughters.

●

Tom Bäckström is professor of Speech Coding at the International Audio Laboratories Erlangen, which is a joint research unit between the Friedrich-Alexander University and Fraunhofer IIS, Erlangen, Germany. He received his M.Sc. and D.Sc. (tech.) degrees from Helsinki University of Technology, Finland, in 2001 and 2004, respectively, and continued working there until 2008. Since then he has worked at the AudioLabs Erlangen, first as a researcher and from 2013 as professor. His research interests include speech and audio coding and perception, speech production, time-frequency analysis, matrix and polynomial algebra.

●

Achim Kuntz received his diploma degree in electrical engineering from the University of Erlangen-Nuremberg in 2002. He then joined the Telecommunications Laboratory at the same university carrying out studies on spatial sound reproduction, multidimensional signals and wave field analysis techniques, resulting in his doctoral degree in 2008. He is currently a researcher at the Fraunhofer Institute for Integrated Circuits IIS in Erlangen and member of the International Audio Laboratories Erlangen. His current primary field of interest are perceptually motivated signal processing algorithms for audio applications including the acquisition, manipulation, coding, and reproduction of spatial sound.