Minimization of Decorrelator Artifacts in Directional Audio Coding by Covariance Domain Rendering

JUHA VILKAMO AND VILLE PULKKI, *AES Fellow*
(juha.vilkamo@aalto.fi) (ville.pulkki@aalto.fi)

Aalto University, Espoo, Finland

Directional Audio Coding (DirAC) is a perceptually motivated method for spatial audio reproduction that models the sound field as a combination of a plane wave and a surrounding diffuse field, in a time-frequency resolution that approximates that of the human spatial hearing. The reproduction of the diffuse sound requires incoherence between the loudspeaker signals, which is typically achieved using decorrelators. The decorrelators, however, have the drawback of compromising the overall sound quality in sound scenes that prominently contain signals with impulsive fine structure, such as applause and speech. In this paper, we apply a recently proposed covariance domain spatial sound processing framework to optimize the DirAC synthesis. The framework takes into use the available independent signal components within the microphone channels, and by these means minimizes the necessary amount of the applied decorrelated sound energy. Listening tests showed that the procedure improves the overall perceived quality of DirAC rendering when there are a number of coincident or spaced microphone signals available, and when the signal content is such that is perceivably affected by the decorrelators.

0 INTRODUCTION

The psychoacoustic knowledge in human spatial hearing [1] has given foundation to the field of perceptually motivated spatial audio processing techniques. These include efficient multi-channel transmission [2], flexible audio object-based transmission [3], stereo upmixing [4], and perceptual reproduction of recorded sound fields [5], to name a few. The underlying assumption in these techniques is that the spatial perception is largely based on a set of inter-aural spatial cues in frequency bands: the inter-aural level difference (ILD), the inter-aural time-difference (ITD), the inter-aural coherence (IC), and the monaural energy envelope.

Directional Audio Coding (DirAC) [5] is a perceptually motivated method for spatial sound reproduction that reproduces those sound field characteristics that contribute to the spatial cues. The sound field is modeled in frequency bands as a combination of a plane wave and a diffuse sound field. In the receiver end, a sound field with similar characteristics is reproduced using the available loudspeakers, producing also a spatial perception that is similar to that of the original space. Listening tests have confirmed that the approach provides spatial sound reproduction with high perceptual quality [6].

When DirAC is configured to best preserve the spatial characteristics of the sound field, it employs decorrelators to build the incoherent loudspeaker signals required in reproduction of the diffuse sound elements. The exceptions to this are when there is a set of largely spaced incoherent microphone signals available [7], or that when the spatial accuracy is intentionally compromised by not applying decorrelators [8]. Otherwise, i.e., when accurate spatial sound reproduction is sought and the available set of microphones is compact, the decorrelators are necessary.

The drawback of the decorrelators is that they smear the sound over a short time interval, which can perceivably affect the overall sound quality of the signals with impulsive fine structure, such as speech and applause. In this paper, we optimize the overall sound quality of DirAC reproduction focusing on this aspect, by performing the synthesis using a recently proposed covariance domain spatial sound rendering framework [9]. The framework is characterized by the ability to take into use the available independent signal components within the input channels, and by these means to minimize the necessary amount of the decorrelated sound energy, while preserving the intended spatial image. In [10], it was shown that the quality of the reproduction of the applause signals can also be improved by applying a multi-resolution STFT filterbank and transient bypass decorrelators. Such an approach is signal specific, and applicable in parallel with the proposed method, and is not included in the subsequent comparisons.

The paper is organized as follows. First, we give the definitions and explain the fundamentals of DirAC and the covariance rendering method. Then the two are combined,

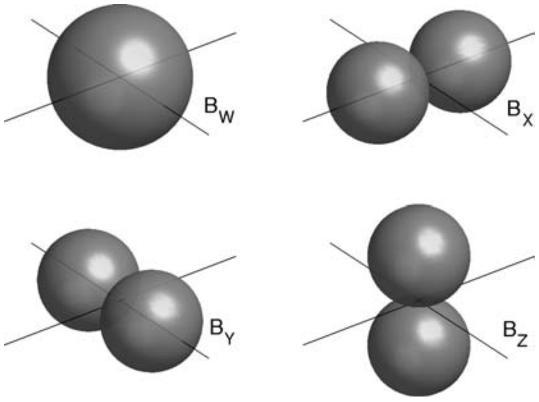

Fig. 1. B-format microphone directivity patterns and alignment.

followed by a description of the implementation, simulations, listening test, results, discussion, and finally the conclusion.

1 BACKGROUND

1.1 Notation

Matrices and vectors are denoted with bold faced symbols, where uppercase denotes a matrix. Matrix elements are marked with row and column subindices i and j, such that $c_{\mathbf{b}ij}$ is an element of matrix $\mathbf{C_b}$. Let us define a coincident B-format microphone signal that contains an omnidirectional microphone signal $b_{\mathrm{W}}(t)$ and the figure-of-eight signals $b_{\mathrm{X}}(t)$, $b_{\mathrm{Y}}(t)$ and $b_{\mathrm{Z}}(t)$ oriented orthogonally as in Fig. 1. The three figure-of-eight components are scaled to have the gain of $\sqrt{2}$ in the look direction. The complex valued short-time Fourier transform (STFT) spectra of these signals are denoted $B_{\mathrm{W}}(k,l)$, $B_{\mathrm{X}}(k,l)$, $B_{\mathrm{Y}}(k,l)$ and $B_{\mathrm{Z}}(k,l)$, where k is the downsampled time index and l is the frequency index. The B-format time-frequency sample vector is

$$\mathbf{b}(k,l) = \begin{bmatrix} B_{\mathrm{W}}(k,l) \\ B_{\mathrm{X}}(k,l) \\ B_{\mathrm{Y}}(k,l) \\ B_{\mathrm{Z}}(k,l) \end{bmatrix}. \tag{1}$$

With a horizontal B-format microphone, the vertical component $B_{\mathrm{Z}}(k,l)$ is omitted. Let us similarly define the signals $s_i(t)$ from an array of spaced omnidirectional microphones, where $i = 1, \ldots, N_s$ and N_s is the number of microphone capsules. The frequency transformed signals are $S_i(k,l)$, and the time-frequency sample vector is

$$\mathbf{s}(k,l) = \begin{bmatrix} S_1(k,l) \\ S_2(k,l) \\ \vdots \\ S_{N_s}(k,l) \end{bmatrix}. \tag{2}$$

The N_y-channel output time-frequency sample vector is

$$\mathbf{y}(k,l) = \begin{bmatrix} Y_1(k,l) \\ Y_2(k,l) \\ \vdots \\ Y_{N_y}(k,l) \end{bmatrix}. \tag{3}$$

The time and frequency indices (k, l) are now omitted for brevity of notation. The covariance matrices of the signals in Eqs. (1)–(3) are

$$\begin{aligned} \mathbf{C_b} &= \mathrm{E}\left[\mathbf{b}\mathbf{b}^H\right] \\ \mathbf{C_s} &= \mathrm{E}\left[\mathbf{s}\mathbf{s}^H\right] \\ \mathbf{C_y} &= \mathrm{E}\left[\mathbf{y}\mathbf{y}^H\right] \end{aligned} \tag{4}$$

where $\mathbf{b}^H$ is the conjugate transpose of $\mathbf{b}$, and E[ ] is the expectation operator. In practical implementations the expectation is replaced with the mean operator over several time-frequency samples over time and/or frequency, in a resolution that approximates that of the human spatial hearing. Such an averaging area is here referred to as time-frequency tile.

1.2 DirAC analysis

DirAC models the sound field in frequency bands dynamically as a combination of a plane wave and a diffuse field. These are expressed by the parameters direction-of-arrival *DOA* of the plane wave, and diffuseness ψ, which is a ratio between 0 and 1 expressing the proportion of the overall sound field energy that is diffuse. To obtain these parameters, let us first define

$$\mathbf{i} = \frac{1}{\sqrt{2}}\Re\left\{\begin{bmatrix} c_{\mathbf{b}21} \\ c_{\mathbf{b}31} \\ c_{\mathbf{b}41} \end{bmatrix}\right\} \tag{5}$$

$$e = \frac{1}{2}c_{\mathbf{b}11} + \frac{1}{4}\sum_{i=2}^{4} c_{\mathbf{b}ii}. \tag{6}$$

The vector $\mathbf{i}$ is linearly related to the sound field intensity, but with an opposite sign, and e is linearly related to the sound field energy density. The linear relation is the result of omitting the constant multipliers of the acoustic impedance, the speed of sound and the mean density of the air from Eqs. (5) and (6). These constants do not affect the final DirAC parameters. *DOA* is the direction of the vector $\mathbf{i}$, and

$$\psi = 1 - \frac{\|\mathbf{i}\|}{e}. \tag{7}$$

The parameters *DOA* and ψ can be derived also with other microphone setups than B-format, such as the spaced microphone arrays [8, 11] and coincident stereo first order microphones [12].

1.3 DirAC legacy synthesis

In the synthesis part, the sound field corresponding to the analyzed *DOA* and ψ is reproduced using the available loudspeakers. The legacy method [5,6] is to divide the audio signal into the nondiffuse and diffuse streams using multipliers $\sqrt{1-\psi}$ and $\sqrt{\psi}$, amplitude panning the nondiffuse part of the sound energy to the analyzed *DOA*, and reproducing the diffuse part of the sound in all directions incoherently using decorrelators.

Additionally, there have been two distinct types of DirAC synthesis. The first performs the rendering from only an omnidirectional microphone component [8], and the second uses virtual directional microphones that are derivable from the B-format microphone signal [6]. In informal testing, the latter has provided better overall sound quality due to the option of implementing more subtle decorrelators since the channel signals are inherently partly incoherent. Furthermore, better source stability in scenarios with concurrent sources has been observed with B-format input due to the better initial source separation.

1.4 Covariance rendering method

The covariance rendering method [9] is essentially an optimized and versatile framework for time-frequency spatial sound processing. Its principle is to translate the processing task into the covariance matrix domain, and provide an optimized mixing solution to perform the processing task with the highest quality. The practical benefit for DirAC is its ability to take into use the available independent signal components in the input channels, allowing by these means the minimization of the amount of the applied decorrelated sound energy, while preserving the intended spatial sound image.

Let us define the N_x channel input signal vector $\mathbf{x}$ and its covariance matrix $\mathbf{C_x}$. The covariance rendering method provides the mixing solution as function of the following parameter matrices in each time-frequency tile:

- $\mathbf{C_x}$ - the input signal covariance matrix,
- $\mathbf{C_y}$ - the target covariance matrix, containing the target spatial sound characteristics, and
- $\mathbf{Q}$ - the $N_y \times N_x$ prototype matrix that defines the prototype signal $\hat{\mathbf{y}} = \mathbf{Qx}$. The matrix $\mathbf{Q}$ is usually defined time-invariant.

Typically $\mathbf{C}_{\hat{\mathbf{y}}} = \mathrm{E}\left[\hat{\mathbf{y}}\hat{\mathbf{y}}^H\right] \neq \mathbf{C_y}$. The covariance rendering method formulates a mixing solution with the input–output relation

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

where $\mathbf{M}$ is the optimal mixing matrix and $\mathbf{r}$ is a residual output signal. Matrix $\mathbf{M}$ is formulated by setting a boundary condition that the output signal $\mathbf{y}$ has the target covariance $\mathbf{C_y}$, and while so, the error energy $e = \|\mathbf{G}_{\hat{\mathbf{y}}}\hat{\mathbf{y}} - \mathbf{y}\|^2$ is minimized, where $\mathbf{G}_{\hat{\mathbf{y}}}$ is a diagonal matrix that normalizes the channel energies of $\hat{\mathbf{y}}$ to those of $\mathbf{y}$. The method also formulates a residual covariance matrix $\mathbf{C_r} = \mathrm{E}\left[\mathbf{rr}^H\right]$ of the residual signal $\mathbf{r}$, which compensates for the remaining difference if the target covariance $\mathbf{C_y}$ is not achievable by only channel mixing. This occurs when there are too few independent signal components available in the input signals. The residual signal is adaptively mixed from the decorrelated versions of the input signals, as is described in the end of this section.

The steps to formulate the mixing solution according to the covariance rendering method are now given in a compact form. A detailed explanation, derivation, and the example code can be found in [9]. The task is to solve $\mathbf{M}$ and $\mathbf{C_r}$ based on the input parameter matrices $\mathbf{C_x}$, $\mathbf{C_y}$ and $\mathbf{Q}$. This is achieved with the following steps:

1. Decomposition of the input and target covariance matrices to matrices $\mathbf{K_x}$ and $\mathbf{K_y}$ fulfilling the following property (e.g., Cholesky decomposition)

$$\begin{aligned}\mathbf{C_x} &= \mathbf{K_x}\mathbf{K_x}^H\\ \mathbf{C_y} &= \mathbf{K_y}\mathbf{K_y}^H.\end{aligned} \tag{9}$$

2. Formulation of the diagonal normalizer matrix $\mathbf{G}_{\hat{\mathbf{y}}}$, with elements

$$g_{\hat{\mathbf{y}}ii} = \sqrt{\frac{c_{\mathbf{y}ii}}{c_{\hat{\mathbf{y}}ii}}},\ i = 1, \ldots, N_y, \tag{10}$$

where $c_{\hat{\mathbf{y}}ii}$ are the diagonal elements of $\mathbf{C}_{\hat{\mathbf{y}}} = \mathbf{QC_xQ}^H$.

3. Singular value decompositions

$$\begin{aligned}\mathbf{USV}^H &= \mathbf{K_x}^H\mathbf{Q}^H\mathbf{G}_{\hat{\mathbf{y}}}^H\mathbf{K_y}\\ \mathbf{U_xS_xV_x}^H &= \mathbf{K_x}.\end{aligned} \tag{11}$$

4. For stability, the regularization of the real-valued diagonal nonnegative matrix $\mathbf{S_x}$ using a manually tuned function, forming the regularized diagonal matrix $\mathbf{S}'_\mathbf{x}$. In the test implementation, the applied function is

$$s'_{\mathbf{x}_{ii}} = \max\left[s_{\mathbf{x}_{ii}}, \alpha s_{\mathbf{x}_{\max}}\right],\ i = 1, \ldots, N_x, \tag{12}$$

where $\alpha = \frac{1}{5}$ and $s_{\mathbf{x}_{\max}}$ is the maximum value of $\mathbf{S_x}$.

5. Formulation of the regularized inverse

$$\mathbf{K}'^{-1}_\mathbf{x} = \mathbf{V_x}\mathbf{S}'^{-1}_\mathbf{x}\mathbf{U_x}^H. \tag{13}$$

6. Definition of a matrix $\mathbf{\Lambda}$ which is an identity matrix zero-padded to dimension $N_y \times N_x$, and

$$\mathbf{P} = \mathbf{V\Lambda U}^H. \tag{14}$$

7. Computation of the mixing matrix

$$\mathbf{M} = \mathbf{K_y}\mathbf{P}\mathbf{K}'^{-1}_\mathbf{x}. \tag{15}$$

8. Finally, the computation of the residual covariance matrix

$$\mathbf{C_r} = \mathbf{C_y} - \mathbf{MC_xM}^H. \tag{16}$$

The residual signal $\mathbf{r}$ with covariance $\mathbf{C_r}$ is generated with the following steps:

1. Generation of the prototype signal $\hat{\mathbf{y}} = \mathbf{Qx}$.
2. Decorrelation of each channel of $\hat{\mathbf{y}}$ with an incoherent decorrelator that has unity frequency response.
3. Estimation of the covariance matrix of the decorrelated signal. A computationally efficient estimate is the diagonal of $\mathbf{C}_{\hat{\mathbf{y}}}$.

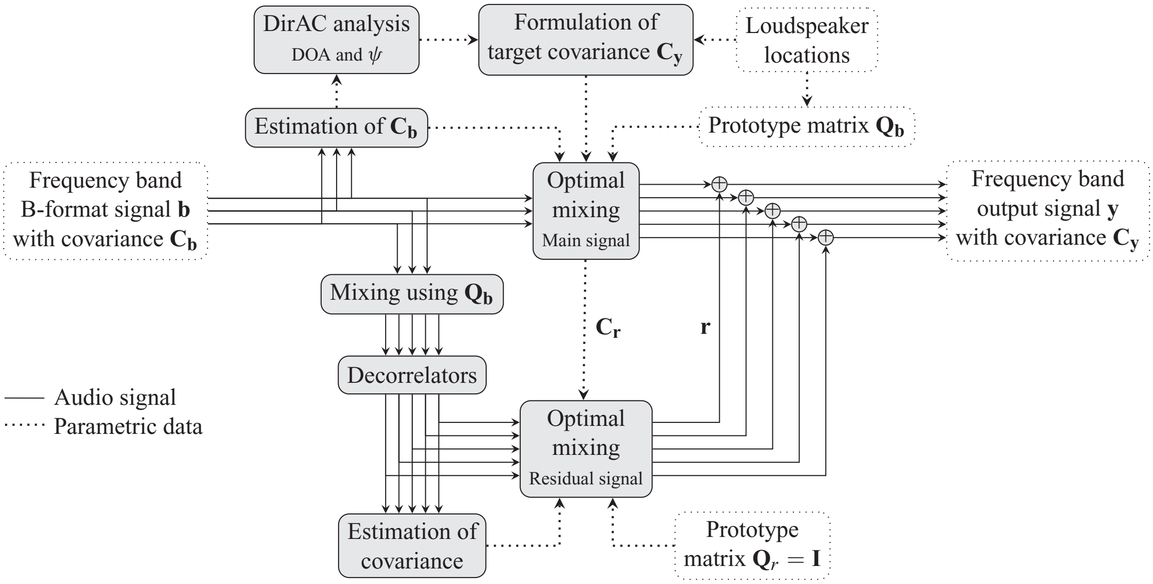
DirAC analysis
DOA and ψ
Formulation of target covariance $\mathbf{C_y}$
Loudspeaker locations
Estimation of $\mathbf{C_b}$
Prototype matrix $\mathbf{Q_b}$
Frequency band B-format signal **b** with covariance $\mathbf{C_b}$
Optimal mixing
Main signal
Frequency band output signal **y** with covariance $\mathbf{C_y}$
Mixing using $\mathbf{Q_b}$
$\mathbf{C_r}$
r
Decorrelators
Audio signal
Parametric data
Optimal mixing
Residual signal
Estimation of covariance
Prototype matrix $\mathbf{Q}_r = \mathbf{I}$

Fig. 2. DirAC processing from a B-format signal using the covariance rendering method. The main optimal mixing block takes into use the available signal components in the input channels. The residual signal **r** compensates for the remaining difference when there are not enough independent signal components available otherwise.

4. Setting $\mathbf{Q} = \mathbf{I}$ and $\mathbf{C_y} = \mathbf{C_r}$, and generation of a mixing solution using steps in Eqs. (9)–(15). Eq. (16) is now unnecessary since there are enough independent signal components in the input.
5. Processing of the decorrelated signals with the resulting mixing matrix, and mixing to the overall output as in Eq. (8).

2. COMBINING DIRAC AND THE COVARIANCE RENDERING

The block diagram of the combined process of performing DirAC rendering using the covariance rendering method is shown in Fig. 2. DirAC analysis is applied to the input signal covariance matrix $\mathbf{C_b}$ to obtain *DOA* and ψ, which in turn control the target covariance matrix $\mathbf{C_y}$, as described in the next section. The selection of the prototype matrix $\mathbf{Q}$ is described in Section 2.2.

2.1 Target covariance matrix in DirAC synthesis

The target covariance matrix $\mathbf{C_y}$ corresponding to the analyzed *DOA* and ψ can be built by formulating the non-diffuse and the diffuse target covariance matrices $\mathbf{C_y}^{\mathrm{ND}}$ and $\mathbf{C_y}^{\mathrm{D}}$ separately, and adding them together. This is possible since the direct and diffuse components are according to the applied sound field model incoherent in respect to each other.

The matrix $\mathbf{C_y}^{\mathrm{ND}}$ is built with the following steps. Let us assume a column vector $\mathbf{v}$(DOA) that contains the vector base amplitude panning (VBAP) [13] gains corresponding to the analyzed *DOA*. $\mathbf{v}$(DOA) has maximum two nonzero values in case of a horizontal loudspeaker setup, and three in case of a 3D loudspeaker setup. The panning covariance matrix is formulated $\mathbf{C_v}(\mathrm{DOA}) = \mathbf{v}(\mathrm{DOA})\mathbf{v}^T(\mathrm{DOA})$, and

$$\mathbf{C_y}^{\mathrm{ND}} = (1-\psi)R\mathbf{C_v}(\mathrm{DOA}), \tag{17}$$

where R is an estimate of the total sound energy of the time-frequency tile. With B-format rendering the estimate can be $R = c_{\mathbf{b}11}$, and with spaced omni microphone array, e.g., $R = \frac{1}{N_s}\sum_{i=1}^{N_s} c_{\mathbf{s}ii}$.

The matrix $\mathbf{C_y}^{\mathrm{D}}$ is built by defining a diagonal $N_y \times N_y$ energy distributor matrix $\mathbf{D}$ with nonnegative entries d_{ii} on the diagonal, so that $\sum_{i=1}^{Ny} d_{ii} = 1$, and

$$\mathbf{C_y}^{\mathrm{D}} = \psi R\mathbf{D}. \tag{18}$$

The matrix $\mathbf{C_y}^{\mathrm{D}}$ is diagonal, which means that the diffuse part of the target covariance matrix is defined incoherent between all channel pairs. The distributor matrix $\mathbf{D}$ can be adjusted to fit to the loudspeaker layout by setting values d_{ii} for each loudspeaker to a value that is inversely proportional to the loudspeaker density at the corresponding direction. For an even loudspeaker layout all values $d_{ii} = \frac{1}{N_y}$. Finally, the target covariance matrix for the time-frequency tile is

$$\mathbf{C_y} = \mathbf{C_y}^{\mathrm{ND}} + \mathbf{C_y}^{\mathrm{D}}. \tag{19}$$

2.2 Prototype signal in DirAC synthesis

The best choice for a prototype matrix $\mathbf{Q}$ defining the prototype signal $\hat{\mathbf{y}} = \mathbf{Qx}$ depends on the microphone and the loudspeaker configurations. If these configurations are static, $\mathbf{Q}$ can be defined time-invariant. For B-format rendering let us define $\mathbf{Q}_{\mathrm{B}}$ for a loudspeaker layout with azimuths θ_i and elevations ϕ_i, where $1 \le i \le N_y$ is the channel index. A reasonable prototype matrix is a virtual directional

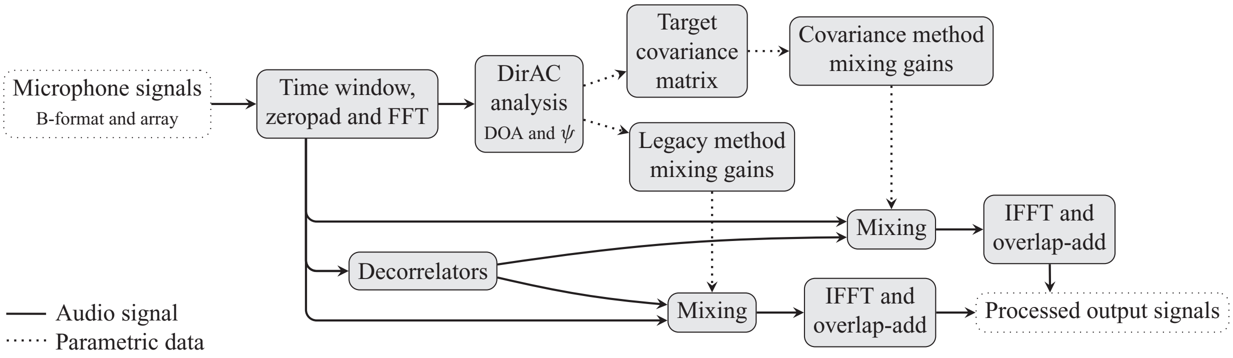

Fig. 3. High-level block diagram of the implementation applied in processing the test items. All reproduction modes share the same resolution, spatial analysis, and decorrelators, but differ in terms of the input signal configuration and the applied mixing gains.

microphone matrix with the look directions towards the directions of the loudspeakers

$$\mathbf{Q}_{\mathrm{B}} = \begin{bmatrix} \mathbf{q}_1^T \\ \mathbf{q}_2^T \\ \vdots \\ \mathbf{q}_{N_y}^T \end{bmatrix}, \tag{20}$$

where

$$\mathbf{q}_i = \begin{bmatrix} \frac{2-\kappa}{2} \\ \frac{\kappa}{2\sqrt{2}} \cos(\theta_i)\cos(\phi_i) \\ \frac{\kappa}{2\sqrt{2}} \sin(\theta_i)\cos(\phi_i) \\ \frac{\kappa}{2\sqrt{2}} \sin(\phi_i) \end{bmatrix}, \tag{21}$$

where κ is a constant between 0 and 2 defining the directivity pattern of the virtual directional microphone. A good choice is $\kappa = 1.5$ since it corresponds to a hypercardioid that maximizes the energy ratio of the look direction in respect to the other directions. The prototype signal in B-format rendering is then $\hat{\mathbf{y}}_{\mathrm{B}} = \mathbf{Q}_{\mathrm{B}}\mathbf{b}$.

When the input signal is from a spaced omnidirectional microphone array with a radius of a few centimeters, a reasonable $\mathbf{Q}_{\mathrm{S}}$ is a complex-valued delay-sum beamforming matrix, defined independently for each frequency band, again with look directions towards the loudspeakers. For simplicity, let us assume horizontal sound field analysis and that the microphones are in a ring with azimuths ε_j, where j is the microphone index. The $N_y \times N_s$ prototype matrix $\mathbf{Q}_{\mathrm{S}}$ has entries

$$q_{\mathrm{S}ij} = \frac{1}{N_s} e^{-i2\pi f \frac{r}{c} \cos(\theta_i - \varepsilon_j)}, \tag{22}$$

where f is the center frequency of the frequency band, c is the speed of sound and r is the microphone ring radius.

3 IMPLEMENTATION

An offline DirAC Matlab rendering software was implemented to compare the legacy and the covariance rendering methods. The high level block diagram is shown in Fig. 3. All reproduction modes, described in Section 3.1, receive identical spatial analysis and processing in all aspects except for how the output channels are mixed from the available input channels and the decorrelated channels. In detail, all reproduction modes use the same *DOA* and ψ parameters analyzed using the B-format input, thus allowing best comparability of only the differing channel mixing techniques. This is an idealized procedure, since in practice the *DOA* and ψ analysis accuracy with the spaced microphone array would not be as robust in the highest frequencies due to the spatial aliasing, nor in the lowest frequencies due to the microphone self-noise.

The input signals with the sampling rate of 44.1 kHz were processed every 512 samples with a 1024 sample length Hanning window, zero-padded to the length of 8192 samples, and processed with the fast Fourier transform (FFT). The covariance matrices of the FFT frequency lines were combined to form 38 perceptual bands approximating the equivalent rectangular bandwidth (ERB) scale [14], which means larger absolute bandwidths toward the higher frequencies. The mixing matrices were formulated in these 38 bands, and applied to the signal in the original FFT resolution.

The decorrelators were inter-channel incoherent 4096 sample noise bursts. They were generated in FFT domain by setting the amplitudes to unity, and randomizing the phases. The selected length of the decorrelator response was based on informal tests, with the design principle to use the shortest response that ensured high reverberation quality also with the omnidirectional input. The responses were zero-padded to the length of 8192 samples, and applied to the signal frames as FFT multiplications. The same decorrelator responses were applied in all reproduction modes.

The time-average was performed using a recursive averaging window with per-band time constants providing window lengths that are inversely proportional to the number of the FFT lines within the perceptual band. This approach provided similar stochastic accuracy in all perceptual bands, and longer time averaging windows toward the lower frequencies.

The long zero-padding enabled avoiding the circular time aliasing effects that otherwise occur in STFT processing. The frequency band mixing gains were also processed to correspond to 1024 sample time domain responses followed by a sequence of zeros. Since the combined length of the nonzero parts of the time domain counterparts of

Table 1. Reproduction modes.

Input format	Covariance rendering	Legacy rendering
B-format (3ch)	**Q** defined as a virtual microphone matrix as in Eq. (20)	Rendering using virtual directional microphones
Omni (1ch)	**Q** defined as $N_y \times 1$ vector of ones	Rendering using an omni microphone
Array (4ch)	**Q** defined as a beamforming matrix as in Eq. (22)	Not included in the test

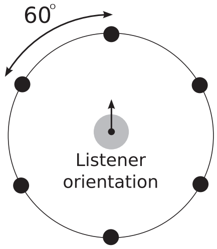

Fig. 4. Loudspeaker setup applied in the listening test and the simulations.

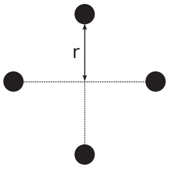

Fig. 5. Array of omnidirectional microphones in the test, with $r = 3$ cm.

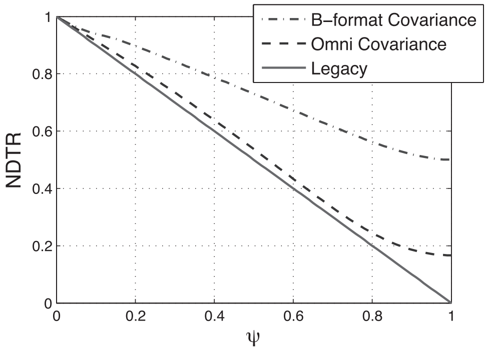

Fig. 6. Nondecorrelated to total energy ratio (NDTR) with rendering using B-format and omnidirectional input. The legacy rendering provides the same NDTR with both input types.

all FFT operations was smaller than 8192 samples, the circular time aliasing effects were completely avoided. Consequently, also no synthesis window was necessary. The approach provided improved robustness against processing artifacts in informal tests in respect to the typical approach of applying a synthesis window to suppress the circular aliasing effects.

In conditions with both diffuse and nondiffuse sound, and with processing based on virtual microphone signals, the legacy rendering method applies a compensating gain as function of ψ and the microphone directivity [6]. The covariance rendering method on the other hand produces the output signal energy according to the target covariance matrix definitions. There are special cases where these two could produce slightly differentiable output spectrum, for example, when the sound field is not exactly a combination of diffuse sound and a plane wave. Since this is more a feature of the implementation rather than an integral property of the rendering methods, the same spectrum of all reproduction modes was ensured using an adaptive post-equalizer in ERB bands, with the gain limits of ±2dB.

3.1 Reproduction modes and setting

The reproduction modes are listed in Table 1. The applied loudspeaker layout was six channels equally distributed as in Fig. 4. This layout is sufficient for reproducing the perception of a horizontally surrounding diffuse sound field [15]. The omnidirectional and the B-format rendering was performed with both the legacy and the covariance method, but the spaced omni array rendering was performed only with the covariance method. The implementation of the beamforming approach for the legacy synthesis was not considered relevant since the advantage of the better use of the independent signal components became obvious already for the B-format case, and the phenomenon is equivalent in the array rendering. The microphone array was assumed so that $N_s = 4$ and $r = 3$ cm, organized as in Fig. 5. All microphones were assumed noiseless.

4 SIMULATIONS

A set of simulations was designed to measure the portion of the sound energy that originates from the decorrelators with the different reproduction modes. The covariance matrices of the microphone signals were formulated in Matlab in different diffuseness conditions assuming a wide band source in front, spherically evenly distributed diffuse field and noiseless microphones. Furthermore, unbiased *DOA* and ψ estimates were assumed. The mixing matrices of the legacy and the covariance rendering methods were formulated using the functions within the test implementation, and were monitored in terms of the nondecorrelated to total energy ratio (NDTR), which describes the fraction of the total output sound energy that is not processed with decorrelators. It is assumed that higher NDTR entails better sound quality.

Fig. 6 shows the NDTR using the omnidirectional and B-format input and Fig. 7 shows the frequency dependent NDTR with the spaced array. Since the amount of the decorrelation with the legacy rendering depends solely on ψ, it provides the same NDTR with both the omni and the B-format inputs. With the B-format input the covariance rendering provides clearly higher NDTR than the legacy rendering, with the largest difference in the fully diffuse

Table 2. Reference sound scenes.

Item	Source material	Item preparation
Applause (3 items)	Three different 5.1 spaced microphone recordings, two of which from [17]	6 channel manually crafted mix using selected incoherent and stable signal excerpts
Speech (3 items)	Items 4, 5, and 9 from [18]	Simulated diffuse reverberation
Classical guitar (3 items)	Three arpeggio excerpts of item 12 from [18]	Simulated diffuse reverberation

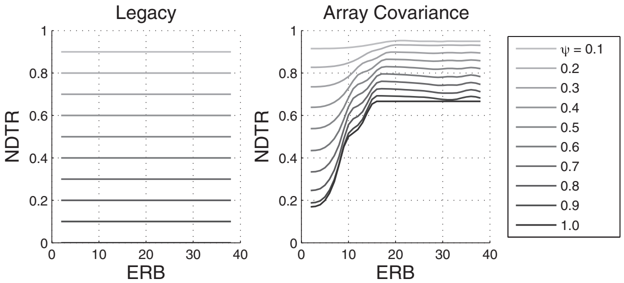

Fig. 7. Nondecorrelated to total energy ratio (NDTR) across frequency with the array input. The covariance rendering method takes into use the incoherent signal components in the medium and high frequencies.

condition. Similar benefit of the covariance rendering is seen also with the spaced array input in the medium and high frequencies, in which there are also a number of independent signal components available. In the low frequencies, the NDTR with the array input becomes similar to rendering using the omnidirectional input since the signals from the spaced microphones become coherent. With the omnidirectional input, the NDTR of the covariance and the legacy rendering are similar in respect to each other since there is only one independent signal component available.

5 LISTENING TEST

A listening test was performed in an anechoic chamber to evaluate the perceptual impact of the improved NDTR. The test design was similar to MUSHRA [16], with a known and a hidden reference, and multiple hidden stimuli compared side-by-side. The lower anchor was the reference signal convolved with the same decorrelator responses that were applied in the test implementation. The task was to rate the overall similarity of the test items in respect to the reference. The difference to MUSHRA was that no verbal labels were displayed other than to indicate that the top of the scale means that the item is not differentiable from the reference, and the values toward the low mean increasing difference. The listeners were urged to use the scale in the maximum extent in each of the test cases. Nine subjects participated in the test, all of which were researchers in the field, and not authors of the paper. The listeners were first introduced to the hidden items and the listening test interface by a short practice period. The listeners were allowed to rotate the head, although asked to keep the general orientation toward the front loudspeaker with which the discrete sources were reproduced.

5.1 Reference stimuli

A total of nine reference sound scenes were generated for the test, listed in Table 2. The item categories were selected based on informal testing on the audibility of the decorrelation to the overall sound quality, where a particularly obvious effect was noticed with the applause items, and particularly subtle effect with the classical guitar items. Speech items were intermediate in this respect. The reference scenes were reproduced over the same six-channel loudspeaker layout as shown in Fig. 4. The B-format and array recordings were generated in Matlab by assuming free field conditions and noiseless microphones.

The surround applause signal is critical since the transients from different directions can fall within the same analysis window. This causes the analyzed ψ to be artificially high, and thus a large amount of decorrelation, and thus smearing of the transients, is produced especially by the legacy rendering method. The applause reference scenes were manually crafted from recorded multi-channel applause signals, by selecting six stable, mutually incoherent and perceptually similar mono channel excerpts from different parts and mostly from different channels of the original mixture. Building such a manual mix was motivated by the signal analysis of the original items, which revealed high inter-channel coherences in the low frequencies, which is a feature typical to the spaced microphone recording techniques. The item preparation ensured wide band incoherence between the channels, which generates to the measurement point the sound field condition that is more similar to that of the original acoustic space.

The speech items were also clearly affected by the decorrelators, however, less than the applause items. The sensitivity is assumed to root from the impulsive fine structure produced by the glottal source, which is then smeared by the decorrelators. In each of the reference scenes in the test set, an anechoic mono speech recording was placed in the front, and incoherent reverberation with a 10 ms pre-delay was added to all channels. The reverberation response was a multi-channel Gaussian noise response decaying exponentially in frequency bands according to specified band-wise reverberation times. The reverberation time was 1.5 s in the lowest frequency band, 0.5 s at the highest frequency band, and linearly interpolated in the ERB bands in between. The overall response gain was adjusted so that the resulting direct-to-total energy ratios ranged from 0.3 in the low frequencies to 0.67 in the high frequencies. Note that this simplified room effect was designed merely to exhibit the semi-diffuse acoustic condition, and does not model all aspects of natural acoustic environments such as discrete early reflections or buildup of the diffuseness over the time.

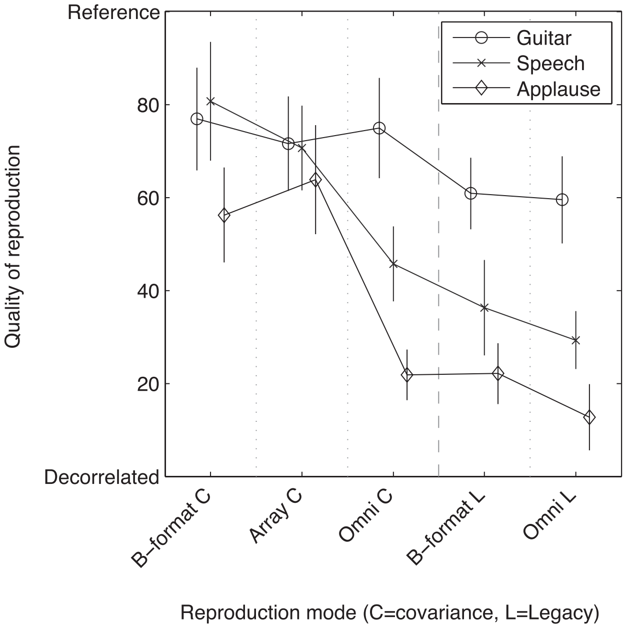

Fig. 8. The means and 95% confidence intervals of the scores of the reproduction modes normalized in respect to the scores given to the reference (top) and to the fully decorrelated anchor (bottom).

The classical guitar items were perceptually clearly least affected by the decorrelators. The corresponding reference sound scenes were generated with the same procedure as the speech reference scenes, i.e., by processing anechoic mono guitar recordings with the above-described room response.

5.2 Results

The score data was first normalized in each reference scene and for each subject according to the scores given to the hidden reference and to the anchor. This data representation illustrates only the relative qualities of the reproduction modes, and not the absolute quality. A two-way repeated measures analysis of variance (RM-ANOVA) was applied to the normalized result data, with factors *Item* and *Reproduction_mode*. Mauchly's test revealed that sphericity was violated in two cases: *Reproduction_mode* in the applause category, and the interaction *Item***Reproduction_mode* in the guitar category. In both $\varepsilon < 0.75$, and the Greenhouse–Geisser correction was applied. Significant effects were found in the factor *Reproduction_mode*, in all three categories. The mean values and the 95% confidence intervals are shown in Fig. 8, and the statistically differing means are shown in Fig. 9. The most relevant result is that significant improvement was provided by the covariance rendering method when there were several microphone signals available, and the input signal was either applause or speech, i.e., such that is more affected by the decorrelators.

6 DISCUSSION

The simulations showed that the covariance rendering method can reduce the amount of the decorrelated sound energy in comparison to the legacy DirAC rendering techniques when there are several fully or partly independent microphone signals available. The listening tests showed that the perceptual benefit depends much on how the perceived characteristics of the particular audio content is affected by the decorrelators.

Furthermore, the listening test results also suggested that high quality DirAC synthesis would be possible also with a low cost spaced microphone array with a radius of a few centimeters. This is reasonable, since in the middle and high frequencies there are several independent signal components available, while in the lowest frequencies the decorrelators seem least harmful to the overall sound quality, as is discussed also in [7].

The legacy rendering with both input signal types provided similar performance, since the same decorrelator response was applied in all reproduction modes. On the other hand, informal tests have suggested that the B-format legacy rendering can be adjusted to provide in average better quality than the omni-based reproduction modes. This is since if the decorrelator responses are made shorter, the quality of the reverberation degrades faster with the omnidirectional input than with the B-format input, since in the latter the input channels are partly incoherent to start with. It is nevertheless reasonable to expect that with all decorrelator types, the average best performance with the B-format input is achieved with the covariance rendering method, since it also reduces the amount of the overall decorrelated sound energy.

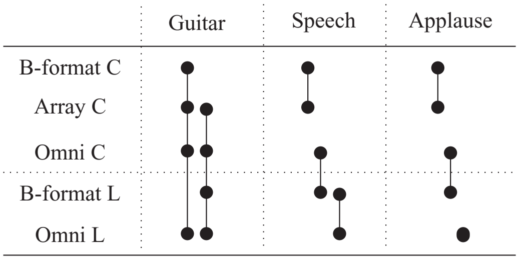

Fig. 9. The mean scores of the reproduction modes that are not significantly different from each other are marked with line connections.

7 CONCLUSION

In this paper, a recently proposed covariance domain spatial sound rendering method was applied to optimize the DirAC reproduction by minimizing the amount of the applied decorrelated sound energy. The procedure was shown to improve the overall perceived sound quality, especially with audio content that has impulsive fine structure such as applause and speech, and when several semi-independent microphone signals were available. The covariance rendering method performed in all tests similarly or better than the legacy rendering method, making it the preferred choice for performing DirAC synthesis.

8 ACKNOWLEDGMENTS

The work has been funded by Fraunhofer IIS, Germany. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [240453]. The authors would like to thank Tapani Pihlajamäki and Achim Kuntz for the valuable feedback.

9 REFERENCES

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization (rev. ed.)*. Cambridge, MA : 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, no. 11, pp. 932–955, 2008 Nov.

[3] 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, no. 9, pp. 655–673, 2012 Sept.

[4] C. Faller, "Multiple-Loudspeaker Payback of Stereo Signals," *J. Audio Eng. Soc.*, vol. 54, no. 11, pp. 1051–1064, 2006 Nov.

[5] V. Pulkki, "Spatial Sound Reproduction with Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007 June.

[6] J. Vilkamo, T. Lokki, and V. Pulkki, "Directional Audio Coding: Virtual Microphone-Based Synthesis and Subjective Evaluation," *J. Audio Eng. Soc.*, vol. 57, no. 9, pp. 709–724, 2009 Sept.

[7] M.-V. Laitinen, F. Küch, and V. Pulkki, "Using Spaced Microphones with Directional Audio Coding," in *Audio Engineering Society 130th Convention*, 2011 May.

[8] J. Ahonen, "Microphone Configurations for Teleconference Application of Directional Audio Coding and Subjective Evaluation," in *Audio Engineering Society 40th International Conference*, 2010 Oct.

[9] J. Vilkamo, T. Bäckström, and A. Kuntz, "Optimized Covariance Domain Framework for Time-Frequency Processing of Spatial Audio," *J. Audio. Eng. Soc.* vol. 61, pp. 403–411, 2013 June.

[10] M.-V. Laitinen, F. Küch, S. Disch, and V. Pulkki, "Reproducing Applause-Type Signals with Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 59, no. 1/2, pp. 29–43, 2011 Jan.

[11] J. Ahonen, G. Del Galdo, F. Küch, and V. Pulkki, "Directional Analysis with Microphone Array Mounted on Rigid Cylinder for Directional Audio Coding," *J. Audio Eng. Soc*, vol. 60, no. 5, pp. 311–324, 2012 May.

[12] C. Tournery, C. Faller, F. Küch, and J. Herre, "Converting Stereo Microphone Signals Directly to MPEG-Surround," in *Audio Engineering Society 128th Convention*, 2010 May.

[13] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997 June.

[14] 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 Aug.

[15] K. Hiyama, S. Komiyama, and K. Hamasaki, "The Minimum Number of Loudspeakers and its Arrangement for Reproducing the Spatial Impression of Diffuse Sound Field," in *Audio Engineering Society 113rd Convention*, 2002 Oct.

[16] "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems." Recommendation ITU-RBS.1534-1.

[17] "Call for Proposals on Spatial Audio Coding." ISO/IEC JTC1/SC29/WG11 (MPEG), Document N6455, 2004 March.

[18] "Music for Archimedes." Compact Disc, Bang & Olufsen, 1992.

THE AUTHORS

Juha Vilkamo

Ville Pulkki

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.

●

Ville Pulkki received his M. Sc. and D. Sc. (Tech) degrees from Helsinki University of Technology in 1994 and 2001, respectively. He majored in acoustics, audio signal processing and information sciences. Between 1994 and 1997 he was a full time student at the Department of Musical Education in Sibelius Academy. In his doctoral dissertation he developed Vector Base Amplitude Panning (VBAP), which is a method for positioning virtual sources to multichannel loudspeaker configurations. In addition, he studied the performance of VBAP with psychoacoustic listening tests and with modeling of auditory localization mechanisms. The VBAP method is now widely used in multi-channel virtual auditory environments, and in computer music installations. Later, he developed with his group a non-linear time-frequency-domain method for spatial sound reproduction and coding, Directional Audio Coding (DirAC). DirAC takes coincident first-order microphone signals as input, and processes output to arbitrary loudspeaker layouts or to headphones. He also researches computational functional model of the brain organs devoted to binaural hearing. He is leading a research group in Aalto University (earlier: Helsinki University of Technology, TKK or HUT), which consists of 18 researchers. The group conducts research also on head-related acoustics measurements, and conducts psychoacoustical experiments to better understand spatial sound perception. Prof Pulkki enjoys being with his family (wife and two children), playing various musical instruments, building his summer place and dancing hip hop. He is the chair of AES Publication Policy Committee, and member of the Technical Committee of Spatial Audio and the board of AES Finnish Section.