Spatial Sound Reproduction with Directional Audio Coding*

VILLE PULKKI, *AES Member*
(ville.pulkki@hut.fi)

Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, FI-02015 TKK, Finland

Directional audio coding (DirAC) is a method for spatial sound representation, applicable for different sound reproduction systems. In the analysis part the diffuseness and direction of arrival of sound are estimated in a single location depending on time and frequency. In the synthesis part microphone signals are first divided into nondiffuse and diffuse parts, and are then reproduced using different strategies. DirAC is developed from an existing technology for impulse response reproduction, spatial impulse response rendering (SIRR), and implementations of DirAC for different applications are described.

0 INTRODUCTION

Methods for multichannel sound reproduction of spatial sound have been discussed for years. None of the traditional methods reproduces a natural spatial audio environment faithfully and is applicable for different loudspeaker configurations. The coincident-microphone approaches [1], such as first-order Ambisonics [2], can in theory utilize any loudspeaker setup. Unfortunately, since the directional patterns of current high-quality microphones are only of zeroth or first order, the resulting loudspeaker signals of a multichannel setup are more coherent than desired. High coherence results in coloring and distortion of the spatial image, especially outside the best listening position. Recently there has been research on higher order Ambisonics [3], [4]. The sound field is captured with a large number of microphones, whose signals are processed to provide virtual microphones with narrower directional patterns than with first-order Ambisonics. This provides better quality; however, the larger number of microphones needed increases the cost.

In spaced-microphone techniques for multichannel loudspeaker systems [1], [5] the microphones are spatially separated from each other with distances from a few centimeters to a few meters. Due to temporal differences in the arrival of sound at the microphones, the resulting loudspeaker signals are less coherent than in the coincident approach. This circumvents some of the problems of Ambisonics, although the perceived source directions might differ from the original source directions. Also, spaced-microphone techniques do not allow playback of the recorded sound over different loudspeaker setups.

The goal of the proposed directional audio coding (DirAC) is to reproduce the spatial properties of sound recorded with existing microphone systems as realistically as possible with different sound reproduction systems. In the analysis part of DirAC, the sound direction and diffuseness are estimated depending on time and frequency. In synthesis these data are used to enhance the quality of reproduction. The sound is divided into nondiffuse and diffuse parts in frequency bands, and then reproduced using different strategies. In the current implementation of DirAC with B-format microphones and multichannel reproduction, the diffuse part of sound is processed differently for each loudspeaker to reduce coherence between loudspeakers, and the nondiffuse part of sound is applied to a small subset of loudspeakers at one time, which also decreases the coherence. The major problem of coincident-microphone techniques is thus eliminated in both branches of processing.

DirAC is based on the same principles and partly the same methods as the recently proposed spatial impulse response rendering (SIRR) technique [6]. SIRR is a technique to reproduce room impulse responses for application in convolving reverberators. The first attempts to reproduce other sound than impulse responses with SIRR were presented in [7], which later led to a system that is currently called directional audio coding. Although the differences between SIRR and DirAC methods are moderately small, the new name is suggested for the method since the name SIRR includes the term "impulse re-

*Manuscript received 2006 November 8; revised 2007 April 5.

sponse," which would be misleading as a name for a method for processing continuous sound.

Although DirAC was initially designed for high-quality applications, such as music reproduction, there are also some other applications where the technique can be used. As will be discussed, the technique can be used to transmit the spatial aspects of audio analyzed from a low-cost microphone grid as metadata, while the transmitted audio would only be a monophonic channel. This would enable teleconferencing applications with spatial sound, as the transmitted data rate would not increase prominently from current solutions using monophonic audio transmission. In addition the system can be used in stereo upmixing.

This engineering report is organized as follows. The assumptions about the perception of spatial sound, based on which the SIRR and DirAC techniques were developed, are first reviewed. Next the theory and implementation of DirAC is presented, and finally some numerical examples and informal listening test results are discussed.

1 ASSUMPTIONS ABOUT PERCEPTION OF SPATIAL SOUND

The design of the SIRR and DirAC techniques is based on four assumptions about the interaction between sound field properties and the perceptual attributes that they produce [6]. The assumptions are repeated briefly here. The interested reader is referred to [8] for a discussion & their validity and for an introduction to the human mechanisms for spatial sound perception relevant in this case. For a more concise understanding of the perceptual mechanisms the reader is referred, for example, to [9], [10]. The assumptions are as follows.

1) The direction of the arrival of sound transforms into interaural time difference (ITD), interaural level difference (ILD), and monaural localization cues.

2) The diffuseness of sound transforms into interaural coherence cues.

3) Timbre depends on the monaural (time-dependent) spectrum together with ITD, ILD, and interaural coherence of sound.

4) The direction of arrival, diffuseness, and spectrum of sound measured in one position with the temporal and spectral resolution of human hearing determines the auditory spatial image the listener perceives.

It has to be noted, that in this study the term "diffuse" denotes only a quality of the physical sound field. It is not used to quantify any perceptual aspect.

In SIRR the number of sources used to measure the impulse response was always one, and the signal content of it was known to be an impulse. In DirAC multiple sources are allowed, and their content is not a priori known. However, the assumptions are not changed. It is still assumed that the listener cannot decode separate in one time instant cues for multiple wavefronts from different directions within a critical band. This is bolstered by psychoacoustic results, where it has been found that two sinusoids with a small spectral difference presented over spatially separated loudspeakers cannot be localized correctly [11]. The listener typically localizes the sinusoids as a single fused event fluctuating inside the head between the ears.

This leads us to assume that humans decode at one time only single cues per each critical band from the ear canal signals. This assumption is in line with a recently proposed auditory model for source localization [12], which hypothesizes that the auditory system obtains the source direction by considering the binaural directional cues only at the time instances when they correspond to one of the source directions.

2 DirAC PROCESSING

In this section the processing parts of DirAC are presented. The general idea is first reviewed, after which different parts of processing are considered separately. Since there are multiple applications for DirAC, there are multiple variations of some of the parts.

2.1 General Idea of DirAC

The DirAC design is based on the assumptions stated in Section 1, as was the SIRR design. In assumption 4) it was stated that the temporal and spectral resolution of the processing should mimic the temporal and spectral resolution the auditory system is using for spatial hearing. For this purpose the microphone signals are divided into frequency bands, following the frequency decomposition of the inner ear, as shown in Fig. 1.

The assumptions also imply that the direction of arrival, the diffuseness, and the spectrum of the sound field should be synthesized correctly, which would lead to the generation of correct spatial cues to the listener. Thus the direction of arrival and the diffuseness are analyzed with a temporal accuracy comparable to the accuracy of the human auditory system at each frequency band. This information is later used in DirAC synthesis to generate these cues correctly. In practice, in synthesis the sound is divided dynamically into two streams, one that corresponds to nondiffuse sound, and another that corresponds to the diffuse sound stream. The nondiffuse sound stream is reproduced with a technique aiming at pointlike sound sources, and the diffuse sound stream with a technique aiming at the perception of sound lacking prominent direction, which is denoted as surrounding reproduction in Fig. 1.

Between analysis and synthesis, the sound may be transmitted over any medium. This possibility is exploited in teleconferencing application, where the audio is transmitted as monophonic channel with the analysis parameters as metadata (see Section 4.2).

2.2 Main Implementations of DirAC

Two main implementations for the DirAC are investigated in this engineering report—high-quality reproduction and teleconferencing. The flowchart of high-quality reproduction is shown in Fig. 2. In this implementation, for simplicity, it is assumed that both analysis and synthesis are performed in the listening stage, or that the synthesized sound is transmitted or stored as DirAC-processed multichannel loudspeaker signals. Thus there is no transmission between analysis and synthesis in this case.

The flowchart for the telecommunication implementation is shown in Fig. 3. In this case the latency and the computational complexity are minimized. To minimize the data flow of transmission, only one audio channel is transmitted together with the direction and optional diffuseness data. The flow diagrams are discussed in detail in the following sections.

2.3 Direction and Diffuseness Analysis

In this section the computational issues of DirAC analysis are discussed. The possible methods for time–frequency analysis are presented first, after which the physics for directional analysis are reviewed, and finally the need for temporal averaging in analysis is presented.

2.3.1 Time–Frequency Analysis

Two different techniques have been used to divide signals into frequency bands. In the teleconferencing implementation the short-time Fourier transform (STFT) was used, where the microphone signals are windowed and transferred into the spectral domain using the fast Fourier transform (FFT). STFT thus divides signals in both time and frequency. Both analysis and synthesis are computed in the spectral domain, and the synthesized signals are transferred into the temporal domain with the inverse FFT (see Fig. 3). However, in its basic form this approach implies the same temporal accuracy at all frequencies, which does not match the frequency-dependent temporal resolution of humans. On the one hand the window length has to be sufficiently long for correct low-frequency reproduction, and on the other hand the length should be short enough to be able to reproduce sound events lasting only for a short time. In informal listening of DirAC implemented with STFT using 20-ms time windows, it was found that the directions of rapid transients in the presence of other sound events were reproduced erroneously [13]. If shorter windows are utilized, some low-frequency artifacts will emerge. However, some parts of DirAC analysis and synthesis can be implemented effi-

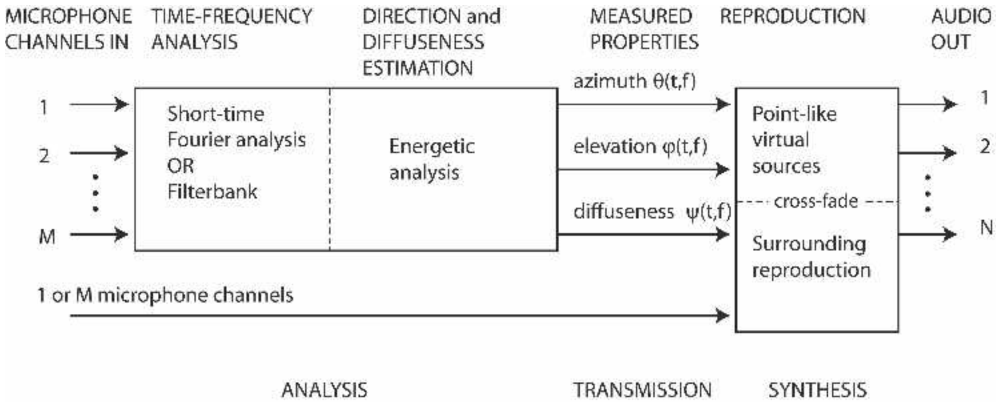

Fig. 1. Flow diagram of DirAC.

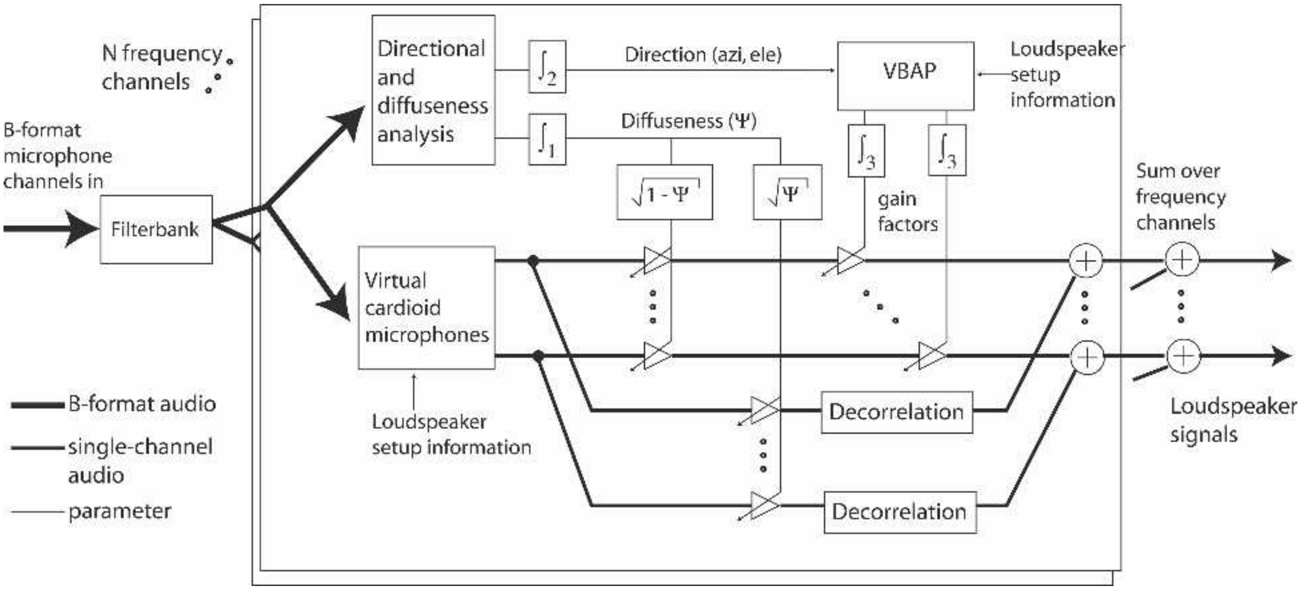

Fig. 2. Flow diagram of DirAC for high-quality reproduction of B-format microphone signal. Integral symbols with subscripts denote temporal averaging with corresponding window functions.

ciently with the STFT approach, which is beneficial in real-time applications.

The other method used for time–frequency analysis is a filter bank, where the microphone signals are filtered with a number of narrow-band filters, resulting in a time signal for each frequency band as used in high-quality applications (see Fig. 2). Optimally the spacing of the center frequencies and the passband widths of the filters in the bank should correspond to the human frequency resolution. This approach supports the human auditory resolution in processing better, since the temporal resolution for each frequency channel can be set to be different. However, the analysis and synthesis in DirAC are computationally more complex with the filter-bank implementation than with the STFT-based approach.

2.3.2 Directional Analysis

The microphone array used with DirAC must enable the analysis of the direction and diffuseness in a broad frequency region. So far B-format microphones have been used exclusively. B-format stands for a coincident microphone array that produces four microphone channels with different directional characteristics: one omnidirectional and three figure-of-eight channels directed toward each orthogonal axis of the Cartesian coordinates. The corresponding signals are $w(n)$, $x(n)$, $y(n)$, and $z(n)$, respectively, where n is the time index. The omnidirectional signal has been scaled by $1/\sqrt{2}$. There are a number of such microphones commercially available. Also, such a microphone can be constructed simply by placing one omnidirectional microphone and three figure-of-eight microphones in a coincident position. As will be considered briefly in Section 4.2, it is also possible to build a low-cost horizontal-only B-format microphone of few omnidirectional capsules. In time–frequency analysis the microphone channels are divided into frequency bands, which are denoted, for example, by $w(n,i)$, where i denotes the frequency channel index. However, for simplicity of notation, the indices are dropped where appropriate.

With B-format input the directional analysis can be performed relatively simply based on an energetic analysis of the sound field, which is now shortly reviewed theoretically, and then with B-format input. Instantaneous energy density can be computed as

$$E = \frac{1}{2}\rho_0\left(\frac{p^2}{Z_0^2} + u^2\right) \tag{1}$$

where p is the sound pressure, u is the particle velocity, ρ_0 is the mean density of air, and Z_0 is the characteristic acoustic impedance of air, $Z_0 = \rho_0 c$, with c being the speed of sound [14]. The instantaneous intensity vector $\boldsymbol{I}$ is defined as

$$\boldsymbol{I} = p\boldsymbol{u} \tag{2}$$

where $\boldsymbol{u}$ is the particle velocity vector [14]. The intensity vector points to the direction of the net flow of energy, and the magnitude denotes the strength of the energy flow. Using these equations, it is possible to define diffuseness,

$$\psi = 1 - \frac{\|\langle \boldsymbol{I}\rangle\|}{c\langle E\rangle} \tag{3}$$

where $\langle\cdot\rangle$ denotes short-time average. Diffuseness gets a value of zero with plane waves from a single direction, where the net flow of energy corresponds to the total energy. It reaches the value of 1 in a field where there is no net transport of acoustic energy, as with ideal standing waves or reverberation.

With B-format input, p is approximated by w, and the particle velocity vector by

$$\boldsymbol{v} = x\boldsymbol{e}_x + y\boldsymbol{e}_y + z\boldsymbol{e}_z \tag{4}$$

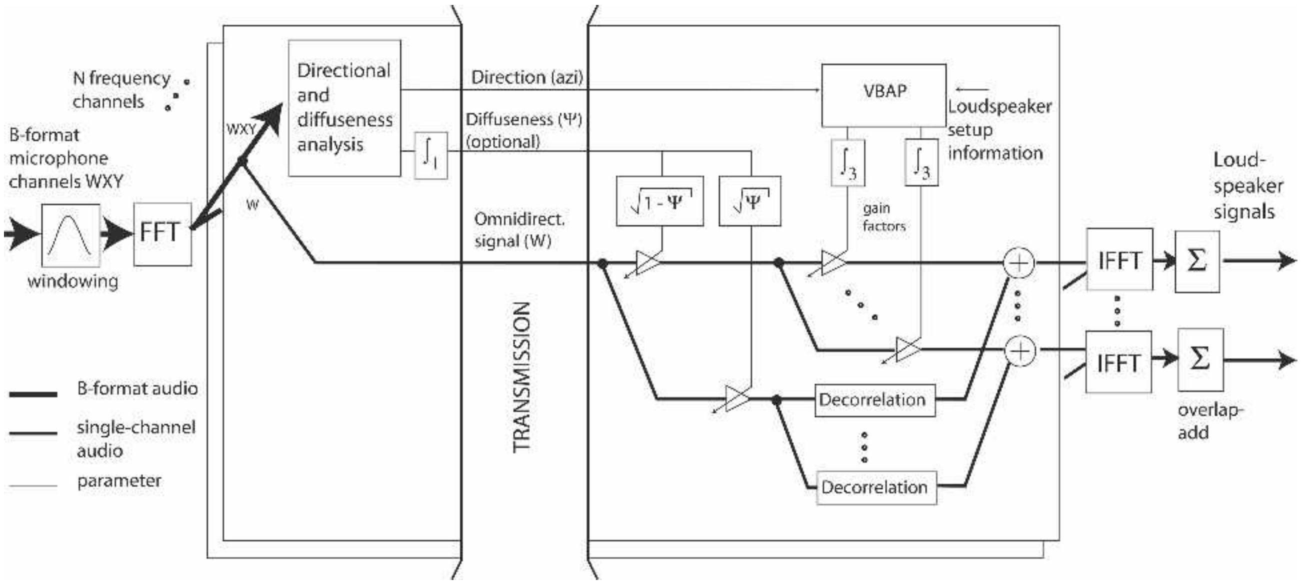

Fig. 3. Flow diagram of DirAC for telecommunications. Integral symbols with subscripts denote temporal averaging with corresponding window functions.

where $\boldsymbol{e}_x$, $\boldsymbol{e}_y$, and $\boldsymbol{e}_z$ represent Cartesian unit vectors. The diffuseness is computed from B-format input with

$$\boldsymbol{\psi}(n) = 1 - \frac{\sqrt{2}\left\| \sum_{m=a_1}^{b_1} w(n+m)\boldsymbol{v}(n+m)W_1(m) \right\|}{\sum_{m=a_1}^{b_1} [|w(n+m)|^2 + |\boldsymbol{v}(n+m)|^2/2]W_1(m)} \qquad (5)$$

where $W_1(m)$ is a window function defined between constant time values $a_1 \leq 0$ and $b_1 > 0$ for short-time averaging explained in the next section. The direction vector is defined to be the opposite direction of the intensity vector. In this case the length of the intensity vector is not of interest, which simplifies the formula a bit. The vector is computed at each frequency channel as

$$\boldsymbol{D}(n) = -\sum_{m=a_2}^{b_2} w(n+m, m)\boldsymbol{v}(n+M)W_2(m) \qquad (6)$$

where W_2 is a window function for short-time averaging $\boldsymbol{D}$, and a_2 and b_2 are defined similarly as a_1 and b_1, respectively. For a more detailed presentation of the energetic analysis with B-format input, see [15].

2.3.3 Temporal Averaging

In SIRR the temporal variations of the analyzed direction and diffuseness were not limited. If such variations were applied in DirAC for the reproduction of continuous sound, there would be audible artifacts. Such fast variations could be allowed in SIRR, since the signal was a priori known to be an impulse response. The relatively flat spectrum of the signal masked such nonidealities. However, in DirAC signals with narrow band may occur; thus such distortion cannot be allowed. In this section the strategies needed for slowing down the temporal variations are discussed separately for diffuseness and direction.

In Eqs. (5) and (6) a window function was used for short-time averaging. With DirAC implementations using STFT time–frequency analysis, the window has been implemented with a first-order low-pass IIR. With off-line processing the window function can also be defined to be symmetric in time, as in the current high-quality implementation the Hanning window has been used. In this case different window lengths are applied at different frequency bands.

The variables from which the diffuseness is computed are temporally averaged, as shown in Eq. (5). With the current filter-bank implementation the length of $W_1(m)$ in Eq. (5) has 10–50 times the period of the center frequency at the corresponding frequency band, but is limited to 3 ms minimum and 150 ms maximum. With the teleconferencing application the time constant of the first-order IIR filter has been about 50 ms. These values have been chosen by informal testing. A more detailed analysis is left as a future study.

The slowing down of the direction vector $\boldsymbol{D}$ is a slightly more complicated task. The reproduction of the direction has to be accurate. In particular, during the onset of continuous sounds, the direction analyzed has to correspond accurately to the direction of the arrival of sound. However, the changes in direction may be rapid in many cases. If such fast changes are applied directly in synthesis, temporal artifacts, such as clicks and bubbling, emerge.

If the direction vector $\boldsymbol{D}$ itself is temporally averaged, fast changes in the direction analyzed are replaced by slow changes. Most of the time the vector will then point to a direction that does not correspond to the instantaneous direction of arrival of sound. In informal tests averaging of $\boldsymbol{D}$ was found suboptimal, as the temporal artifacts were not completely removed, and the perceived sound source directions were instable. The effect of temporal averaging of the direction vector is shown with a synthetic example in Section 1.

A solution to this problem is to apply for temporal averaging of the direction vector with as short a time window as is feasible to compute the direction of arrival reliably. In addition, to avoid audible clicks and other artifacts, the parameters of the reproduction method are averaged with a longer time window. In the filter-bank implementation the length of the Hanning window W_2 in Eq. (6) has been only three times the period of the center frequency of the corresponding frequency band, but limited to 1 ms minimum. With STFT time–frequency analysis the direction vector is automatically averaged over the STFT time window utilized. The window is typically long enough so that no further averaging is needed. However, in some cases it can be necessary to apply further averaging at low frequencies. The second stage of temporal averaging performed in the synthesis stage is discussed in Section 4.2.

2.4 Synthesis Techniques

In SIRR and DirAC synthesis audio is first divided into nondiffuse and diffuse parts in each frequency channel, and then reproduced with different strategies. In this section the techniques currently used in synthesis are described, and other possible methods are discussed.

2.4.1 Dividing Sound Signals into Diffuse and Nondiffuse Streams

In DirAC synthesis the microphone signal(s) are divided into nondiffuse and diffuse streams by multiplying the input signal(s) with two time-variant factors derived from diffuseness. The diffuse stream will be reproduced in random phase when compared to the nondiffuse stream, and the cross fading between the streams has to conserve energy. This can be simply implemented by multiplying the input signal(s) with $\sqrt{1-\psi}$ and $\sqrt{\psi}$, respectively, as shown in Figs. 2 and 3. In practice the analyzed diffuseness has never a value of 1 or 0, but one that lies in between. Thus the produced loudspeaker signals cross fade at each frequency band between the signals reproduced with diffuse and nondiffuse methods.

2.4.2 Reproduction of Pointlike Virtual Sources

The choice of the method to create pointlike sources is dependent on the sound reproduction equipment targeted. In this work relatively sparse two- and three-dimensional loudspeaker systems have been available. Thus amplitude panning formulated with vector base amplitude panning (VBAP) [16] has been considered sufficient. With amplitude panning the computation is straightforward and the two- or three-dimensional directional information is given to the VBAP algorithm, which computes the gain factors $g(n, i, k)$ for each time instant n, frequency channel i, and loudspeaker channel k, as shown schematically in Figs. 2 and 3.

In principle DirAC could be formulated for other reproduction methods, such as wave field synthesis [17], higher order Ambisonics [3], binaural headphone [18], or crosstalk canceled loudspeaker techniques [19]. However, this is left as a subject for future studies.

As mentioned in Section 2.3.3, the second phase of temporal averaging in directional reproduction has to be computed in the synthesis phase. The averaging is now described for amplitude panning. The artifacts mentioned in Section 2.3.3 are avoided by computing the weighted temporal-domain average of the gain factors from VBAP,

$$g(n, i, k) = \frac{\sum_{m=-M/2}^{M/2} g(n+m, i, k)[1 - \psi(n+m, i)]W_3(m)}{\sum_{m=a_3}^{b_3} [1 - \psi(n+m, i)]W_3(m)} \tag{7}$$

where $g(n, i, k)$ is the gain factor at subband i, time n, and loudspeaker channel k. With filter-bank time–frequency analysis the length of the Hanning window $W_3(m)$ has been as high as 100 times the period of the center frequency of the corresponding channel, but limited to 1000 ms at maximum. However, the weighting with $(1 - \psi)$ effectively shortens the time window when an onset with low diffuseness occurs. The weighting factor in Eq. (7) can also be instantaneous energy E, for example, instead of $(1 - \psi)$. If the diffuseness is not transmitted in the teleconferencing application, instantaneous energy is the only choice in that case. The effect of gain factor averaging is shown with a synthetic example in Section 3.1.

2.4.3 Audio Signal for Pointlike Virtual Sources

In the teleconferencing implementation the only channel that is transmitted is w, and it is the audio signal for pointlike virtual sources after optional multiplication by $\sqrt{1 - \psi}$, as shown in Fig. 3. This has one shortcoming: only in the case when diffuseness reaches the value of zero, which means that in the recording phase there was sound from only one direction in this time instant, w includes only nondiffuse sound. When diffuseness is greater than 0, which is generally the case, w is a superposition of diffuse and nondiffuse sound. Thus the sound that is applied to the pointlike virtual source includes also some diffuse sound, which is not desired.

This defect can be partly avoided when more channels are transmitted. In high-quality implementation B-format signals are transmitted, thus the audio signal for a pointlike virtual source can be derived from a virtual cardioid microphone directed toward the direction analyzed. This will reduce the amount of diffuse sound energy on average by 4.8 dB [20]. In practice this is implemented by computing a signal for each loudspeaker $y(n, i, k)$ with the equation

$$\begin{aligned} y(n, i, k) = \frac{1}{2} g(n, i, k)\big[& w(n, i)\sqrt{2} + x(n, i) \cos\theta_k \cos\phi_k \\ & + y(n, i) \sin\theta_k \cos\phi_k + z(n, i) \sin\phi_k\big] \end{aligned} \tag{8}$$

where θ_k is the azimuth and ϕ_k is the elevation of loudspeaker k. A system implementing this is illustrated in Fig. 2.

The signal used in the synthesis of nondiffuse sound is thus different in different loudspeakers, though all of them have been captured with a virtual cardioid microphone. The total signal used to create nondiffuse sound is the sum of these signals, and it is interpreted to be captured with a virtual microphone, whose directional characteristics depend on computed gain factors. If only a single gain factor has nonzero value, the directionality of the virtual microphone is naturally also cardioid. This can happen if the direction analyzed corresponds stably to a loudspeaker direction. If few gain factors of particular loudspeakers near each other have nonzero values, the resulting directivity of the virtual microphone is subcardioid. This occurs when the direction analyzed points stably between the loudspeakers. If the analyzed direction varies randomly and fast enough with time, the temporal averaging of gain factors produces equal values for all loudspeakers, and the directionality of the virtual microphone is close to omnidirectional. However, such fast changing sound direction typically produces high diffuseness, and most of the sound is produced with diffuse synthesis.

2.4.4 Diffuse Synthesis

The aim of diffuse synthesis is to produce a perception of surrounding sound lacking prominent direction. Also, the processing should not introduce any coloration to the sound spectrum. With multichannel listening, such perception can be produced by applying a sound signal that is decorrelated to all loudspeakers. In that case the magnitude spectrum of the sound should not be changed, but the phase spectrum should be random.

In SIRR different techniques were tested for decorrelation, and finally a method was implemented that randomized the phase spectrum in each STFT window [8]. It is unclear whether the method can be used with continuous sound, due to possible distortion and preechos produced by the method. In this project a simple, although computationally quite complex method is to convolve sound for each loudspeaker with a random sequence, which was also one of the methods used with SIRR. The sequences must be different for each loudspeaker, and they should be composed in a way that the convolution process would change the phase at all frequencies. For example, if 1-ms burst of white noise were used as the sequences, the high frequencies would be decorrelated properly; however, the loudspeaker signals would be in the relative phase of 0° or 180° below about 200 Hz. This would produce strong coloration artifacts. Setting the noise bursts longer would make the distribution of relative phases more random at low frequencies. However, the convolution with longer noise bursts would add length to impulse-like sounds, which is not desired. These prerequisites are achieved by having short decay times for high frequencies in the sequence, and longer decay times for low frequencies.

In practice, in the current high-quality implementation, sound is decorrelated at three frequency bands with exponentially decaying white-noise bursts with different time constants. At frequencies below 400 Hz the time constant has been on the order of 100 ms, but at higher frequencies, below 1300 Hz, the time constant has been on the order of 40 ms. At the highest frequency band the time constant has been about 10 ms. The length of the sequence has been three times the time constant. These constant values have been chosen by informal listening to DirAC reproduction over multiple loudspeakers in an anechoic chamber, and it has to be noted that a more formal definition of them is a subject of future studies.

2.4.5 Audio Signal for Diffuse Synthesis

When only a single channel is transmitted, as in the teleconferencing implementation of DirAC, the diffuse stream may be composed of only the single channel multiplied by $\sqrt{\psi}$. In the current implementation for multichannel loudspeaker setups, the audio channel is simply decorrelated for each loudspeaker, as shown in Fig. 3. Using only one signal as a source for diffuse sound may be suboptimal. The diffuse sound may have a directional distribution, which should be reproduced, and in some cases the sound arriving from different directions has different content.

If a B-format signal is transmitted, as in the high-quality implementation of DirAC, these factors are reproduced more faithfully to the original case. A virtual cardioid microphone is computed pointing toward each loudspeaker direction, and the signals are applied after decorrelation to corresponding loudspeakers. The decorrelation is needed to decrease the coherence between loudspeaker signals, which is due to the broadness of applied first-order microphone directional patterns.

As different reproduction methods are applied to nondiffuse and diffuse sound, their outputs may have to be scaled with a constant factor to maintain an even balance between the methods. This scaling is not presented in the figures, since it can be performed by adjusting the reproduction methods, such as by scaling the gain factors of amplitude panning, or by scaling the random sequences used in decorrelation. In the current implementation a factor $1/\sqrt{N}$, where N is the number of loudspeakers, is used as the factor for random sequences used in decorrelation.

3 NUMERICAL EXAMPLES

In this section some features of the DirAC technique are demonstrated with numerical examples. The effect of different temporal averaging methods in directional analysis and synthesis is studied first. Second, a case with two sinusoidal sounds arriving from different directions with slightly different frequencies is investigated, and finally a more complicated scene is studied, where two concurrent speech sources are processed with DirAC.

3.1 Temporal Averaging in Directional Analysis and Synthesis

To illustrate the different strategies for temporal averaging of the direction vector $\boldsymbol{D}$ defined in Eq. (6) and of the gain factors defined in Eq. (7), a synthetic example is given in the following. In the example a continuous tone arrives from the direction corresponding to loudspeaker 1, shown in left panel of Fig. 4(a). The situation is stable for the first 49 ms; thus $\boldsymbol{D}$ points stably toward loudspeaker 1. Starting at time instant 49 ms, the vector turns quickly during 2 ms toward the direction corresponding to loudspeaker 3 shown in the middle panel of Fig. 4(a), and the situation remains constant for the last 49 ms, as shown in right panel of Fig. 4(a). The gain factors are now computed with three different strategies for temporal averaging. In the first case the time windows for averaging both $\boldsymbol{D}$ and the gain factors are very short. The computed gain factors are shown in Fig. 4(b), where the gain factor for loudspeaker 1 has a value of 1 for the first 48 ms, and fades after that quickly to zero. Correspondingly the gain factor for loudspeaker 3 ramps up quickly after a 50-ms time instant, and keeps a unity value for the last 48 ms. The quick fades can produce audible clicks in synthesis. Moreover the gain factor for loudspeaker 2 has a high value for a very short time when the direction analyzed turns over it; thus in synthesis a sinusoidal burst will be present in loudspeaker 2 in synthesis, which may be perceived as a click.

In the second case the direction vector $\boldsymbol{D}$ was temporally averaged with a Hanning window with a length of 60 ms. The computed gain factors are shown in Fig. 4(c), where all fades are smooth, and no artifacts will be caused by them. Unfortunately the gain factor for loudspeaker 2 has a high value for a relatively long time, which would produce a perception of sound arriving from its direction, which does not correspond to the direction analyzed.

In the third case $\boldsymbol{D}$ is averaged with short time windows, and the gain factors are averaged with a Hanning window with a length of 60 ms. It can be seen in Fig. 4(d) that the short bump in the gain factor function of loudspeaker 2 smears in time and loses its amplitude. The gain factors for loudspeakers 1 and 3 are faded smoothly, and no artifacts would be audible. However, it has to be taken into account that the averaging window should be as short as possible, since if the gain factors are averaged for too long a time period, the coherence of the loudspeaker signals rises, which blurs the spatial image and can be perceived as coloration.

3.2 Slightly Mistuned Sinusoids from Different Directions

DirAC analyzes the direction of arrival of sound at the frequency bands and uses the analyzed data in the synthesis of sound. The analysis reflects the physical direction of arrival accurately only in the case when all frequency components arrive from a single direction at each analyzed

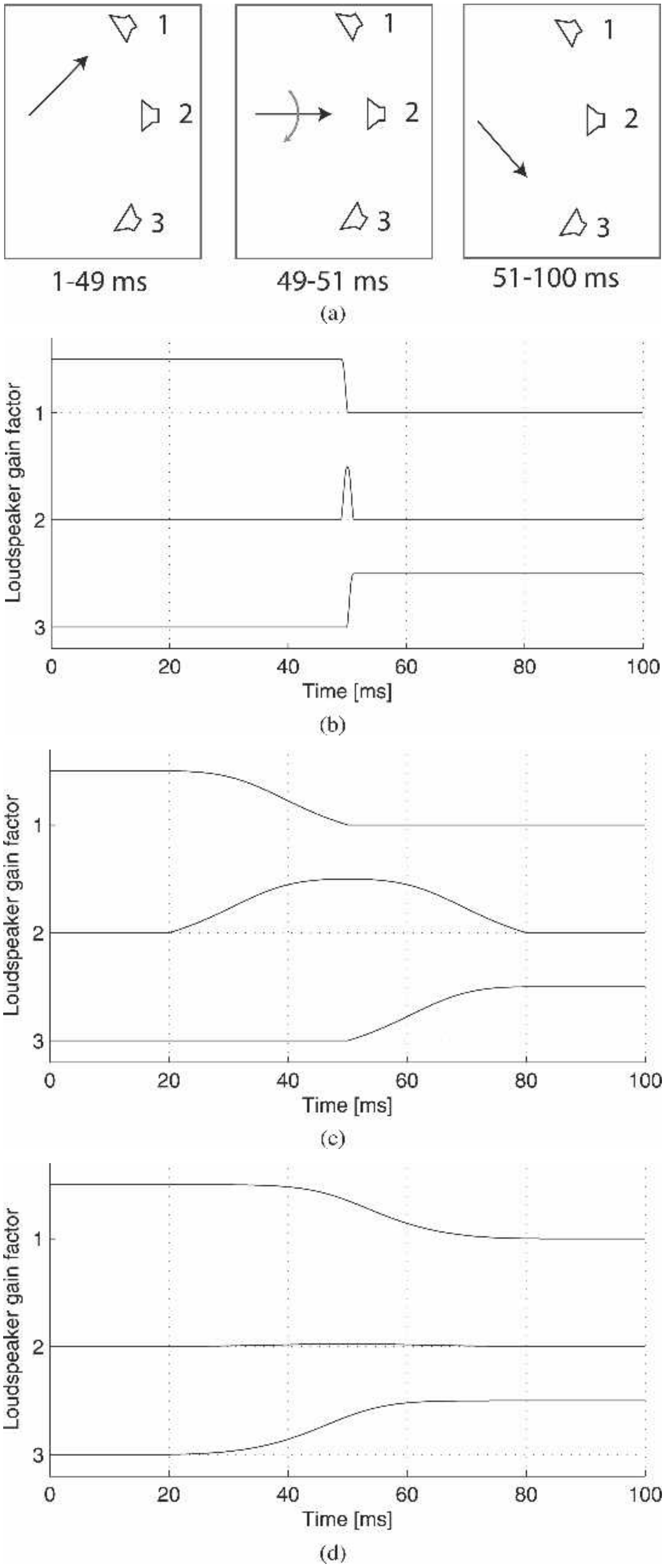

Fig. 4. Example illustrating different temporal averaging methods for directional analysis and synthesis. (a) Analyzed direction vector ***D*** points toward direction of loudspeaker 1 for 49 ms (left panel), turns rapidly over loudspeaker 2 (middle panel), and points toward loudspeaker 3 for last 49 ms (right panel). (b)–(d) Gain factor functions for loudspeakers 1–3. (b) Very short temporal averaging of ***D*** (c) ***D*** averaged prominently. (d) Temporal averaging of gain factors.

frequency band. In practice this situation occurs only after silence during the onset of a single sound source and before reflections arrive. It has to be noted that this is a very important special case, since localization relies largely on the cues decoded during this time period. This effect is called the precedence effect [21]. However, in many cases there are multiple frequency components, which arrive from different directions toward a microphone at each frequency band.

In this numerical example it is studied what happens when there are two frequency components at a single analysis band arriving from different directions. An anechoic listening and recording condition was simulated, which included a sound source in the direction of 45° producing a continuous sinusoid with frequency $f_1 = 563$ Hz, and a sound source of the direction of −45° producing a continous sinusoid with frequency f_2 having 0.125, 0.25, 0.5, or 1 ERBs higher frequency than f_1. In Fig. 5 the following time-dependent functions are presented: signal *w,* signal *wchn,* which is present in the frequency channel of 563 Hz, computed diffuseness, and azimuth angle.

When slightly mistuned sinusoids arrive from different directions toward a B-format microphone, a sinusoid modulated with frequency $f_2 - f_1$ is present in all microphone channels. However, the phase of modulation and the resulting signal may be different in the microphones. With this spatial arrangement, the modulation in the *w* and *x* signals is in the same phase, and in the *y* signal in the opposite phase. Thus when the sinusoids cancel each other out in the *w* and *x* signals, there will still be energy in the *y* signal. This will produce high instantaneous values for diffuseness, which are, however, smoothed due to temporal averaging, as shown in Fig. 5. When the sinusoids cancel each other in the *y* signal, the diffuseness function has its local mimimum. The cancellations of sinusoids have a similar effect on the azimuth function.

When $f_2 - f_1$ increases, there are more frequent oscillations in the azimuth and diffuseness functions. The oscillations are less prominent in the diffuseness function than in the azimuth function, since the short-time averaging window length is longer in diffuseness analysis than in direction analysis. It can be seen that as the frequency separation increases, the direction analyzed moves towards 45° on average, whereas diffuseness decreases toward zero. Also the modulation is less prominent in *wchn* with increasing frequency separation, as opposed to the

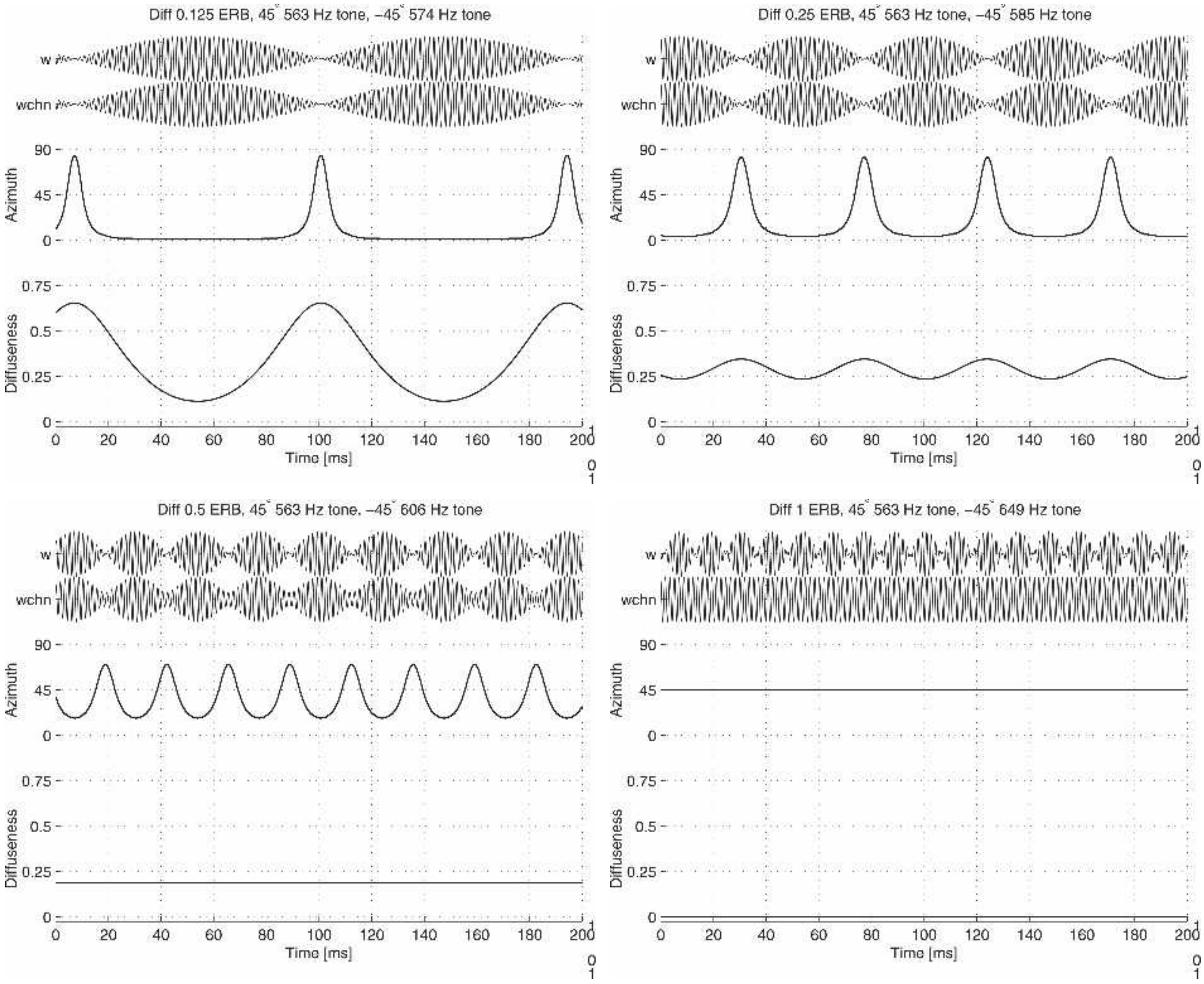

Fig. 5. DirAC parameters analyzed at frequency band with center frequency of 563 Hz. Simulated sound scene included a 563-Hz tone from 45°. Another tone with variable frequency was produced by sound source in direction of −45°.

fact that the depth of modulation is constant in *w*. These changes occur since the interfering frequency f_2 is less and less present in the frequency channel corresponding to f_1. The analysis results thus are more and more based on a single sinusoid instead of two sinusoids.

The gain factors were synthesized for a five-loudspeaker setup, having loudspeakers C (center, 0°), R (right, 45°), L (left, −45°), RS (right side, 110°), and LS (left side, −110°). The computed gain factors remain constant with time due to long time windows in the averaging process. Thus they are discussed only verbally, since their plotting would give no extra information. In the 0.125-ERB difference case, the C gain factor has constantly a value of almost unity, and R and RS have constant low values. As frequency separation is increased, the C gain factor decreases, the R gain factor increases prominently, and RS increases slightly. In the 1-ERB difference case the frequencies are separated perfectly, as the modulation is nonexistent in signal *wchn.* As the gain factor R has a unity value for this signal, it can be concluded that in this case DirAC reproduces the signals in space perfectly, as expected, due to the large enough spectral separation between the frequency components.

It is clear that energetic analysis cannot decode the directions of the frequency components correctly, when the components are so narrowly spaced in frequency that they are analyzed within the same frequency channel. However, we assume that the separation is not necessary in sound reproduction, since it is known that humans also cannot do it. When two slightly mistuned sinusoids are presented with loudspeakers in an anechoic space, the listener perceives a broad image, the properties of which are dependent on spatial and spectral separation between loudspeakers and sinusoids, respectively [11].

The case simulated in this section was reproduced with two loudspeakers in an anechoic chamber as a reference case, and then informally compared to the DirAC reproduction with a similar five-loudspeaker setup as used in the simulation. The resulting spatial images were similar, and both reference and DirAC reproduction were perceived as broad spatial images when the difference $f_2 - f_1$ was below approximately 0.5 ERB. The perceptual difference was small, however audible. In all cases the DirAC reproduction was perceived to produce a slightly wider sound image than the reference. Also, the timbre was slightly different for the reference and the reproduction.

3.3 Two Concurrent Sources

Finally a more complex scenario was tested. A male speech source was positioned at 45° and a female speech source at −45° of azimuth in a simulated presentation with respect to a simulated B-format microphone. This virtual recording was processed with the high-quality implementation of DirAC with 2-ERB-wide frequency channels. The computed parameters are presented in Fig. 6 for a

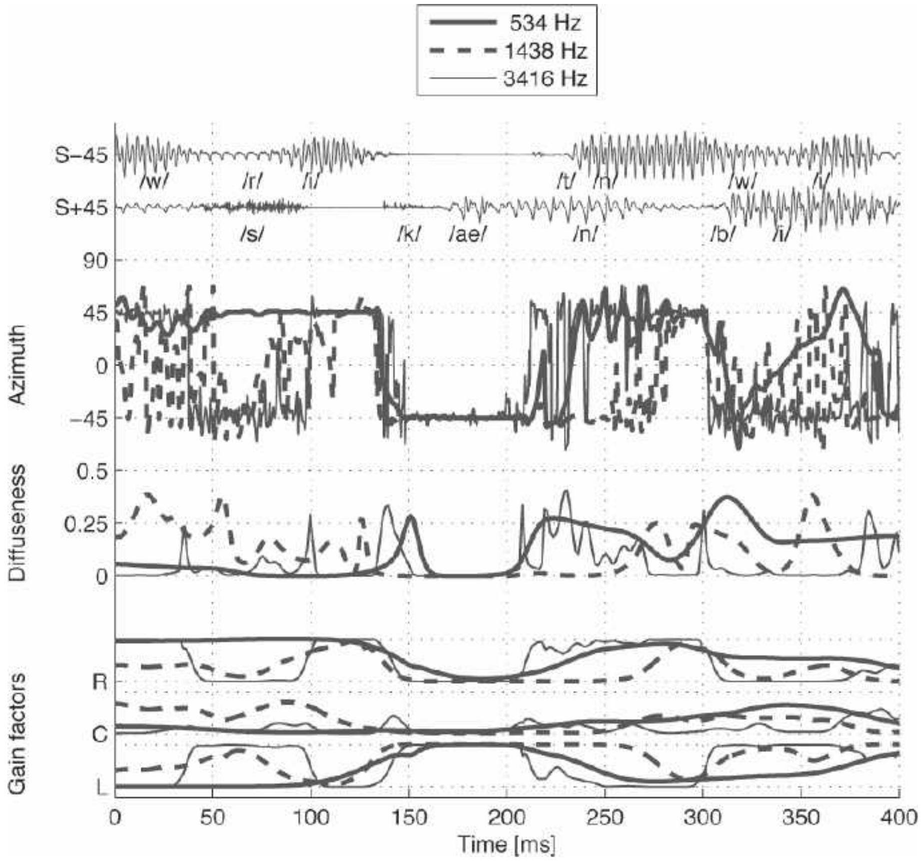

Fig. 6. DirAC parameters analyzed at a frequency band with center frequencies shown. In simulated sound presentation a male speech source was positioned at 45° and a female speech source at −45° of azimuth. Presentation was captured with a simulated B-format microphone.

short segment of speech. It can be seen that the direction analyzed fluctuates between ±45°, lower frequency bands being slower and higher frequency bands faster. The analyzed diffuseness function gets relatively low values at all time instants, since there were no reverberations in the recording. However, in most cases where the direction changes rapidly, the diffuseness gets larger values, as, for example, with a 1438-Hz band in the beginning of the segment. Such rapid changes and increased diffuseness may occur when both signals arriving at a microphone have a nearly equal amount of sound energy at the same frequency band.

As longer time windows are used in the gain factor averaging than in the direction vector averaging, the change of gain factors with time is slower than the corresponding temporal change of direction. The effect of gain factor averaging can also be seen in the gain factor function of the 3416-Hz channel. Although the analyzed direction fluctuates rapidly between ±45°, most of the time either L or R gains have values near unity, and most of the time the C gain factor is close to zero. If the direction vector had been averaged, the C gain factor would have had larger values than the L or R gain factor.

However, based on graphical observations in this figure nothing can be said about how the sound is perceived in this case. The original case reproduced with L and R loudspeakers of the 5.0 setup was compared to the DirAC processed version in a ITU-R BS.1116 listening room with loudspeakers at 2.5-m distance from the best listening position. According to informal observations of multiple sound researchers it was concluded that the difference is not audible without very careful listening when the listeners are not closer than about 1 m from the loudspeakers. The corresponding audio files for 5.0 listening are available at [22].

4 APPLICATIONS

In this section different applications DirAC are discussed.

4.1 High-Quality Versatile Reproduction of Spatial Sound

A natural application for DirAC is the delivery of spatial sound for arbitrary domestic listening conditions. In the production of audio material, DirAC provides new types of processing of B-format recordings. The spatial image can be modified after directional analysis by modifying the directions analyzed, or amplifying sound depending on the direction. Also, as the diffuse and nondiffuse streams are separated, they can be weighted differently to emphasize either dry or reverberant sound. This also enables different effects on different streams. For example, diffuse sound, which is mostly reverberant sound, can be equalized differently than nondiffuse sound.

The audio delivery chain for domestic applications could also be based on DirAC processing. The audio could be stored either as a mono channel with metadata of direction and diffuseness, or as a three-or four-channel B-format signal with or without metadata. The production of such an audio format could be based on either real recordings or mixes. The existing stereo and multichannel material could be decoded into B-format with simple matrixing, as discussed in Section 4.3.

The domestic audio device would then decode the sound into the desired loudspeaker setup, or to headphones. The systems could be made self-calibrating, equipped with a small microphone array capable of measuring the directions and distances of the loudspeakers.

The high-quality version of DirAC was demonstrated with recorded B-format material and 12-channel three-dimensional loudspeaker setup at the AES 28th International Conference. The DirAC reproduction was acknowledged to provide better "openness of reverberation, depth of soundfield, and image stability when moving around or sitting off center" than the corresponding virtual cardioid decoding in the conference report [23]. Different B-format recordings have also been processed for 5.0 listening with DirAC and virtual hypercardioids corresponding to Ambisonics processing, and the files are available in [22].

4.2 Teleconferencing

In teleconferencing the application of DirAC is interesting when, at least at one end of the transmission line, there are multiple persons. The aim would be that the directions of the talkers are reproduced to increase the intelligibility in multitalker or noisy environments, and to increase the naturalness of the connection. As the sound quality requirements are not as high for teleconferencing as for music listening, the quality can be compromised for shorter latency in encoding and decoding, for narrower bandwidth in transmission, for lower computational complexity, and for cheaper microphone devices. The DirAC implementation has already been discussed, and its flow diagram is shown in Fig. 3.

The principles of DirAC teleconferencing were published in [13], and further details and a full-duplex C-language implementation appear in [24]. The frequency analysis was based on STFT, and a mono PCM audio channel was transmitted with a narrow sideband with azimuth and optionally diffuseness values averaged over each ERB band. The microphones used were composed of three or four omnidirectional electret capsules arranged as a rectangular triangle or a square, with the distances between microphones being approximately 20 to 40 mm. This enabled the computation of dipole signals x and y by subtracting pairs of the microphone signals. As the phase difference of a signal is very small at low frequencies with only a 20–40-mm microphone distance, the internal noise of the capsules is effectively amplified in the computed dipole signals at low frequencies. However, as in this case, the dipole signals are used only to steer the omnidirectional signal spatially, this noise is not audible as an audio signal. It could be perceived as changing localization at low frequencies. However, the temporal averaging at low frequencies removes its effects in a great deal.

The system has been tested with several concurrent speech sources. The sources were perceived stable at the original locations. Demonstrations decoded for 5.0 listening are available at [22]. It was also demonstrated at the AES 30th International Conference with full-duplex connection between two rooms equipped with 5.0 loudspeaker layouts and grids with four omnidirectional microphones [24].

4.3 Stereo Upmixing

Generally in upmixing, a two-channel stereophonic audio file is taken as input, and it is processed to yield loudspeaker signals for a desired multichannel loudspeaker setup. There exist a variety of techniques for this, some of which have been discussed in [25]. Avendano and Jot proposed a method where the stereophonic sound is analyzed and synthesized with similar principles as in DirAC [25]. Their method computes the coherence between left and right channels, which is then used to divide sound into diffuse and nondiffuse sound. The direction of nondiffuse sound is computed using a similarity index.

The DirAC method can also be used to upmix stereo files to multichannel setups or headphones. The main idea is to simulate an anechoic B-format recording, and then use DirAC to reproduce the computed signals. The recording can be simulated by applying left- and right-channel signals l and r to simulated loudspeakers in the directions of ±30° with respective to a simulated B-format microphone in the far field. In practice this is implemented as simple matrixing of left and right signals as $w = (l + r)/\sqrt{2}$, $x = (l + r)/\cos\alpha$, and $y = l/\cos(\pi + \alpha) + r/\cos(\pi - \alpha)$, where α is the angle between the frontal direction and one of the loudspeakers in the simulated stereophonic layout. The DirAC method can then be used to produce sound for different reproduction methods.

This application has been implemented [26], and informal testing of the system suggested that the concept of using DirAC in upmixing is valid and usable in the upmixing of two-channel sound to a 5.0 system. If there existed some amplitude panned sources, or in general signals in the same phase, the direction analyzed varied between the left and right loudspeakers. Thus in DirAC synthesis these signals were pairwise panned between only left, center, and right loudspeakers. In contrast, when the signals were out of phase, such as reverberation or time-panned signals, the direction analyzed could also be outside the directions spanned between the left and right loudspeakers. In these cases the signal was also analyzed to be prominently diffuse. Thus such out-of-phase signals were also applied to side channels, both as diffuse sound and as pointlike virtual sources.

5 DISCUSSION

5.1 Relation between Analyzed Diffuseness and Direct-to-Reverberant Ratio

DirAC analysis could be interpreted as a mechanism to separate between direct and reverberant sound, where the term "direct sound" denotes the sound via direct path from sources to the microphone, and the term "reverberant sound" denotes the sound via the room to the microphone. However, the measure of diffuseness in DirAC analysis does not necessarily reflect the ratio between direct and reverberant sound. In many cases this may be the true, but not in all of them. This is illustrated by a few examples. If a single source is recorded in an anechoic environment, the sound will be analyzed totally nondiffuse. However, when there are spatially separated sound sources in an anechoic space, the analyzed diffuseness will depend on the signal content and the spatial positions of the sources. If, for example, two sources are on the opposite sides of a microphone and produce sound in a specific frequency channel at the same time, the recorded sound will be analyzed with Eq. (3) to be diffuse, as the net transport of energy is smaller than the total energy. On the other hand, the effect of the room will increase diffuseness in most cases, but in some cases it will not. For example, if a single source presents a short sound, distinct echoes will be analyzed to be totally nondiffuse. Also, although the reverberant sound field is diffuse as the long-term average, the relatively short time average of diffuseness used in DirAC analysis fluctuates with time between the values of 0 and 1, as can be seen in [6, Fig. 4(c)]. Thus in general the diffuseness of sound does not reflect the reverberance of sound directly, and nondiffuseness of sound does not reflect the dryness of sound directly.

5.2 Future Work

So far all DirAC parameter values, such as the lengths of time windows for temporal averaging and the parameters for time–frequency analysis have been defined by informal listening during the development. The possibility of using some more advanced methods, such as formal listening tests or auditory modeling to find optimal parameter values, is reserved for future work.

The functioning was illustrated only with some numerical examples, with results of informal listening tests, by providing some references to multichannel listening material, and by referring to a conference demo report [23]. Publishing the method without rigorous perceptual testing was deemed sufficient, because it was considered that including the perceptual tests would have delayed the publication for a long time period, and that the amount of information would have been too large for a single publication. The measurement of perceptual quality obtainable with the methods is thus reserved for future work.

5.3 Relation between Parametric Multichannel Audio Coding and DirAC

DirAC technology is closely related to recently developed technologies for coding multichannel audio files to one or two audio tracks with metadata, and then back to multichannel presentation [27], [28], called parametric multichannel audio coding. The input audio channels are decomposed into frequency channels, and the differences between audio and channels are analyzed depending on time. The differences include such properties as interchannel level, temporal differences, and interchannel coherence.

As can be seen, DirAC processing resembles parametric multichannel audio coding methods. However, there are two main points which make DirAC not directly comparable to them.

- In audio coding techniques the input is a multichannel file containing signals meant to be reproduced with loudspeakers. This is different in DirAC, where the input is microphone signals, which may not be applied directly to loudspeakers. However, the output of DirAC is meant to be produced using loudspeakers. Thus besides being a coding method, DirAC must also be seen as a microphone technique.
- In the multichannel audio coding techniques, the metadata consist typically of differences between loudspeaker signals, such as level and temporal differences. In DirAC the metadata consist of estimates of physical quantities, such as sound direction and diffuseness. Thus DirAC is less tied to certain loudspeaker setups.

It has to be noted that a version of the audio coding techniques has been proposed which largely resembles, DirAC processing in that instead of analyzing interchannel differences the direction of sound is analyzed based on loudspeaker signals [29].

In principle DirAC can be used to perform the parametric multichannel surround coding to encode multichannel audio files into a stream with one or three audio tracks similarly as in [29]. The encoding could be performed in a fashion similar to stereo upmixing. The listening setup would be simulated in anechoic B-format recording, and DirAC would be applied to estimate the directions and diffuseness in the frequency bands. However, this has not been tested.

5.4 Differences between First-Order Ambisonics and DirAC

In first-order Ambisonics the major problem is the high coherence between loudspeaker signals. The DirAC implementation presented can be interpreted to be a dynamic enhancement of first-order Ambisonics. In the flow diagram of DirAC, presented in Fig. 2, virtual cardioid microphone signals are first computed, which can be interpreted as Ambisonics-style processing. In DirAC the loudspeaker signals are divided dynamically into diffuse and nondiffuse streams, which makes it possible to avoid the coherence-related problems in both streams. The diffuse part of sound is decorrelated for each loudspeaker to reduce the coherence, and the nondiffuse part of sound is pairwise or tripletwise panned, which also decreases the coherence.

6 CONCLUSIONS

The directional audio coding (DirAC) technique is presented in this engineering report. DirAC is a technique to reproduce spatial sound over different reproduction systems, and it can also be used to encode spatial aspects of sound as metadata which are transmitted or stored along with a single or several audio channels. The technique is based on analyzing the sound direction and diffuseness depending on time at narrow frequency bands, and further using these parameters in sound reproduction with appropriate techniques. DirAC is based on spatial impulse response rendering (SIRR) [6]. However, new methods for parameter temporal averaging are developed in this study. Also, new techniques to gain better spatial selectivity for both diffuse and nondiffuse synthesis are suggested. The main applications, which are addressed, are high-fidelity reproduction of spatial sound recordings and teleconferencing. The system is demonstrated by means of numerical examples, and with results from informal listening tests.

7 ACKNOWLEDGMENT

This work has received funding from the Academy of Finland under project 105780 and from the Emil Aaltonen foundation. The author wishes to thank Juha Merimaa and Christof Faller for discussions during the work. Jukka Ahonen is thanked for help with microphone tests in telecommunication applications.

8 REFERENCES

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

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

[3] J. Daniel, R. Nicol, and S. Moreau, “Further Investigations of High-Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging,” presented at the 114th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 51, p. 425 (2003 May), convention paper 5788.

[4] A. Laborie, R. Bruno, and S. Montoya, “Designing High Spatial Resolution Microphones, “presented at the 117th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 53, p. 99 (2005 Jan./Feb.), convention paper 6231.

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

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

[7] V. Pulkki and J. Merimaa, “Spatial Impulse Response Rendering: Listening Tests and Applications to Continuous Sound,” presented at the 118th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 53, p. 674 (2005 July/Aug.), convention paper 6371.

[8] V. Pulkki and J. Merimaa, “Spatial Impulse Response Rendering II: Reproduction of Diffuse Sound and Listening Tests,” *J. Audio Eng. Soc.,* vol. 54, pp. 3–20 (2006 Jan./Feb.).

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

[10] D. W. Grantham, “Spatial Hearing and Related Phenomena,” in *Hearing,* B. C. J. Moore, Ed. (Academic Press, San Diego, CA, 1995), pp. 297–345.

[11] D. R. Perrott, “Discrimination of the Spatial Dis-

tribution of Concurrently Active Sound Sources: Some Experiments with Stereophonic Arrays," *J. Acoust. Soc. Am.,* vol. 76, pp. 1704–1712 (1984 Dec.).

[12] C. Faller and J. Merimaa, "Source Localization in Complex Listening Situations: Selection of Binaural Cues Based on Interaural Coherence," *J. Acoust. Soc. Am.,* vol. 116, pp. 3075–3089 (2004 Nov.).

[13] V. Pulkki and C. Faller, "Directional Audio Coding: Filterbank and STFT-Based Design," presented at the 120th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 54, p. 670 (2006 July/Aug.), convention paper 6658.

[14] F. J. Fahy, *Sound Intensity* (Elsevier Science Publ., Essex, UK, 1989.)

[15] J. Merimaa, "Analysis, Synthesis, and Perception of Spatial Sound—Binaural Localization Modeling and Multichannel Loudspeaker Reproduction," Ph.D. thesis, Helsinki University Technology (2006). Available at http://lib.tkk.fi/Diss/2006/isbn9512282917/.

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

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

[18] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-Related Transfer Functions of Human Subjects," *J. Audio Eng. Soc.,* vol. 43, pp. 300–321 (1995 May).

[19] D. Begault, "3-D Sound for Virtual Reality and Multimedia" (NASA, 2000).

[20] R. B. Schulein, "Microphone Considerations in Feedback-Prone Environments, "*J. Audio Eng. Soc.,* vol. 24, pp. 434–445 (1976 July/Aug.)

[21] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The Precedence Effect," *J. Acoust. Soc. Am.,* vol. 106, pp. 1633–1654 (1999 Oct.).

[22] V. Pulkki, "Directional Audio Coding Web Pages," http://www.acoustics.hut.fi/research/cat/DirAC/,2006.

[23] "AES 28th International Conference—The Future of Audio Technology—Surround and Beyond," *J. Audio Eng. Soc.* (*Features*), vol. 54, pp. 858–864 (2006 Sept.).

[24] J. Ahonen, V. Pulkki, and T. Lokki, "Teleconference Application and b-Format Microphone Array for Directional Audio Coding," in *Proc. AES 30th Int. Conf.* (Saariselkä, Finland, 2007), CD-ROM proceedings.

[25] C. Avendano and J. M. Jot, "A Frequency-Domain Approach to Multichannel Upmix, "*J. Audio Eng. Soc. (Engineering Reports),* vol. 52, pp. 740–749 (2004 July/Aug.).

[26] V. Pulkki, "Directional Audio Coding in Spatial Sound Reproduction and Stereo Upmixing," in *Proc. AES 28th Int. Conf.* (Piteå, Sweden, 2006), pp. 251–258.

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

[28] L. Villemoes, J. Herre, J. Breebaart, G. Hotho, S. Disch, H. Purnhagen, and K. Kjörling, "MPEG Surround: The Forthcoming ISO Standard for Spatial Audio Coding," in *Proc. AES 28th Int. Conf.* (Piteå, Sweden, 2006), pp. 213–230.

[29] M. Goodwin and J. M. Jot, "A Frequency-Domain Framework for Spatial Audio Coding Based on Universal Spatial Cues," presented at the 120th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 54, pp. 700, 701 (2006 July/Aug.), convention paper 6751.

THE AUTHOR

Ville Pulkki was born in 1969 in Jyväskylä, Finland. He received M.Sc. and D.Sc. (Tech.) degrees from the Helsinki University of Technology, Finland, in 1994 and 2001, respectively. He majored in acoustics, audio signal processing, and information sciences.

From 1994 to 1997 he was a full-time student in the Musical Education department at the Sibelius Academy. His thesis describes his development of vector base amplitude panning (VBAP), which is a method to position virtual sources to any loudspeaker configuration, and the study of its performance with psychoacoustic listening tests and with modeling of auditory localization mechanisms. The VBAP method is widely used in multichannel virtual auditory environments and in computer music installations. It is utilized in two commercial and multiple noncommercial software.

Dr. Pulkki works in the Laboratory of Acoustics and Audio Signal Processing at the Helsinki University of Technology, Finland. His research activities include methods to reproduce spatial audio and methods to evaluate quality of spatial audio reproduction. He has also worked on diffraction modeling in interactive models of room acoustics. He enjoys playing various musical instruments and singing.