Directional audio coding in spatial sound reproduction and stereo upmixing

Ville Pulkki

Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland

Correspondence should be addressed to Ville Pulkki (Ville.Pulkki@tkk.fi)

ABSTRACT

Directional audio coding (DirAC) is a method for spatial sound representation, applicable to arbitrary audio reproduction methods. In the analysis part, the diffuseness and direction of arrival of sound is estimated in a single point depending on time and frequency. In the synthesis part, microphone signals are first divided into non-diffuse and diffuse parts, and are then reproduced using different strategies. In this paper, the applications of DirAC to reproduce recorded B-format sound and to upmix stereo in multi-channel listening are described.

1. INTRODUCTION

Methods for multichannel audio reproduction of spatial sound have been debated for years. None of the traditional methods reproduce faithfully the reality and is applicable to 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 pattern of current high-quality microphones are only of zeroth or first order, the resulting loudspeaker signals are more coherent than desired. This results in coloring and blur of the spatial image. Recently, there has been research on Higher-order Ambisonics [3, 4]. The sound field is captured with a large number of microphones, which 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 costs.

In spaced microphone techniques the microphones are spatially separated from each other with distances between few centimeters to few meters. They also solve some problems of first-order coincident microphone techniques. Due to temporal differences in arrival of sound at the microphones, the resulting loudspeaker signals are less coherent. This circumvents some of the problems of Ambisonics, but 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 as Ambisonics does.

The goal of the proposed Directional Audio Coding (DirAC) is to reproduce spatial sound as realistically as possible in arbitrary sound reproduction systems. Also, it is targeted to make it possible to transmit spatial aspects of audio as side information to monophonic audio channel with minimal bit rate. DirAC is based on the same principles and partly the same methods as the Spatial impulse response rendering (SIRR) technique proposed recently [5]. SIRR is a technique to reproduce room impulse responses over multi-loudspeaker arrays for application in convolving reverberators. In analysis part, the sound direction and diffuseness is estimated depending on time and frequency, and in synthesis a single channel of audio is rendered to whatever reproduction system using the analyzed data. The technique can be in principle applied also to reproduce continuous sound over an arbitrary reproduction method, as suggested in [6]. In [7] the DirAC was developed from SIRR, and the applications for teleconferencing and B-format reproduction were presented.

In this paper, the DirAC method is further elaborated, and tuned for the two main applications considered in this paper, the reproduction of recorded spatial sound and the upmixing of stereophonic audio files. The paper is organized as follows: The assumptions about perception of spatial sound are first reviewed based on which the SIRR and DirAC techniques were developed. Next, the theory and implementation of DirAC is presented, after which some informal listening test results are discussed. The application of DirAC in stereo upmixing is considered last.

2. ASSUMPTIONS ABOUT PERCEPTION OF SPATIAL SOUND

The design of the SIRR and DirAC techniques is based on five assumptions about the interaction between sound field properties and perceptual attributes that they produce [5]. The assumptions are repeated here, and their validity is discussed. The assumptions are:

1. Direction of arrival of sound will transform into interaural time difference (ITD), interaural level difference (ILD), and monaural localization cues.
2. Diffuseness of sound will transform into interaural coherence cues.
3. Timbre depends on the monaural (time-dependent) spectrum together with ITD, ILD, and interaural coherence of sound.

Assumption 1 has been proven in many listening tests [8] and the cues are produced by interaction between the sound field and the listener. Coherence as a measure of diffuseness, as implied in Assumption 2, has been discussed in [9]. In concert hall acoustics it is common to use interaural cross-correlation to measure spatial impression [10]. The assumption was also verified with an auditory model simulation in [11]. In Assumption 3 the auditory cues affecting binaural timbre are considered. It is known that timbre depends on the short-time monaural spectrum [12]. However, the knowledge on the influence of binaural properties on timbre perception is quite sparse. Frequency-dependent manipulations of the phase difference in headphone listening of broad-band sound alter timbre and produce pitch-like effects [13, 14], thus ITD affects timbre. The effect of ILD on binaural timbre has not been studied. However, it is known that ILD affects the binaural speech reception threshold [15] and it is assumed assumed that it also affects timbre. Timbre perception has been studied also in room acoustics. Room reflections may affect the timbre but listeners are at least partially insensitive to the coloration caused by the room [16, 17].

To create correct localization cues, interaural coherence, and timbre, the physical attributes that are to be reproduced are thus direction of arrival, diffuseness, and the spectral properties of the sound. The temporal and spectral resolution of human hearing must also be taken into account in the reproduction. Thus the fourth assumption is:

4. The direction of arrival, diffuseness, and spectrum of sound measured in a point with the temporal and spectral resolution of human hearing determines the auditory spatial image the listener is perceiving.

The implication of this is that a reproduced sound field is perceived similarly as long as the direction of arrival, diffuseness, and spectrum of sound are similar in the listening point.

In SIRR, the number of sources was always one. Thus estimated directions and diffuseness were always evoked by that source. In DirAC, multiple sources are also allowed. However, the assumptions are not changed. It is assumed that the listener can not decode separate cues for separate sources, but only single cues which have been produced by ear canal signals summed from different sources. This assumption is in line with a recently proposed auditory model for source localization [18], which hypothesizes that the auditory system obtains the source direction by considering the binaural directional cues only at time instants when they correspond to one of the source directions.

3. BASIC PRINCIPLES OF DIRECTIONAL AUDIO CODING (DirAC)

The flow diagram of DirAC is presented in Fig. 1. The principles in DirAC design are based on the assumptions stated in Section 2. That is, 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. The time-frequency decompositions which have been used in DirAC are short-time Fourier transform and a filterbank [7], as indicated in the figure.

The assumptions also imply that the direction of arrival, diffuseness, and spectrum of the sound field should be synthesized correctly. Thus, the direction of arrival and diffuseness are analyzed with a temporal accuracy comparable to the accuracy of the human auditory system at each frequency band. In this work energetic analysis of the sound field is used, as shown in the figure, although different techniques for this could be applied.

The next step is to transmit the audio and estimated properties of the sound field. Depending on the application,

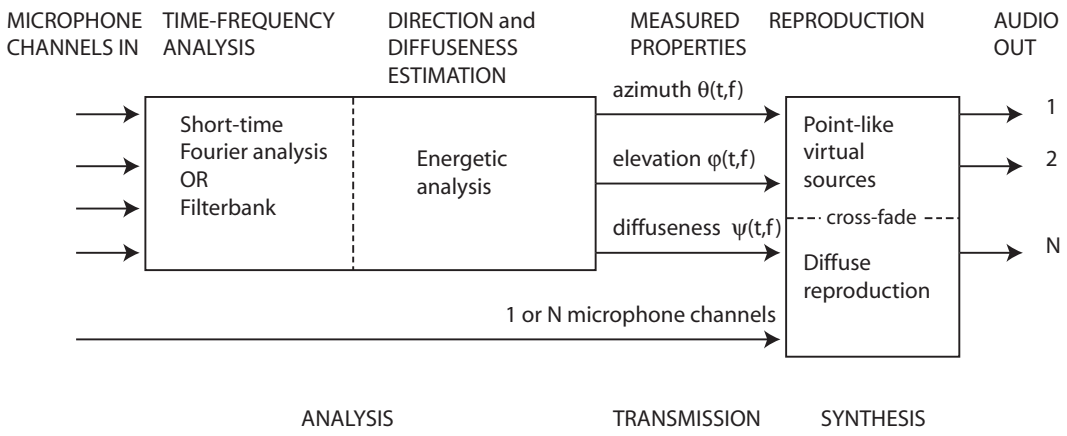

Fig. 1: Overall flow diagram of Directional audio coding.

different parameters with different accuracy can be transmitted. The number of audio channels to be transmitted can also vary, depending on the application.

The DirAC decoder receives the audio channel(s) and estimated properties, and generates loudspeaker signals. To reproduce the auditory cues of the recording situation, the decoder tries to generate the correct spectrum by using reproduction methods which color the sound minimally and also targets to reproduce spatial aspects of sound as well as possible to produce correct auditory cues for the listener.

4. DirAC FOR SPATIAL SOUND REPRODUCTION AND STEREO UPMIXING

In this study, the target is to reproduce sound with as good fidelity as can be achieved, without considering restrictions in computational complexity. In this case DirAC is also considered to be simply a method for reproducing B-format audio signals, or a method for stereo upmixing. In these applications the transmission of B-format or stereo signal is performed before analysis, or alternatively the reproduced sound is transmitted after DirAC processing, thus the transmission needs not to be taken into account in the method. The flow diagram of the implementation described in this paper is shown in Fig. 2.

4.1. Direction and diffuseness estimation

In the current implementation, B-format microphones have been solely used. B-format stands for a coincident microphone array capable producing four microphone channels with different directional characteristics: one omnidirectional, and three figure-of-eight channels directed towards orthogonal axis. The corresponding signals are $w(n)$, $x(n)$, $y(n)$ and $z(n)$, respectively, where n is the time index. 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.

4.1.1. Time-frequency analysis

Two different techniques have been used to divide signals in time and frequency. One is the same technique as utilized by SIRR, the short time Fourier transform (STFT). However, this approach has some shortcomings, such as uniform temporal accuracy with frequency. In informal listening reported in [7] it was found that the directions of rapid transients in presence of other sound events were reproduced erroneously with STFT. Thus, in this paper this technique is not used, and instead the filterbank solution is applied.

The filterbank was implemented simply by transforming the signal to spectral domain with FFT, by windowing the complex spectrum with a window for each frequency channel, and by transferring the windowed spectra back into temporal domain. The window spacing is based on the ERB frequency scale, in this implementation the distance between adjacent bands was two ERBs. Since the scale is not linear with frequency, the rise and decay of a single window originate from Hanning windows with different lengths. The length of rise and decay was

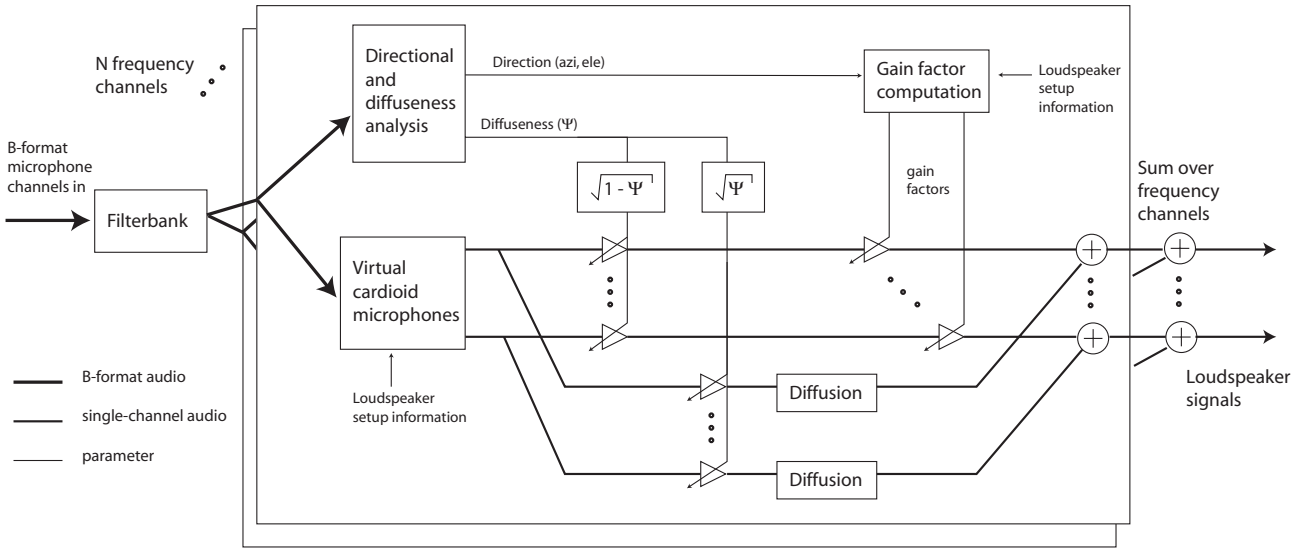

Fig. 2: The flow diagram of Directional audio coding for B-format sound reproduction.

the frequency distance between adjacent frequency channels. The filtered signals are denoted with an additional parameter i, for example $w(n,i)$ denotes the omnidirectional signal at frequency channel i.

4.1.2. Directional analysis

The directional analysis is based on energetic analysis of sound field, which is now shortly reviewed. The instantaneous velocity vector is composed as

$$\mathbf{v}(n,i) = x(n,i)\mathbf{e}_x + y(n,i)\mathbf{e}_y + z(n,i)\mathbf{e}_z\,, \qquad (1)$$

where $\mathbf{e}_x$, $\mathbf{e}_y$ and $\mathbf{e}_z$ represent Cartesian unit vectors. Instantaneous intensity $\mathbf{I}$ is computed as

$$\mathbf{I}(n,i) = w(n,i)\mathbf{v}(n,i)\,, \qquad (2)$$

and instantaneous energy as

$$E(n,i) = w^2(n,i) + ||\mathbf{v}||^2(n,i)\,, \qquad (3)$$

where $||\cdot||$ denotes vector norm. The diffuseness is computed with

$$\psi(n,i) = 1 - \frac{\sqrt{\sum_{m=-M/2}^{M/2} ||\mathbf{I}(n+m,i)||^2 W_1(m)}}{\sum_{m=-M/2}^{M/2} E(n+m,i)W_1(m)}, \qquad (4)$$

where $W_1(m)$ is a window function defined between $-M/2$ and $M/2$ for short-time averaging. Hanning windows are used with different lengths at different frequency bands. The instantaneous direction is computed as

$$\mathbf{D}(n,i) = -\sum_{m=-M/2}^{M/2} \mathbf{I}(n+m,i)W_2(m), \qquad (5)$$

where W_2 is the Hanning window for short-time averaging $\mathbf{D}$.

4.2. Sluggishness

In SIRR, the temporal variations of the analyzed direction and diffuseness were not limited. If such variations are applied in DirAC for reproduction of continuous sound, there will be audible artifacts and distortion. Such fast variations could be allowed in SIRR, since the signal was a priori known to be an impulse, the flat spectrum of the signal masked such non-idealities. However, with DirAC they can not be allowed. In this section, the strategies needed for slowing down the temporal variables are discussed separately for diffuseness and direction.

The slowing down of diffuseness is quite straightforward task, only the variables from which the diffuseness is computed are temporally integrated. Different values were used for the length of Hanning window $W_1(m)$ in Eq. 4. Typically the length was 10–50 times the cycle length of center frequency at corresponding frequency band, however, 3 ms at minimum and 150 ms at maximum.

The slowing down of direction is a slightly more com-

plicated task. The synthesis technique of point-like virtual sources has to be taken into account. In the analysis part, the directional information is slowed down only slightly, the length of Hanning window $W_2(m)$ was only three times the cycle length of center frequency corresponding frequency band, however, 1 ms at minimum. It is desired, that the estimated direction vector would follow the direction of sound without latency or sluggishness. The second stage of temporal averaging related to reproduction of directional information is performed in the synthesis stage, and it will be discussed in Sec. 4.3.1.

4.3. Synthesis techniques

In this section, the techniques currently used to create point-like virtual sources and to synthesize diffuse sound are described.

4.3.1. Point-like virtual sources

The choice of the optimal method to create point-like sources is dependent on the selected reproduction method. In this work, the typical horizontal loudspeaker systems has been considered, such as 5.1 surround setups and an octagonal setup. Simple pair-wise amplitude panning has been used with vector base amplitude panning (VBAP) formulation [19]. Low-frequency effects channels are not discussed here, thus in the following the 5.1 system is referred as 5.0.

As mentioned in Sec. 4.2, the second phase of temporal averaging in directional reproduction has to be computed in the synthesis phase. The **D** vector varies rapidly in time. If it is applied directly as panning direction, there will be audible artifacts. If a frequency band is reproduced only a short time with one loudspeaker due to rapid change of panning direction, a click may be audible. If this artifact is removed by temporally averaging the direction vector, the directional information will be wrong, and fast sweeps are replaced by slow sweeps. This can produce artifacts which sound like “bubbling”.

Such artifacts are avoided by temporally averaging the gain factors instead of temporally averaging the direction:

$$g(n,i,k) = \frac{\sum_{m=-M/2}^{M/2} g(n+m,i,k)\psi(n+m,i)W_3(m)}{\sum_{m=-M/2}^{M/2} \psi(n+m,i)W_3(m)}, \quad (6)$$

where $g(n,i,k)$ is the gain factor at subband i, time n and loudspeaker channel k. In the current tests, the length of Hanning window $W_3(m)$ was as high as 100 times the cycle length of center frequency of corresponding channel, however, with maximum of 1000 ms. However, the weighting with ψ ensures that if an onset with low diffuseness occurs, it will be correctly localized in reproduction phase.

In this way, the rapid sweeps over a loudspeaker will smear over time, and lose their peak amplitude. However, the transients will be synthesized from the correct direction. For example, if in the 5.0 setup the onset of continuous tone is synthesized with the center loudspeaker, and when the first reflection arrives, the analyzed direction turns rapidly towards to the direction corresponding to the left surround loudspeaker. In this case, if the gain factors are not averaged, a sinusoidal burst will be audible in the left loudspeaker, while the direction points towards it. When the gain factors are averaged, the sound energy of the burst will smear in time, and no artifacts will be audible. If the gain factors are averaged for a too long time period, the loudspeaker signal coherence rises, which can be perceived as coloration.

In earlier studies, the $w(n,i)$ signal was directly multiplied with each gain factor to yield the non-diffuse sound. This has one shortcoming: only in the case when diffuseness reaches value zero, $w(n,i)$ includes only non-diffuse sound. When diffuseness is greater than zero, which is generally the case, $w(n,i)$ is a superposition of diffuse and non-diffuse sound. Thus the sound which is applied to point-like virtual source will also have some diffuse sound within, which is not desired.

This defect can be partly avoided with a technique where the sound for point-like virtual source is not $w(n,i)$ but a cardioid signal directed towards the analyzed direction. This will reduce the amount of diffuse sound energy in average by 4.8 dB [20]. In practice, this can be implemented by computing the sound for each loudspeaker with equation

$$y(n,i,k) = \frac{1}{2}g(n,i,k)\{w(n,i) + x(n,i)\cos\theta_k\cos\phi_k + \\ + y(n,i)\sin\theta_k\cos\phi_k + z(n,i)\sin\phi_k\}, \quad (7)$$

where θ_k is the azimuth and ϕ_k is the elevation of loudspeaker k. This can be interpreted as a situation where a virtual cardioid microphone is directed to each loudspeaker direction, and each channel is multiplied by the corresponding gain factor. This is quite close to Ambisonics decoding with cardioid characteristics, although

in this case the crosstalk between loudspeakers is minimized adaptively, which reduces the coloring and blurring artifacts.

4.3.2. Diffuse synthesis

In DirAC, two alternative strategies have been used to produce diffuse sound [7]. One is to simply decorrelate the $w(n,i)$ signal by convolving it with exponentially decaying white noise with a time constant of 20 ms. Decorrelation is performed for each loudspeaker with a different noise sample. This is called single-channel diffusion.

This method might produce artifacts, since the $w(n,i)$ signal may be colored due to e.g. comb-filter effects. Also, in some cases, the diffuse sound may have a directional distribution, which should be reproduced. An alternative method to this is to transmit all channels of the B-format signal and to obtain virtual cardioid microphones pointing towards each loudspeaker direction. To further increase the diffuseness of to so-obtained signals convolution diffusion is applied. This method is called as cardioid diffusion. The advantage of this approach is that the colorations present in $w(n,i)$ signal may be less present in virtual cardioid microphones, and also that the directional distribution of diffuse sound is preserved somehow. This method is used in the present work.

Note that the first order Ambisonics uses similar processing as the described technique. The difference of applying this in DirAC is, that the loudspeaker signals are diffused, since this Ambisonics-type of decoding is only used for generating the diffuse sound. The diffusion also avoids some coloration problems in Ambisonics, since after it the signals are not anymore in the same phase in loudspeakers.

4.4. Discussion: Differences between Ambisonics and DirAC

In first-order Ambisonics, the major problem is the high coherence between loudspeaker signals. The DirAC implementation presented in this paper can be interpreted to be adaptive 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. However, in DirAC, the loudspeaker signals are then adaptively divided into diffuse and non-diffuse parts. The diffuse part of sound is further diffused to avoid too high coherence, and the non-diffuse part of sound is applied to a small subset of loudspeakers, which also decreases the coherence. The major problem of Ambisonics is thus eliminated in both branches of processing.

5. REPRODUCTION OF B-FORMAT SOUND

The methods described here were discussed using different listening methods. The first, and most critical one is listening in anechoic conditions to 16-channel 3-D reproduction of virtual reality generated with 16 loudspeakers. The system was also tested in an ITU recommendation BS.1116-1 listening room with a 5.0 system. The results of the informal listening are described in the following.

In the first tests, a virtual acoustical environment generated with 16 loudspeakers in an anechoic chamber was reproduced with different versions of DirAC. The environments were simulated with the image source method [21] with three different geometries, a large room, a medium room, and a small room. The details of reference generation can be found in [11].

A sound sample, either snare drum shots or male speech, was convolved with the virtual acoustical environment impulse responses, resulting in natural-sounding samples. These samples were considered as references. The B-format recording was simulated and the samples were reproduced using the current implementation of DirAC.

A general impression is, that DirAC reproduces faithfully the spatial properties of sound. The sound sources are perceived in the same directions with the original ones. Also, the distinct reflections in the reference material were reproduced, however, the timbre of the reflections did change a bit. There is also a small overall timbral change between the reference and reproduction. In the best case with a speech signal, it was barely audible, however, with drum signal it was more prominent.

As a second test, different B-format recordings were reproduced with DirAC in anechoic chamber with 16-channel 3-D system, and in a listening room with 5.0 system. The diffuse and non-diffuse parts were listened also separately. It was found, that the diffuse reproduction lacked transients and the directions of sound sources could be perceived only faintly, as supposed. The effect of the room was also present prominently there. Although different values for diffuseness averaging window length were tried, no artifacts were audible.

The non-diffuse sound was a bit more problematic in terms of finding proper window lengths for averaging both direction and diffuseness. If the constants were

set wrongly, there could be some artifacts, such as bubbling or perception of movement. However, in all cases, such values were found that the artifacts were not audible in the signal where both diffuse and non-diffuse sound were summed. The DirAC presentations were also compared to hypercardioid Ambisonics presentations of the same B-format sound files in 16-channel and 5.0 systems. The largest difference was that the directions of sound sources were perceived more clearly, and that the directions stayed consistently in the same loudspeakers in a large listening area. Also, a difference was perceived in reverberation, the rooms sounded a bit wider with DirAC than with Ambisonics. Demonstrations with four different source material are available for 5.0 listening in [22].

6. USING DIRAC IN UPMIXING

In upmixing, a two-channel stereophonic audio file is taken as input file, and it is processed to yield loudspeaker signals for desired multichannel loudspeaker setup. There exists a variety of techniques for this, some of which have been discussed in [23]. Avendano and Jot proposed a method [23] where the stereophonic sound is analyzed and synthesized with similar principles with DirAC. Their method computes coherence between left and right channels, which is then used to divide sound into diffuse and non-diffuse sound. The direction of non-diffuse sound is computed using a similarity index.

DirAC method can also be used to upmix stereo files to multichannel setups. Anechoic B-format recording of stereophonic representation of two-channel audio file can be simulated easily. Left and right channels are applied to simulated loudspeakers in directions $\pm 30^\circ$, and a B-format microphone is simulated in the optimal far-field listening position. The DirAC method can then be used to produce sound for arbitrary reproduction methods.

The DirAC method has to be modified a bit, though. In current implementation, the diffuse sound and non-diffuse sound are composed of virtual cardioid microphones pointing to loudspeaker directions. This is not feasible in upmixing, since there are only two point-like sound sources in the recorded scenario. In the current implementation, the sound for point-like virtual sources is a sum of left and right channels. The sound for diffuse reproduction is left channel for loudspeakers in left hemisphere, and right channel for loudspeakers in right hemisphere.

The estimated direction behaves also a bit differently than in case of B-format recordings of real spaces, also depending on the material which is upmixed. In testing with different audio files it was found that the analyzed direction tended to be biased towards the center. In the processing $w(n)$ and $x(n)$ are both a direct sum of left and right stereophonic channels with equal weights, thus $w(n)$ is always in the same phase with $x(n)$, which implies that direction vector $\mathbf{D}$ points always to frontal hemisphere. In contrast, $w(n)$ and $y(n)$ may be in different phase, thus directions between $\pm 90^\circ$ of azimuth are possible.

Since the analyzed direction is distributed in the front, and since there is some averaging in directional analysis and gain factor computation, the analyzed directions are biased towards the center. This can be avoided by multiplying the analyzed azimuth values by a constant factor.

In testing of the system it was found, that the concept of using DirAC in upmixing is valid, and usable. In DirAC processing the material which included incoherent reverberant sound, produced sound to surround channels which gave a perception that the room surrounded the listener better than in stereo listening. If the material included sound with phase differences between loudspeakers, e.g., if time panning or phase reversal was applied, the sound was applied to directions exceeding the stereophonic setup.

7. CONCLUSIONS

The Directional audio coding (DirAC) technique has been presented, which is a technique to reproduce spatial sound over arbitrary reproduction techniques. It can also be used to encode spatial aspects of sound as side information which is transmitted/stored along with a single or more audio channels. The technique is based on analyzing the sound direction and diffuseness depending on time at narrow frequency bands, and further decoding these parameters with appropriate techniques. In this study, the methods for analysis and synthesis were derived with B-format microphones. The implementation was tuned for high fidelity reproduction of spatial sound recordings, and for upmixing of stereophonic material to 5.0 setups. The system was informally tested with comparisons to virtual auditory environments and available B-format recordings, and with upmixing stereo material.

8. ACKNOWLEDGMENTS

Ville Pulkki has received funding from the Academy of

Finland (project 105780) and from Emil Aaltonen foundation.

9. REFERENCES

[1] S. P. Lipshitz, "Stereo microphone techniques... Are the purists wrong?," *J. Audio Eng. Soc.*, vol. 34, no. 9, pp. 716–744, 1986.

[2] M. A. Gerzon, "Periphony: Width-Height Sound Reproduction," *J. Aud. Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.

[3] J. Daniel, R. Nicol, and S. Moreau, "Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging," in *AES 114th Convention*, 2003, Paper # 5788.

[4] R. Bruno, A. Laborie, and S. Montoya, "Designing high spatial resolution microphones," in *AES 117th Convention*, 2004, Paper # 6231.

[5] J. Merimaa and V. Pulkki, "Spatial Impulse Response Rendering I: Analysis and synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127, 2005.

[6] V. Pulkki and J. Merimaa, "Spatial impulse response rendering: Listening tests and applications to continuous sound," in *AES 118th Convention*, Barcelona, Spain, 2005, Preprint 6371.

[7] V. Pulkki and C. Faller, ," in *AES 120th Convention*, Paris, France, 2006, Preprint 6658.

[8] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, The MIT Press, Cambridge, Massachusetts, USA, revised edition, 1997.

[9] H. Nélisse and J. Nicolas, "Characterization of a diffuse field in a reverberant room," *J. Acoust. Soc. Am.*, vol. 101, no. 6, pp. 3517–3524, 1997.

[10] L. L. Beranek, *Concert and Opera Halls — How They Sound*, Acoustical Society of America, Woodbury, NY, USA, 1996.

[11] V. Pulkki and J. Merimaa, "Spatial Impulse Response Rendering II: Parameters and listening tests," *J. Audio Eng. Soc.*, vol. 54, no. 1, 2006.

[12] B.S. Atal and M.R. Schroeder, "Perception of coloration in filtered gaussian noise - short-time spcetral analysis by the ear," in *Fourth Int. Congr. Acoust.*, Copenhagen, Sweden, August 1962.

[13] F. A. Bilsen, "Pitch of noise signals: Evidence for a "central spectrum"," *J. Acoust. Soc. Am.*, vol. 61, no. 1, pp. 150–161, 1977.

[14] J. F. Culling, A. Q. Summerfield, and D. H. Marshall, "Dichotic pitches as illusions of binaural unmasking. I. Huggins' pitch and the binaural edge pitch," *J. Acoust. Soc. Am.*, vol. 103, no. 6, pp. 3509–3526, Jun. 1998.

[15] A.W. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 83, pp. 1508–1516, 1988.

[16] S. Bech, "Timbral aspects of reproduced sound in small rooms. I," *J. Acoust. Soc. Am.*, vol. 97, no. 3, pp. 1717–1726, March 1995.

[17] M. Brüggen, "Coloration and binaural decoloration in natural environments," *acta acustica - ACUSTICA*, vol. 87, pp. 400–406, 2000.

[18] 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, no. 5, pp. 3075–3089, Nov. 2004.

[19] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.

[20] R. B. Schulein, "Microphone considerations in feedback-prone environments," *J. Audio Eng. Soc.*, vol. 24, no. 6, pp. 434–445, July/August 1976.

[21] J. B. Allen and D. A. Berkeley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.

[22] V. Pulkki, "Directional audio coding demonstrations for 5.0," http://www.acoustics.hut.fi/demos/DirAC, April 2006.

[23] C. Avendano and J-M. Jot, "A frequency-domain approach to multichannel upmix," *J. Audio Eng. Soc.*, vol. 52, no. 7/8, pp. 740–749, 2004.