International Workshop
on the Principles and Applications of Spatial Hearing
November 11-13, 2009, Zao, Miyagi, Japan

Directional audio coding - perception-based reproduction of spatial sound

V. Pulkki[1], M-V Laitinen[1], J Vilkamo[1], J Ahonen[1], T Lokki[2] and T Pihlajamäki[1] *

[1] *Dept Signal Processing and Acoustics,* [2] *Dept of Media Technology*
Helsinki University of Technology, TKK, Finland

Abstract

Directional Audio Coding (DirAC) is a perceptually motivated technique for spatial audio processing. DirAC analyzes in short time windows the sound spectrum together with direction and diffuseness in frequency bands of human hearing, and uses this information in synthesis. It has applications in capturing, coding and resynthesis of spatial sound, in teleconferencing, in directional filtering, and in virtual auditory environments.

1 Introduction

The spatial properties of sound perceivable by humans are the directions and distances of sound sources in three dimensions, and the effect of the room to sound. In addition, the spatial arrangement of sound sources affects the timbre. The directional resolution of spatial hearing is limited within auditory frequency band [1]. In principle, all sound within one critical band can be only perceived as a single source with broader or narrower extent. In some special cases binaural narrow-band sound stimulus can be perceived as two distinct auditory objects, but the perception of three or more sources is generally not possible simultaneously. This is different from visual perception, where already one eye can detect the spatial locations of a large number of visual objects sharing the same color. The limitations of spatial auditory perception imply that such spatial realism needed in visual reproduction is not needed in audio. In other words, the spatial accuracy in reproduction of acoustical wave field can be compromized without decrease in perceptual quality. A recent technology for spatial audio, Directional Audio Coding (DirAC) [2], explores the possibilities to exploit the frequency-band resolution of human sound perception in audio. In this paper, the basic DirAC processing principles are overviewed, and different applications are presented.

2 Directional Audio Coding

In DirAC, it is assumed that at one time instant and at one critical band the spatial resolution of auditory system is limited to decoding one cue for direction and another for inter-aural coherence. It is further assumed, that if the direction and diffuseness of sound field is measured and reproduced correctly, a human listener will perceive the directional and coherence cues correctly. In practise, the DirAC processing is performed in two phases: the analysis of directional metadata and the synthesis of sound, where the directional metadata is used actively in reproduction. The processing is performed separately for each frequency band. The analysis phase is shown in Fig.1 a, and synthesis in Fig.1 b.

2.1 Division into frequency bands

In DirAC, both analysis and synthesis are performed in the frequency domain. There are several methods for dividing the sound into frequency bands, with distinct properties each. The most commonly used frequency transforms include short time Fourier transform (STFT), and quadrature mirror filterbank (QMF). In addition to these, there is a full liberty to design a filterbank with arbitrary filters that are optimized to any specific purposes. Regardless of the selected time-frequency transform, the design goal is to mimic the resolution of the human spatial hearing.

In the first implementations of DirAC, a filterbank with arbitrary subband filters alternatively with STFT with 20 ms time windows was used [2]. The even time resolution at all frequencies is a drawback for STFT implementation, which

*E-mail address: Ville.Pulkki@tkk.fi

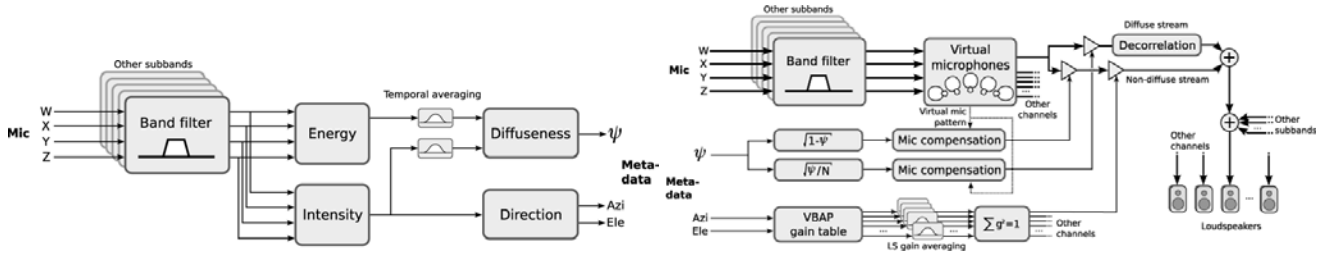

Figure 1. a) DirAC analysis b) DirAC synthesis

may produce some artifacts at high frequencies with some critical signals due to too long temporal windows. The filterbank implementation solves this problem, however, it may have constraints in computational complexity. In [3] a linear-phase filterbank was used, and in [4] a multi-resolution version of STFT was utilized, where the input sound was divided into few frequency channels and processed with different STFTs having window lengths suited to each frequency band. It is quite clear that the choise of time-frequency transfer does not have a major impact in audio quality in DirAC reproduction. The differences are typically audible only with input material having about 100-1000 Hz modulations in signal envelopes at high frequencies, an example of such sound is the snare drum sound.

2.2 Directional analysis

The target of directional analysis is to estimate at each frequency band the direction of arrival of sound, together with an estimate if the sound is arriving from one or multiple directions at the same time. In principle this can be performed with a number of techniques, however, the energetic analysis of sound field has been found to be suitable, which is shown in Fig.1 a. The energetic analysis can be performed, when the pressure signal and velocity signals in 1-3 dimensions are captured from a single position. In first-order B-format signals, the omnidirectional signal is called W-signal, which has been scaled down by $\sqrt{2}$. The sound pressure can be estimated as $P = \sqrt{2}W$, expressed in SFTF domain. The X-, Y- and Z-channels have the directional pattern of a dipole directed along the Cartesian axis, which form together a vector $\mathbf{U} = [X, Y, Z]$. The vector estimates the sound field velocity vector, and it is also expressed in STFT domain. The energy E of sound field can be computed as $E = (\rho_0/4)||\mathbf{U}||^2 + \left(1/(4\rho_0 c^2)\right)|P|^2$, where ρ_0 is the mean density of air, and c is the speed of sound. The capturing of B-format signals can be obtained with either coincident positioning of directional microphones, or with closely-spaced set of omnidirectional microphones. In some applications, the microphone signals may be formed in computational domain, i.e., simulated. The analysis is repeated as frequently as is needed for the application, typically with the update frequency of value 100Hz-1000Hz.

The direction of sound is defined to be the opposite direction of intensity vector $\mathbf{I} = \overline{P}\mathbf{U}$, where $\overline{(\cdot)}$ denotes complex conjugation. The direction is denoted as corresponding angular azimuth and elevation values in the transmitted metadata. The diffuseness of sound field is computed as

$$\psi = 1 - \frac{\|\mathrm{E}\{\mathbf{I}\}\|}{c\mathrm{E}\{E\}}, \tag{1}$$

where E is the expectation operator. The outcome of this equation is a real-valued number between zero and one, characterizing if the sound energy is arriving from a single direction, or from all directions. This equation is appropriate in the case when the full 3D velocity information is available. If the microphone setup delivers velocity only in 1D or 2D, the equation

$$\psi_{cv} = \sqrt{1 - \frac{||\mathrm{E}\{\mathbf{I}\}||}{\mathrm{E}\{||\mathbf{I}||\}}} \tag{2}$$

yields estimates that are closer to the actual diffuseness of sound field than Eq.(1) does [5].

2.3 DirAC transmission

In many applications, spatial sound needs to be transmitted from a location to another. In DirAC, this can be performed with a few different approaches. A straightforward technique is to transmit all the signals of B-format. In such case no metadata is needed, and analysis can be performed in receiving end. However, in low-bit-rate version only one channel of audio is transmitted, which provides a large reduction in data rate, and the drawback is a slight decrease in timbral quality of reverberant sound, and a decrease in directional accuracy in multi-source scenarios.

In some cases it is beneficial to merge multiple mono or stereo DirAC streams together. This is not a trivial task, as there is no simple way to merge directional metadata. However, two methods have been proposed, which provide artifact-free and efficient merging [6].

2.4 DirAC synthesis with loudspeakers

The high-quality version of DirAC synthesis, which is shown in Fig. 1 b, receives all B-format signals, from which a virtual microphone signal is computed for each loudspeaker direction. The utilized directional pattern is typically a dipole. The virtual microphone signals are then modified in non-linear fashion, depending on the metadata. The low-bit-rate version of DirAC is not shown in the figure, however, in it only one channel of audio is transmitted. The difference in processing is that all virtual microphone signals would be replaced by the single channel of audio received. The virtual microphone signals are divided into two streams: the diffuse and the non-diffuse streams, which are processed separately.

The non-diffuse sound is reproduced as point sources by using vector base amplitude panning (VBAP) [7]. In panning, a monophonic sound signal is applied to a subset of loudspeakers after multiplication with loudspeaker-specific gain factors. The gain factors are computed using the information of loudspeaker setup, and specified panning direction. In the low-bit-rate version, the input signal is simply panned to the directions implied by the metadata. In the high-quality version, each virtual microphone signal is multiplied with the corresponding gain factor, which produces the same effect with panning, however it is less prone to any nonlinear artifacts.

In many cases the direction in metadata is subject to abrupt temporal changes. To avoid artifacts, the gain factors for loudspeakers computed with VBAP are smoothed by temporal integration with frequency-dependent time constant equaling

to about 50 cycle periods at each band. This removes effectively the artifacts, however the changes in direction are not perceived to be slower than without averaging in most of the cases.

The aim of the synthesis of **the diffuse sound** is to create perception of sound that surrounds the listener. In the low-bit-rate version, the diffuse stream is reproduced by decorrelating the input signal and reproducing it from every loudspeaker. In the high-quality version, the virtual microphone signals of diffuse stream are already incoherent in some degree, and they need to be decorrelated only mildly. This approach provides better spatial quality for surrounding reverberation and ambient sound than the low-bit-rate version.

2.5 DirAC synthesis with headphones

In [8] different approaches to reproduce spatial audio over headphones with and without head tracking were researched in the context of DirAC. After testing different versions, it was found, that the best quality is obtained when the DirAC is formulated with about 40 virtual loudspeakers around the listener for non-diffuse stream, and 16 loudspeakers for diffuse stream. The virtual loudspeakers are implemented as convolution of input signal with measured HRTFs.

A common problem in headphone reproduction is that the reproduced auditory space moves with the head of the listener, which causes internalized sound events. To prevent this, a method to utilize head tracking information in DirAC was also developed [8]. A simple and effective method was found to be to update the metadata, and to transform the velocity signals, according to the head tracking information [8]. No specific treatment was needed to avoid temporal artifacts.

2.6 Similarity to channel-based coding

DirAC shares many processing principles and challenges with existing spatial audio technologies in coding of multi-channel audio [9, 10]. DirAC can be used similarly in processing of multi-channel audio files. A difference is, that DirAC is also applicable for recording real spatial sound environments.

3 DirAC Applications

Convolving reverberators. The first implementation of DirAC was reproduction of measured B-format impulse responses over arbitrary loudspeaker setups, Spatial impulse response rendering (SIRR) [11]. The application was in convolving reverberators. An acoustically dry monophonic recording can be processed for multichannel listening to sound like performed in the hall where the B-format impulse response was measured.

Teleconferencing. In basic telecommunication application of DirAC, two groups of people want to have a meeting with each other. Both of the groups gather in the vicinity of typically 1D or 2D B-format microphone. The directional metadata is encoded from the microphone signals, and the transmitted signal is mono DirAC stream. In the receiving end the mono DirAC stream can be rendered to whatever loudspeaker layout, which effectively spatializes the talkers in reproduction. Different microphone setups have been tested, and it has been found that the metadata can be encoded already from closely spaced pair of low-end omnidirectional or directional microphones [12]. It was also shown, that a basic quality can be obtained from a stereo microphone, although the directional patterns are not known in the processing.

High-quality reproduction. As already mentioned, the initial target for DirAC development was to reproduce a recorded sound scenario as realistically as possible. This is interesting at least in academic level, and in some cases also in audio industry.

The DirAC metadata seems also a good starting point for a **generic audio format**, which would tolerate different loudspeaker setups, and headphone listening. The methods to transform existing audio content into DirAC formats are under construction, and the audio synthesis in DirAC has to be tested in various conditions.

Spatial filtering. In spatial filtering, a signal is formed which is sensitive only to a direction defined by the user. The DirAC method can be used also for this, in simplest form by listening only one loudspeaker of a multi-loudspeaker setup. This already emphasizes greatly the sound arriving from the direction of the loudspeaker. A fine-tuned method to use directional analysis parameters to suppress diffuse sound and sound originating from arbitrary directions has already been suggested [13], where it was found to outperform some traditional beam forming methods in some aspects.

Source localization. Another application of directional analysis data is in localization of sound sources, where the usage of diffuseness and direction parameters provides an efficient method. For example, in teleconferencing scenario, where the energetic analysis is already performed from the microphone inputs, they can also be used to steer camera towards active talkers. In [14] it is shown that the localization method based on directional analysis data provides reliable estimates even in reverberant environments and with multiple concurrent talkers. The approach also allows for trade off between localization accuracy and tracking performance of moving sound sources.

Applications in mixing, game sound, and in virtual realities. The previous applications concentrated on the cases where the directional properties of sound are captured with real microphones from a live situation. It is also possible to use DirAC in virtual realities, without connection to any real acoustical situation [15]. In these applications, the directional metadata connected to a DirAC stream is defined by the user. For example, a single channel of audio is spatialized as a point-like virtual source with DirAC, when the same direction for all frequency channels is added as metadata to the signal. In some cases it would be beneficial to control the perceived extent or width of the sound source. A simple and effective method for this is to use a different direction value for each frequency band, where the values are distributed inside the desired directional extent of the virtual source.

A common task in virtual worlds is the generation of room effect or reverberation. In [15] it is suggested, that the number of single-output reverberators can be only two for any horizontal loudspeaker setup, and three for any three-dimensional

loudspeaker setup. According to informal testing, this generates diffuse reverberation with good quality. If distinct localizable reflections are needed, feasible results are obtained, if the reflections are reproduced as individual virtual sources with mono DirAC streams.

4 Subjective evaluation

The loudspeaker versions of DirAC and SIRR in reproduction of real sound scenes were researched by comparing reference scenarios to their reproductions. The reference scenarios were produced with more than 20 loudspeakers, which generate the direct sounds from sources, and also reflections and reverberation [3, 16]. An ideal or real B-format microphone was used to capture the signal in the center of the setup, and the recorded signal was then reproduced using a high-quality implementation of DirAC or SIRR. The perceptual quality of reproduction of room impulse responses with SIRR was tested in anechoic listening. The result was that SIRR reproduction was very close to the original, rated "perceptible, difference not annoying" at the worst case [16]. To measure the perceptual audio quality of DirAC, two distinct listening tests were conducted, one with 3D reproduction in anechoic environment, and one with horizontal-only reproduction in a multi-channel listening room. Different loudspeaker layouts with 5 to 16 loudspeakers were utilized in both conditions for DirAC reproduction. It was proven, that DirAC produces better perceptual quality in loudspeaker listening than other available techniques using the same microphone input. In such direct comparison to reference, DirAC reproduction was rated in almost all conditions either "excellent" or "good" in ITU scale. Only in off-sweet-spot 5.1 listening the quality was scored "fair".

The binaural reproduction of DirAC was investigated in [8] with and without head-tracking in a test which did not utilize reference scenarios for practical reasons. The results with binaural head-tracked reproduction were very good. In ITU scale the overall quality was rated "excellent", and the spatial impression was rated "truly believable". For the telecommunication application, the speech reception level with two concurrent talkers has been measured with DirAC reproduction with 1-D and 2-D microphone setups [17]. It was found, that the reception of speech was almost as good with DirAC with mono DirAC stream, than with two-channel transmission of dipole signals, which served as a reference. In [18] the data rate of directional metadata was studied in the context of teleconferencing application. It was found, that the rate can be as low as 2-3 kBit/s, and the directions of talkers are perceived still correctly.

5 Summary

DirAC is a perceptually motivated signal-dependent spatial sound reproduction method based on frequency-band processing and directional sound field analysis. The time-dependent data originating from the analysis is used in the synthesis of sound. The method uses existing first-order microphones for recording and any surround loudspeaker setup for reproducing the spatial sound. In this paper, a review to different variants of the method was made, and the applications of the method were discussed. The results from subjective tests conducted so far were also summarized.

6 Acknowledgements

The Academy of Finland (Projects #105780 and #119092) and Fraunhofer Gesellshaft IIS have supported this work. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreements n^o [240453] and n^o [203636].

References

[1] J. Blauert, *Spatial Hearing*. The MIT Press, 1983.
[2] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, June 2007.
[3] J. Vilkamo, T. Lokki, and V.Pulkki, "Directional audio coding: Virtual microphone based synthesis and subjective evaluation," *J. Audio Eng. Soc.*, vol. 57, no. 9, 2009.
[4] T. Pihlajamäki, "Multi-resolution short-time fourier transform implementation of directional audio coding," Master's thesis, Helsinki Univ. Tech., 2009.
[5] J. Ahonen and V. Pulkki, "Diffuseness estimation using temporal variation of intensity vectors," in *Workshop on Applications of Signal Processing to Audio and Acoustics WASPAA*, Mohonk Mountain House, New Paltz, 2009.
[6] G. D. Galdo, V. Pulkki, F. Kuech, M.-V. Laitinen, R. Schultz-Amling, and M. Kallinger, "Efficient methods for high quality merging of spatial audio streams in directional audio coding," in *AES 126th Convention*, Munich, Germany, 2009, paper 7733.
[7] V. Pulkki, "Virtual source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, June 1997.
[8] M.-V. Laitinen and V. Pulkki, "Binaural reproduction for directional audio coding," in *Workshop on Applications of Signal Processing to Audio and Acoustics WASPAA*. New Paltz, NY: IEEE, 2009.
[9] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, and W. Oomen, "MPEG surround-the ISO/MPEG standard for efficient and compatible multichannel audio coding," *J. Audio Eng. Soc.*, vol. 56, pp. 932–955, 2008.
[10] M. M. Goodwin and J.-M. Jot, "A frequency-domain framework for spatial audio coding based on universal spatial cues," in *120th AES Convention*, Paris, May 2006, paper # 6751.
[11] J. Merimaa and V. Pulkki, "Spatial impulse response rendering 1: Analysis and synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127, December 2005.
[12] J. Ahonen, V. Pulkki, F. Kuech, G. D. Galdo, M. Kallinger, and R. Schultz-Amling, "Directional audio coding with stereo microphone input," in *AES 126th Convention*, Munich, Germany, 2009, paper 7708.
[13] M. Kallinger, G. D. Galdo, F. Kuech, D. Mahne, and R. Schultz-Amling, "Spatial filtering using directional audio coding parameters," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE Computer Society, 2009, pp. 217–220.
[14] O. Thiergart, R. Schultz-Amling, G. D. Galdo, D. Mahne, and F. Kuech, "Localization of sound sources in reverberant environments based on directional audio coding parameters," in *127th AES Convention*, New York, Oct 2009, paper # 7853.
[15] V. Pulkki, M.-V. Laitinen, and C. Erkut, "Efficient spatial sound synthesis for virtual worlds," in *AES 35th Conf. Audio for Games*, London, UK, 2009.
[16] V. Pulkki and J. Merimaa, "Spatial impulse response rendering 2: Reproduction of diffuse sound and listening tests," *J. Audio Eng. Soc.*, vol. 54, no. 1/2, pp. 3–20, January/February 2006.
[17] J. Ahonen, V. Pulkki, F. Kuech, M. Kallinger, and R. Schultz-Amling, "Directional analysis of sound field with linear microphone array and applications in sound reproduction," in *the 124th AES Convention*, Amsterdam, Netherlands, May 17-20 2008, paper 7329.
[18] T. Hirvonen, J. Ahonen, and V. Pulkki, "Perceptual compression methods for metadata in directional audio coding applied to audiovisual teleconference," in *the 126th AES Convention*, Munich, Germany, May 7-10 2009, paper 7706.