BINAURAL REPRODUCTION FOR DIRECTIONAL AUDIO CODING

Mikko-Ville Laitinen and Ville Pulkki

Helsinki University of Technology, Dept Signal Processing and Acoustics
Otakaari 5, 02150, Espoo, Finland.
mikko-ville.laitinen@tkk.fi, ville.pulkki@tkk.fi

ABSTRACT

Directional audio coding (DirAC) is a recently proposed method for spatial sound reproduction. So far it has been used only in loudspeaker reproduction, and a method for headphone reproduction is presented in this article. In principle, the method uses virtual loudspeakers simulated with head related transfer functions (HRTFs), and head tracking. The method was evaluated subjectively, and the results are presented. The results show that a plausible spatial impression can be reproduced using the binaural realization of DirAC.

Index Terms— Spatial audio, binaural reproduction, HRTF

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 also the timbre. A task in spatial reproduction of sound is, that these properties should be relayed from the original space to the reproduction phase.

A recently proposed method for spatial sound reproduction is Directional Audio Coding (DirAC) [1]. In DirAC spatial sound is recorded typically using a B-format microphone, and reproduced using an arbitrary number of loudspeakers. The direction and diffuseness is analyzed in frequency bands depending on time from recorded B-format signals, and this information is used actively in reproduction. Previously only the loudspeaker reproduction has been presented. A method for binaural reproduction is developed in this article.

DirAC shares many processing principles and challenges with existing spatial audio technologies in coding of multi-channel audio [2, 3]. 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.

2. DIRECTIONAL AUDIO CODING

The general idea of DirAC is that there is no need to reproduce sound pressure field physically perfectly. 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. Based on these assumptions, the principle of DirAC is that the sound in one frequency band is simply presented with two cross-fading streams: a non-directional diffuse stream, and a directional non-diffuse stream.

The DirAC processing is performed in two phases: the analysis and the synthesis. The processing is performed separately for each frequency band. The signals are divided in time and frequency, in this implementation using the short-time Fourier transform (STFT). In following equations, the dependency on frequency and time is dropped for notational simplicity.

2.1. Energetic analysis in DirAC

The input signals for DirAC analysis are B-format signals in STFT domain, which consists of four channels: W, X, Y and Z. The W-channel has been measured with omnidirectional characteristics and scaled down by $\sqrt{2}$. The X-, Y- and Z-channels have the directional pattern of a dipole directed along the Cartesian axis, which form together a vector $\mathbf{V} = [X, Y, Z]$ relative to sound field velocity vector.

The direction of sound is defined to be the opposite direction of intensity vector $\mathbf{I} = (1/\sqrt{2})\mathrm{Re}(W)^*\mathbf{V}'$, and is denoted as corresponding angular azimuth and elevation values in the transmitted metadata. The diffuseness is computed as

$$\psi = 1 - \frac{\sqrt{2}||\mathrm{Re}\{W^*\mathbf{V}'\}||}{|W|^2 + |\mathbf{V}|^2/2} \quad (1)$$

which 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 is repeated for each 20 ms STFT frame, which yields the metadata to be transmitted with audio signals.

2.2. Spatial sound synthesis in DirAC

In the low-bit-rate version of DirAC only one channel of audio is transmitted, which is used as the signal applied to all loudspeakers, after non-linear manipulations according to the metadata, as shown in Fig. 1 a. The high-quality version, which is shown in Fig. 1 b, transmits B-format signals, from which a virtual microphone signal is computed as a weighted sum of B-format signals for each loudpeaker direction. The virtual microphone signals are then manipulated similarly as in the low-bit-rate version.

The signal in the low-bit-rate version, and the virtual microphone signals in the high-quality version are divided into two streams: the diffuse and the non-diffuse stream. The non-diffuse stream includes mostly the part of sound that has a certain direction. The diffuse stream includes mostly the reverberant and ambient parts, and is reproduced with a technique producing surrounding perception of sound.

The non-diffuse sound is reproduced as point sources by using vector base amplitude panning (VBAP) [4]. 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,

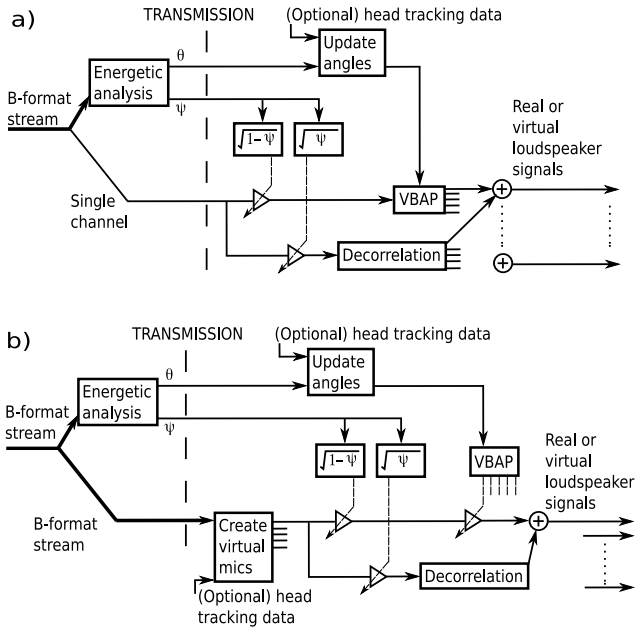

Figure 1: a) The low-bit-rate version of DirAC, where single channel of audio is transmitted with metadata. b) The high-quality version of DirAC, where up to four channels of audio are transmitted with metadata. The metadata consists of direction (θ) and diffuseness (ψ) of sound in each frequency band depending on time.

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. The dipole was used as the directional pattern for the virtual microphones.

3. BINAURAL REPRODUCTION FOR DIRAC

Binaural reproduction techniques are often based on head-related trasfer functions (HRTFs). It is a function that, for a certain angle of incidence, describes the sound transmission from a free field to a point in the ear canal of a human subject [5]. In practise, HRTFs

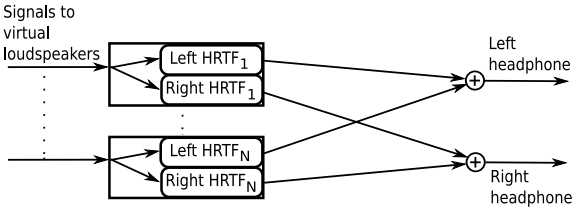

Figure 2: Virtual loudspeakers are created by convolving input signal with HRTFs measured from the corresponding direction.

are measured by placing a microphone to the entrance of the ear canal of the subject and measuring impulse responses from desired directions.

In headphone reproduction, a convolution is computed between the input signal and the HRTF for each ear. The auditory object is perceived to be positioned to the direction from which the HRTFs have been measured. A problem with HRTFs is that especially frontal auditory objects are often perceived to be inside the head, or behind the listener. However, this problem is more severe when HRTF processing is not used at all.

3.1. Synthesis of Non-Diffuse Sound

Non-diffuse stream includes mostly the part of signal that originated from a certain direction, which direction is transmitted in the metadata. A straightforward way to synthesize this stream would be to filter the signal with HRTFs corresponding to the directions in metadata. Unfortunately, this would lead to complicated solutions, as either a very large catalogue of HRTFs, or a HRTF interpolation method would be needed. Also, the direction of each frequency band can be different and it can change rapidly, which can cause artifacts. The averaging of the direction on the other hand can cause sluggishness in the perceived direction [1].

A straightforward solution was used in this work, where the loudspeaker realization of DirAC was utilized with virtual implementation of the loudspeakers using HRTFs. The signal entering to a virtual loudspeaker is convolved with the HRTF of the left and right ear of the corresponding direction. The left and right outputs for every virtual loudspeaker are summed to form the signals for the left and right headphone channels, as shown in Fig. 2. The smoothing of the gain factors of the virtual loudspeakers is used to avoid artifacts caused by abrupt temporal changes in the metadata.

3.2. Synthesis of Diffuse Sound

In principle, the generation of perception of perfectly surrounding sound field for a single audio signal for binaural presentation could be computed by applying the signals to both ears after sufficient decorrelation. Unfortunately, the application of such simple method was not found feasible, as the virtual loudspeaker processing in non-diffuse sound produces some spectral changes in sound which have to be imitated also in this part of processing.

Thus, the diffuse sound is reproduced analogously as in loudspeaker-based DirAC reproduction. The decorrelated loudspeaker signals are created, and applied to virtual loudspeakers. The directions of the virtual loudspeakers should be selected so that they evenly cover the whole sphere around the listener. In

this way the diffuse sound does not appear to originate from any distinct direction but rather from everywhere around the listener. According to informal listening, an adequate number of virtual loudspeakers used for diffuse sound was found to be about 12 - 20 in three dimensions, which agrees with a study on the spatial impression of diffuse sound field [6].

The decorrelation was implemented using frequency-dependent delays, which are static with time and different for different virtual loudspeakers. The result is a signal that has a random delay at each frequency band but the magnitude response has not been changed. If delays are within eligible boundaries they do not contribute to the perception of direction or spaciousness. If they are too long, sound is perceived to become more spacious or distinct echoes are heard. On the other hand, if delays are limited to be too short, the signal is not decorrelated and affects the perceived direction. At 100 Hz, the range of delays was 5 ms — 60 ms. The range changes linearly with frequency, and was 3 ms — 30 ms at frequencies above 10 kHz.

3.3. Head Tracking

In head tracking the spatial position of the head of the listener is monitored depending on time. When this information is available, it can be used in HRTF processing. If the HRTFs in the virtual loudspeaker are updated dynamically according to the head tracking data, an auditory object is perceived to the same position even though a listener moves his/her head. Head tracking also reduces the front-back confusions [7].

There are two possibilities to include head tracking information in DirAC reproduction. Either the directions of virtual loudspeakers are updated with the information, or the directional metadata and virtual microphone directions are transformed with the tracking information. The first alternative was soon rejected, as it would require updating filters with relatively long impulse responses, which is prone to artifacts. The latter alternative requires only updating of virtual microphone coefficients, and some algebraic manipulations of metadata, and was utilized in this work.

In practise, the direction information in metadata and the directions of virtual microphones are updated to match the orientation of the head. For example, if the listener rotates his/her head to -45 degrees in azimuth, the directional information is corrected by shifting all computed azimuth angles by $+45$ degrees. The updating is performed with 50 Hz rate. The STFT processing causes latency of about 10 ms so the total latency was about 30 ms, which is less than the detection threshold of system latency in head tracked binaural rendering, having value of about 75 ms [8].

If the listener rotates his head rapidly, the smoothing of the gain factors of the virtual loudspeakers performed in the synthesis of non-diffuse sound can cause sluggishness to the perceived direction, as there would be a discrepancy between the movement of the head and the perceived direction of sound. This is avoided by making the smoothing faster when the head is rotated rapidly. When the rotation is slow enough, the default smoothing is applied. This way the sound field is updated fast enough to correspond the movement of the head, but no artifacts are perceived due to faster smoothing.

4. LISTENING TESTS

The sound was reproduced using Sennheiser HD600 headphones. Head tracking system is installed in the listening room, so the subject was free to move his/her head and rotate the chair. However, the subject was advised to be seated all the time and not move the chair. A loudspeaker setup with 24 loudspeakers in 3D positioning was visible during the test.

Table 1: Reproduction methods to be evaluated.

DvmicT	DirAC: virtual microphones, head tracking, real HRTFs
DomniT	DirAC: omnidirectional mic only, head tracking, real HRTFs
DHATT	DirAC: virtual microphones, head tracking, head and torso model HRTFs
DvmicNT	DirAC: virtual microphones, no head tracking, real HRTFs
stT	Stereo: head tracking
stNT	Stereo: no head tracking
mono	Mono

Four sound samples recorded in real acoustical environments with real B-format microphones were reproduced in the test: 1) acoustical pop music played in a room, 2) folk music played with a guitar and a violin, which were at opposite sides of the listener, 3) outdoor ambient sounds including barking dogs, singing birds and a car passing by, 4) a symphony orchestra playing in a concert hall.

These samples were reproduced using different techniques. There were three basic techniques: *binaural DirAC*, *stereo* and *mono*. *Stereo* technique denotes a technique having two cardioids pointing to -60 and +60 degrees, one for each ear. These cardioids are created from the B-format signals. *Mono* is a technique where the omnidirectional microphone channel of the B-format microphone is applied unprocessed to both channels of the headphones. DirAC was reproduced using the *virtual microphone* and the *omnidirectional* versions. There were also two different HRTF catalogs used. One was non-individual HRTFs, which were measured from one of the authors of this paper. The other was a simple head-and-torso model created according to [9]. These techniques were used in a different "modes" so there were seven reproduction methods in total, see Table 1.

The test was conducted in three parts. The first part was a training session, where the subject was able to listen to different samples and methods and to get familiarized to the user interface. The actual test consisted of two parts. In the first part the subjects were given the following question: Grade the overall quality of reproduction. The subjects listened to one sample at a time and graded different methods in a multiple-stimulus test. The subjects were able to change between different techniques freely and also to seek wanted scenes in the samples. The order of the methods was randomized for every sample and the subjects did not know which method they were listening to. All four samples were graded twice. The order of the samples was also randomized. The score was given using an integer scale from 1 to 100. The scale was divided into five equal intervals with the following adjectives given: 100-80 Excellent, 80-60 Good, 60-40 Fair, 40-20 Poor and 20-0 Bad.

The second part of the test was identical to the first part, only the question was different: Grade your spatial impression. The adjectives describing the scale were: 100-80 Truly believable and engaging, 80-60 Like being reproduced in a small cinema, 60-40 Better than normal headphone reproduction, 40-20 As normal

headphone reproduction and 20-0 Worse than normal headphone reproduction.

The training session took about 15 minutes and both parts of the actual test took about 25 min each. The subjects had a short break between different parts of the test. There were 8 subjects in the test. The authors of this article did not participate.

The results of the listening tests can be seen in Fig. 3. The results were analyzed using one-way analysis of variance (ANOVA). The analysis showed that a statistically significant difference between means occurs, overall: $F(6, 441) = 53.75, p < .05$, spatial impression: $F(6, 441) = 145.78, p < .05$. To determine which pairs of means are significantly different, multiple comparisons of one-way ANOVA were performed using Tukey's least significant difference procedure, the confidence level was 0.05. The results can be seen in Fig. 3, where the means that have insignificant differences have been grouped using a horizontal line.

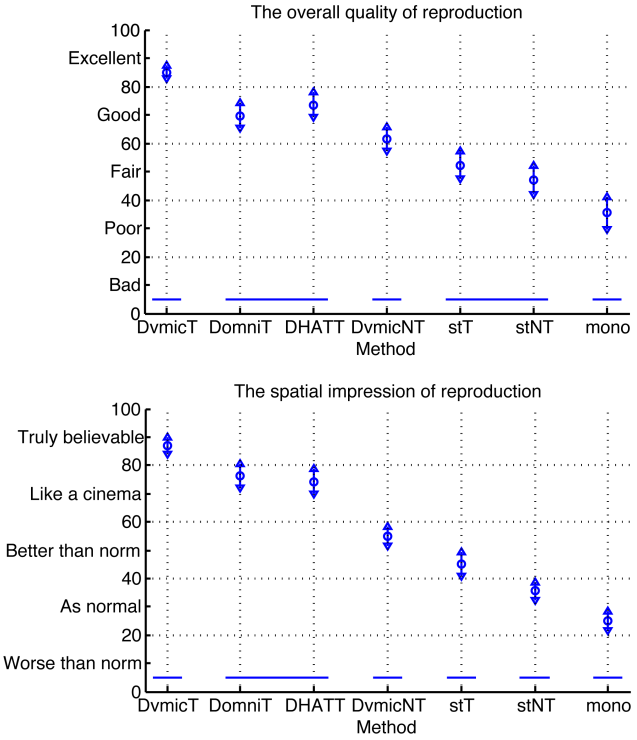

Figure 3: The upper panel shows the overall quality, and lower panel shows the spatial impression of different methods measured in a listening test. The means and the 95% confidence intervals are presented in the figure. Groups with insignificant differences in means have been marked with a horizontal line.

5. RESULTS AND DISCUSSION

It can be seen, that using binaural reproduction of DirAC a realistic reproduction of spatial sound can be achieved, as both overall quality and spatial impression were rated with highest scores in the case when head tracking was involved. An important result is also that the DirAC reproduction without head tracking was preferred to traditional stereophonic reproduction both with or without head tracking. At least according to this listening test, binaural DirAC could be something that people would actually like to use.

It has to be noted, that the room where the listening tests were carried out was quite large, and there were loudspeakers visible in the room. The loudspeakers might have helped in the externalization, because subjects were able to see something from where the sound could be coming from. Further tests are required to measure the perceptual quality in other listening spaces with or without accompanying video display.

Many listeners commented informally, that the externalization worked so well that they did not know whether the sound was coming from the loudspeakers or the headphones. Many of the listeners checked this by taking the headphones off momentarily. This can be seen as a promising result, as the HRTFs which were used for virtual loudspeakers were not individual, but measured from one of the authors.

6. ACKNOWLEDGEMENTS

The Academy of Finland (Projects #105780 and #119092) and Emil Aaltonen foundation have supported this work.

7. REFERENCES

[1] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, June 2007.

[2] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, *et al.*, "MPEG surround-the ISO/MPEG standard for efficient and compatible multichannel audio coding," *J. Audio Eng. Soc.*, vol. 56, pp. 932–955, 2008.

[3] 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.

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

[5] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *J. Audio Eng. Soc.*, vol. 43, no. 5, pp. 300–321, May 1995.

[6] K. Hiyama, S. Komiyama, and K. Hamasaki, "The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field," *AES 113th Convention*, Los Angeles, California, U.S.A., October 2002.

[7] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head-tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *AES 108th Convention*, Paris, France, 2000.

[8] S. Yairi, Y. Iwaya, and Y. Suzuki, "Influence of large system latency of virtual auditory display on behavior of head movement in sound localization task," *Acta Acustica United with Acustica*, vol. 94, pp. 1016–1023, 2008.

[9] V. R. Algazi, R. O. Duda, and D. M. Thomson, "The use of head-and-torso models for improved spatial sound synthesis," *AES 113th Convention*, Los Angeles, U.S.A., October 2002.