Directional Audio Coding -
A Perception-Based Method for
Spatial Sound Reproduction

Mikko-Ville Laitinen

Department of Signal Processing and
Acoustics, Aalto University, Espoo, Finland
Outline

- Introduction
- Background
- Directional Audio Coding (DirAC)
- Applications
- Evaluation
- Summary
Introduction
Background

Possible approaches

• Reproduce spatial sound in a way that
  – The physical sound field is accurate
  – The human perception is accurate
Background

Sound field reconstruction methods

• Methods
  – Wave field synthesis
  – (Higher-order) Ambisonics

• Reproduced sound field is compared to original sound field
  – Perfect reconstruction -> perfect perceptual quality!
  – What is the perceptual impact of possible deviations?
Background

Perception-based methods

• Methods
  – DirAC
  – MPEG Surround
  – Binaural cue coding (BCC)

• Original sound field is not aimed to be reconstructed

• Analyze features from the sound field that are important to spatial sound perception

• Reproduce these features accurately
  – The aim is obtain equal *perception* of the reproduced sound field as the original sound field
Directional audio coding (DirAC)
Assumptions about spatial hearing

• Human hearing analyzes sound in frequency bands
  – Equivalent rectangular bandwidth (ERB)

• At each frequency band can be perceived
  – Interaural time difference (ITD)
  – Interaural level difference (ILD)
  – Interaural coherence

• If two sinusoids are at the same frequency band, they cannot be localized individually
Directional audio coding (DirAC)  
Basic idea

- Analyze features of the sound field that translate into these parameters in hearing:

<table>
<thead>
<tr>
<th>Sound field</th>
<th>Human hearing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direction of arrival (DOA)</td>
<td>ITD, ILD</td>
</tr>
<tr>
<td>Diffuseness</td>
<td>Coherence</td>
</tr>
</tbody>
</table>
Directional audio coding (DirAC) Basic block diagram
Directional audio coding (DirAC) 
Microphone input

- B-format
Directional audio coding (DirAC)
Pressure and velocity approximates

\[ W \sim p \]
\[ [X_{e_x}, Y_{e_y}, Z_{e_z}] \sim u \]
Energetic analysis of sound field

- **Intensity**
  \[ I = pu \]
  - pressure x velocity

- **Energy density**
  \[ E = \frac{1}{2} \rho_0 \left( \frac{p^2}{Z_0^2} + ||u||^2 \right) \]
  - pressure squared + velocity squared
Directional audio coding (DirAC)  
DirAC analysis

• Analyzed separately for each frequency band:
  – Direction of arrival
    \[ \theta = -\mathbf{I} \]
  
  – Diffuseness
    \[ \psi = 1 - \frac{\|E\{\mathbf{I}\}\|}{cE\{E\}} \]
    (ratio of propagating and total energies)
Directional audio coding (DirAC)
DirAC analysis

Fig. 3. DirAC analysis.

The target of directional analysis, which is shown in Figure 3, is to estimate the direction of arrival of sound together with an estimate of whether the sound is arriving from one or multiple directions at the same time. The time-frequency transformation used in this article is the short-time Fourier transform (STFT). The following equations are given in the frequency domain. The first-order B-format stream [Furness 1990] consists of an omnidirectional signal $W$ and $X$, $Y$, $Z$-signals having the directional pattern of a dipole directed along the Cartesian axes. The signal $W$ has been scaled down by $\sqrt{2}$. Thus, the sound pressure can be estimated as $P = \sqrt{2}W$, and the dipole signals together form a vector $U = [X, Y, Z]$, which estimates the sound field velocity vector.

The intensity vector $I$ expresses the net flow of sound energy as a 3D vector and can be computed as

$$I(k, n) = \left(\frac{1}{\sqrt{2}}\right)\text{Re}\left\{W(k, n) \ast U(k, n)\right\},$$

where $\ast$ denotes complex conjugation, $k$ is the frequency bin index, and $n$ is the temporal frame index.

The direction of sound is defined to be the opposite of the direction of the intensity vector at each frequency band. The diffuseness of sound field is computed as

$$\psi(k, n) = 1 - \frac{\sqrt{2}||\langle \text{Re}\{W(k, n) \ast U(k, n)\}\rangle ||}{||\langle |W(k, n)|^2 + |U(k, n)|^2 / 2 \rangle ||},$$

where $|| \cdot ||$ denotes the $l_2$-norm of the vector and $\langle \cdot \rangle$ denotes temporal averaging. The outcome of this equation is a real-valued number between zero and one, indicating whether the sound field is approximated to consist of direct sound only ($\psi = 0$), a diffuse field ($\psi = 1$), or a partly direct and partly diffuse sound ($0 < \psi < 1$). The analysis is repeated, typically with an update frequency of about 100Hz. The sound intensity vectors are averaged with energy weighting within ERB frequency bands [Moore 1982], and one direction and diffuseness values for each band is estimated. Hence, the analysis and the synthesis are performed separately for each frequency band.

2.2.2 DirAC Synthesis.
The high-quality version of DirAC synthesis, shown in Figure 4, receives B-format signals from which a virtual microphone signal is computed for each loudspeaker direction. The directional pattern utilized in this study is cardioid. The virtual microphone signals are then modified in a nonlinear fashion, depending on the metadata. The low-bit-rate version of DirAC is not shown in the figure. 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 nondiffuse streams, which are processed separately. The nondiffuse stream is reproduced as point sources by using vector-base amplitude panning (VBAP) [Pulkki 1997]. The aim of the synthesis of the diffuse stream is to...
Directional audio coding (DirAC)
Time-frequency transform

• Short-time Fourier transform (STFT)
• Multi-resolution STFT
• Filterbank
Directional audio coding (DirAC)

DirAC synthesis

Fig. 4. DirAC synthesis.

Fig. 5. Different streams used in DirAC processing. The metadata describes the direction and the diffuseness of sound in the auditory frequency bands, with an update rate of about 100 Hz.

DirAC-based spatial sound reproduction for virtual worlds

The use of DirAC to spatialize virtual sound sources in virtual worlds is presented in this section. Initial versions of the techniques were introduced in Pulkki et al. [2009], where the processing was based on utilization of the mono DirAC stream. In this work the method has been evolved to utilize the B-format DirAC stream, which provides a somewhat higher perceptual quality, as shown later in Section 4. In addition, an evolved version of the method to synthesize the spatial extent of virtual ACMS.
Directional audio coding (DirAC)  
DirAC synthesis
Directional audio coding (DirAC)

DirAC synthesis

- Time-frequency transform
- Virtual microphones
- Decorrelation
- Diffuse stream
- Non-diffuse stream
- Loudspeakers

Metadata

- VBAP
- Gain averaging

Other channels

θ
Ψ

Mono DirAC stream
B-format DirAC stream
Directional audio coding (DirAC) Virtual microphones

- Virtual microphones are created as a weighted sum of the B-format signals (e.g. supercardioids)
- One microphone points to each loudspeaker:
Directional audio coding (DirAC)

DirAC synthesis

- Time-frequency transform
- Virtual microphones
- Gain averaging
- VBAP
- Metadata
- Decorrelation
- Diffuse stream
- Non-diffuse stream
- Loudspeakers
Directional audio coding (DirAC)

DirAC synthesis

Fig. 4. DirAC synthesis.

Fig. 5. Different streams used in DirAC processing. The metadata describes the direction and the diffuseness of sound in the auditory frequency bands, with an update rate of about 100 Hz.

2.2.3 DirAC Glossary.

— DirAC stream: a stream consisting of 1–4 audio channels with metadata for sound direction and diffuseness.
— Mono DirAC stream: a DirAC stream with having one audio channel and metadata, see Figure 5. This is typically used in low-bit-rate applications such as teleconferencing.
— Mono DirAC decoding: a technique to make a mono DirAC stream audible over loudspeakers or headphones. This is used in the low-bit-rate version.
— B-format stream: a stream consisting of an omni signal W and three dipole signals X, Y, and Z. Also known as the B-format audio bus.
— B-format DirAC stream: a stream where B-format audio is transmitted within the DirAC stream.
— B-format DirAC decoding: a technique to make the B-format DirAC stream audible over loudspeakers or headphones. This usually gives better perceptual quality for reverberation than mono DirAC decoding.

3. DIRAC-BASED SPATIAL SOUND REPRODUCTION FOR VIRTUAL WORLDS

The use of DirAC to spatialize virtual sound sources in virtual worlds is presented in this section. Initial versions of the techniques were introduced in Pulkki et al. [2009], where the processing was based on utilization of the mono DirAC stream. In this work the method has been evolved to utilize the B-format DirAC stream, which provides a somewhat higher perceptual quality, as shown later in Section 4. In addition, an evolved version of the method to synthesize the spatial extent of virtual...
Directional audio coding (DirAC)  
Nondiffuse stream

• Includes mostly the part of the sound that has a certain direction  
  – Reproduced as point-sources
• Vector base amplitude panning (VBAP)  
  – According to the direction sent in the metadata
• In many cases the direction in metadata is subject to abrupt temporal changes  
  – Gain smoothing is applied to avoid artifacts
Directional audio coding (DirAC)

DirAC synthesis

![Diagram of DirAC synthesis]

- Time-frequency transform
- Virtual microphones
- Decorrelation
- Diffuse stream
- Non-diffuse stream
- Loudspeakers
- Metadata
- Gain averaging
- VBAP
- Other channels
- Diffuseness
- Direction angle
- Updates about 100 Hz

2.2.3 DirAC Glossary.

- **DirAC stream**: A stream consisting of 1-4 audio channels with metadata for sound direction and diffuseness.
- **Mono DirAC stream**: A DirAC stream with having one channel of audio and metadata, see Figure 5. This is typically used in low-bit-rate applications such as teleconferencing.
- **Mono DirAC decoding**: A technique to make a mono DirAC stream audible over loudspeakers or headphones. This is used in the low-bit-rate version.
- **B-format stream**: A stream consisting of an omni-directional signal and three dipole signals (signals X, Y, and Z). Also known as the B-format audio bus.
- **B-format DirAC stream**: A stream where B-format audio is transmitted with the DirAC stream.
- **B-format DirAC decoding**: A technique to make the B-format DirAC stream audible over loudspeakers or headphones. This usually gives better perceptual quality for reverberation than mono DirAC decoding.

3. DIRAC-BASED SPATIAL SOUND REPRODUCTION FOR VIRTUAL WORLDS

The use of DirAC to spatialize virtual sound sources in virtual worlds is presented in this section. Initial versions of the techniques were introduced in Pulkki et al. [2009], where the processing was based on utilization of the mono DirAC stream. In this work the method has been evolved to utilize the B-format DirAC stream, which provides a somewhat higher perceptual quality, as shown later in Section 4. In addition, an evolved version of the method to synthesize the spatial extent of virtual...
Directional audio coding (DirAC)
Diffuse stream

• Includes mostly the reverberant and the ambient parts
  – Reproduced as surrounding
• Decorrelation
  – E.g. frequency-dependent delays or noise-bursts
Directional audio coding (DirAC)

DirAC synthesis

![Diagram of DirAC synthesis process]

- Time-frequency transform
- Virtual microphones
- Decorrelation
- Diffuse stream
- Non-diffuse stream
- Loudspeakers
- Metadata
Directional audio coding (DirAC)
DirAC synthesis

- Time-frequency transform
- Virtual microphones
- Decorrelation
- Direction angle
- Diffuseness

**Fig. 5. Different streams used in DirAC processing.**

The metadata describes the direction and the diffuseness of sound in the auditory frequency bands, with an update rate of about 100 Hz.

create a perception of a sound that surrounds the listener. The diffuse stream is reproduced by decorrelating the virtual microphone signal and reproducing the result with the corresponding loudspeaker.

**2.2.3 DirAC Glossary.**

- **DirAC stream**: A stream consisting of 1-4 audio channels with metadata such as sound direction and diffuseness.
- **Mono DirAC stream**: A DirAC stream with having one channel and metadata. This is typically used in low-bit-rate applications such as teleconferencing.
- **Mono DirAC decoding**: A technique to make a mono DirAC stream audible over loudspeakers or headphones. This is used in the low-bit-rate version.
- **B-format stream**: A stream consisting of an omni signal (W) and three dipole signals (X, Y, Z). Also known as the B-format audio bus.
- **B-format DirAC stream**: A stream where B-format audio is transmitted within the DirAC stream.
- **B-format DirAC decoding**: A technique to make the B-format DirAC stream audible over loudspeakers or headphones. This usually gives better perceptual quality for reverberation than mono DirAC decoding.

**3. DIRAC-BASED SPATIAL SOUND REPRODUCTION FOR VIRTUAL WORLDS**

The use of DirAC to spatialize virtual sound sources in virtual worlds is presented in this section. Initial versions of the techniques were introduced in Pulkki et al. [2009], where the processing was based on utilization of the mono DirAC stream. In this work, the method has been evolved to utilize the B-format DirAC stream, which provides a somewhat higher perceptual quality, as shown later in Section 4. In addition, an evolved version of the method to synthesize the spatial extent of virtual sources is presented.
Directional audio coding (DirAC)

Basic idea

- Synthesize features that translate into parameters for spatial-sound perception in hearing:

<table>
<thead>
<tr>
<th>Reproduction</th>
<th>Human hearing</th>
</tr>
</thead>
<tbody>
<tr>
<td>VBAP</td>
<td>ITD, ILD</td>
</tr>
<tr>
<td>Decorrelation</td>
<td>Coherence</td>
</tr>
</tbody>
</table>
Applications

- Spatial sound capturing and reproduction
- Teleconferencing
- Spatial impulse response rendering (SIRR)
- Virtual world / game audio

- Combinations of all these techniques!
Applications
Spatial sound reproduction
Applications

Spatial sound reproduction

• Spatial sound is captured with a B-format microphone
• DirAC processing can be applied for arbitrary loudspeaker layouts
• Reproduction is also possible with headphones (with or without head tracking)
Applications
Teleconferencing
Applications
Teleconferencing

- Only mono channel + low-bit-rate metadata is sent
  - Good quality can still be obtained
- Locations of the participants are reproduced
  - Increases immersion and naturalness
  - Increases speech intelligibility
Applications
Spatial impulse response rendering

Multi-channel impulse response measurement and processing
Convolution

5.1 reproduction
Applications
Spatial impulse response rendering

- Basically the same processing as in DirAC
- Smoothing is turned off
- High temporal resolution is important
Applications
Virtual worlds

![Diagram of a virtual world with three sound sources: source 1, source 2, and source 3. The user should auditorily perceive the sources in the directions as they are relative to the avatar, the spatial extents of the sources, in some cases the reflections from nearby surfaces as echoes, and also the surrounding reverberation generated by all sources.](image)

The first task is to reproduce the sources such that the listener will perceive the direction of a virtual sound source in a controlled manner. The transfer function between the source and the listener may also be modeled with proper delaying, gain, and also with filtering, which implements the effects of frequency-dependent directivity of the source and some filtering effects of the atmosphere [Savioja et al. 1999]. In virtual worlds, the sound sources are usually reproduced in the direction where the sound object is relative to the coordinates of the avatar. This is typically performed with amplitude panning [Blumlein 1958] in loudspeaker listening. With a two-channel stereophonic setup, the sources can be panned to directions between the two loudspeakers [Bennett et al. 1985]. If more loudspeakers are used, more directions can be covered [Pulkki 2001]. In headphone listening, head-related transfer-function (HRTF) techniques are used [Møller et al. 1995], where in principle the virtual sources can be positioned to arbitrary directions. Furthermore, the virtual-source direction can be kept constant with the external world by measuring the direction of the listener's head in headphone listening [Begault et al. 2001].

In another task, the spatial extent of sound sources is to be reproduced. In the real world, humans are capable of perceiving its extent to some degree [Blauert 1997]. The current techniques for this are a bit complex: a typical method is to approximate a large sound source with a number of point sources [Tsingos et al. 2004]. In principle, this is a valid method. In practice, it is problematic to obtain the sound signals for each point source, as this demands recording the sound emanating from different surface points of a large sound source. Potard and Burnett [2004] suggested that these different point sources could be obtained using decorrelation. Tsingos et al. [2004] suggested clustering all point sources into a few points sources based on their perceptual importance. Another suggested method is to control the coherence between the loudspeaker or the headphone signals [Jot et al. 2006]. Recently it has also been suggested that the perception of the spatial extent of a virtual source could be reproduced by applying different frequency bands of the signal to different directions [Pulkki et al. 2009; Verron et al. 2010].

ACM Transactions on Applied Perception, Vol. 9, No. 2, Article 8, Publication date: June 2012.
Applications

Virtual worlds

• DirAC can perform all traditional tasks for spatial sound reproduction in virtual worlds
  – Source positioning
  – Reverberation generation

• In addition
  – Spatial extent can be controlled efficiently
  – Reverberation can be generated efficiently
Applications
Combining different applications

• These methods can also be used at the same time
• Audio from various sources can be processed with the same “backend”
  – No significant increase in the computational complexity
Evaluation

- Signal-dependent nonlinear processing
- Does not target to reconstruct wave field
- Must be evaluated with human listeners
Evaluation
Loudspeaker reproduction

Fig. 11. Mean opinion scores for different reproduction methods in listening room with 95% confidence intervals.

Fig. 10. Mean opinion scores in different test cases in anechoic chamber with 95% confidence intervals.
Evaluation
Headphone reproduction

The overall quality of reproduction

Excellent
Good
Fair
Poor
Bad

hqDirAC htrack
telecDirAC htrack
v2DirAC htrack
hqDirAC no_htrack
stereo htrack
stereo
mono
Evaluation
Teleconference

The listening test was divided into three parts, which included five 40-phrase sessions. One session corresponded to one technique. The order of the sessions was randomized in all parts. Also talker calls were randomized in the sessions. Half of the test phrases (20/40) contained the call sign "Arrow" for the listener. The total amount of the test phrases in three parts was 60/120. Two training sessions (2 x 9/25) with the reference reproduction preceded the actual listening test.

4.3. Test results

The results including the mean scores and 95% confidence intervals for each technique are shown in Fig. 9. The effect of the listener was removed from the analysis by normalizing the data of each listener with respect to the mean values and standard deviation according to ITU-R BS.1284-1 [21]. A one-way variance analysis (ANOVA) indicated that the technique has a significant effect on the results:

\[ F(4, 75) = 187.07 \] and \[ p < 0.05 \]. Additionally, multiple comparison (Tukey-Kramer method) was used to determine which pairs of the means were significantly different.

According to the test results in Fig. 9, the reference, the XY stereo, DirAC with both one- and two-dimensional microphone arrays have similar results to each other, whereas the mean value of mono differed significantly from the means of the other techniques. The listeners responded to the simultaneous calls more often wrongly with mono than with the XY stereo and DirAC.

4.3.1. Discussion

Nearly the same number of correct responses were obtained with the XY stereo, DirAC with both one- and two-dimensional microphone arrays, and the reference. The measured mean values for these techniques were close to each other in Fig. 9. It can be thus said that DirAC provides practically similar improvement of speech intelligibility than the XY stereo does when compared to mono transmission, although the increase of the data rate compared to mono is minimal in DirAC. Utilizing the perceptual compression methods presented in [16] and shortly reviewed in Section 2.3, bit-rate for the directional metadata including the azimuth values could be as low as 3 kbit/s in the wideband range of telecommunication. This is very low increase in bit-rate, when compared to required doubling of the transmission bandwidth in the XY stereo. Typical wideband speech codecs have a bit-rate of about 64 kbit/s for a single audio channel [15].

The fact that DirAC provides such result is notable therefore that the XY stereo may not produce such good speech intelligibility than in now presented listening test in all practical purposes. In the XY stereo, the end-fire direction of the dipole microphone was pointed ideally towards the one loudspeaker and the directional null of the dipole microphone towards the other loudspeaker. Thus, sound from one loudspeaker was optimally captured in one microphone and prevented in the other, which might not be the case in a typical recording scenario. Besides, dipole microphones are seldom used in consumer applications, because of the relatively high prices of the microphones. Instead of dipoles, cardioids are typically used. Unfortunately, the directivity pattern of the cardioid microphone is broader than the pattern of the dipole microphone, and the reproduced sound is thus more distorted spatially decreasing also speech intelligibility [17].

The results for the re-conducted listening test, presented in this paper, differs from the results of the earlier test, which are presented in [11]. Simultaneous talkers overlapped one another more efficiently in the presented test than in the earlier test, and less correct responses were achieved in CRM with all techniques. Also, 95% confidence intervals are now much smaller than in the results of the earlier test.

5. SUMMARY

This paper reviews the use of Directional Audio Coding AES 40
Evaluation
Spatial impulse response rendering

![Chart showing evaluation results for different system pairs and room sizes.](chart.png)
Evaluation Challenges

- Temporal resolution is not infinite
  - Smoothing
- Decorrelation smears the signal in time
Summary

• Perception-based reproduction of spatial audio
• Input from first-order microphones
• Directional analysis of sound field in auditory frequency bands
• Nondiffuse sound reproduced as point sources
• Diffuse sound reproduced as surrounding
• Many applications
• Listening tests: difference to reference condition small
References

• J. Ahonen and V. Pulkki. **Speech Intelligibility in Teleconference Application of Directional Audio Coding.** AES 40\textsuperscript{th} international conference, Tokyo, Japan, October 2010.
• M.-V. Laitinen, T. Pihlajamäki, C. Erkut, and V. Pulkki. **Parametric time-frequency representation of spatial sound in virtual worlds.** ACM Transactions on Applied Perception, June 2012.
Thank you!