ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Signal Processing in Acoustics
Session 2aSP: Array Signal Processing for Three-Dimensional Audio Applications II

2aSP3. Spatial audio coding with spaced microphone arrays for music recording and reproduction.

Archontis Politis*, Mikko-Ville Laitinen and Ville Pulkki

*Corresponding author's address: Department of Signal Processing and Acoustics, Aalto University, Espoo, 02150, Espoo, Finland, archontis.politis@aalto.fi

Spaced microphone arrays are commonly used in multichannel recording of music, due to their inherent quality of natural incoherence between the surround channels at reproduction, at the expense of accurate localization. Recent methods in parametric spatial audio coding, such as Directional Audio Coding, exploit coincident microphone patterns to extract directional information and reproduce it in a perceptually optimal way. In this study we present how Directional Audio Coding can be adapted for spaced arrays, offering improved localization cues at reproduction, without compromising the qualities of the spaced microphone recording techniques. Examples are presented for some well-established array configurations.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

After the adoption of stereophony in the music production industry and its evolution to multichannel loudspeaker setups, such as the early quadraphony or the later standardized surround setups, multichannel recording techniques have been an active topic of research. Setting aside individual recording of instruments with spot-microphones and their mixing by the sound engineer, recording of the overall sound scene requires carefully placed microphone arrays, which are intended for direct reproduction over the surround system.

Two main categories of recording methods have evolved for multichannel sound capture and reproduction, coincident and spaced ones, both originating from stereo recording techniques. Coincident techniques assume microphones with spacings small enough that are ‘seen’ by the wavelengths of interest as being in the same point in space. A single impinging plane wave is then captured by all microphones with almost equal phase but with different amplitudes, depending on the orientation of the microphones and their directivity patterns. Examples of coincident techniques are the well-known Blumlein stereo (or XY stereo), mid-side stereophony, its multichannel extension of double mid-side [1], and the first-order B-format signals which form the basis for the ambisonic methods of sound capture and reproduction [2]. Coincident methods distribute the audio to the loudspeakers by some means of matrixing of the microphone channels, and they also permit some control of the reproduced stage-width and other directional characteristics of the perceived sound scene after the recording has been made. Furthermore, due to the fact that they encode primarily inter-channel level differences they can achieve consistent localization of discrete sound events at all directions around the listener. However, in terms of perception, the drawback of coincident recordings is the high correlation of the output channels. Even in the case of purely diffuse sound the reproduced channels are highly correlated for all frequencies resulting in a sense of lack of envelopment and depth, loss of immersion, and comb filter-like effects for small movements of the head.

In the second class of recording methods, that of spaced arrays, the microphones have considerable distances between them, so that an impinging plane wave is captured by the microphones with significant phase differences. In the case of directional microphones, a combination of inter-channel phase-differences and inter-channel level differences occur, which is determined by the direction of incidence, the geometry of the array, the directional patterns of the microphones and their orientation. Reproduction of spaced microphone recordings results in a more spacious enveloping sensation compared to coincident methods, due to a reduced inter-aural coherence for the listener. Hence, they are more commonly used in practice. However, the relation of inter-channel level differences and time-differences to the perceived direction is more complex than the pure level-differences of coincident techniques, which results in sources collapsing to single loudspeakers, sources perceived as spread in space or sources with a fluctuating direction [3].

A recent parametric technique for multichannel audio, Directional Audio Coding (DirAC) [4], is based on analyzing the coincident B-format recording and extracting psychoacoustically relevant directional information, which is subsequently used in the synthesis stage to generate the output channels in a way that, when reproduced, they are perceptually close to the original. From an audio-engineering perspective, DirAC aims to combine the best of both approaches, coincident and spaced recordings, through a separation between the directional and the non-directional/diffuse sounds. This separation enables accurate and stable localization along with good sense of envelopment and depth in reproduction. Using DirAC with spaced microphones was introduced in [5, 6]. In these studies, the methods were presented for a general case, and the example microphones arrays were optimized for DirAC processing. Formal listening tests showed that close-to-reference perceptual quality can be obtained. In the present paper, these methods are extended to existing spaced microphone arrays that are commonly used for music recording. Using the parametric time-frequency processing of DirAC, it is
possible to enhance the spaced recordings with stable localization cues without losing their inherent good sense of envelopment. Furthermore, the proposed parametrization of the directional information in the recording brings further possibilities for creative processing/mixing of the material [7] and for a unified common representation and treatment of various recording formats, coincident and spaced.

**Spaced microphone arrangements**

In the case of spaced array arrangements, the perceived sound scene depends completely on the adjustment of inter-microphone distances, pick-up patterns and orientations of the individual microphones. Since various objectives should be met throughout the design process, such as perceived stage width, clarity of imaging on the front, diffuse reproduction on the back etc., there are many different arrangements in use. In [8], a further categorization is attempted on the various designs in use, with a distinction between “main-microphone techniques” and “techniques with front-rear separation”. Main-microphone techniques are the ones where a single spaced array with front and surround microphones is used for capturing the sound scene, usually placed close to the music stage, approximately on the critical distance from the instruments.

![Diagram of spaced microphone arrangements](image)

**Figure 1:** Standard spaced array setups for multichannel recording of music.

Front-rear techniques use instead a main 3-channel array placed close to the stage, intended to capture mostly the direct sound for the front-channels (L-C-R) and a second multichannel array placed several meters away from the the main one and adjusted to capture as much reverberant sound as possible. Then the total output in reproduction is a mix between the front and back arrays. Examples of front arrays are the popular Fukada tree, the Optimal Cardioid Triangle (OCT) variants, the INA-3 variants. Examples of rear arrays are the Hamasaki square and the IRT-Cross array. The front arrays can also be augmented to full-surround main arrays by the addition of surround microphones, see OCT-surround and INA-5 arrangements in [9]. Examples of the above arrays can be seen in Fig. 1, where all dimensions and orientations are based on [9].
DIRECTIONAL ANALYSIS AND PARAMETRIZATION

The enhancement of spaced microphone recordings proposed in this work is based on the principles of Directional Audio Coding (DirAC) [4]. DirAC is a parametric method for a perceptually motivated representation, transmission, and reproduction of spatial sound. The model of DirAC assumes that with a time-frequency representation similar or finer to the resolution of the auditory system, it is adequate to encode and decode the local sound field with two parameters for each time-frequency tile. These are the direction of arrival (DOA) of incident sound energy and the diffuseness. DOA is assumed to relate to the directional cues of localization, while diffuseness relates to the sense of reverberation or sound source extent represented by interaural coherence. More information on the formulation of the parameters and their physical meaning can be found in [4]. DOA and diffuseness are used in the synthesis stage to recreate perceptually the captured sound scene. The synthesis is based on a decomposition of the input signals to a directional stream, corresponding to the part of the recording analyzed as clearly directional, and a diffuse stream, which corresponds mostly to reverberant and spatially extended sounds.

The parameters are extracted in the time-frequency domain directly from the input signals. A short-time Fourier Transform (STFT) is used in this implementation. Assuming a combination of a single dominant sound source and diffuse sound for each time-frequency frame, a DOA vector $\mathbf{I}_{\text{DOA}}$ is computed pointing to the direction of the sound source. Considering an STFT implementation, the azimuth parameter is then computed as:

$$\theta(k, n) = \arctan \left( \frac{I_y(k, n)}{I_x(k, n)} \right)$$

where $k, n$ are the frequency and time indices of the STFT and $I_x, I_y$ are the components of the DOA vector. The diffuseness parameter $\psi$ is estimated by the temporal variation of the DOA vector, as described in [10]:

$$\psi = \sqrt{1 - \frac{\mathbb{E}[\|\mathbf{I}_{\text{DOA}}\|]}{\mathbb{E}[\|\mathbf{I}_{\text{DOA}}\|]}}.$$  \hspace{1cm} (2)

Diffuseness is bounded by $\psi \in [0, 1]$ with a zero value for a single plane wave and a unity value for a perfectly diffuse field. In practice the expectation operators are approximated with time-averaging.

DOA vector computation

The DOA estimation is conducted in different ways for the low and the mid/high frequency range. The method can be adapted to use all the microphones in a 5-channel main microphone array, or to utilize only the front triangle array. Here, we demonstrate the formulation for the front triangle only, since this approach can be used for both main surround arrays and front-rear separated arrays.

At low frequencies, with a wavelength significantly longer than the maximum distance between the microphones in the triangle, the microphones can be assumed to be effectively coincident. The low-frequency range for coincidence in our case is taken as $f_{\text{coinc}} \leq \frac{c}{d_{\text{max}}}$, where $c$ is the speed of sound and $d_{\text{max}}$ is the maximum distance between the microphones in the array. The B-format microphone signals are captured by an omnidirectional component $W(\theta) = 1$ and two orthogonal dipole components $X(\theta) = \cos \theta$ and $Y(\theta) = \sin \theta$, where $\theta$ is the angle of incidence. Then with the coincidence assumption, the microphone signals can be used for estimating the B-format signals, as follows. Any signal captured with a first-order directional pattern $D(\theta) = a + (1-a)\cos \theta$, where $0 < a < 1$, can be expressed as a linear combination of the B-format signals. Let as assume that the microphone placement and orientations follow the
Using the B-format signals, the DOA vector $\mathbf{I}_a$ can be expressed as

$$\mathbf{I}_a(k,n) = \left[ \begin{array}{c} \Re \left\{ W(k,n) \cdot \overline{X(k,n)} \right\} \\ \Re \left\{ W(k,n) \cdot \overline{Y(k,n)} \right\} \end{array} \right],$$

where $\overline{\cdot}$ denotes complex conjugation. $\mathbf{I}_a$ is opposite and proportional to the active intensity vector, expressing the mean flow of sound energy.

For the frequency range $f_{\text{coinc}} \geq \frac{c}{4d_{\text{max}}}$, an alternative DOA estimation method is used. By exploiting the assumption of a single dominant plane wave at each time-frequency tile, we can get an estimate of each DOA simply from the magnitude differences between the microphones of the array without the assumption of coincidence. Then the magnitude spectrum of the microphone signals can be expressed as:

$$\left| S_L(k,n) \right| = \left| A(k,n) \right| \left| e^{-j\phi_L D(\theta - \theta_0)} \right| \right| D(\theta - \theta_0) \right| \right| D(\theta - \theta_0) \right|$$

$$\left| S_C(k,n) \right| = \left| A(k,n) \right| \left| e^{-j\phi_C D(\theta)} \right| \right| D(\theta) \right| \right| D(\theta) \right|$$

$$\left| S_R(k,n) \right| = \left| A(k,n) \right| \left| e^{-j\phi_R D(\theta + \theta_0)} \right| \right| D(\theta + \theta_0) \right| \right| D(\theta + \theta_0) \right|$$

where $A(k,n)$ is the complex amplitude of the plane wave, $\phi_L, \phi_C, \phi_R$ are respective phase differences due to the position of each microphone and DOA of the plane wave. Then a DOA vector $\mathbf{I}_m$ can be expressed in a similar way to Eq. 4 as

$$\mathbf{I}_m = \left| A(k,n) \right| \left[ \begin{array}{c} \cos \theta \\ \sin \theta \end{array} \right] = \left| A(k,n) \right| \left[ \begin{array}{ccc} 1/(\cos \theta_0 - 1) & -2/(\cos \theta_0 - 1) & 1/(\cos \theta_0 - 1) \\ 1/\sin \theta_0 & 0 & -1/\sin \theta_0 \end{array} \right] \cdot \left[ \begin{array}{c} D(\theta - \theta_0) \\ D(\theta) \\ D(\theta + \theta_0) \end{array} \right]$$

$$= \mathbf{M}_{\text{card2xy}} \cdot \left[ \begin{array}{c} \left| S_L(k,n) \right| \\ \left| S_C(k,n) \right| \\ \left| S_R(k,n) \right| \end{array} \right]$$

**Figure 2**: Front triangle arrangements for directional analysis.
The final DOA vector is then

\[ \mathbf{I}_{\text{DOA}}(k, n) = \begin{cases} I_a(k, n), & \text{for } f_k < f_{\text{coinc}} \\ I_m(k, n), & \text{for } f_k \geq f_{\text{coinc}} \end{cases} \] (8)

To investigate the performance of the DOA estimation, impulse responses for the three front arrays of Fig. 1 are simulated for plane wave propagation on 36 directions in the horizontal plane. Ideal cardioid capsules are assumed. The output to a broadband plane wave is then simulated for each direction by convolving 1 second of white noise with the impulse responses. To simulate more realistic conditions, internal noise is introduced on the analysis signals by adding uncorrelated random noise to each microphone signal, corresponding to a signal-to-noise ratio (SNR) of 20dB with respect to the plane wave. Finally, the L, C, R signals are analysed according to the procedure described above and the estimates are averaged across all time frames for each target angle. The results are presented in Fig. 3. The estimated mean DOAs are close to the true ones for all arrays and the RMS error for all directions is less than 20 degrees.

**SYNTHESIS OF OUTPUT CHANNELS**

The synthesis method is similar to B-format DirAC synthesis but is adapted to exploit the characteristics of the spaced arrangements. The separation of the microphone channels into the directional and the diffuse stream enhances the spatial imaging by applying amplitude panning gains on adjacent channels according to the analyzed direction. The mixing between diffuse and directional stream is adjusted by the diffuseness separately for each frequency bin. Contrary to the B-format DirAC, where decorrelation is necessary for all frequencies due to the high coherence between the output channels, the microphone signals of the spaced arrays are naturally partially incoherent between them, and decorrelation is not required at all frequencies [5]. Thus, possible decorrelation artifacts can be avoided. For a main surround array the microphone channels are only partially decorrelated, depending on the frequency and the geometry of the array. The proposed processing scheme for these arrays is presented in the block diagram of Fig. 4a. The gains \( \epsilon, \gamma \) are used to tune the amount of the signals that are decorrelated and the amount that is directly routed to the outputs. These gains are computed according to the ideal coherence of cardioid patterns in an ideal diffuse field [11]. A detailed presentation of their derivation is presented in [6] for a custom spaced microphone array. Furthermore, it suggested that only the L, R, LS, and RS channels are used for the diffuse stream outputs as they are more uniformly arranged and less coherent between them than with the center channel included.

Fig. 4b shows the proposed scheme for a front-rear separated array. In this case, only the rear array is used for the diffuse stream. Since the rear arrangements are positioned far enough from the stage, the reverberant sound dominates. It is assumed that the signals are completely incoherent with the directional stream, which originates from the front array close to the stage. Hence, decorrelation is not used in this case. Furthermore, the amplitude panning gains for the directional stream are limited in the range of the front-center speakers, since it is not possible to achieve spatial imaging on the side with panning due to low coherence between the L/R and the LS/RS channels. This is achieved simply by limiting the DOA parameter to the range between the L-R speakers even when the analyzed directions fall outside of it. A delay control can be also used for the front triangle signals to tune the time-alignment between the directional stream and the diffuse stream.
FIGURE 3: Directional analysis results for 36 directions of a plane wave from 0deg to 180deg at every 5deg with SNR=20dB. First column: true angle and mean analyzed angle with frequency, second column: corresponding absolute error between true and analyzed angle, bottom figure: RMS error averaged across all intended directions. The vertical lines denote the frequency $f_{\text{coinc}}$ where the estimation method is changed.

CONCLUSIONS

Directional Audio Coding is a perceptually-motivated parametric method to reproduce spatial sound. The input is typically a B-format recording, but recently, the use of custom spaced
microphone arrays has been suggested and shown to enhance the spatial imaging of directional sound events, while preserving the spacious sensation of reproduction of spaced recordings. In this work, these methods are extended to existing spaced arrays that are commonly used for music recording. The performance of the proposed processing depends on a proper estimation of the directional parameters. A novel formulation of this estimator is presented, suitable for various microphone geometries. The estimator is evaluated with simulations and the results show adequate accuracy with the arrays under test. Furthermore, two variations of the processing flow are proposed for main surround and front-rear combined arrays, that consider the natural incoherence of the microphone channels in reproduction.

ACKNOWLEDGMENTS

The Fraunhofer IIS and the GETA Graduate School 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 agreement n° [240453].

REFERENCES


