Capturing and Reproduction of a Crowded Sound Scene Using a Circular Microphone Array

Nikolaos Stefanakis* and Athanasios Mouchtaris*,†
*FORTH-ICS, Heraklion, Crete, Greece, GR-70013
†University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-70013

***Abstract*—Over the years, different spatial audio techniques have been proposed as the means to capture, encode and reproduce the spatial properties of acoustic fields, yet specific issues need to be modified each time in accordance to the type of microphone array used as well as with the technology used for reproduction. Using a circular array of omnidirectional microphones, we formulate in this paper a parametric and a non-parametric approach for capturing and reproduction of the crowded acoustic environment of a football stadium. A listening test performed reveals the advantages and disadvantages of each approach in connection to the particularities of the acoustic environment.**

I. Introduction

In recent years a lot of research has been produced about technologies for capturing and reproducing the spatial properties of sound. More and more human activities, such as watching movies, listening to music, playing games and communicating through the web, rely on multichannel sound reproduction facilities for improving realism, sensation and ineligibility of communication. One of the most challenging tasks is to deliver multichannel sound related to large scale events, such as in the capturing and reproduction of athletic events. For example, the broadcaster can capture and transmit a panoramic image of the spectators responses during a football game so that home users can enjoy an immersive experience of the athletic event.

Typically, capturing of large sport events is accomplished with several microphones placed around the pitch or inside the crowd, so that each microphone focuses on a particular segment of the event [1]. A great amount of equipment needs to be carefully distributed all around the playing field, requiring a lot of preparation time and attendance during the game. Then, it depends on the experience and subjective judgement of the sound engineer to mix all the signals into the final stereo or surround format that is transmitted by the broadcaster. The approach, to use one or just a few compact sensor arrays to capture and reproduce sound from such large scale events presents an interesting alternative; it may reduce the cost of equipment and implementation, allow flexibility in the processing and manipulation of the captured spatial information and allow for efficient encoding of the data to reduce bandwidth requirements during transmission.

The project leading to this application has received funding partly from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687605, Project COGNITUS.

Directional Audio Coding (DirAC) [2] represents an important paradigm in the family of parametric approaches, providing an efficient description of spatial sound in terms of a few audio downmix signals and parametric side information, namely the Direction-of-Arrival (DOA) and diffuseness of the sound. While originally designed for differential microphone signals, an adaptation of DirAC to compact planar microphone arrays with omnidirectional sensors has been described in [3], [4] and an adaptation to spaced microphone arrangements in [5]. In the same direction, the approach in [6] presents an example about how the principles of parametric spatial audio can be exploited for the case of a linear microphone array.

In the context of binaural reproduction, Cobos et al. presented an approach based on a fast 3-D DOA estimation technique in [7]. While not explicitly calculating any parameter related to diffuseness, the authors claimed that diffuseness information is inherently encoded by the variance in the DOA estimates. Capturing and reproduction of an acoustic scene using a circular microphone array has been presented in [8], [9]. The authors were able to demonstrate an advantage in terms of perceived spatial impression and sound quality, but in a scenario with limited number of discrete sound sources whose number and direction is provided by the DOA estimation and counting technique described in [10].

Demonstrating the applicability of techniques of this family to large-scale sport events is certainly interesting, not only because of the great potential for commercial exploitation, but also because of the inherent technical challenges which such acoustic environments introduce. At each time instant there are hundreds of spectators cheering and applauding simultaneously from many different directions, and therefore, the source sparseness and disjointness conditions which are assumed for DOA estimation are most of the time not met. Yet, more conventional techniques which attempt accurate physical reconstruction of the sound field, such as Ambisonics [11] and Wave Field Synthesis [12], do not present ideal solutions either. The former technique suffers from a very narrow optimal listening area, while the latter requires a prohibitively large number of loudspeakers which is impractical for commercial use. An interesting approach to tackle these problems has been presented in [13], using a circular array of first-order differential microphones. Essentially, the method proposes to use linear array processing in order to emulate microphones with directivity responses which conform to stereophonic pan-

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

ning laws. As it does not require a DOA estimation step, this non-parametric approach presents an interesting alternative, as opposed to the for-mentioned parametric techniques.

Borrowing ideas from this last approach as well as from the work of Cobos et al. in [7], we formulate in this paper a non-parametric and a parametric approach respectively for capturing and reproduction of a sound scene in 2-D using a circular sensor array of omnidirectional sensors. We then apply both techniques to a recording of a crowded football stadium including thousands of spectators. A listening test performed through a square loudspeaker configuration illustrates the advantages and drawbacks of each approach with respect to the particular sensor array topology and acoustic environment.

II. Technical background

Assume an array of M sensors and a reproduction system comprised of L loudspeakers. In what follows $\mathbf{x}(\tau,\omega) = [X_1(\tau,f),...,X_M(\tau,f)]^T$ is the observed signal at the M microphones at time-fame with index τ and at frequency index f. At each time-frequency (TF) point let $Y_l(\tau,f)$ denote the signal which is sent to lth loudspeaker. In terms of the parametric approach, the process of capturing and reproduction may be written in a generic form as

$$Y_l(\tau,f) = F(\mathbf{x}(\tau,f),\theta_{\tau,f},\psi_{\tau,f}), \tag{1}$$

where $F(\cdot)$ denotes a non linear process to synthesize the lth loudspeaker signal based on the single estimated direction $\theta_{\tau,f}$ and diffuseness value $\psi_{\tau,f}$. On the other hand, the non-parametric approach provides the simplest way for capturing the acoustic scene, relying on the linear process of beamforming for obtaining the loudspeaker signals as

$$Y_l(\tau,f) = \mathbf{w}_l(f)^H\mathbf{x}(\tau,f), \tag{2}$$

where $\mathbf{w}_l(f) = [w_{1l}(f),...,w_{Ml}(f)]^T$ is the vector with the M fixed complex beamformer weights for the lth channel at frequency f.

In this paper we use Vector Base Amplitude Panning (VBAP) [14] in order to define a mapping between the incident angle θ and the loudspeaker gains $\mathbf{g}(\theta) = [g_1(\theta),...,g_L(\theta)]^T$, which is in accordance to the physical loudspeaker distribution around the listening area. VBAP requires that the loudspeakers are distributed along a circle of fixed radious around the listening area but their number and direction might be arbitrary, a fact that provides important flexibility for multichannel audio reproduction. In Fig. 1 we show an example of how such panning gains would look like for a 2-D loudspeaker setup in the case of 4 and 8 uniformly distributed loudspeakers in (a) and (c) and for 5 non-uniform loudspeakers in (b).

Similar to the requirements stated in [13], we may rely on VBAP to dictate optimal panning gains with the following characteristics;

- the problem of unwanted inter-channel crosstalk is efficiently addressed by ensuring that given a single plane wave incident at a certain angle (in the azimuth plane), only two loudspeakers will be activated during reproduction (the two ones which are adjacent to the estimated angle),
- the sum of the squares of all loudspeaker gains along θ is equal, meaning that there is no information loss.

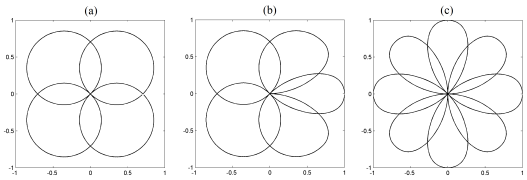

Fig. 1. Desired directivity patterns for (a) a 4-channel, (b) a 5-channel and (c) an 8-channel system.

This implies that we can use the loudspeaker gains provided by VBAP as the desired directivity response for a set of beamformers, which we can then use in order to capture the acoustic environment at different directions. The signal at the output of each beamformer can then be sent directly to the corresponding loudspeaker without further processing [13], supporting thus use of Eq. (2).

III. Non-parametric approach

Using the panning gains dictated by VBAP, we present in this section an approach for calculating the beamformer weights assuming a circular microphone array of M sensors. Consider a grid of N uniformly distributed directions θ_n in $[-180^o, 180^o)$ and keeping in mind that the desired response for any of the L beamformers is real, the problem becomes to find the weights $\mathbf{w}_l(f)$ in order to satisfy

$$\mathbf{D}^H(f)\mathbf{w}_l(f) \approx \mathbf{G}_l,\ l = 1,...,L. \tag{3}$$

Here, $\mathbf{G}_l = [g_l(\theta_1),...,g_l(\theta_N)]^T$ is the desired response provided by VBAP and $\mathbf{D}(f) = [\mathbf{d}(f,\theta_1),...,\mathbf{d}(f,\theta_N)]$ is the matrix with the array steering vectors which model the array response to a plane wave incident at angle θ_n. For the case of a circular array of radius R the propagation model can be written as $d_m(f,\theta) = e^{jkR\cos(\phi_m-\theta)}$ [9], where ϕ_m denotes the angle of the mth sensor with respect to the sensor array center and k is the wavenumber. Assuming that $L < N$, the linear problem of (3) is overdetermined and the solution can be found by minimizing, in the Least Squares (LSQ) sense, the cost function

$$J = \left\|\mathbf{G}_l - \mathbf{D}(f)^H\mathbf{w}_l(f)\right\|_2^2. \tag{4}$$

However, unconstrained minimization involves inversion of matrix $\mathbf{D}(f)\mathbf{D}(f)^H$ which is ill-conditioned at low frequencies as well as other distinct frequencies. An example of this ill-behaviour is shown in Fig. 2 considering a circular array of 8 uniformly distributed microphones with radius $R = 0.05$. Direct inversion of $\mathbf{D}(f)\mathbf{D}(f)^H$ might thus lead to severe amplification of noise at certain frequencies which is perceived as unwanted spectral colouration. In order to avoid such a problem, we propose to use Tikhonov regularization

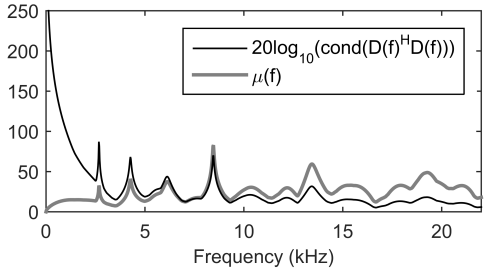

Fig. 2. Condition number of matrix $\mathbf{D}(f)^H\mathbf{D}(f)$ in dB (black) and variation of the proposed regularization parameter μ (gray) as a function of frequency.

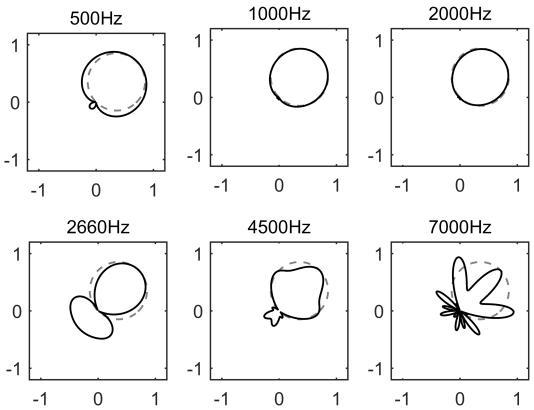

Fig. 3. Actual directivity patterns (solid-black) versus desired directivity patterns (dashed-gray) at different frequencies for an 8-element circular sensor array. The directivity patterns shown correspond to the first loudspeaker at 45 degrees, considering a symmetric arrangement of 4 loudspeakers at angles of 45, 135, -135 and -45 degrees.

by adding a penalty term proportional to the noise response in the previous cost function as

$$J = \left\|\mathbf{G}_l - \mathbf{D}(f)^H\mathbf{w}_l(f)\right\|_2^2 + \mu(f)\mathbf{w}_l(f)^H\mathbf{w}_l(f), \quad (5)$$

with $\mu(f)$ implying that the value of the regularization parameter varies with frequency. We have observed that this approach achieves a better trade-off between the noise gain and the array gain, as opposed to a constant value of the regularization parameter. In this paper, we propose a varying value of the regularization parameter of the form

$$\mu(f) = \lambda f 20\log_{10}(\text{cond}(\mathbf{D}(f)\mathbf{D}(f)^H)), \quad (6)$$

where λ is a fixed scalar and $\text{cond}(\cdot)$ represents the condition number of a matrix, e.g. the ratio of its largest eigenvalue to its smallest one. The beamformer weights can then be found through LSQ minimization as

$$\mathbf{w}_l^o(f) = (\mathbf{D}(f)\mathbf{D}(f)^H + \mu(f)\mathbf{I})^{-1}\mathbf{D}(f)\mathbf{G}_l, \quad (7)$$

where $\mathbf{I}$ is the $M \times M$ identity matrix.

Finally, we consider an additional normalization step, which aims to ensure unit gain and zero phase shift at the direction of maximum response for each beamformer. Letting θ_l^0 denote this direction for the lth beam at frequency f, the final weights are calculated as

$$\hat{\mathbf{w}}_l^o(f) = \frac{\mathbf{w}_l^o(f)}{\mathbf{w}_l^o(f)^H\mathbf{d}(f,\theta_l^0)}, \quad (8)$$

and the signal for the lth loudspeaker is obtained as in (2). The beamformer weights are calculated once for each frequency and stored to be used in the application phase. In Fig. 3 we present plots of the actual directivity versus the desired directivity pattern for the case of an 8-element sensor array considering four uniformly distributed loudspeakers on the azimuth plane. Observe the increment in the amplitude of the side-lobes at 2660 Hz which is close to a problematic frequency according to Fig. 2. Also, the subplot corresponding to 7 kHz is indicative of spatial aliasing problems which occur at higher frequencies.

IV. Parametric approach

As shown in Fig. 3 it is difficult to obtain exactly the desired directivity patterns relying on simple beamforming. Looking for example at the subfigure corresponding to 2660 Hz, we see that a sound source at -135 degrees would be also played-back by the loudspeaker at 45 degrees, something that may blur the sense of direction transmitted to the listener. On the other hand, a parametric approach avoids this problem by defining the loudspeaker response as a function of the estimated DOA.

However, it is questionable what type of processing is applicable to the particular type of array that we focus on this paper, an 8-sensor circular array of radius of 5 cm. This array has been used for capturing and reproduction of multichannel audio in [8], [9], but in a scenario with limited number of discrete sound sources. While the method works sufficiently well for applications such as a teleconference, it is inappropriate for the considered acoustic conditions due to the enormous amount of potential sound sources that participate in the sound scene. On the other hand, DiRAC does not pose a limitation on the number of sound sources comprising the sound scene, but the particular sensor array is impractical for such an approach. Even if we throw away information from 4 out of the 8 available sensors in order to approximate the 4-sensor planar array described in [3], the radius of this array implies that the maximum frequency free from spatial aliasing would be equal to 1715 Hz, far below the frequency limit that the particular array is designed for.

In this paper, we decided to consider an adaptation of the technique described by Cobos et al. in [7]. This approach estimates the DOA for each TF element, based on the phase differences between the microphones and a reference microphone of the array. Each time-frequency element of the signal from an arbitrary microphone is then filtered with the head-related transfer function (HRTF) according to its corresponding DOA estimate. Although originally designed for binaural reproduction, the approach can be straightforwardly adapted to loudspeaker reproduction using the mapping between the estimated DOA and loudspeaker gains provided by VBAP. How-

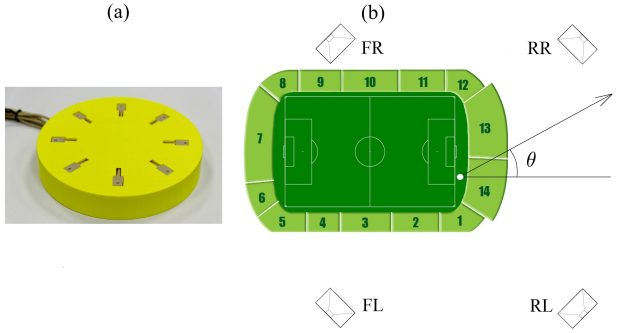

Fig. 4. Picture of the sensor array in (a) and sketch of the football stadium with respective loudspeaker setup used for evaluation in (b). The big white dot on the lower right corner of the football field denotes the array location.

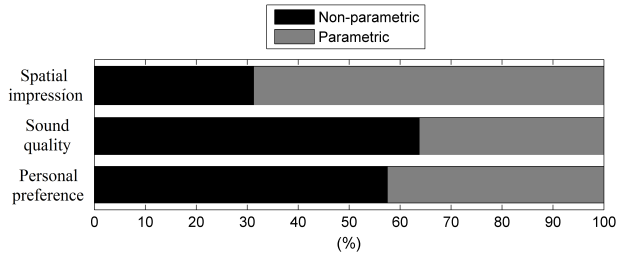

Fig. 5. Preference listening test results.

ever, instead of calculating the phase difference with respect to one reference microphone only, in our implementation we exploit all 28 pairwise microphone combinations. Certainly, we expect that localization of the TF points is correct only up to a maximum frequency due to spatial aliasing, expecting that the DOA estimates are distributed disorderly at frequencies above the aliasing limit. The authors in [7] however state that the variance in the DOA estimates encodes diffuseness information.

It is interesting to note some contradictions between the parametric and the non-parametric approach. As the parametric one does not explicitly consider a diffuse component, it activates at most two loudspeakers at each TF point, a fact that may significantly increase the sense of direction in comparison to the beamforming approach, but may significantly reduce the sense of envelopment. Also, the parametric method avoids the spectral colouration problem related to beamforming but it is prone to musical noise, especially when considering the high likelihood of spectrally overlapping sources in a typical crowded acoustic environment. These concepts illustrate that the two methods have different characteristics in terms of spatial impression and sound quality. The listening test results shown in the next section further support this argument.

V. Evaluation

The recording took place in a crowded open stadium during a football match of the Greek Super League. The acoustic environment was recorded at a sampling rate of 44100 Hz using the circular array of 8 sensors and radius of 5 cm shown in Fig. 4(a). Figure 4(b) presents a sketch of the football stadium with the location of the array represented by a big white dot. The array was placed at a height of 0.8 m. in front of Gates 13 and 14 which were populated with the organized fans of the team which was hosting the game. These fans were cheering and singing constantly throughout the entire duration of the recording, providing thus most of the acoustic information captured by the array.

The listening tests took place at the FORTH-ICS reference listening room which has been built following the ITU-R BS.1116 specifications. The reproduction system that we used for evaluation consisted of 4 loudspeakers uniformly distributed around the azimuth plane and specifically at 45, 135, -135 and -45 degrees (see Fig. 4), at a radius of 2.10 m. With respect to the listener's orientation, these loudspeakers were located Rear-Right (RR), Front-Right (FR), Front-Left (FL) and Rear-Left (RL). The considered configuration is of particular interest as it can be easily extended with a 5.1 surround system, adding one more channel to be used for the commentator. The panning gains derived from VBAP for this setup are identical to those depicted in Fig. 1(a).

The recorded signals were processed with both the parametric and the non-parametric technique, using the overlap-add method. For the STFT we used a Hanning window of 2048 samples length and hop size of 1024 samples (50% overlap). For the non-parametric method, we used the varying with frequency regularization parameter of Eq. (6) with $\lambda = 0.003$. The variation of μ as a function of the frequency is illustrated with the gray line in Fig. 2, while the polar plots of Fig. 3 are illustrative of the deviation between the desired and the actual beamformer directivities.

Preference listening tests were performed, asking 20 subjects to compare the two methods in terms of "spatial impression", "sound quality" and "personal preference". For this evaluation, we chose a segment of 2 min length starting right after a goal that the hosting team scored. This part was particularly interesting as it included cheering and applause from the crowd all around the stadium (the files used for the listening test are available online at http://users.ics.forth.gr/nstefana/Eusipco2016/). Instead of asking the subjects to listen to each mixture separately, we provided them the ability to switch in real time from one mix to the other by simply using the mouse to move the cursor from one colour-coded predefined area to the other on the screen. This allowed the differences between the methods to become audible instantaneously. The listeners were able to indicate preference towards a particular technique selecting the field "Better" or "Slightly better", or to indicate "No difference perceived". The results of the test are depicted in Fig. 5.

A map of the stadium like the one depicted in Fig. 4 was shown to the listeners, and we furthermore provided a description of the locations of the most prominent acoustic sources inside the stadium. Although there was no reference, the listeners were asked to grade spatial impression accord-

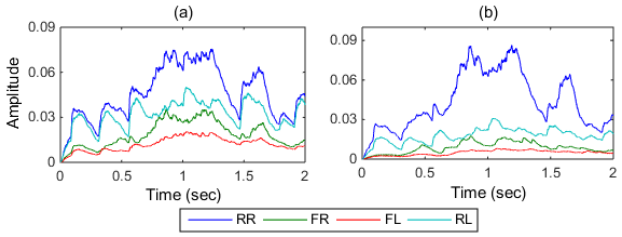

Fig. 6. Rectified loudspeaker signal amplitudes as a function of time for the non-parametric method in (a) and the parametric method in (b).

ing to how well the perceived directions agreed with their expectation. The results of the listening test indicate a clear superiority for the parametric method. The opinion of the authors, who were also present inside the stadium during the recording, is that the parametric method indeed produced a better sense of direction in comparison to beamforming, and was also more consistent with respect to changes in the orientation of the listeners head inside the listening area, as well as with reasonable displacements from the sweet spot. On the other hand, the non-parametric method provided a more blurred sense of direction but a much better sense of the reverberation in the Stadium. As additional evidence for this contradiction between the two methods, we have plotted in Fig. 6 the rectified loudspeakers' signal amplitudes in time, as derived by each technique, for a short duration segment where the crowd at Gates 13 and 14 was by far the most dominant acoustic source inside the stadium. As the acoustic energy is concentrated at a particular part of the scene, one should expect an uneven distribution of the signal energy across the different channels, which is what we actually observe for the parametric method in (b). On the other hand, for the non-parametric method we may observe an increased contribution from loudspeakers which are at irrelevant directions.

Despite the clear superiority of the parametric method with respect to source localization, the non-parametric method gathered slightly better score in terms of personal preference. Intuitively, one would assume that the listeners were disappointed by the sound quality of the parametric method and voted indifferently and slightly towards the non-parametric approach, but the truth is that the listeners showed clear preference to one or the other method in terms of personal preference, with 9 choosing the parametric method, 11 choosing the non-parametric one and no one choosing the "No preference" field. Subjects posed reasons related to spatial characteristics for supporting their choice, as for example that while a spreading of the acoustic scene was indeed perceived with the non-parametric method, this was not judged to be a disadvantage, because it resulted to a stronger sense of envelopment, or that they preferred the localization clarity of the parametric approach. To our opinion, this is an indication that the unique spatial character of each technique was indeed perceived and that whether envelopment or clarity of direction is more appreciated, is a matter of personal taste. As a general statement, it can perhaps be proposed that instead of only aiming at accurate reproduction of the spatial properties of a sound field, it would be probably meaningful to provide to the user a control parameter for him or her to select the balance between an increased sense of envelopment and an improved sense of direction, according to his personal preference.

VI. Conclusion

We have formulated a parametric and a non-parametric technique for sound scene spatialization in 2-D and we have used these techniques for rendering a recording produced in the crowded environment of a football stadium. The non-parametric method provided an advantage in terms of sound quality and although it resulted to a spreading of the sound sources locations, this was not necessarily perceived as a disadvantage by the listeners. On the other hand, the parametric method performed more accurate sound source localization by slightly impairing the sound quality. As future work, it would perhaps be interesting to find ways to combine the unique spatial characteristics of the two methods.

References

[1] G. Cengarle, T. Mateos, N. Olaiz, and P. Arumi, "A new technology for the assisted mixing of sport events: Application to live football broadcasting," in *Proc. of 128th Convention of Audio Eng. Soc.*, 2010.

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

[3] F. Kuech, M. Kallinger, R. Schultz-Amling, G. Der Galdo, and V. Pulkki, "Directional audio coding using planar microphone arrays," in *Proc. of Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 37–40.

[4] M. Kallinger, F. Kuech, R. Schultz-Amling, G. Del Galdo, J. Ahonen, and V. Pulkki, "Analysis and adjustment of planar microphone arrays for application in directional audio coding," in *Proc. of 124th Convention of Audio Eng. Soc.*, 2008, paper 7374.

[5] A. Politis, M. Laitinen, J. Ahonen, and V. Pulkki, "Parametric spatial audio processing of spaced microphone array recordings for multichannel reproduction," *J. Audio Eng. Soc*, vol. 63, no. 4, pp. 216–227, 2015.

[6] O. Thiergart, M. Kallinger, G. Del Galdo, and F. Kuech, "Parametric spatial sound processing using linear microphone arrays," in *Microelectronic Systems*. Springer, 2011, pp. 321–329.

[7] M. Cobos, J. Lopez, and S. Spors, "A sparsity-based approach to 3d binaural sound synthesis using time-frequency array processing," *EURASIP Journal on Advances in Signal Processing*, 2010.

[8] A. Alexandridis, A. Griffin, and A. Mouchtaris, "Directional coding of audio using a circular microphone array," in *Proc. of ICASSP*, 2013, pp. 296–300.

[9] ——, "Capturing and reproducing spatial audio based on a circular microphone array," *Journal of Electrical and Computer Engineering*, vol. 2013, 2013, article ID 718574.

[10] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, 2013.

[11] M. Gerzon, "Ambisonics in multichannel broadcasting and video," *J. Audio Eng. Soc*, vol. 33, no. 11, pp. 859–871, 1985.

[12] M. Boone, E. Verheijen, and P. van Tol, "Spatial sound-field reproduction by wave-field synthesis," *J. Audio Eng. Soc*, vol. 43, no. 12, pp. 1003–1012, 1995.

[13] H. Hachabiboğlu and Z. Cvetković, "Panoramic recording and reproduction of mutlichannel audio using a circular microphonoe array," in *Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 117–120.

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