On the Use of Small Microphone Arrays for Wave Field Synthesis Auralization

Maximo Cobos[1], Sascha Spors[2], Jens Ahrens[2] and Jose J. Lopez[3]

[1] *Departamento de Informática, University of Valencia, Burjassot, 46100, Valencia, Spain*

[2] *Deutsche Telekom Laboratories, TU Berlin, D-10587, Berlin, Germany*

[3] *Institute of Telecommunications and Multimedia Applications, Universitat Politécnica de València, 46022, Valencia, Spain*

Correspondence should be addressed to Maximo Cobos (Maximo.Cobos@uv.es)

ABSTRACT

The synthesis of a captured sound field with preservation of its perceptual properties is called auralization. Data-based Wave Field Synthesis (WFS) auralization makes use of a set of measured impulse responses along an array of microphone positions. However, a considerable array size must be employed for having an appropriate angular resolution. In this paper, we explore the possibilities of time-frequency analysis for WFS auralization using the signals acquired from an array of closely-spaced microphones. The aim is to reproduce in WFS a sound scene captured by a small microphone array, while preserving the spatial impression of the original sound. To this end, time-frequency Direction-Of-Arrival (DOA) estimates are used to obtain a set of directional audio components extracted from the recorded microphone signals. These components are later synthesized as spatially distributed plane waves using WFS. Additionally, high-power components are synthesized as point sources to improve localization stability for multiple listeners or when listeners move inside the listening area.

1. INTRODUCTION

Wave Field Synthesis (WFS) is a sound field synthesis technique that is able to create a virtual auditory scene over a large listening area using an array of loudspeakers [1],[2]. The main motivation behind WFS is its capability to overcome some of the classical limitations of stereophonic reproduction techniques. These limitations are usually related to the existence of a preferred listening position, i.e. the sweet-spot [3]. Usually, WFS is applied on dry sources recorded with close-by spot microphones, thus, the corresponding signal processing is just aimed at simulating the true spatial location of each source. Moreover, since WFS is an object-oriented reproduction approach, the sound sources can be positioned in any configuration to produce a desired acoustic atmosphere [4].

Auralization is the process of making the acoustics of a room audible in a different space [5]. Usually, this is done in WFS by reproducing measured impulse responses in the original room, preferably convolved with the anechoic (or almost anechoic) source signals [6]. The data-based approach enables to virtually compose a sound scene with the acoustic properties encoded by the measured impulse responses and allows the listeners to experience the simulated acoustics consistently. The sound field is decomposed from the impulse responses into plane sound waves coming from different directions of incidence. In reproduction, direct sound and dominant reflections are treated separately from the diffuse sound components [5]. The direct sound and first reflections are reproduced as individual point sources, while the diffuse part of the plane-wave-decomposed impulse responses are reproduced through a limited set of plane waves. However, the measurement of impulse responses must be carried out along an array of microphone positions, which must have a considerable size (of the order of 2 meters) for having an appropriate angular resolution [6].

Recently, the authors presented a method to capture and process the spatial characteristics of sound with the aim of providing a real-time 3D audio experience [7]. Instead of using an expensive dummy head setup, a small tetrahedral microphone array was utilized to discriminate among the three spatial dimensions, providing an effective way of constructing a full 3D audio system.

In this work, we present a study of the possibilities offered by small microphone arrays for capturing the spatial information of a sound scene to be reproduced in WFS. Based on the same analysis stage of [7], several alternatives are explored to synthesize the recorded sound over a large listening area. As with *Directional Audio Coding* (DirAC) [8], the presented method is intended to provide enhanced spatial audio features by means of processing in time and frequency the signals recorded by a small microphone array. However, to our knowledge, the direct application of DirAC to WFS has not been covered so far.

As opposed to traditional WFS auralization techniques, the proposed recording and processing scheme is not based on impulse response measurement, but on the directional analysis of the impinging sound in the time-frequency domain. This processing makes the proposed approach a time-variant system capable of managing non-stationary sources. Nevertheless, the reproduction step is based on similar principles, since point sources and plane-wave components are also used to provide the listeners with a perceptually similar recreation of the original recorded scene.

The paper is structured as follows. Section 2 explains the processing used in the analysis stage to perform DOA estimation in the time-frequency domain with a tetrahedral microphone array. Section 3 presents the proposed reproduction approach in WFS. Section 4 discusses several experiments performed in a real WFS system using simulated and real recordings. Finally, the conclusions of this work are summarized in Section 5.

2. TIME-FREQUENCY DOA ANALYSIS

In this section, we describe the geometry of the small tetrahedral array and its associated time-frequency processing for DOA estimation. This analysis, which can be also found in [7], is here discussed with the aim of providing the reader with a reasonably self-contained work.

2.1. Signal Model

The signals recorded by a microphone array, with sensors denoted with indices $m = 1,2,\ldots,M$ in an acoustic environment where N sound sources are present, can be modeled as a finite impulse response (FIR) convolutive mixture, written as

$$x_m(t) = \sum_{n=1}^{N} \sum_{\ell=0}^{L_m-1} h_{mn}(\ell) s_n(t-\ell), \quad m = 1,\ldots,M \quad (1)$$

where $x_m(t)$ is the signal recorded at the m-th microphone at time sample t, $s_n(t)$ is the n-th source signal, $h_{mn}(t)$ is the impulse response of the acoustic path from source n to sensor m, and L_m is the maximum length of all impulse responses.

The model in (1) can also be expressed in the *short-time Fourier transform* (STFT) domain as follows

$$X_m(k,r) = \sum_{n=1}^{N} H_{mn}(k) S(k,r), \quad (2)$$

where $X_m(k,r)$ denotes the STFT of the m-th microphone signal, being k and r the frequency index and time frame index, respectively. $S_n(k,r)$ denotes the STFT of the source signal $s_n(t)$ and $H_{mn}(k)$ is the frequency response from source n to sensor m.

2.1.1. Sparse Sources

In the time-frequency domain, source signals are usually assumed to be sparse. A sparse source has a peaky probability density function: the signal is close to zero at most time-frequency points, and has large values in rare occasions. This property has been widely applied in many works related to source signal localization [9][10] and separation [11][12] in underdetermined situations, i.e. when there are more sources than microphone signals.

If we assume that the sources rarely overlap at each time-frequency point, Equation (2) can be simplified as follows

$$X_m(k,r) \approx H_{ma}(k) S_a(k,r), \quad (3)$$

where $S_a(k,r)$ is the dominant source at time-frequency point (k,r). To simplify, we assume an anechoic model where the sources are sufficiently distant to consider plane wavefront incidence. Then, the frequency response is only a function of the time-delay τ_{mn} between each source and sensor

$$H_{mn}(k) = e^{j2\pi f_k \tau_{mn}}, \quad (4)$$

being f_k the frequency corresponding to frequency index k.

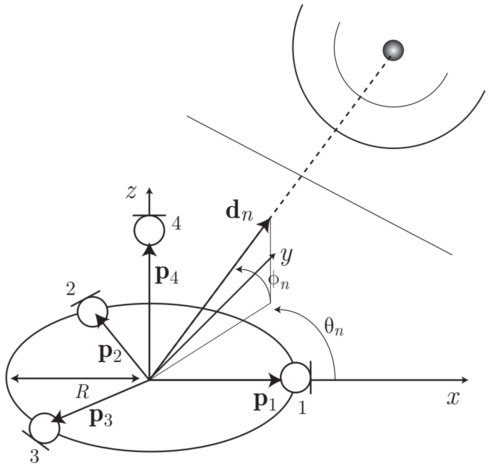


Fig. 1: Tetrahedral microphone array for 3-D DOA estimation.

2.2. Array Geometry and DOA Estimation

Now consider a tetrahedral microphone array $(M = 4)$ with base radius R, as shown in Figure 3. The sensor location vectors in the 3-dimensional space with origin in the array base center, are given by:

$$\begin{aligned} \mathbf{p}_1 &= [R,\ 0,\ 0]^T, \\ \mathbf{p}_2 &= \left[-\frac{R}{2},\ \frac{\sqrt{3}}{2}R,\ 0\right]^T, \\ \mathbf{p}_3 &= \left[-\frac{R}{2},\ -\frac{\sqrt{3}}{2}R,\ 0\right]^T, \\ \mathbf{p}_4 &= \left[0,\ 0,\ R\sqrt{2}\right]^T. \end{aligned} \quad (5)\ (6)$$

The DOA vector of the n-th source as a function of the azimuth θ_n and elevation ϕ_n angles is defined as

$$\mathbf{d}_n = [\cos\theta_n \cos\phi_n,\ \sin\theta_n \cos\phi_n,\ \sin\phi_n]^T. \quad (7)$$

The source to sensor time delay is given by $\tau_{mn} = \mathbf{p}_m^T \mathbf{d}_n / c$, being c the speed of sound. Therefore, the frequency response of Equation (4) can be written as

$$H_{mn}(k,r) \approx e^{j\frac{2\pi f_k}{c}\mathbf{p}_m^T \mathbf{d}_n}. \quad (8)$$

Taking into account this last result and Equation (3), it becomes clear that the phase difference between the microphone pair formed by sensors i and j, is given by

$$\angle\left(\frac{X_j(k,r)}{X_i(k,r)}\right) \approx \frac{2\pi f_k}{c}(\mathbf{p}_j - \mathbf{p}_i)^T \mathbf{d}_n, \quad (9)$$

where $\angle$ denotes the phase of a complex number.

Using a reference microphone q, the phase differences at point (k,r) of $M-1$ microphone pairs $(m,q) \neq (q,q)$ are stored in the vector

$$\mathbf{b}_q(k,r) = \left[\angle\left(\frac{X_1(k,r)}{X_q(k,r)}\right), \ldots, \angle\left(\frac{X_M(k,r)}{X_q(k,r)}\right)\right]^T, \quad (10)$$

forming the following system of equations:

$$\mathbf{b}_q(k,r) = \frac{2\pi f_k}{c}\mathbf{P}\mathbf{d}_n, \quad (11)$$

where

$$\mathbf{P} = \left[\mathbf{p}_{1q}, \ldots, \mathbf{p}_{Mq}\right]^T, \quad \mathbf{p}_{nq} = \mathbf{p}_n - \mathbf{p}_q. \quad (12)$$

Finally, the DOA at time-frequency bin (k,r) is obtained by taking the inverse of the $\mathbf{P}$ matrix

$$\hat{\mathbf{d}}_n(k,r) = \frac{c}{2\pi f_k}\mathbf{P}^{-1}\mathbf{b}_q(k,r). \quad (13)$$

The regular tetrahedral geometry used in this paper leads to the following simple equations for $\mathbf{d}_n(k,r) = [\hat{d}_1,\ \hat{d}_2,\ \hat{d}_3]^T$:

$$\begin{aligned} \hat{d}_1 &= \cos\theta_n \cos\phi_n = \frac{c}{2\pi f_k}\frac{1}{\sqrt{3}}(b_2 + b_3), \quad (14) \\ \hat{d}_2 &= \sin\theta_n \cos\phi_n = \frac{c}{2\pi f_k}(b_3 - b_2), \quad (15) \\ \hat{d}_3 &= \sin\phi_n = \frac{c}{2\pi f_k}\left[\frac{1}{\sqrt{6}}(b_2 + b_3) - \sqrt{\frac{3}{2}}b_4\right], \quad (16) \end{aligned}$$

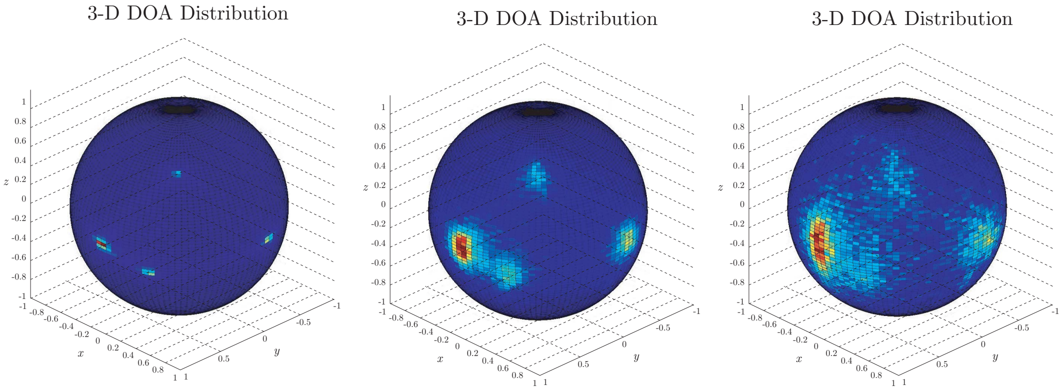
where b_n is the n-th element of the vector $\mathbf{b}_1(k,r)$ (reference microphone $q = 1$). The azimuth angle is obtained using the four quadrant inverse tangent function:

$$\hat{\theta}_n(k,r) = \mathrm{atan}^{360^\circ}(\hat{d}_1, \hat{d}_2). \quad (17)$$

The elevation angle is directly obtained as

$$\hat{\phi}_n(k,r) = \sin^{-1}(\hat{d}_3). \quad (18)$$

Note that for each time-frequency point (k,r), estimating the 3-D DOA is relatively simple, just using the observed



(a) (b) (c)

Fig. 2: Histograms showing the distribution of DOA estimates in the 3-D space calculated from a mixture of 4 speech sources. (a) Anechoic conditions. (b) $T_{60} = 150$ ms. (c) $T_{60} = 300$ ms.

phase differences between 3 microphone pairs of the array. Another aspect to consider is spatial aliasing. The distance between microphones determines the angular aliasing frequency. Due to the 2π ambiguity in the calculation of the phase differences, the maximum ambiguity-free frequency in a microphone pair sub-array would be given by $f_k = c/2d$, where d is the separation distance between the capsules. Beyond this frequency, there is not a one-to-one relationship between phase difference and spatial direction. However, small arrays with $d \approx 1.5$ cm provide an unambiguous bandwidth greater than 11 kHz, covering a perceptually important frequency range.

Figure 2 shows the 3-D histograms that represent the amount of estimates produced in a given direction for a recording of four speech sources. Note how in the anechoic case (a), the sources appear as localized peaky zones corresponding to their real DOAs. The diffuseness added by room reflections can be clearly seen in (b)-(c), where the estimates, although clustered around the real DOA directions, have been highly spread.

3. PROPOSED WFS REPRODUCTION

The proposed WFS reproduction scheme is based on the traditional approach of using plane waves and point sources to reproduce separately direct sound and room reflections [5], as depicted in Figure 3. In the approach here presented, instead of using measured impulse responses to separate direct and diffuse components, a spatial filtering procedure is performed to divide the recorded sound into different directional signals. These signals are extracted as a function of the estimated DOAs in the time-frequency domain and are later synthesized as plane waves in WFS. Directions where direct sound is dominant result in high-power components that are synthesized as point sources to improve source localization over a larger area.

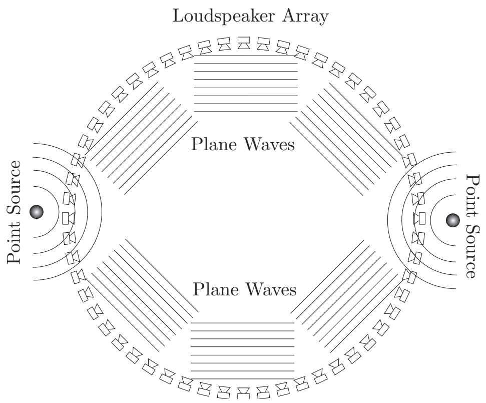


Fig. 3: WFS reproduction by means of combined plane waves and point sources.

Next, we describe how to extract the directional signals used to reproduce the scene in WFS from the DOA information obtained in the previous analysis stage.

3.1. Plane-Wave Components

The analysis stage described in Section 2 is based on

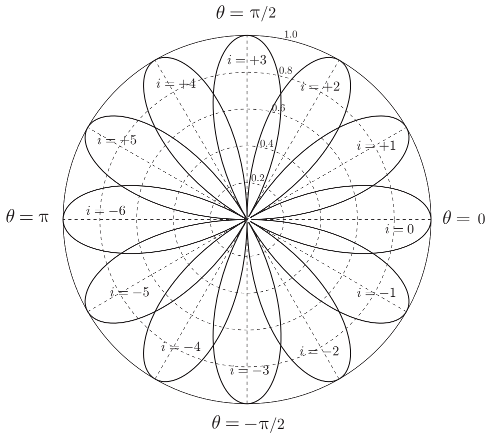


Fig. 4: Spatial filters for the extraction of plane-wave components.

the fundamental assumption that the arriving wave fronts correspond to plane waves originated by sources located in the far-field. As a result, the idea of synthesizing again the recorded sound as plane waves with WFS emerges naturally from this assumption. However, whereas any estimated direction could result from the analysis stage, in reproduction, the number of plane-wave components is usually limited due to computational constraints. This issue forces to group time-frequency bins having similar DOA into single plane wave components by following a spatial filtering procedure. The different plane-wave components are extracted from a set of directional patterns that are functions of the estimated DOA vector. Moreover, since most WFS systems are designed to work only in the horizontal plane, only the estimated azimuth $\hat{\theta}(k,r)$ is here considered. In any case, the generalization to 3-D is straightforward [2].

Given a number N_{pw} of reproduced plane waves, the different spatial filters used to extract the azimuthal components are defined as follows. To facilitate notation, the variable $\theta_i' = \theta - i\frac{2\pi}{N_{pw}}$ is introduced. Hence, the filters are given by

$$C_i(\theta) = \begin{cases} \left|\cos\left(\frac{N_{pw}\theta_i'}{4}\right)\right|, & \left||\theta_i'| - \pi\right| \geq (\pi - \frac{2\pi}{N}) \\ 0, & \text{elsewhere} \end{cases}, \quad (19)$$

where $i = -\frac{N_{pw}}{2} \ldots (\frac{N_{pw}}{2} - 1)$.

Figure 4 shows an example of the resulting spatial filters for $N_{pw} = 12$. Note that these filters have been designed to be uniformly distributed in azimuth, being centered at angles $\theta_{ci} = i\frac{2\pi}{N_{pw}}$. Moreover, a cosine shape has been chosen to preserve the total power among the extracted components:

$$\sum_{i=-\frac{N_{pw}}{2}}^{\frac{N_{pw}}{2}-1} C_i^2(\theta) = 1. \quad (20)$$

The components are simply obtained by weighting the center microphone signal according to the estimated directions in the time-frequency domain:

$$Y_i(k,r) = X_4(k,r)C_i(\hat{\theta}(k,r)). \quad (21)$$

The filtered signals, $Y_i(k,r)$ are transformed back to the time domain by using the inverse STFT operator. In reproduction, the i-th plane wave must be configured to arrive from the corresponding direction θ_{ci}.

Notice that the data-based WFS auralization approach in [5] can only be applied in the case of stationary scenarios where source locations do not change with time. In the proposed method, the directional components are extracted in real-time on a frame-by-frame basis, thus, the reproduction stage has the capability to follow source movements in real-time. The number of plane waves selected is usually limited by the system computational capabilities (both in the analysis and synthesis stage). Some works have suggested that 8 plane waves are sufficient to provide a convincing listening experience [6]. In [13], it was shown that 10 plane waves are sufficient to evoke a perceptually diffuse sound field for a static listener. Although increasing the number of plane waves could result in better localization, further work is needed to examine the perceptual benefits of introducing a large amount of waves regarding different spatial sound attributes. Moreover, a high number of plane waves results in very narrow spatial filters that lead to artifacts in the extracted signals, reducing the overall sound quality of the sound scene.

3.2. Point Sources

In most situations, the extracted plane-wave components corresponding to actual source directions have significantly more power than the rest. These high-power components can be assumed to be dominated by direct sound and, therefore, it seems reasonable to synthesize them as point sources instead of using plane waves. Point sources support stable localization for large auditoria and when listeners move within the listening area.

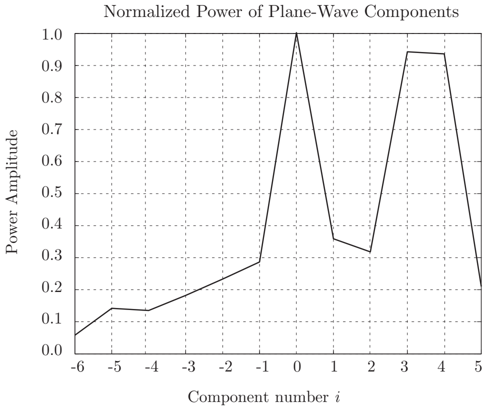


Fig. 5: Normalized power for each plane-wave component in an example mixture with 2 speech sources and $N_{pw} = 12$.

The power of each component is computed as

$$P_i = \sum_{k,r} |Y_i(k,r)|^2. \tag{22}$$

To illustrate this fact, Figure 5 shows the normalized power of an example recording of two simultaneous speech sources in a room with reverberation time $T_{60} = 0.2$ s. The real source directions are 0° and 105°. Note that there is a prominent peak in the component $i = 0$ corresponding to 0°, while two high components $(i = 3,4)$ share significant energy, since the source is located between their centers (90° and 120°).

As a first approach, a threshold can be defined to select the signals to be synthesized as point sources, where neighboring signals having power above the threshold are added up to form a single source signal. This would be the case of the example in Figure 5, where the component $(i = 0)$ would be selected as a point source and the components $(i = 3,4)$ would be added and synthesized as another point source. However, depending on the acoustic environment and its diffuseness, peaks corresponding to sources could be hardly identifiable, what would make more reasonable to carry out a plane-wave only synthesis.

4. EXPERIMENTS

A set of experiments using both simulated and real recordings were conducted to evaluate informally some of the reproduction aspects above considered. First, 10 s long mixtures with different number of speech sources were artificially generated by means of the source-image method [14]. Recordings using a real microphone array were performed in a room with reverberation time $T_{60} = 0.2$ s, using four instrumentation quality microphones from Bruël & Kjaer (model 4958) with inter-sensor distance $d = 1.5$ cm. The sampling frequency used was 16 kHz, although the final signals were resampled to 44.1 kHz before reproduction for the system convenience. Time-frequency processing was performed using Hann windows of 1024 samples of length, with an overlap factor of $2/3$. The loudspeaker array installed at the Usability Laboratory at Deutsche Telekom Laboratories was used to test the proposed reproduction method. It is composed of 56 equiangularly spaced loudspeakers on a circle with a nominal radius of 1.495 m. The reproduced scenes were not formally evaluated due to the preliminary character of the test. However, some interesting observation remarks are here presented, which are quite representative of the system performance.

4.1. Plane-wave Only Reproduction

Simulated and real recordings with one and two speech sources were processed and reproduced in WFS using only plane waves. The signals corresponding to each plane wave were extracted as explained in Section 3.1. In the simulated recordings, two different reverberation times were considered ($T_{60} = 0$ s and $T_{60} = 0.5$ s) to analyze the effects of room reflections both in the analysis and synthesis stage. In the experiments using simulated recordings, the source directions were 45° and 135° in the simulated environment. In the experiments with real recordings the sources were located at 0° and 105°. Reproduction was first performed using only 8 plane waves and then using 24 plane waves to investigate the dependance on the chosen number of plane waves.

According to the authors' judgement, the overall sound quality was very good, as well as the spatial impression. Besides the fact that real recordings tend to be a bit noisier than simulations, there were no significant differences between the simulated and real scenes. However, a relevant observation was that the location of the sources becomes slightly confusing when the listener moves around the listening area. The effect was less noticeable when 24 plane waves were used in reproduction, although a higher amount of artifacts were present due to the excessively narrow beampatterns used to extract the plane-wave signals, specially in the case of higher reverberation time.

It is also worth to mention that the system we used was quite small, thus, localization issues could become more critical when using larger systems.

4.2. Plane-wave + Point Source Reproduction

In order to improve localization stability within the listening area, the experiment above was repeated by using point-sources to reproduce in WFS high-power directional components. The directional components to be reproduced as point sources were those having a normalized power above the mean. The direction of the selected high-power components were in accordance with the actual source locations.

The overall sound quality in this experiment was very similar to the one using only plane waves, however, localization stability was significantly improved when adding point sources in the reproduction set-up. This fact reveals the usefulness of using point-sources and encourages to develop further research to improve the selection of point-source signals in complex environments.

4.3. Point Source Only Reproduction

Despite the fact that plane-wave propagation can be assumed in the analysis stage of the method, this assumption can be hardly taken in reproduction when the listener moves around the listening area. In fact, the perspective of the sound scene perceived by the listener changes with respect to the one analyzed by the microphone array. As a result, plane-wave only reproduction is not justified anymore in this case.

A possible solution could be to reproduce all the directional components as point sources arranged at a similar distance, ensuring that the components are synchronized. In this case, the radius of the arrangement could be selected according to the size of the recorded scene and the size of the reproduction system. This reproduction option will be further explored in future experiments.

5. CONCLUSION

In this paper, we presented a method to reproduce in WFS complex sound scenes captured my means of a small tetrahedral microphone array. The method allows to reproduce the recorded sound scene over a large listening area by means of WFS, while preserving the spatial impression of the original sound. The proposed WFS reproduction scheme is based on traditional WFS auralization approaches that use plane waves and point sources to reproduce separately direct sound and room reflections. However, instead of using measured impulse responses to separate direct and diffuse components, a spatial filtering procedure using time-frequency DOA information is performed to divide the recorded sound into different directional signals. This approach makes the system time-variant, having the capability to manage changes in the source locations in real-time. High-power components are assumed to be dominated by direct sound and are reproduced using point sources, while the rest of directional components are reproduced as plane waves that are equiangularly distributed. Although preliminary experiments are very encouraging, further work is needed to better understand the limitations of the method and to improve the performance of the system. To this end, further reproduction options and formal listening experiments will be conducted in the near future.

6. REFERENCES

[1] A. J. Berkhout. A holographic approach to acoustic control. *Journal of the Audio Engineering Society*, 36:977-995, December 1988.

[2] S. Spors, R. Rabenstein and J. Ahrens, “The theory of Wave Field Synthesis revisited,” presented at the AES 124th convention, Amsterdam, The Netherlands, 2008 May 17–20.

[3] H. Wittek. *Perceptual differences between Wave Field Synthesis and stereophony*. PhD thesis, University of Surrey, 2007.

[4] A. Wagnet, A. Walther, F. Melchior and M. Strauss, “Generation of highly immsersive atmospheres for Wave Field Synthesis reproduction,” presented at the AES 116th Convention, Berlin, Germany, 2004.

[5] E. Huselbos. *Auralization using Wave Field Synthesis*. PhD thesis, Delft University of Technology, 2004.

[6] D. de Vries and E. Huselbos, “Auralization of room acoustics by Wave Field Synthesis based on array measurements of impulse responses,” presented at the 12th European Signal Processing Conference (EUSIPCO 2004), Vienna, Austria, September 2004.

[7] M. Cobos, J. J. Lopez and S. Spors. A sparsity-based approach to 3-D binaural sound synthesis

using time-frequency array processing. *EURASIP Journal on Advances in Signal Processing*, Vol. 2010, Article ID 415840, 2010.

[8] V. Pulkki. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, vol. 55, no. 6, 503–516, 2007.

[9] S. Araki, H. Sawada, R. Mukai and S. Makino. DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors. *Journal of Signal Processing Systems*, 2009.

[10] S. Rickard and F. Dietrich,"DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET," presented at the 10th IEEE Workshop on Statistical Signal and Array Processing (SSAP2000), Pocono Manor, PA, August 2000.

[11] O. Yilmaz and S. Rickard. Blind estimation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, vol. 52, no. 7, 1830–1847, 2004.

[12] P. Bofill and M. Zibulevski. Underdetermined blind source separation using sparse representations. *Signal Processing*, vol. 81, 2353–2362, 2001.

[13] J. J. Sonke. *Variable acoustics by wave field synthesis*. PhD thesis, Delft University of Technology, 2000.

[14] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, vol.65, no. 4, 943–950, 1979.