Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2013, Article ID 718574, 16 pages
http://dx.doi.org/10.1155/2013/718574

Research Article

Capturing and Reproducing Spatial Audio Based on a Circular Microphone Array

Anastasios Alexandridis,[1,2] Anthony Griffin,[1] and Athanasios Mouchtaris[1,2]

[1] *Foundation for Research and Technology-Hellas, Institute of Computer Science (FORTH-ICS), 70013 Heraklion, Crete, Greece*
[2] *Department of Computer Science, University of Crete, 71409 Heraklion, Crete, Greece*

Correspondence should be addressed to Anastasios Alexandridis; analexan@ics.forth.gr

Received 11 October 2012; Revised 10 February 2013; Accepted 19 February 2013

Academic Editor: Moo Young Kim

This paper proposes a real-time method for capturing and reproducing spatial audio based on a circular microphone array. Following a different approach than other recently proposed array-based methods for spatial audio, the proposed method estimates the directions of arrival of the active sound sources on a per time-frame basis and performs source separation with a fixed superdirective beamformer, which results in more accurate modelling and reproduction of the recorded acoustic environment. The separated source signals are downmixed into one monophonic audio signal, which, along with side information, is transmitted to the reproduction side. Reproduction is possible using either headphones or an arbitrary loudspeaker configuration. The method is compared with other recently proposed array-based spatial audio methods through a series of listening tests for both simulated and real microphone array recordings. Reproduction using both loudspeakers and headphones is considered in the listening tests. As the results indicate, the proposed method achieves excellent spatialization and sound quality.

1. Introduction

Spatial audio refers to the multichannel or stereophonic sound reproduction by preserving the spatial information of an acoustic environment. Multiple loudspeakers or headphones are employed to enable the listener to perceive the direction of each sound source, preserving the original sound scene. In the last few years, methods to extract, transmit, and reproduce the spatial characteristics of a sound field have attracted great attention from audio researchers for several reasons. They allow the development of entertainment systems that enable listeners to listen to multichannel music, record and reproduce a concert together with the spatial characteristics of the orchestra, or watch movies that feature surround sound. Gaming can also benefit from multichannel audio by providing a more realistic sensation of the game environment and a more immersive gaming experience. Moreover, teleconferencing applications can use spatial audio to create an immersive and more natural way of communication between two or more parties.

Humans utilize a complicated hearing system, which enables us to perceive the direction of each sound source in a sound field. Human perception of spatial information has been found to be associated with the inter-aural time differences (ITDs) and inter-aural level differences (ILDs), while the source width and diffuseness properties of the sound field are linked to the inter-aural coherence (IC) [1].

In this paper, we propose a real-time array-based method, for capturing and reproducing the directional information of a sound field, based on microphone arrays and beamforming. The acoustic scene is recorded using a uniform circular microphone array. Directional information is extracted through direction-of-arrival (DOA) estimation and spatial sound is delivered to the reproduction side using one signal and side information. We consider microphone arrays—particularly circular arrays—for spatial audio as they are often used in several applications, such as teleconferencing and providing noise-robust speech capture.

Over the years, many different spatial audio methods have been proposed. Wave field synthesis (WFS) [2] is a powerful technique, capable of reconstructing the entire recorded sound field with high accuracy. However, the required number of loudspeakers makes it unsuitable for

many practical deployments. Recent work in [3] combines microphone arrays and WFS, but, again, the number of required loudspeakers for accurate reproduction is high.

In [4], high-order differential microphone arrays are employed to design directivity patterns that emulate stereophonic panning laws. However, the loudspeaker configuration depends on the microphone configuration, strictly linking the reproduction and recording methods. Moreover, the required number of microphones is again high. The use of blind source separation for creating binaural spatial audio was investigated in [5]. The authors use signal subspace analysis to estimate the number of sources and their DOAs, and separation is achieved using Frequency-domain independent component analysis (FD-ICA). However, FD-ICA carries a significant computational burden, making the system impractical for real-time applications.

Directional audio coding (DirAC) [6] is a system for recording and reproducing spatial audio, based on B-format signals. ITDs and ILDs are extracted from the estimated DOAs, while IC is extracted from estimation of sound diffuseness. DOA and diffuseness estimation can be carried out in individual time-frequency (TF) elements or frequency subbands. Recently, versions of DirAC based on microphone array signals have been proposed [7, 8]. In [7] differential microphone array techniques are employed to convert the signals recorded from a planar microphone array to B-format. DirAC is applied to the estimated B-format signals. However—as illustrated in [9]—there is a bias in the B-format approximation by the differential array, that leads to biased DOA and diffuseness estimates which can reduce the overall quality of reproduction. An alternative approach is discussed in [8]. The authors use a linear array and propose a modified real-time version of the ESPRIT algorithm—especially designed for linear arrays—to estimate the DOA in each TF element. The estimation of diffuseness is based on the magnitude squared coherence (MSC) between the two outer microphones of the array.

Another approach to spatial audio reproduction using time-frequency array processing is the binaural one presented in [10]. This approach also estimates the DOA for each time-frequency 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.

The individual DOA estimation procedure in each TF element of the aforementioned methods is susceptible to many practical and theoretical considerations. Spatial aliasing occurring in microphone arrays—due to the discrete sampling in space—makes it very challenging to accurately estimate the DOAs across the whole spectrum of frequencies. As a result, if significant content of the signal lies above the spatial-aliasing cutoff frequency—determined by the array geometry—then errors in the spatial impression of sources will become noticeable.

The assumption of signal sparsity and disjointedness, inherent in the methods described previously, plays a crucial role in the overall quality of a spatial audio system. A basic assumption in estimating a different DOA for each time-frequency element is that in each such element there is only one dominant source. As a result, these methods require strong W-disjoint orthogonality (WDO) conditions [11] (which is a measure of source disjointedness); otherwise source localization will result in significant errors. When there are many sound sources active or when the source signals overlap significantly in the frequency domain, the WDO hypothesis is weakened and the variance of the DOA estimates—even in adjacent frequencies—will dramatically degrade the spatial impression. This may also result in quality degradation (especially in binaural reproduction) since the HRTF filters will change rapidly over time and frequency, creating distortions, such as musical or metallic noise.

We propose an approach—which combines conventional methods for estimating the DOAs and beamforming—that overcomes some of the problems mentioned previously. It is based on a per time-frame DOA estimation and source separation through spatial filtering. For the DOA estimation procedure, we employ our recently proposed algorithm of [12–14]. This method considers only the spatial aliasing-free part of the spectrum to estimate the DOAs, so spatial aliasing does not affect our estimates. Spatial aliasing may affect the beamformer performance, degrading source separation. However, as our experimental results indicate, such degradation in source separation is unnoticeable to listeners. Moreover, since we do not estimate a different DOA for each TF element, our method does not suffer from erroneous estimates that occurred due to the weakened WDO hypothesis when multiple sound sources are active. Our listening test results show that this approach to modelling the acoustic environment is more effective than other array-based approaches that have been recently proposed.

Moreover, based on a novel downmixing process, the separated source signals—and thus the entire sound field—are encoded into one monophonic signal and side information. During downmixing our method assumes WDO conditions, but as our listening test results indicate, in this stage the WDO assumption does not affect the spatial impression and quality of the reconstructed sound field. The reason is that, compared to other methods, we do not rely on WDO conditions to extract the directional information of the sound field, but only to downmix the resulting separated source signals.

Another important issue is that source separation through spatial filtering results in musical noise in the filtered signals, a problem which is evident in almost all blind source separation methods. However, since our goal is to recreate spatial audio, the separated signals are rendered simultaneously from different directions, which eliminates the musical distortion. This is an important result of our work supported by listening tests.

The rest of this paper is organized as follows: Section 2 explains the signal model used in this work. The recording, analysis, and reproduction sides of the proposed method are discussed in Section 3. Section 4 presents the listening test methodology, and experimental results of simulated and real microphone array recordings for both loudspeaker and

binaural reproduction are presented in Sections 5 and 6. Finally, Section 7 concludes the work and discusses future plans.

2. Signal Model

For an acoustic environment where P sound sources are active, the signal recorded by the mth microphone of a circular microphone array with M sensors can be modelled as

$$x_m(t) = \sum_{p=1}^{P} h_{mp}(t) \star s_p(t), \quad (1)$$

where $s_p(t)$ is the pth source signal, $h_{mp}(t)$ is the impulse response of the acoustic path from source p to sensor m, and $\star$ denotes the convolution operation.

With the use of a time-frequency transform, such as the short-time fourier transform (STFT), the model above can be expressed in the time-frequency domain as

$$X_m(k,\omega) = \sum_{p=1}^{P} H_{mp}(k,\omega) S_p(k,\omega), \quad (2)$$

where k and ω are the time frame and frequency indices, respectively, and $X_m(k,\omega)$, $H_{mp}(k,\omega)$, and $S_p(k,\omega)$ are the Fourier transforms of the signals $x_m(t)$, $h_{mp}(t)$, and $s_p(t)$, respectively.

If we assume an anechoic model for the sound propagation and that the room characteristics do not change over time, then the time dependency of the frequency response H_{mp} can be omitted. Moreover, under the far-field assumption that the sound sources are distant enough, the wavefronts impinging on the microphones are planar, and the frequency response can be written as

$$H_{mp}(\omega) = e^{j2\pi f_\omega \tau_m(\theta_p)}, \quad (3)$$

where $\tau_m(\theta_p)$ is the time delay from source p to the mth microphone, θ_p is the DOA of source p with respect to a predefined microphone array coordinate system, and f_ω denotes the frequency in Hertz that corresponds to frequency index ω.

Note that although the model is simplified, the experiments are performed using signals recorded in a reverberant environment (real or simulated) and the localization method used has been tested in such environments [12–14].

3. Proposed Method

We propose an array-based approach for the recording and reproduction of spatial audio, using a uniform circular array of microphones. Our method is divided into two parts: the analysis and the synthesis stages. In the analysis stage, the sound sources that are present in the environment are identified and the DOA of each source is estimated. With the use of spatial filtering, we separate the source signals that come from different directions. One audio signal and additional side information are used in the synthesis stage for reproduction using either headphones or an arbitrary loudspeaker configuration. Each estimated source signal is processed as an individual entity. The signals are then played back together from different directions, eliminating musical distortions resulting from spatial filtering and beamforming. Both stages are real time, with the analysis stage consuming approximately 50% of the available processing time on a standard PC (Intel 2.53 GHz Core i5, 4 GB RAM). The synthesis stage can also be easily implemented in real time since its main operation is amplitude panning (or HRTF filtering for binaural reproduction).

3.1. Analysis Stage. The main operations of the analysis stage are depicted in Figure 1. The microphone array signals $x_m(t)$, $m = 1,\ldots,M$, are divided into small overlapping time frames and transformed into the STFT domain. A DOA estimation method is applied in the frequency domain, to give an estimate of the number of active sources $\widehat{P}$ (with P being the true number of sources) and their DOAs. Source separation is then carried out through spatial filtering with a fixed superdirective beamformer, to yield $\widehat{P}$ source signals that are downmixed into one signal. Side information is also retained, namely, the DOAs in each frequency element of the downmixed signal. During downmixing, a certain frequency range from one of the microphone signals (e.g., microphone number one in Figure 1) may optionally be used as diffuse sound, as explained in Section 3.1.5 (dashed line in Figure 1).

3.1.1. DOA Estimation. For DOA estimation we utilize the method of [12–14]. This method is capable of estimating the direction of arrival, as well as the number of active sources $\widehat{P}$, in real time and with high accuracy even when the number of sources is high. The output of the DOA estimation procedure is the estimated number of sources $\widehat{P}_k$ and a vector with the estimated DOA for each source $\boldsymbol{\theta}_k = [\theta_1 \cdots \theta_{\widehat{P}_k}]$ per time frame k. To estimate the DOAs, a history block that can range from 0.5 sec to 1 sec is used. As well as improving the accuracy of DOA estimations, the history block smooths the source trajectories, which is beneficial—especially in the case of moving sources—as it avoids abrupt changes in source movements that can result in unnatural or loss of perception of the source direction.

To illustrate the efficiency of our DOA estimation method in the context of spatial audio capturing and reproduction, Figure 2 presents some examples of the estimated DOAs for the recordings used in our listening tests. The microphone array signals were created in a reverberant simulated environment. The recordings include classical and rock music with both impulsive and nonimpulsive instruments, as well as speech signals of simultaneously active moving speakers. More details about the recordings are given in Section 4.1.

It is clear that the method can detect all the sources, resulting in sufficiently accurate and smooth source trajectories. In the case of moving sources (Figure 2(c)) some problems occur before and after the two sources meet and cross each other. Erroneous estimates in some individual frames are also evident in the figure. However, since there are no active

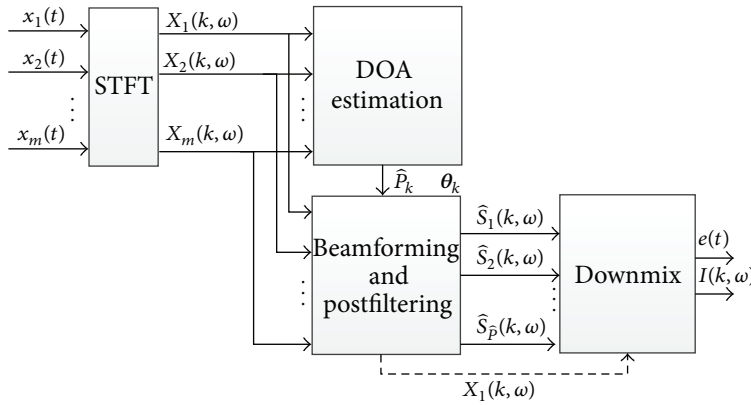
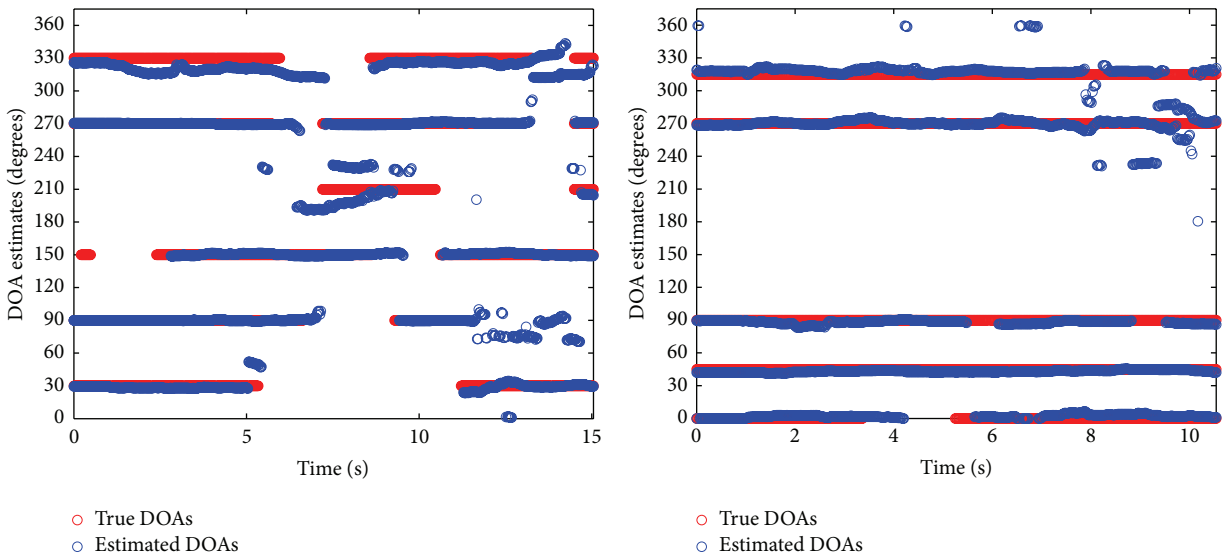
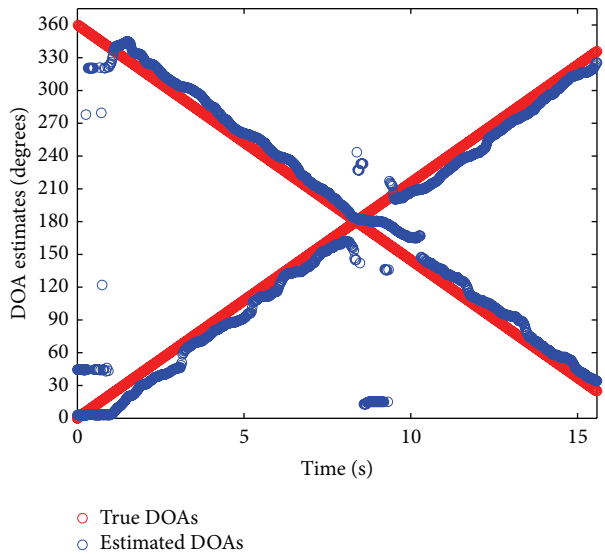


Figure 1: Analysis stage of the proposed method.



(a) Classical music recording (6 sources)

(b) Rock music recording (5 sources)



(c) Speech recording (2 moving sources)

Figure 2: Estimated DOAs for the simulated microphone array recordings.

sources present in these directions, the beamforming and downmixing operations in the next stages are expected to cancel the reproduction of the signals from these erroneous directions. Thus, we expect that as long as all the active sources are identified, these individual erroneous estimates—caused by an overestimation of the number of active sound sources—will not degrade the spatial audio reproduction, which is validated by our listening tests.

3.1.2. Superdirective Beamforming. In the next step, spatial filtering with a fixed filter-sum superdirective beamformer separates the source signals. The frequency domain output of a filter-sum beamformer is given by

$$Y(\omega) = \sum_{m=1}^{M} w_m^*(\omega, \theta_s) X_m(\omega), \tag{4}$$

where $w_m(\omega, \theta_s)$ is a complex filter coefficient for the mth microphone to steer the beam to the desired steering direction θ_s and $(\cdot)^*$ denotes the complex conjugate operation.

Superdirective beamformers aim to maximize the directivity factor or array gain, which measures the beamformer's ability to suppress spherically isotropic noise (diffuse noise). The array gain is defined as [15]

$$G_a(\omega) = \frac{\left|\mathbf{w}(\omega, \theta_s)^H \mathbf{d}(\omega, \theta_s)\right|^2}{\mathbf{w}(\omega, \theta_s)^H \boldsymbol{\Gamma}(\omega) \mathbf{w}(\omega, \theta_s)}, \tag{5}$$

where $\mathbf{w}(\omega, \theta_s) = [w_1(\omega, \theta_s) \cdots w_M(\omega, \theta_s)]^T$ is the vector of filter coefficients for all sensors, $\mathbf{d}(\omega, \theta_s) = [e^{-j2\pi f_\omega \tau_1(\theta_s)} \cdots e^{-j2\pi f_\omega \tau_M(\theta_s)}]^T$ is the steering vector of the array, $\boldsymbol{\Gamma}(\omega)$ is the $M \times M$ noise coherence matrix, $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and the Hermitian transpose operation, respectively, and j is the imaginary unit.

Under the assumption of a diffuse noise field, $\boldsymbol{\Gamma}(\omega)$ can be modelled as [16]

$$\Gamma_{ij}(\omega) = B_0\left(\frac{2\pi f_\omega d_{ij}}{c}\right), \tag{6}$$

with $B_0(\cdot)$ being the zeroth-order Bessel function of the first kind, c the speed of sound, and d_{ij} the distance between microphones i and j, which in the case of a uniform circular array with radius r is given by

$$d_{ij} = 2r\left|\sin\left(\frac{2\pi(i-j)}{2M}\right)\right|. \tag{7}$$

The optimal filter coefficients for the superdirective beamformer can be found by maximizing (5), while maintaining a unit-gain constraint on the signal from the steering direction θ_s; that is,

$$\mathbf{w}(\omega, \theta_s)^H \mathbf{d}(\omega, \theta_s) = 1. \tag{8}$$

Moreover, since superdirective beamformers are susceptible to extensive amplification of noise at low frequencies, a constraint is placed on the white noise gain (WNG), which expresses the beamformer's ability to suppress spatially white noise. The WNG is a measure of the beamformer's robustness and is defined as the array gain when $\boldsymbol{\Gamma}(\omega) = \mathbf{I}$, where $\mathbf{I}$ is the $M \times M$ identity matrix. Thus, the WNG constraint is expressed as

$$\frac{\left|\mathbf{w}(\omega, \theta_s)^H \mathbf{d}(\omega, \theta_s)\right|^2}{\mathbf{w}(\omega, \theta_s)^H \mathbf{w}(\omega, \theta_s)} \geq \gamma, \tag{9}$$

where γ represents the minimum desired WNG.

According to [15], the optimal filters given the constraints of (8) and (9) are given by

$$\mathbf{w}(\omega, \theta_s) = \frac{[\epsilon \mathbf{I} + \boldsymbol{\Gamma}(\omega)]^{-1} \mathbf{d}(\omega, \theta_s)}{\mathbf{d}(\omega, \theta_s)^H [\epsilon \mathbf{I} + \boldsymbol{\Gamma}(\omega)]^{-1} \mathbf{d}(\omega, \theta_s)}, \tag{10}$$

where the constant ϵ is used to control the WNG constraint and is associated with γ in the sense that the WNG increases monotonically with increasing ϵ [15]. However, there is a trade-off between robustness and spatial selectivity of the beamformer, as increasing the WNG decreases the directivity factor.

To calculate the beamformer filter coefficients, we used an iterative procedure to determine ϵ in a frequency-dependent manner. Starting from $\epsilon = 0$ we iteratively increase ϵ by 0.005, until the WNG becomes equal or greater than γ. The resulting Directivity Factor and WNG for different values of γ, namely, $\gamma = -10$ dB, $\gamma = 0$ dB, and $\gamma = 10$ dB, are shown in Figure 3, which also indicates the trade-off between WNG and directivity of the beamformer. The directivity factor and WNG when no constraint is applied to the WNG are also shown in Figure 3. The results in Figure 3 have been calculated using a uniform circular array of 8 microphones and a radius of 0.05 m, for a steering direction of 90°. Note that these are the same microphone array specifications used in our listening tests. In our implementation we set $\gamma = -10$ dB. To illustrate the expected directivity of our beamformer, Figure 4 shows the power spectrum of the normalized far-field directivity pattern across frequency for the same setup as Figure 3. It is evident that the beamformer maintains a good directivity pattern across frequency. Spatial aliasing in the directivity pattern is also evident above the spatial-aliasing cutoff frequency, which is 4 kHz for the specific array geometry. However, as our listening test results indicate, spatial aliasing in the beamforming process does not affect the spatial audio capturing and reproduction.

Fixed beamformers are signal independent, so they are computationally efficient to implement, since the filter coefficients for all steering directions need to be estimated only once and then stored offline. An adaptive version of a beamformer would increase the computational burden, since the filter coefficients would have to be estimated at run time. Moreover, the performance of adaptive beamformers has been shown to be susceptible to correlated noise [17].

In each time frame k, the beamforming process employs $\widehat{P}_k$ concurrent beamformers. Each beamformer steers its beam to one of the directions specified by vector $\boldsymbol{\theta}_k$, yielding in total $\widehat{P}_k$ signals $B_s(\omega)$, $s = 1, \ldots, \widehat{P}_k$ in the frequency domain, according to (4).

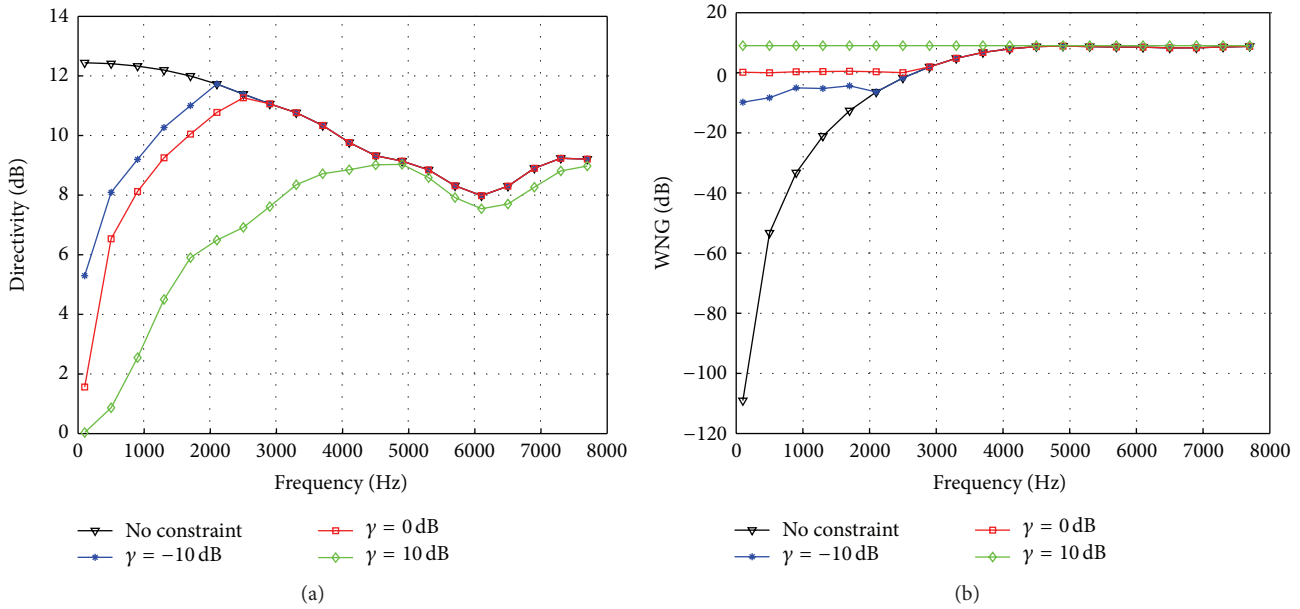


FIGURE 3: (a) Directivity factor and (b) white noise gain (WNG) across frequency for different values for the WNG constraint, for a circular array with 8 microphones and 0.05 m radius and a steering direction of 90°.

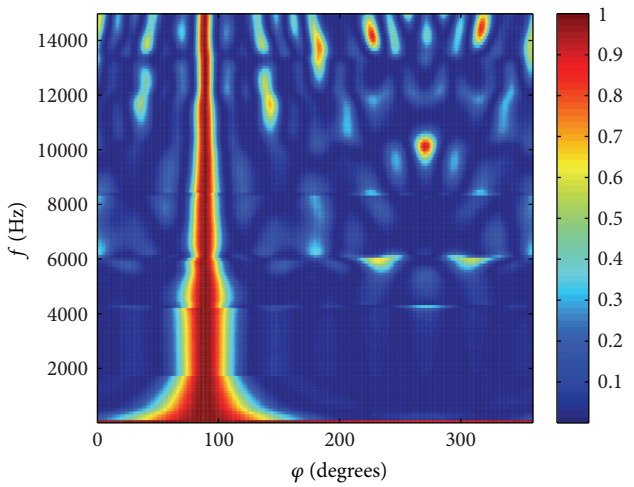


FIGURE 4: Power spectrum of the normalized far-field directivity pattern of an 8-element uniform circular array with 0.05 m radius, at a steering direction of 90°.

3.1.3. Postfiltering. A postfilter following the beamformer output can result in significant cancellation of interference from other directions. The design of Wiener filters that are based on the auto and cross-power spectral densities between microphones and applied to the output of the beamformer has been extensively investigated (see [18–20] and the references therein). In this work, we utilize a postfilter especially designed for overlapped speech [21], in order to cancel interfering speakers from the target speakers' signals.

We assume that in each time-frequency element there is only one dominant sound source (i.e., there is one source with significantly higher energy than the other sources). In speech signals this is a reasonable assumption, since the sparse and varying nature of speech makes it unlikely that two or more speakers will carry significant energy in the same time-frequency element (when the number of active speakers is relatively low). Moreover, it has been shown that the spectrogram of the additive combination of two or more speech signals is almost the same as the spectrogram formed by taking the maximum of the individual spectrograms in each time-frequency element (see [22] for a discussion).

Under this assumption, we construct $\widehat{P}_k$ binary masks as follows:

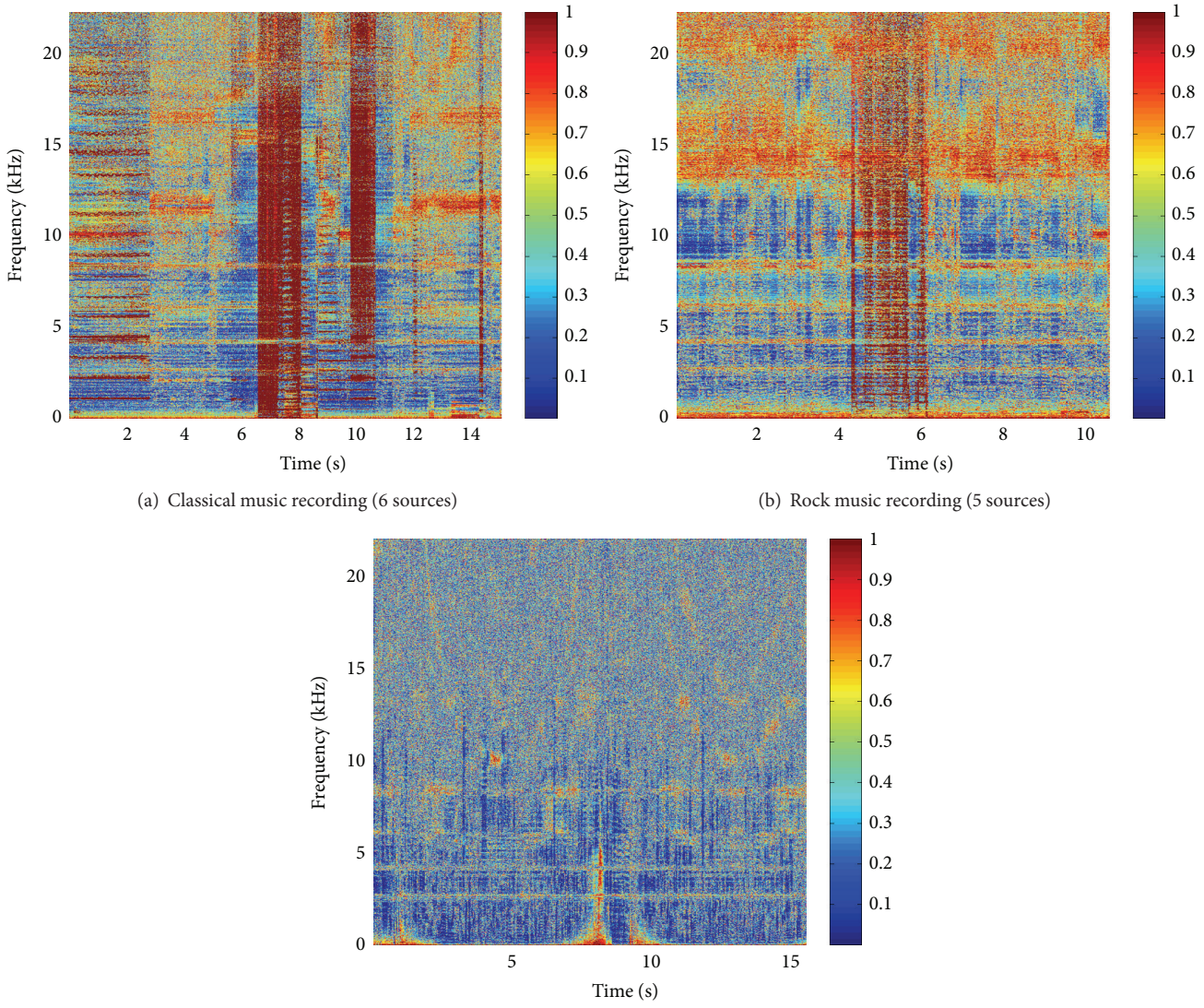
$$U_s(\omega) = \begin{cases} 1, & \text{if } s = \arg\max_p \left|B_p(\omega)\right|^2, \ p = 1, \ldots, \widehat{P}_k \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

Equation (11) implies that for each frequency element, only the corresponding element from one of the beamformed signals is retained, that is, the one with the highest energy with respect to the other signals at that frequency element. Each mask is applied to the corresponding beamformer output signal to yield the estimated source signals:

$$\widehat{S}_s(\omega) = U_s(\omega) B_s(\omega), \quad s = 1, \ldots, \widehat{P}_k. \quad (12)$$

The post-filter can also be viewed as a classification procedure, as it assigns a time-frequency element to a specific source, based on the energy of the signals B_s.

3.1.4. Downmixing. From (11), it can be seen that the masks are orthogonal with respect to each other. This means that if $U_s(\omega) = 1$ for some frequency index ω, then $U_{s'}(\omega) = 0$ for $s' \neq s$, which is also the case for the signals $\widehat{S}_s$. Using this



(a) Classical music recording (6 sources)

(b) Rock music recording (5 sources)

(c) Speech recording (2 moving sources)

FIGURE 5: Energy ratio in the time-frequency domain of the beamformer output signals of the source with the second largest energy over the source with the largest energy for the simulated microphone array recordings.

property, we can downmix the signals $\widehat{S}_s$ into one signal, by summing them up in the frequency domain to form a full spectrum signal. Furthermore, by encoding the DOA of each bin as side information, we can easily separate the source signals at the decoder.

The downmixed signal $E(\omega)$ can be written as:

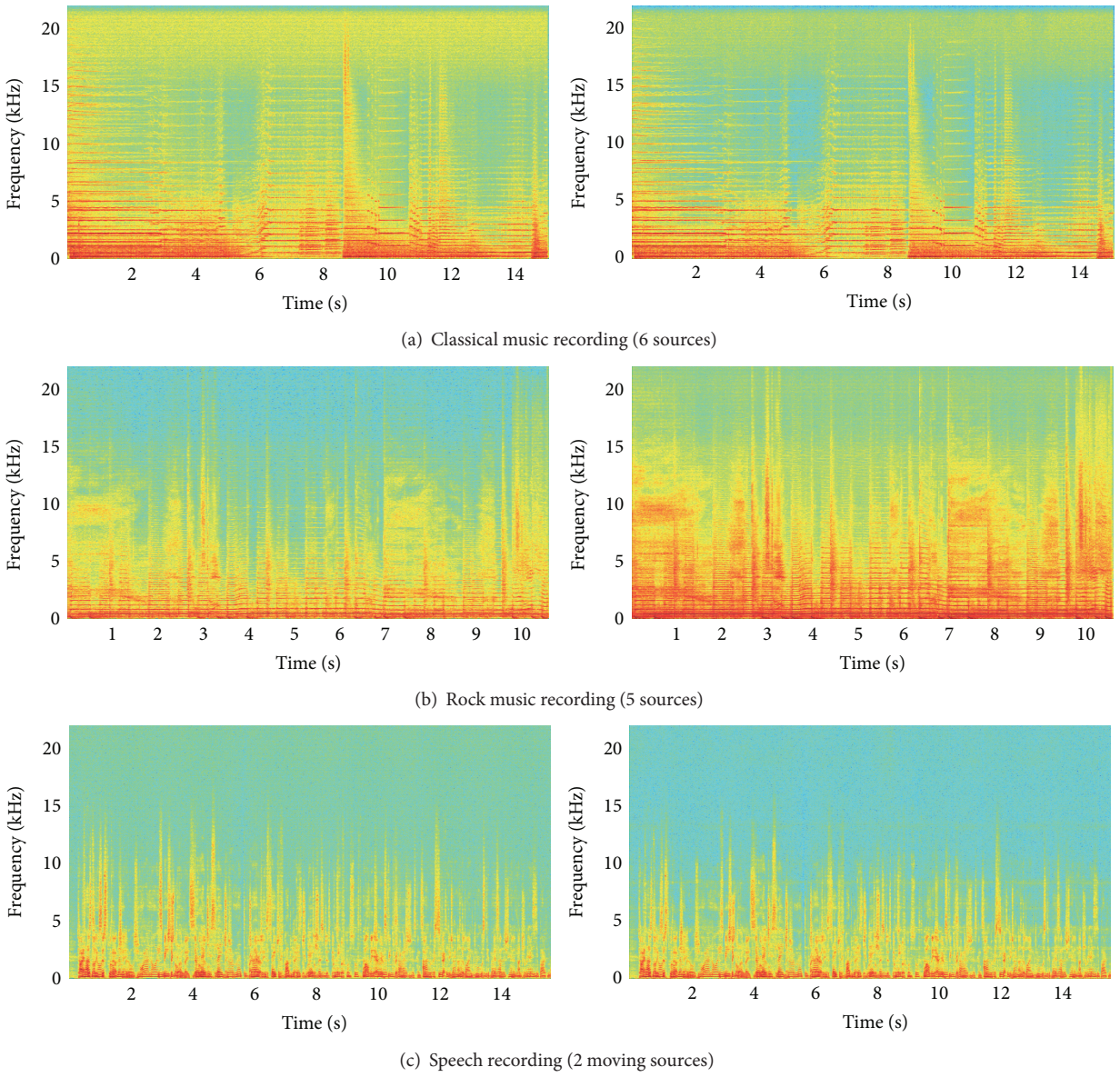
$$E(\omega) = \sum_{s=1}^{\widehat{P}_k} \widehat{S}_s(\omega), \tag{13}$$

together with sideinformation for each frequency element:

$$I(\omega) = \theta_s, \quad \text{for the } s \text{ such that } \widehat{S}_s(\omega) \neq 0. \tag{14}$$

The downmixed signal $E(\omega)$ is transformed back to the time domain and is transmitted to the decoder, along with side information as specified by (14). Note that the signal $e(t)$ can also be encoded as monophonic sound with the use of some coder (e.g., MP3) in order to reduce bitrate. However, encoding and bitrate aspects are part of our future work.

During post-filtering the WDO assumption is made, which also allows the separated source signals to be downmixed into one audio signal. In order to provide some examples of this assumption, Figure 5 depicts the energy ratio in the time-frequency domain of the beamformer output signals of the source with the second largest energy after the source with the highest energy, for the microphone array recordings in a reverberant simulated environment that were used in our listening tests. The recordings include classical and rock music with tonal and percussive sounds, as well as simultaneously active moving speakers (Section 4.1).



(a) Classical music recording (6 sources)

(b) Rock music recording (5 sources)

(c) Speech recording (2 moving sources)

Figure 6: Spectrogram of (left column) the signal at the first microphone of the array and (right column) the downmixed signal for the simulated recordings.

The energy ratios for the classical music, rock music, and speech recordings are shown in Figures 5(a), 5(b), and 5(c), respectively.

When a source dominates in a time-frequency element, we expect the ratio to be small. When both sources carry the same amount of energy, we expect the ratio to be close to one. It is clear that the assumption of sparsity and disjointedness of the source signals is more evident in speech and when the number of active sources is low (Figure 5(c)). The sparsity assumption weakens when music signals with many active sound sources are considered (Figures 5(a) and 5(b)). To get an intuition of how close the downmixed signal is to the originally recorded sound field, Figure 6 compares the spectrograms of the downmixed signals and the signal received at the first microphone of the array for the same simulated recordings (Section 4.1). The example signals in Figure 6 demonstrate that there are strong similarities between the spectrograms of the downmixed signal and the signal received at the first microphone, in all three types of signals used in the experiments (classical music, rock music, and speech). This indicates that the downmixing process results in a valid reconstruction of the sound signals and retains the major time-frequency elements in the original recording. The effectiveness of the downmixing approach is validated through listening tests in both simulated and real environments and with sounds that include both speech and

instruments of different types (tonal and percussive). Our results reveal that even when the WDO hypothesis is weak, the reconstructed acoustic environment is not degraded in terms of spatial impression and overall quality.

3.1.5. Incorporating Diffuse Sound. The beamforming and post-filtering procedure can be realized across the whole spectrum of frequencies or up to a specific *beamformer cutoff frequency*. Processing only a certain range of the frequency spectrum may have several advantages, such as reduction in the computational complexity, especially when the sampling frequency is high, and reduction in the side information that needs to be transmitted, since DOA estimates are available only up to the beamformer cutoff frequency. Moreover, issues related to spatial aliasing may be avoided if the beamformer is applied only to the frequency range which is free from spatial aliasing. While the DOA estimation process does not suffer from spatial aliasing—since it only considers frequencies below the spatial-aliasing cutoff frequency (for details, see [12–14])—the beamformer's performance may theoretically be degraded.

There are spatial audio applications, that would tolerate this suboptimal approach. For example, a teleconferencing application, where the signal content is mostly speech and there is no need for very high audio quality, could tolerate using only the frequency spectrum up to 4 kHz (treating the rest of the spectrum as diffuse sound), without significant degradation in source spatialization.

For the frequencies above the beamformer cutoff frequency, the spectrum from an arbitrary microphone is included in the downmixed signal, without additional processing. As there are no DOA estimates available for this frequency range, it is treated as diffuse sound in the decoder and reproduced by all loudspeakers, in order to create a sense of immersion for the listener. However, extracting information from a limited frequency range can degrade the spatial impression of the sound. For this reason, including a diffuse part is offered as an optional choice, and we also consider the case where the beamformer cutoff frequency is set to $f_s/2$; that is, there is no diffuse sound. In this case, the beamformer's performance may be affected by spatial aliasing; however, as our listening tests indicate, such degradation is not audible.

3.2. Synthesis Stage

3.2.1. Loudspeaker Reproduction. In the synthesis stage (Figure 7), the downmixed signal and side information are used in order to create spatial audio with an arbitrary loudspeaker setup. The nondiffuse and the diffuse parts (if the latter exists) of the spectrum are treated separately. First, the signal is divided into small overlapping frames and transformed to the STFT domain, as in the analysis stage. The diffuse signal option corresponds to the dashed line in Figure 7.

The nondiffuse part of the spectrum is synthesized using vector-base amplitude panning (VBAP) [23] at each frequency index, according to its corresponding DOA from $I(\omega)$. By adjusting the gains of a set of loudspeakers, VBAP

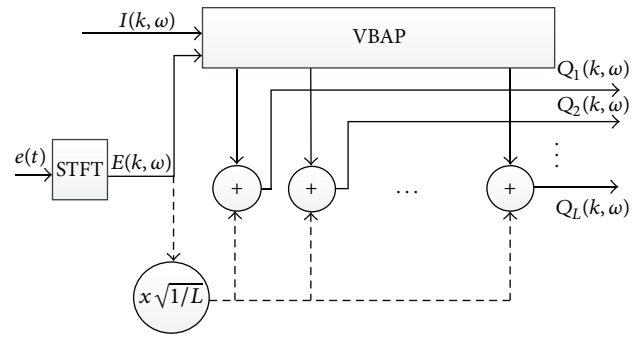


Figure 7: Synthesis stage of the proposed method for loudspeaker reproduction.

can position a sound source anywhere across an arc defined by two adjacent loudspeakers (2-dimensional VBAP) or inside a triangle defined by three loudspeakers in the 3-dimensional case. In this work, only 2-dimensional reproduction is considered. If a diffuse part is included, then it is played back from all loudspeakers.

Assuming a loudspeaker configuration with L loudspeakers, the lth loudspeaker signal is given by

$$Q_l(\omega) = \begin{cases} g_l(\omega) E(\omega) & \text{for } \omega \le \omega_{\text{cutoff}} \\ \dfrac{1}{\sqrt{L}} E(\omega) & \text{for } \omega > \omega_{\text{cutoff}}, \end{cases} \tag{15}$$

where ω_{cutoff} is the beamformer cutoff frequency index, as discussed in Section 3.1.5, $g_l(\omega)$ is the gain for the lth loudspeaker at frequency index ω, as computed from VBAP, and the diffuse part is divided by the square root of the number of loudspeakers to preserve the total energy. If $\omega_{\text{cutoff}} = f_s/2$, then the full spectrum processing method is applied and no diffuse part is included.

3.2.2. Binaural Reproduction. The binaural version of the proposed method utilizes HRTFs in order to position each source in a certain direction. The nondiffuse and the diffuse parts (if the latter exists) of the spectrum are again treated separately. After transforming the downmixed signal $e(t)$ into the STFT domain, the nondiffuse part is filtered in each time-frequency element with the HRTF, according to the side information available in $I(\omega)$. Thus, the left and right output channels for the nondiffuse part, at a given time frame, are produced by

$$\begin{aligned} Y_L(\omega) &= E(\omega)\,\text{HRTF}_L(\omega, I(\omega)), \quad \omega \le \omega_{\text{cutoff}} \\ Y_R(\omega) &= E(\omega)\,\text{HRTF}_R(\omega, I(\omega)), \quad \omega \le \omega_{\text{cutoff}}, \end{aligned} \tag{16}$$

where $\text{HRTF}_{\{L,R\}}$ is the head-related transfer function for the left or right channel, as a function of frequency and direction.

The optional diffuse part is filtered with a diffuse field HRTF, in order to make its magnitude response similar to the nondiffuse part. Diffuse field HRTFs can be produced by averaging HRTFs from different directions across the whole circle around the listener. The filtering process in this case becomes the following:

$$Y_L(\omega) = E(\omega)\,\mathrm{HRTF}_L^{\mathrm{diff}}(\omega), \quad \omega > \omega_{\mathrm{cutoff}}$$
$$Y_R(\omega) = E(\omega)\,\mathrm{HRTF}_R^{\mathrm{diff}}(\omega), \quad \omega > \omega_{\mathrm{cutoff}}. \tag{17}$$

4. Listening Test Procedure

To evaluate the efficiency of our proposed method we conducted listening tests on simulated and real microphone array recordings. Listening tests were performed in a quiet office environment. For loudspeaker reproduction we used $L = 8$ uniformly spaced loudspeakers (Genelec 8050) arranged in a circular configuration. During the tests the subject was sitting in the "sweet spot" of the configuration and the distance of the loudspeakers from the subject was set to one meter. Ten volunteers participated in each test (authors not included).

The circular array consisted of $M = 8$ microphones and a radius $r = 0.05$ m. The sampling frequency was $f_s = 44100$ Hz. The signals were divided into frames of 2048 samples with 50% overlap and windowed with a Hann window. The FFT size was 4096.

The proposed method was compared against other state-of-the-art array-based methods, namely, the method of [10] and the microphone array version of DirAC, as presented in [8]. For our proposed method, three versions with different beamformer cutoff frequencies, $B = 4$ kHz, $B = 8$ kHz, and $B = f_s/2$, were included in the test. While the authors in [10] discuss only binaural reproduction, the extension to loudspeaker reproduction is straightforward by applying VBAP in each time-frequency element, using its corresponding DOA estimate. The DOA estimation method of [8] is based on a linear array geometry, so we used the localization procedure from [10], combining it with the diffuseness estimation and synthesis method of [8]. The microphone array, loudspeaker configuration, and the method's parameters were the same as specified previously.

4.1. Test Samples. Both simulated and real recordings were included in the tests. To produce simulated recordings, we used the image-source method (ISM) [24]. The room dimensions were set to 6 m, 4 m, and 3 m in length, width, and height, respectively. The reverberation time was $T_{60} = 250$ ms and the walls were characterized by a uniform reflection coefficient of 0.5. The microphone array was placed in the center of the room. The array center defines the origin of a two-dimensional Cartesian coordinate system, and the DOA of each source is defined by the azimuth angle formed by the y-axis and the line connecting the array center with the point where the source is located. The same coordinate system is used for reproduction; thus, 0° corresponds to the direction in front of the listener with the DOAs increasing clockwise, similar to a compass bearing.

Three test samples were used for the listening test with the simulated recordings:

(i) a 10-second rock music recording with one male singer at 0° and 4 instruments at 45°, 90°, 270°, and 315°;

(ii) a 15-second classical music recording with 6 sources at 30°, 90°, 150°, 210°, 270°, and 330°;

(iii) a 16-second recording with two speakers, one male and one female, starting from 0° and walking the entire circle in opposite directions.

The recordings were multitrack with simultaneously active sound sources. The music tracks included impulsive and nonimpulsive instruments. The tracks for the classical recording were obtained from [25], and the rock music tracks are publicly available from the band "Nine Inch Nails." The speech recordings were produced in our laboratory.

Each track of the recording was filtered with the room impulse response from its corresponding direction as estimated from the ISM [24], to simulate the microphone array signals for this track. The tracks of the entire recording from each microphone were then added together to form the final microphone array recording.

The real recordings were recorded with a microphone array in an office room. The room dimensions and microphone specifications were the same as in the simulated case. We used Shure SM93 omnidirectional microphones and a TASCAM US2000 USB sound card with 8 channels.

The real recorded test samples are

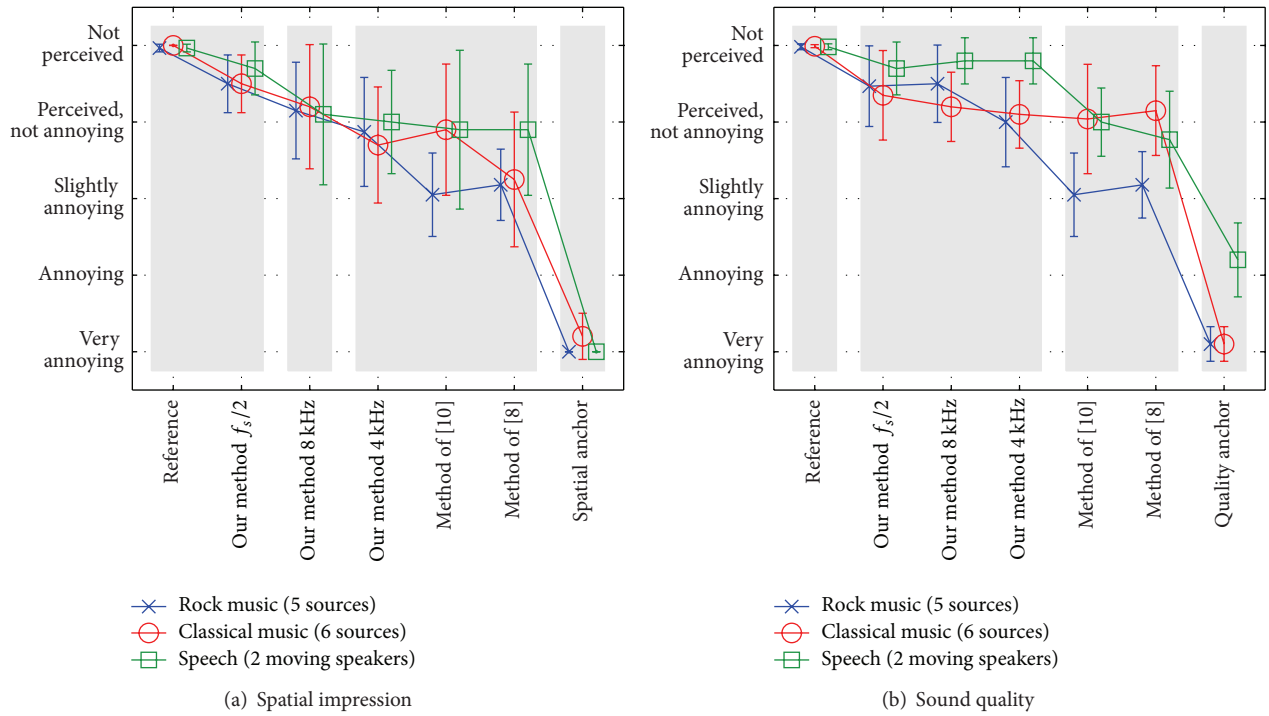
(i) a 10-second rock music recording with 5 sources at 0°, 45°, 90°, 270°, and 315°;

(ii) a 15-second classical music recording with 4 sources at 0°, 45°, 90°, and 270°;

(iii) a 10-second recording with two male speakers, one stationary at 240° and one moving clockwise from about 320° to 50°.

Each source signal was reproduced from a loudspeaker located at the corresponding direction and at a distance of 1.5 m from the array. The sound signals were reproduced simultaneously and recorded with the microphone array. The moving speech recording was produced with two male persons (one stationary and one moving) speaking while being recorded from the microphone array. The classical and rock music signals were the same multitrack recordings as in the simulated recordings.

Some audio samples of our method are available at http://www.ics.forth.gr/~mouchtar/scar/.



(a) Spatial impression

(b) Sound quality

Figure 8: Listening test results for simulated recordings with loudspeaker reproduction.

4.2. Listening Test Methodology. The listening tests with the simulated recordings were based on the ITU-R BS.1116 methodology [26]. Each simulated track of a recording at an arbitrary microphone was panned using VBAP and positioned in its corresponding direction. The panned tracks were played simultaneously and served as reference signal for this recording. The output of each method was compared against the reference recording, using a 5-scale grading system, with 1 being "very annoying" difference compared to the reference and 5 being "not perceived" difference from the reference. A mixed signal with all the sources as recorded from the first microphone, played back from all loudspeakers served as spatial anchors. The low-pass filtered (with 4 kHz cutoff frequency) reference recordings served as quality anchors. The listening tests were carried out in two sessions: in the first session the subjects were asked to rate the recordings in terms of spatial impression, while in the second session the rating was based on sound quality.

For the listening test with the real recordings, a reference recording was not available; thus we employed a preference test (forced choice). A reference would only be possible if an in-ear recording was made for each listener's head. All possible combinations between the proposed methods and the methods of [8, 10] were included in pairs, and the listeners were asked to indicate their preference among each pair, judging again for spatial impression and sound quality in two different sessions. For this test only the versions with $B = f_s/2$ (no diffuse) and $B = 4$ kHz of our method were included.

5. Listening Test Results for Loudspeaker Reproduction

5.1. Simulated Recordings. The mean scores across all subjects and 95% confidence intervals, for the spatial impression and quality sessions, are presented in Figure 8. The results were analyzed using one-way analysis of variance (ANOVA). All the recordings were analyzed together. The analysis showed that a statistically significant difference between the methods occurs, in spatial impression $F = 66.95$, $p < 0.01$ and in quality $F = 82.62$, $p < 0.01$. To determine which pairs of methods are significantly different, multiple comparison tests were performed on the ANOVA results using Tukey's least significant distance. The methods with statistically insignificant differences have been grouped together in gray shading in Figure 8.

As expected, the ratings for the reference and anchor signals are at the opposite ends of the scale. The higher grading of the quality anchor for the speech recording compared with the other anchors can be explained by the fact that speech content lies mostly below 4 kHz, which is the cutoff frequency of the low-pass filter. Our proposed methods outperform the other methods—in terms of spatial impression—while sustaining the audio quality at very high levels, for both music and speech. The best results are achieved with our proposed method when $B = f_s/2$ (i.e., no diffuse). While beamforming across all the frequency spectra can theoretically degrade

the beamformer's performance to separate the sources—due to spatial aliasing—our listening test results indicate that such degradation is not audible. The beamformer cutoff frequency seems to play a key role in the spatialization of sound, since a degradation in spatial impression is evident as the beamformer cutoff frequency decreases, while the quality seems to be less affected by this parameter. This can be explained by the fact that the beamformer cutoff frequency specifies the frequency range for which directional information will be extracted. Thus a degradation in spatial impression with decreasing beamformer cutoff frequency is expected. However, in all versions of our method, the full frequency spectrum is reproduced either from a specific loudspeaker pair or from all loudspeakers (for the diffuse part) according to the beamformer cutoff frequency. Thus, the quality is less affected. However, even with the cutoff frequency set as low as 4 kHz, our proposed methods still receive better ratings than the other methods.

5.2. Real Recordings. Figures 9, 10, and 11 show the spatial impression and quality results for the classical, rock, and moving speech recordings, respectively, indicating the efficiency of our proposed method (for both $B = 4\,\text{kHz}$ and $B = f_s/2$) to provide audio with good spatialization and quality, according to listeners' preferences. All versions of our proposed method achieve better results than the other methods. The best comparative results in favor of our method were achieved for the speech recording. Real recordings of moving speakers are by themselves a challenge to localize accurately and spatially reproduce. We speculate that poor localization performance of the per time-frequency element approach of [10] is the reason that this method obtained poor results. Moreover, an overestimation in the diffuseness in [8] causes most of the sound to be played back from all loudspeakers, significantly degrading the spatial impression.

The proposed method with the beamformer cutoff frequency set to $B = 4\,\text{kHz}$ appears to work significantly better with speech than with music, a result which is also evident in the listening test with the simulated recordings. Speech signals contain most of their information in the frequency range up to 4 kHz, something that is not the case with music. Thus, the 4 kHz version seems to be rather suitable for speech signals, providing a good trade-off between reproduction accuracy and amount of side information that needs to be stored or transmitted. Overall results for all recordings (Figure 12) show a clear preference in favor of our method (for both $B = 4\,\text{kHz}$ and $B = f_s/2$), both in spatial impression and quality.

In terms of quality, the full spectrum signal is reproduced either from a specific loudspeaker or from all loudspeakers (for the frequencies in the diffuse part), which explains why our proposed method with $B = 4\,\text{kHz}$ sometimes achieves better ratings compared to our method with $B = f_s/2$. The sound quality not being affected by the beamformer cutoff frequency for loudspeaker reproduction is also evident in the listening test with the simulated recordings (Section 5.1).

6. Listening Test Results for Binaural Reproduction

6.1. Simulated Recordings. The binaural version of the proposed method was also tested through another set of listening tests. The methods of [10] and the array-based DirAC were again included for comparison. The analysis stage of the array-based DirAC was again implemented based on [8], while the synthesis stage for binaural reproduction was implemented according to [27].

For HRTF filtering, we utilized the HRTF database provided by [28]. The measurement resolution in this database is 5° at an elevation angle of 0°. For the DOAs where an HRTF was not available, we simply used the HRTF of the nearest DOA.

We again considered both simulated and real recordings, and the test samples and methodology were the same as specified in Section 4.

The binaural reference recording for the simulated case was created by filtering each simulated track of the recording from an arbitrary microphone with the HRTF corresponding to its direction and then playing all the tracks together via headphones. Low-pass filtered (with 4 kHz cutoff frequency) versions of the reference recordings were used as quality anchors, while the spatial anchor consisted of the monophonic recording (with all tracks) at an arbitrary microphone.

The spatial impression and quality ratings (mean scores and 95% confidence intervals) for the simulated recordings are depicted in Figure 13. The results were analyzed again using one-way ANOVA, which showed that a statistically significant difference between the methods occurs, in spatial impression $F = 42.48$, $p < 0.01$ and in quality $F = 61.15$, $p < 0.01$. All the recordings were analyzed together. To determine which pairs of methods are significantly different, multiple comparison tests were performed on the ANOVA results using Tukey's least significant distance. The methods with statistically insignificant differences have been grouped together in gray shading in Figure 13.

Again, the reference and anchor signals have the highest and lowest ratings, respectively, while our proposed method with $B = f_s/2$ outperforms all the other approaches both in spatial impression and quality. The degradation in spatial impression when reducing the beamformer cutoff frequency is also evident in the binaural case. However, a significant degradation in quality is also observed in this case. High-quality headphone reproduction allows the listener to notice small impairments and degradation in quality more easily than in loudspeaker reproduction, which explains this discrepancy in the quality ratings between the two reproduction types. Moreover, distortions from the HRTF filtering (especially when the number of sources is high) are another factor that degrades quality. In some cases, the method of [8, 10] achieves better quality than the proposed method when $B = 8\,\text{kHz}$ or $B = 4\,\text{kHz}$, although this difference is statistically insignificant.

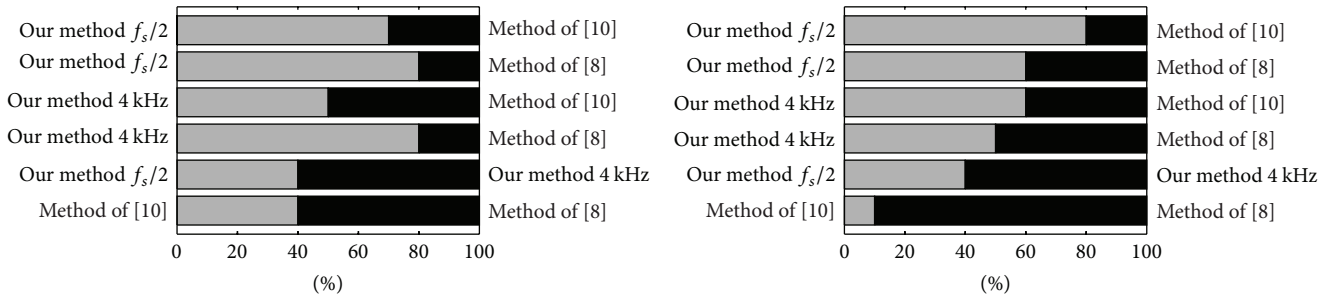


Figure 9: Preference listening test results for the classical music recording with loudspeaker reproduction.

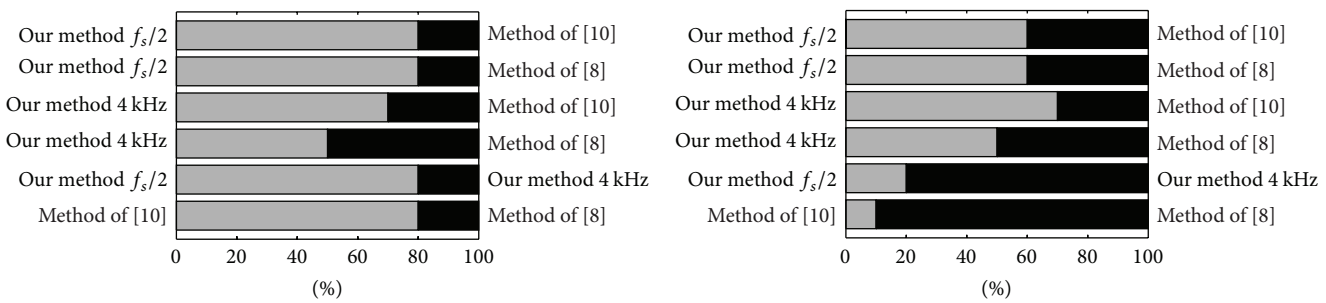


Figure 10: Preference listening test results for the rock music recording with loudspeaker reproduction.

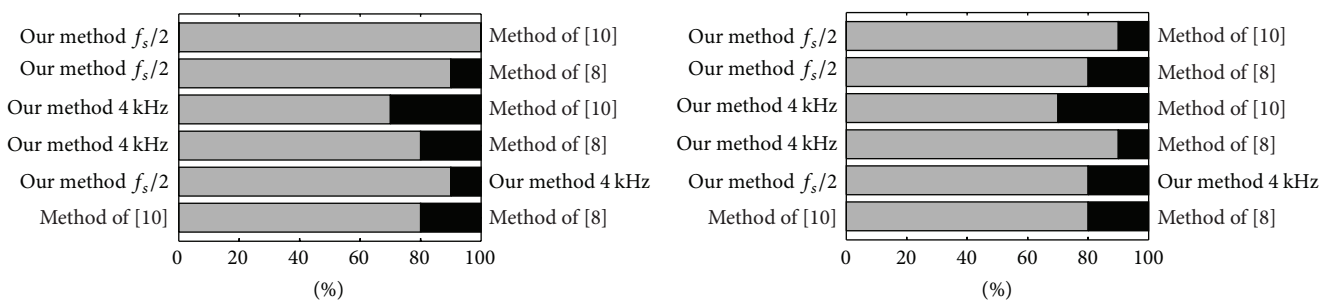


Figure 11: Preference listening test results for the speech recording with loudspeaker reproduction.

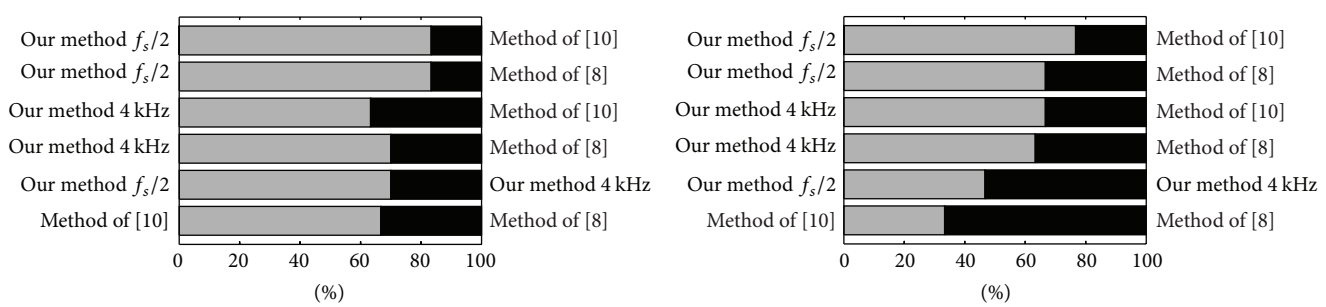
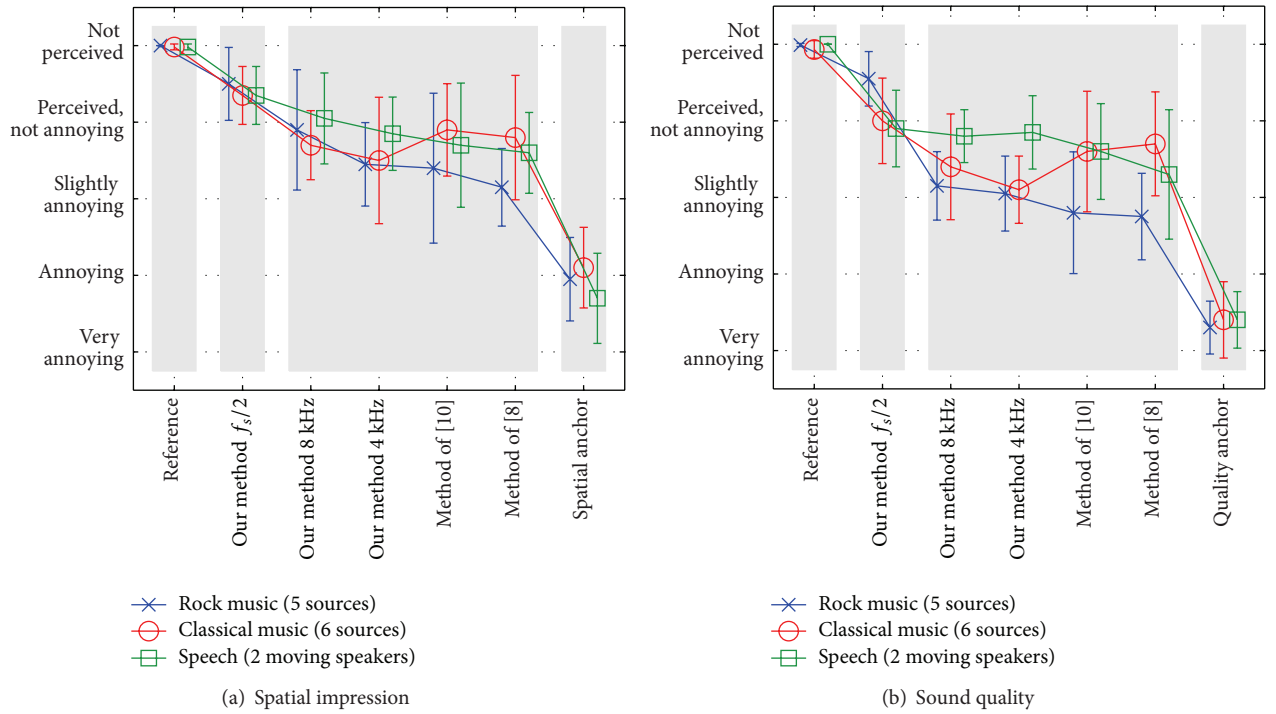
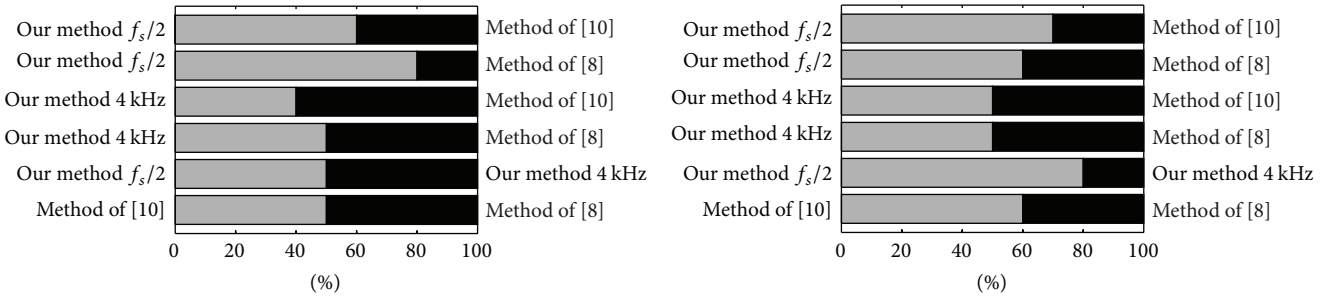


Figure 12: Overall results for the preference listening test for loudspeaker reproduction.



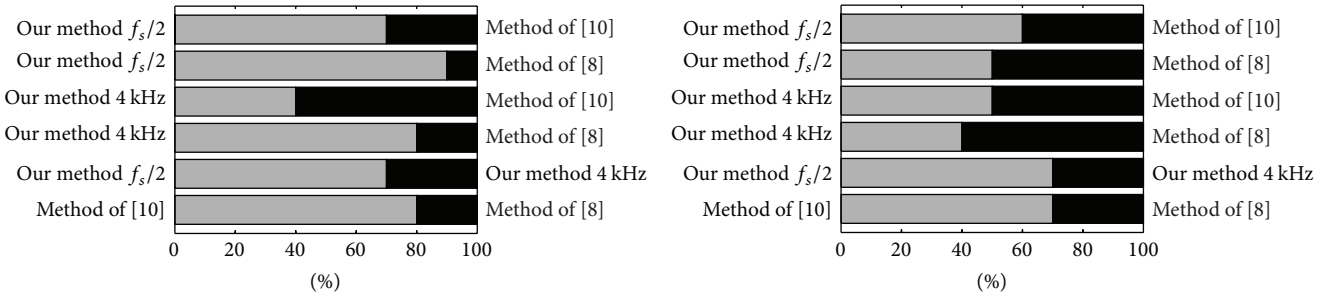
(a) Spatial impression (b) Sound quality

Figure 13: Listening test results for simulated recordings with binaural reproduction.



(a) Spatial impression (b) Sound quality

Figure 14: Preference listening test results for the classical music recording with binaural reproduction.



(a) Spatial impression (b) Sound quality

Figure 15: Preference listening test results for the rock music recording with binaural reproduction.

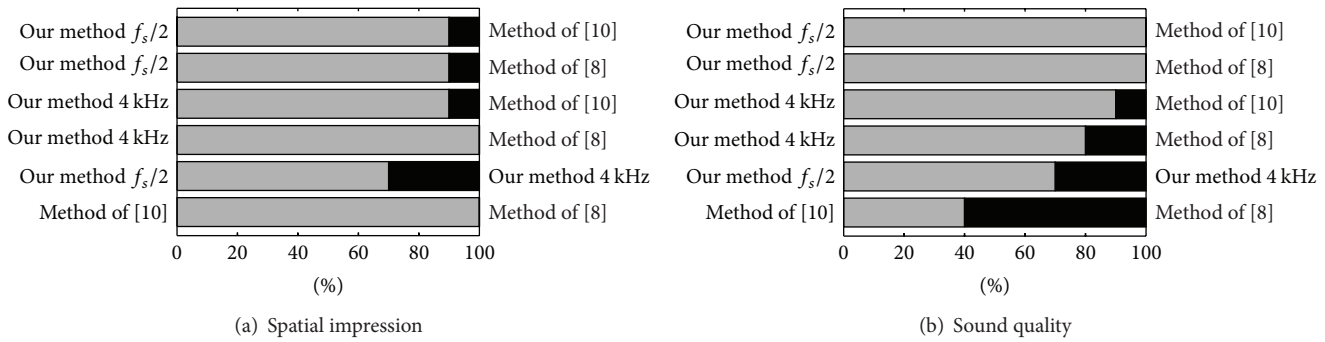


Figure 16: Preference listening test results for the speech recording with binaural reproduction.

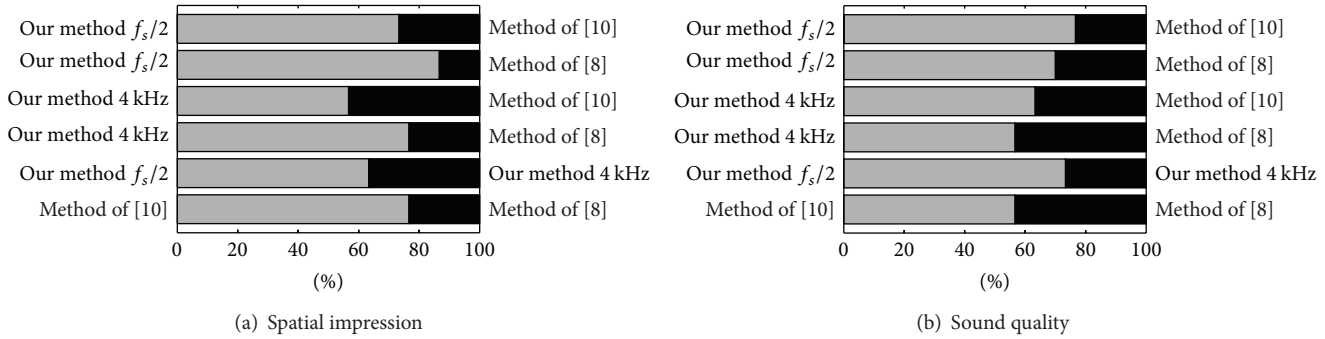


Figure 17: Overall results for the preference listening test for binaural reproduction.

6.2. Real Recordings. Figures 14, 15, and 16 show the spatial impression and quality preference results for the classical, rock, and moving speech recordings, respectively, while Figure 17 presents overall results across all the recordings. In terms of spatial impression, our proposed method either equals or surpasses the other methods. The clear advantage of our method for the speech recording, is again due to the reasons discussed in Section 5, since only the reproduction method differs between the binaural and loudspeaker versions of the methods. Our proposed method with $B = 4$ kHz achieves the best results in the speech recording—for binaural reproduction too—validating our claim that this version of our method can provide excellent results for speech applications, while using the least side information.

Conclusive results about the most preferred method in terms of quality cannot be drawn, since for the music recordings all methods seem to be equally preferred most of the time. For the moving speech recording an advantage in favor of our method is evident (Figure 16), although we believe that poor spatialization of sound in the case of the other methods may have also affected the quality ratings.

7. Conclusions

In this paper, a real-time microphone array-based approach to creating spatial audio using headphones or an arbitrary loudspeaker configuration was proposed. Based on a circular microphone array our method can reproduce the recorded acoustic environment in real time. Through a novel downmixing procedure, we encode the acoustic environment using only one monophonic audio signal and side information. Moreover, by estimating the DOAs in each time frame we overcome some of the problems related to spatial aliasing and the need for strong WDO conditions which are encountered when processing each time-frequency element individually.

The efficiency of our proposed method for capturing and reproducing spatial audio was validated by listening tests using microphone array recordings in both simulated and real environments. Our results indicate that our method achieves excellent reconstruction of the recorded acoustic environment both in terms of spatial impression and sound quality.

An important aspect for future work is the encoding of the downmixed signal, using an encoder such as MP3. Psychoacoustic models and frequency masking phenomena need to be taken into account in MP3 coding, in order to reduce the size of the signal. We plan to investigate the effects of such encoding in the multichannel case. We also plan to investigate encoding schemes to efficiently encode the side-information at low bitrates.

Acknowledgments

The authors would like to thank all the volunteers who participated in the listening tests. This work has been funded in part by the PEOPLE-IAPP "AVID-MODE" Grant within the 7th European Community Framework Program.

References

[1] J. Blauert, *Spatial Hearing*, MIT Press, Cambridge, UK, 1997.

[2] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, 1993.

[3] M. Cobos, S. Spors, J. Ahrens, and J. J. Lopez, "On the use of small microphone arrays for wave field synthesis auralization," in *Proceedings of the 45th International Conference: Applications of Time-Frequency Processing in Audio Engineering Society Conference*, 2012.

[4] H. Hacihabiboğlu and Z. Cvetković, "Panoramic recording and reproduction of multichannel audio using a circular microphone array," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '09)*, pp. 117–120, New Paltz, NY, USA, October 2009.

[5] K. Niwa, T. Nishino, and K. Takeda, "Encoding large array signals into a 3D sound field representation for selective listening point audio based on blind source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 181–184, Las Vegas, Nev, USA, April 2008.

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

[7] F. Kuech, M. Kallinger, R. Schultz-Amling, G. Del Galdo, J. Ahonen, and V. Pulkki, "Directional audio coding using planar microphone arrays," in *Proceedings of the Hands-free Speech Communication and Microphone Arrays (HSCMA '08)*, pp. 37–40, Trento, Italy, May 2008.

[8] O. Thiergart, M. Kallinger, G. D. Galdo, and F. Kuech, "Parametric spatial sound processing using linear microphone arrays," in *Proceedings of the Microelectronic Systems*, A. Heuberger, G. Elst, and R. Hanke, Eds., pp. 321–329, Springer, Berlin, Germany, 2011.

[9] M. Kallinger, F. Kuech, R. Schultz-Amling, G. Del Galdo, J. Ahonen, and V. Pulkki, "Enhanced direction estimation using microphone arrays for directional audio coding," in *Proceedings of the Hands-free Speech Communication and Microphone Arrays (HSCMA '08)*, pp. 45–48, Trento, Italy, May 2008.

[10] M. Cobos, J. 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*, vol. 2010, Article ID 415840, 2010.

[11] S. Rickard and Ö. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proceedings of the IEEE International Conference on Acustics, Speech, and Signal Processing*, pp. 529–532, Orlando, Fla, USA, May 2002.

[12] D. Pavlidi, M. Puigt, A. Griffin, and A. Mouchtaris, "Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '12)*, pp. 2625–2628, March 2012.

[13] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Source counting in real-time sound source localization using a circular microphone array," in *Proceedings of the Sensor Array and Multichannel Signal Processing (SAM '12)*, pp. 529–532, Hoboken, NJ, USA, June 2012.

[14] A. Griffin, D. Pavlidi, M. Puigt, and A. Mouchtaris, "Real-time multiple speaker DOA estimation in a circular microphone array based on matching pursuit," in *Proceedings of the European Signal Processing Conference (EUSIPCO '12)*, Bucharest, Romania, August 2012.

[15] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[16] B. F. Cron and C. H. Sherman, "Spatial-correlation functions for various noise models," *The Journal of the Acoustical Society of America*, vol. 34, no. 11, pp. 1732–1736, 1962.

[17] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, 2001.

[18] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, 1998.

[19] S. Leukimmiatis, D. Dimitriadis, and P. Maragos, "An optimum microphone array post-filter for speech applications," in *Proceedings of the INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing (INTERSPEECH '06) (ICSLP '06)*, pp. 2142–2145, September 2006.

[20] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[21] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2257–2269, 2007.

[22] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proceedings of the European Conference on Speech Communication (EUROSPEECH '03)*, 2003.

[23] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[24] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1429–1439, 2010.

[25] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica United with Acustica*, vol. 94, no. 6, pp. 856–865, 2008.

[26] ITU-R, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.

[27] M. V. Laitinen and V. Pulkki, "Binaural reproduction for directional audio coding," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '09)*, pp. 337–340, New Paltz, NY, USA, October 2009.

[28] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummyhead microphone," MIT Media Lab, 1994.