Author(s): Symeon Delikaris-Manias and Ville Pulkki

Title: Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays

Year: 2013

Version: Author accepted / Post print version

Please cite the original version:

Symeon Delikaris-Manias and Ville Pulkki. Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays. IEEE Transactions on Audio, Speech, and Language Processing, Volume 21, issue 11, pages 2356-2367, November 2013. DOI: 10.1109/TASL.2013.2277928

Cross Pattern Coherence Algorithm for Spatial Filtering Applications Utilizing Microphone Arrays

Symeon Delikaris-Manias, *Member, IEEE,* and Ville Pulkki

Abstract

A parametric spatial filtering algorithm with a fixed beam direction is proposed in this paper. The algorithm utilizes the normalized cross-spectral density between signals from microphones of different orders as a criterion for focusing in specific directions. The correlation between microphone signals is estimated in the time-frequency domain. A post-filter is calculated from a multichannel input and is used to assign attenuation values to a coincidentally captured audio signal. The proposed algorithm is simple to implement and offers the capability of coping with interfering sources at different azimuthal locations with or without the presence of diffuse sound. It is implemented by using directional microphones placed in the same look direction and have the same magnitude and phase response. Experiments are conducted with simulated and real microphone arrays employing the proposed post-filter and compared to previous coherence-based approaches, such as the McCowan post-filter. A significant improvement is demonstrated in terms of objective quality measures. Formal listening tests conducted to assess the audibility of artifacts of the proposed algorithm in real acoustical scenarios show that no annoying artifacts existed with certain spectral floor values. Examples of the proposed algorithm can be found online at **http://www.acoustics.hut.fi/projects/cropac/soundExamples**.

Index Terms

Array signal processing, microphone arrays, beamforming, spatial filtering, cross-pattern spectral density, coherence.

Manuscript submitted December 12, 2012. Resubmitted May 2, 2013 and July 31, 2013. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° [240453]. The Academy of Finland has supported this work.

S. Delikaris-Manias is with the Department of Signal Processing and Acoustics, School of Electrical Engineering, Aalto University, Espoo, Finland (email: symeon.delikaris-manias@aalto.fi).

V. Pulkki is with the Department of Signal Processing and Acoustics, School of Electrical Engineering, Aalto University, Espoo, Finland (email: ville.pulkki@aalto.fi)

I. Introduction

Microphone arrays allow the design of spatial filters that can focus in one specific direction while suppressing noise or interfering sources from other directions. Such spatial filtering techniques are commonly referred to as beamforming. The most basic beamforming approaches are the conventional delay-and-sum and the filter-and-sum utilizing spaced arrays. The delay-and-sum beamformer algorithm estimates the time delays of signals received by each microphone of an array and compensates for the time difference of arrival [1]. By aligning and summing the microphone input signals, the directionality of the microphone array can be adjusted in order to create a constructive interference for the desired propagating sound wave and a destructive interference for sound waves originating from all other directions. Narrow directivity patterns can be obtained, but this requires a large spacing between the microphones and a large number of microphones.

Adaptive beamforming methods have been proposed to optimally combine microphone signals from an array to minimize the level of noise while retaining the signal arriving from the desired direction. One of the well known techniques in adaptive beamforming is the Minimum Variance Distortionless Response (MVDR), where the underlying principle is to track the variation of the spatial noise field and adaptively search for the optimum location of nulls that can significantly reduce noise under the constraint that the desired signal is not distorted at the output [2]. Although MVDR beamforming provides the optimal solution, it does not provide sufficient noise reduction of diffuse noise and reverberation. To further improve the signal-to-noise ratio (SNR) for broadband input signals in a noisy environment, a Wiener filter can be added at the output [3]. Multichannel Wiener filtering, calculated from the microphone input signals, is part of a well established class of spatial filtering or signal enhancement approaches which are known as post-filtering algorithms. Post-filters usually employ coherence-based measures between microphone channels, and they are used to modulate the output of a beamformer. The premise underlying coherence-based algorithms is that audio signals between microphone channels are correlated while noise is uncorrelated. However, limitations arise when the noise signals in different microphone channels are correlated.

A multichannel post-filter based on Wiener filtering is introduced by Zelinski in [4] assuming that noise received by different microphones is uncorrelated. In this technique the output of a delay-and-sum beamformer is modulated with a post-filter based on auto- and cross-spectral densities of the omnidirectional microphone signals. For the case of correlated noise in the microphone signals, the McCowan post-filter is proposed by employing a model of the coherence for a spherically isotropic field in order to identify the correlated noise [5]. Another technique for highly correlated noise is proposed in [6] based on a generalized sidelobe canceler. Unfortunately, these methods are characterized by poor performance at low frequencies when the correlation between microphone signals is high [7].

Another related class of multichannel signal enhancement methods are the blind source separation (BSS) algorithms. The term blind refers to the fact that the source signals and the mixing system are assumed to be unknown and that the source signals are statistically independent, a condition also referred to as W-disjoint orthogonality [8]. The target of a BSS algorithm is to find a de-mixing system with statistically independent outputs. The performance

of BSS algorithms has been assessed with standard datasets [9], [10], [11]. A common set of performance measures for such methods are proposed which can also be used to evaluate post-filtering algorithms [12]. The main difference between traditional post-filtering algorithms and BSS is that post-filtering algorithms require usually a beamformer with a specific look direction. BSS do not require any prior knowledge of a look direction. Additionally, while BSS algorithms focus on separating signals from a given mixture of signals, post-filtering algorithms focus on adjusting the level of the output signal depending on the direction-of-arrival (DOA) of incoming sound. Most BSS algorithms are non real-time as the whole mixture of signals is required to be processed before applying the algorithms. The existing real-time approaches are divided into three main categories: the block-wise, step-size and combinatory approaches. Block-wise approaches apply a BSS algorithm to a set of time frames before calculating the output, resulting in a computationally expensive approach [13]. Step-size approaches apply a BSS algorithm for each incoming time frame, which is computationally efficient but less accurate [14]. Combinatory approaches are also proposed in [15], employing both block- and step-size approaches, offering a trade-off between computational complexity and performance accuracy.

In the class of time-invariant methods, a closely-spaced microphone array technique has been proposed and can be applied to beamforming [16]. In this technique, the microphone signals are summed together in the same or opposite phase with different gains and frequency equalization, where the target is microphone signals with directional patterns following the spherical harmonics of different orders. The resulting response has tolerable quality only in a limited frequency window; at low frequencies the system suffers from amplification of the self noise of microphones, and at high frequencies the directional patterns are deformed due to spatial aliasing. These beamforming techniques do not assume anything about the signals of the sources.

Recently, some techniques have been proposed, which assume that the signals arriving from different directions to the microphone array are sparse in time-frequency domain, i.e., one of the sources is dominant at one time-frequency position [17]. Each time-frequency frame is then attenuated or amplified according to spatial parameters analyzed for the corresponding time-frequency position, which leads to the formation of the beam. It is clear that such methods might produce distortion at the output; however, the assumption is that the distortion is most prominent with weakest time-frequency slots of the signals making the artifact inaudible or at least tolerable.

A microphone array consisting of two cardioid capsules in opposite directions has been proposed in [18] for such a technique. Correlation measures between the cardioid capsules and Wiener filtering are used to reduce the level of coherent sound in one of the microphone signals. This produces a directive microphone, whose beamwidth can be controlled. An inherent result is that the width varies depending on the sound field. For example, with few speech sources in relatively anechoic conditions, a prominent narrowing of the cardioid pattern is obtained. However, with many uncorrelated sources, and in a diffuse field, the method does not change the directional pattern of the cardioid microphone at all. The method is still advantageous, as the number of microphones is low, and the setup does not require a large space.

The assumption of the sparsity of the source signals is also utilized in another technique, directional audio coding (DirAC) [19], which is a method to capture, process and reproduce spatial sound over different reproduction setups.

The most prominent DOA and the diffuseness of the sound field are measured as spatial parameters for each time-frequency position of sound. The DOA is estimated as the opposite direction of the intensity vector, and the diffuseness is estimated by comparing the magnitude of the intensity vector with the total energy. In the original version of DirAC, the parameters are utilized in reproduction to enhance audio quality. A variant of DirAC has been used for beamforming [20], where each time-frequency position of sound is gained or attenuated depending on the spatial parameters and a specified spatial filter pattern. In practice, if the DOA of a time-frequency position is far from the desired direction, it is attenuated. Additionally, if the diffuseness is high, the attenuation is reduced since the DOA is considered to be less certain. However, in cases when the assumption of W-disjoint orthogonality is violated and two audio signals are active in the same time-frequency position, the analyzed DOA provides erroneous data, and artifacts may occur.

The main limitation of coherence-based methods is their weakness in suppressing noise at low frequencies, since the signals between microphone channels become highly correlated when the sensor distance is low compared to the wavelength and the calculated post-filter might be inaccurate. This motivated the current research to employ coherence-based measures between directional microphones. In this paper, we propose a post-filtering technique employing a microphone array, where the input consists of microphones having three arbitrary-order directional patterns. This technique measures the cross-spectral density in each time-frequency position between signals originating from directional microphones having the positive-phase of the maxima directivity in the desired direction. A time-variant post filter is then computed, based on the time-averaged normalized cross-spectral density. The corresponding time-frequency positions in the third modulated signal are then attenuated when the signals from the directional microphones are uncorrelated. The application of the proposed method is feasible with any order of microphone inputs, and the directional shape of the beam can be altered by changing the formation of the directional patterns of the microphones from which the post-filter is computed.

The paper is organized as follows. Section II describes the general encoding process in order to derive directional microphones from a microphone array and provides a review of coherence-based post-filters such as the Zelinski and the McCowan. In Section III the proposed approach is presented and the calculation of the post-filter metadata is derived. An objective and subjective evaluation with a simulated and a prototype array conducted to verify the performance of the algorithm in a multi-speaker scenario is discussed in Section IV, and Section V concludes the paper.

II. Background

A. Spatial Encoding Utilizing Pressure Microphone Arrays

This section reviews the derivation of directional microphones employing a microphone array. A theoretical approach to this problem has been addressed by using the spherical or cylindrical harmonic framework for matrixing the microphone signals [16], [21], and equalizing the output using regularization measures [22], [23]. The spherical and cylindrical harmonic functions are discussed in [24]. For a comprehensive analysis of modal microphone-array processing, the reader is referred to [25] and [26]. Direct synthesis of directional microphone patterns from a

set of measurements is suggested in [27]. Spatial encoding for microphone arrays employing such frameworks is associated with errors caused by capsule misalignment and/or capsule mismatch, and thus, in the present study, a least squares approach employing the Fourier series is preferred.

The least squares approach is a common approach for antenna radiation pattern synthesis [28] and has been used for synthesizing directional microphones from arbitrary microphone arrays [29]. Although the directional pattern synthesis is shown in this section for the two-dimensional problem, the three-dimensional problem is a straightforward extension.

A directional microphone pattern $B(\phi, f)$ for azimuth angle $\phi \in [0, 360)$ and frequency f can be expressed as a weighted sum of microphone input signals $X_n(\phi, f)$:

$$B(\phi, f) = \sum_{n=1}^{N} w_n(f) X_n(\phi, f), \tag{1}$$

where $X_n(\phi, f)$ is the frequency response of the n^{th} microphone signal at angle ϕ of an arbitrary microphone array for $n = 1, \ldots, N$, with N being the total number of microphones and $w_n(f)$ are the frequency-dependent weights. The set of weights w is applied to each microphone to approximate the directional pattern B. Employing the Fourier series, limited to a number of harmonics U, both the target directional pattern B and the microphone input signals X_n can be be decomposed into

$$B(\phi, f) = \sum_{u=-U}^{+U} b_u(f) e^{iu\phi}, \tag{2}$$

$$X_n(\phi, f) = \sum_{u=-U}^{+U} x_{un}(f) e^{iu\phi}, \tag{3}$$

where $b_u(f)$ and $x_{un}(f)$ are the complex Fourier coefficients. Substituting (2) and (3) in (1) give

$$\sum_{n=1}^{N} x_{un}(f) w_n(f) \quad = \quad b_u(f), \tag{4}$$

and in matrix form

$$\mathbf{x}(f)\mathbf{w}(f) = \mathbf{b}(f), \tag{5}$$

where $\mathbf{x}(f)$ is a $(2U + 1) \times N$ matrix, $\mathbf{w}(f)$ a $N \times 1$ vector and $\mathbf{b}(f)$ a $(2U + 1) \times 1$ vector. The weights for a given directional pattern $B(\phi, f)$ can be calculated from

$$\mathbf{w}(f) = x^{+}(f)\mathbf{b}(f), \tag{6}$$

where $\mathbf{x}^{+}(f)$ is the Moore-Penrose inverse matrix of $\mathbf{x}(f)$.

A general frequency-domain framework for deriving the weights that can be applied to a microphone array and to obtain a desired microphone directional pattern is shown in (6). The resulting microphone directional pattern for an arbitrary microphone array with signals $X_n(f)$ is

$$S_p^{\sigma}(f) = \sum_{n=1}^{N} w_{p\,n}^{\sigma}(f) X_n(f) \tag{7}$$

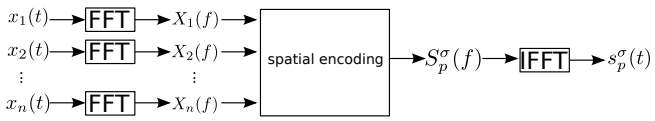

Fig. 1. Encoding process to obtain the directional microphone signal $s_p^\sigma(t)$ from a microphone array.

where the subscript $p = 0, 1, \ldots, M$ denotes the order of the directional pattern and the superscript σ the orientation. For $\sigma = 1$ the directional pattern is in its original orientation while for $\sigma = -1$ the pattern is rotated $90°$. An inverse FFT is then applied to obtain the time-domain signal $s_p^\sigma(t)$. The encoding process is shown in Fig. 1.

To demonstrate the performance of this approach, a virtual cylindrical array consisting of five equidistant sensors at $0.03\,\text{m}$ radius is employed. The target pattern is a second-order directional pattern S_2^1. The set of weights is calculated from (6) for each microphone, and (7) provides the resulting directional microphone. In Fig. 2, the magnitude of the resulting pattern is shown with the corresponding total number of Fourier coefficients used per frequency. Spatial aliasing is evident at approximately $8\,\text{kHz}$ and is due to the radius of the array. For the case of a real microphone array, depending on the internal microphone noise levels, the spatial encoding results in low-frequency noise amplification.

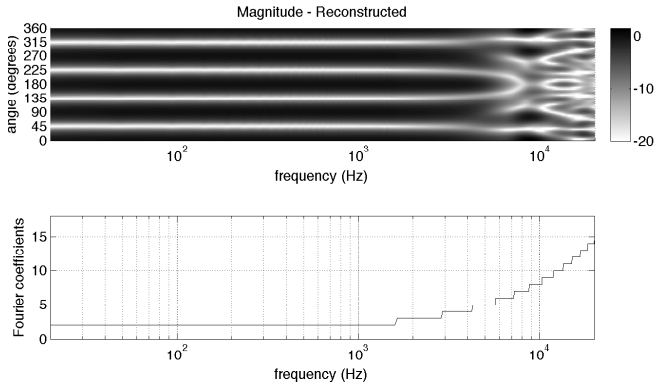

Fig. 2. Magnitude of a reconstructed second-order directional pattern from the virtual microphone array (top) and the number of Fourier coefficients U employed per frequency (bottom).

B. Time-Frequency Processing

In this work, the directional microphone signals obtained from a microphone array are transformed to the time-frequency domain through a short-time Fourier transform (STFT). Given a directional microphone signal $s_p^\sigma(t)$, the corresponding complex time-frequency representation is denoted as $S_p^\sigma(k, i)$, where k is the frequency frame and i the time frame.

C. Coherence-Based Post-Filtering Algorithms

The complex coherence for two microphone input signals $X_i(k,i)$ and $X_j(k,i)$ is defined as

$$\Gamma_{ij}(k,i) = \frac{\Phi_{ij}(k,i)}{\sqrt{\Phi_{ii}(k,i)\Phi_{jj}(k,i))}}, \tag{8}$$

where $\Phi_{ij}(k,i) = E[X_i^*(k,i)X_j(k,i)]$ is the cross-spectral density and $\Phi_{ii}(k,i) = E[|X_i^2(k,i)|]$ the auto-spectral density, E an expectation operator, and $*$ denotes complex conjugation. An overview on the coherence estimation is found in [30]. The absolute value of the coherence function $|\Gamma_{ij}|$ is bounded in $[0,1]$ and is a measure of similarity between signals at two discrete points in a noise field. In a diffuse noise field, coherence is real valued and can be modeled as

$$\Gamma_{ij}(k,i) = \mathrm{si}\left(\frac{2\pi f_\mathrm{s} d_{ij}}{c}\right), \tag{9}$$

where d_{ij} is the microphone spacing, f_s the sampling frequency, and c the speed of sound. One can see from (9) that the values of the modeled coherence converges to 1 as the sensor distance decreases.

The magnitude of coherence has been employed previously to indicate the activity of a target signal at each time-frequency frame. One of the earlier coherence-based approaches was introduced by Zelinski in [4] under two assumptions: that noise and the desired signal are uncorrelated and that the microphone signals are also uncorrelated in a perfectly incoherent noise field. The principle of operation is based on post-filtering a delay-and-sum beamformer output with a Wiener filter based on the estimated auto- and power-spectra densities of each microphone signal. The post-filter is given by

$$G_\mathrm{ze}(k,i) = \frac{\frac{2}{N(N-1)}\sum_{n=1}^{N-1}\sum_{m=n+1}^{N}\Re[\Phi_{X_nX_m}(k,i)]}{\frac{1}{N}\sum_{n=1}^{N}[\Phi_{X_nX_n}(k,i)]}, \tag{10}$$

where $\Phi_{X_nX_n}$ is the auto-spectral density of each microphone signal X_n, $\Phi_{X_nX_m}$ is the cross-spectral density of microphone signals X_n and X_m, and $\Re$ denotes the real part operator.

An extension of the post-filter in (10) was presented by McCowan in [5] based on the fact that in a practical application, diffuse noise received from each pressure microphone is correlated through a complex coherence function. The McCowan post-filter is calculated as

$$G_\mathrm{mc}(k,i) = \frac{\frac{2}{N(N-1)}\sum_{n=1}^{N-1}\sum_{m=n+1}^{N}\Phi_{X_{ss}^{nm}}(k,i)}{\frac{1}{N}\sum_{n=1}^{N}[\Phi_{X_nX_n}(k,i)]}, \tag{11}$$

where

$$\Phi_{X_{ss}^{nm}}(k,i) = \frac{\Re[\Phi_{X_nX_n}(k,i)] - 0.5\Re[\hat{\Gamma}_{d_nd_m}\Phi_{X_nX_n}(k,i)]}{1-\Re[\hat{\Gamma}_{d_nd_m}]}, \tag{12}$$

and $\Gamma_{d_nd_m}$ is the complex coherence. The Zelinski algorithm provides a post-filter without taking into consideration that noise received by microphone signals are correlated in a diffuse sound field at low frequencies. Although the McCowan algorithm provides a more accurate estimate by introducing the complex coherence function and an improved performance over Zelinski's algorithm, its performance depends on the accurate estimation of the coherence function.

III. Cross Pattern Coherence (CroPaC) Algorithm

In this study we propose the use of a coherence-based post-filter, computed between signals of higher-order directional microphones, that can be used to focus in the direction of a target signal while attenuating signals from other directions. The main idea behind this approach is that the cross-spectral density between two signals, captured by microphones of different orders, achieves its maximum value when the directional patterns of the microphones have equal phase and high sensitivity in the desired direction. In other words, a plane-wave signal is captured coherently by such directional microphones only when the DOA of the plane wave coincides with that direction. In all other cases the cross-spectral density between the signals is reduced. Due to the directional characteristics of higher order microphones, such a post-filter obtains low values also in the low frequency region in a diffuse sound field, which addresses the main drawback of the Zelinski and McCowan post-filters, presented in Section II-C.

A. Proposed Algorithm

The initial step in the proposed algorithm is to compute the cross-spectral density Φ_{pq} between two directional microphone signals of different orders p and q:

$$\Phi_{pq}(k,i) = E[S_p^{1^*}(k,i) S_q^1(k,i)], \tag{13}$$

where $S_p^1(k,i)$ and $S_q^1(k,i)$ are the time-frequency representation of the signals from microphones with directional patterns of different order p and q that are in the same look direction. While in the McCowan algorithm the microphone signals are typically scaled and aligned before the calculation of the post-filter, in the present case this is not necessary as the directional microphones are coincident.

From (13), it is clear that Φ_{pq} depends on the magnitudes of the microphone signals, which is not desired as the post-filter should depend only on the DOA of sound. This is circumvented by normalizing Φ_{pq}:

$$G(k,i) = \frac{2\Re[\Phi_{pq}(k,i)]}{\sum_{\sigma=}^{1,-1} \Phi_{pp}^{\sigma}(k,i) + \sum_{\sigma=}^{1,-1} \Phi_{qq}^{\sigma}}, \tag{14}$$

where $\Phi_{pp}^{\sigma}(k,i) = E[|(S_p^{\sigma})^2(k,i)|]$ and $\Phi_{qq}^{\sigma}(k,i) = E[|(S_q^{\sigma})^2(k,i)|]$ are the auto-power spectral densities of the microphones S_p^{σ} and S_q^{σ} with directional patterns selected such that

$$\sum_{\sigma=}^{1,-1} S_p^{\sigma}(k,i) = \sum_{\sigma=}^{1,-1} S_q^{\sigma}(k,i) = S_0(k,i), \tag{15}$$

where S_0 is a signal from a microphone with omnidirectional characteristics and should be satisfied for all plane waves with DOA of $\phi \in [0, 360)$. The normalization process in (14) ensures that with all inputs the calculated post-filter value is bound in the interval $[-1, 1]$, and that values near unity are obtained only when the signals $S_p^1(k,i)$ and $S_q^1(k,i)$ are equivalent in both phase and magnitude. In this study, G is a normalized cross-pattern spectral density and it is referred as the Cross-Pattern Coherence (CroPaC) post filter.

B. Half-Wave Rectification

Following the definition of the magnitude square coherence function [30], normalized so that its value is in $[0, 1]$, a similar scheme is adapted for the normalized cross-spectral density in (14). As only the G values near unity imply that there is sound arriving from the look direction, the values that are below zero indicate that the sound of the analyzed time-frequency frame does not originate from the look direction. By taking this into consideration, a rectification process can be used. Waveform rectification has been expressed in [31] as

$$G_{\mathrm{r}}(k,i) = \frac{(1+\beta)|G(k,i)| + (1-\beta)G(k,i)}{2}. \tag{16}$$

For $\beta = 0$, (16) corresponds to half-wave rectification and ensures that only non-negative values are used. In particular, the part of the lobe that is chosen results in a unique beamformer in the look direction.

So far we have introduced an attenuation value G_{r} that can be used to synthesize the output signal of the proposed spatial filtering technique. The output signal is computed by multiplying the half-wave rectified post-filter $G_{\mathrm{r}}(k,i)$ and a microphone signal $S_p^{\sigma}(k,i)$.

C. Temporal Averaging

The value of G_{r} is calculated according to the cross-spectral densities between microphone signals for each time frequency frame. In a real recording scenario, the levels of sound sources with different directions of arrival may fluctuate rapidly and result in rapid changes in the calculated values of G_{r}. By modulating a directional input signal $S_p^{\sigma}(k,i)$ with the post-filter $G_{\mathrm{r}}(k,i)$, clearly audible artifacts are produced. The main cause is the relatively fast fluctuations of the post-filter estimates which introduces a high variance in the G values in the interval $[0, 1]$ at each time-frequency frame. The specific artifact is referred to as the bubbling effect or musical noise. Similar effects have been reported in adaptive feedback cancellation processors used in hearing aids [32], [33], and intensity-based spatial filtering techiques [34].

In order to mitigate these artifacts, additional temporal averaging is performed in the post-filter G_{r}. This type of averaging or smoothing, which is essentially a single-pole recursive filter, is defined as

$$\hat{G}(k,i) = \alpha(k)G_r(k,i) - (1-\alpha(k))\hat{G}(k,i-1), \tag{17}$$

where $\hat{G}(k,i)$ are the smoothed gain coefficients for frequency k and time i and $\alpha(k)$ are the smoothing coefficients for each frequency k.

D. Spectral Floor

In real acoustical conditions with one and many talkers the fluctuations of $\hat{G}$ may vary significantly especially in the presence of background noise. In spite of the time averaging process, these fluctuation may still produce audible musical noise. The use of a spectral floor has been used in speech enhancement applications to overcome such artifacts when noise is present [35]. Therefore, a lower bound λ is imposed on $\hat{G}$ to prevent the resulting

values from reaching below a certain level:

$$\hat{G}^{+}(k,i) = \begin{cases} \hat{G}(k,i), & \text{if } \hat{G}(k,i) \geq \lambda, \\ \lambda, & \text{if } \hat{G}(k,i) < \lambda. \end{cases} \tag{18}$$

The spectral floor λ of the derived parameter $\hat{G}^{+}$ can be adjusted according to the application, and it is a trade-off between the effectiveness of the spatial filtering method and the preservation of the quality of the output signal. The effect of λ on the annoyance caused by the artifacts is shown later in this study in Section IV-C.

E. Synthesis of Output Signal

The output Y of the CroPaC algorithm is

$$Y(k,i) = \hat{G}^{+}(k,i) S_p^{\sigma}(k,i), \tag{19}$$

in which an inverse STFT (ISTFT) is applied to obtain the time-domain signal $y(t)$.

The signal $S_p^{\sigma}(k,i)$ being selectively attenuated by the single channel post-filter $\hat{G}^{+}(k,i)$, calculated from a multichannel input, should originate from a microphone with directional characteristics of a low-order and a spectrally flat response, not suffering from amplified low-frequency noise. In practice, when the microphone array allows decoding higher-order microphones up to order $p \geq 2$, S_{p-2}^{σ} should be selected, where $p-2$ is the order, and G should be computed with signals S_{p-1}^{σ} and S_p^{σ}. In this way, the higher-order microphones S_{p-1}^{σ} and S_p^{σ} result in better spatial resolution of the output Y without introducing audible noise. Depending on the internal noise level of the microphones, the low-frequency noise in higher-order microphones might produce some erroneous analysis in the computation of the post-filter, but the temporal averaging in (17) mitigates the effect. An exemplary solution for the signal to be modulated is to use the zeroth-order microphone S_0 for this purpose, as available pressure microphones typically have a flat magnitude response with a tolerable noise level. The output of a super-directive beamformer, such as the MVDR under the constraint of white noise gain (WNG), can also be modulated with the proposed post-filter. The constraint of WNG ensures that low-frequency noise amplification is not boosted in cases of uncorrelated noise in the microphones by dynamically adjusting the sensor noise level [36].

IV. Performance Evaluation

In this section we demonstrate the performance of the CroPaC post-filter in various scenarios. At first, an ideal virtual microphone array is employed to illustrate the performance in optimal conditions. The second part describes a real microphone array built to illustrate the directivity of the beamformer and the performance of the algorithm in real acoustical scenarios. Objective criteria are employed to compare the proposed algorithm with previous coherence-based approaches. The last part of the evaluation discusses listening tests performed to show the effect of the spectral floor of CroPaC in various real acoustical scenarios.

A. Optimal Conditions: An Ideal Virtual Microphone Array

The performance of the CroPaC algorithm is demonstrated by deriving the directional attenuation patterns in different sound scenarios in ideal conditions. A similar method for assessing the performance of a real-weighted beamformer has been used in [37] by employing the ratio of the power of the beamformer output in the steering direction to the average power of the system. The directional patterns in this case are derived by steering the beamformer every 5° and calculating the $\hat{G}^+$ value for each position while maintaining the sound sources at their initial position. In this example, a scenario with single and multiple sound sources has been simulated. Sound sources with and without background noise and different SNRs are positioned at various angles around an ideal virtual microphone array. Figs. 3 and 4 show the directional patterns of the algorithm for the various cases.

In Fig. 3, a single sound source is positioned at 0° with added diffuse noise. The diffuse noise has been generated with 73 noise sources positioned equidistantly around the virtual microphone array. The directional pattern shows the performance of the beamformer under different SNR values for the single sound source and the sum of the noise sources. When the beam is steered towards the target source at 0°, the attenuation is $4\,$dB with an SNR of $20\,$dB. As the beam is steered away from the target source, there is a noticeable attenuation for angles of $\pm 30^\circ$ or more which reaches $12\,$dB at $\pm 60^\circ$. Outside the sector of $\pm 60^\circ$ the attenuation level varies between 15 to $19\,$dB. With an SNR of $10\,$dB, the beamformer assigns a value of $-10\,$dB and attenuates the output to $18\,$dB outside the sector of $\pm 30^\circ$. For lower SNR values, from 0 to $-\infty$, in diffuse field conditions the beamformer assigns a uniform attenuation of $18\,$dB for all directions. This part of the simulation thus suggests that in diffuse conditions the SNR has to be 10–20 dB in a given time-frequency frame for CroPaC to be effective.

The directional attenuation patterns in double sound source scenarios are illustrated in Fig. 4 (a), (b) and (c). The main sound source is positioned at 0° and the interferer is positioned at 60°, 120° and 180° respectively. The patterns are calculated for different SNRs for the main and interfering sources. With SNR values of 20 and $10\,$dB the beamformer provides an attenuation of 1–3 dB when it is steered towards the main sound source and an attenuation greater than $10\,$dB for all other directions. When the level of the two sound sources is equal the attenuation is between $4\,$dB and $12\,$dB when the beam is steered towards the source in Fig. 4 (a), (b) and (c) and less than $5\,$dB for all other directions. In this case the attenuation should be $3\,$dB since the signal S_0 contains both uncorrelated signals at equal levels.

In the multiple-talker scenario in Fig. 4 (d), three sound sources are present simultaneously with the target source at 0° and two interferers at 90° and 180°. The level provided by the beamformer is approximately the same as in the two sound source scenario for all beam directions and for the SNR of 20 and $10\,$dB. As expected from the previous cases in Fig. 4 (a), (b) and (c), when all sources receive the same level, the attenuation level that the beamformer applies is much lower: $10\,$dB for 0°, $11\,$dB for -90° and $18\,$dB for 180°.

It is thus evident that in the case of one or two interfering sources the performance of CroPaC is consistent and provides consistent filtering results, not only for the cases of high SNR (20 and $10\,$dB), but also for some cases where the SNR is $0\,$dB. The advantages shown through this simulation are that the algorithm provides a high

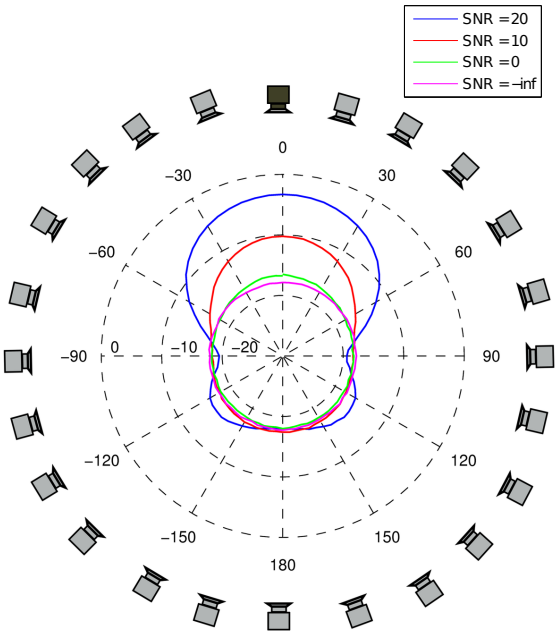

Fig. 3. Directional attenuation patterns $\hat{G}^{+}$ of the CroPaC algorithm with a single source and diffuse noise in dB. The directional attenuation pattern is calculated under different SNRs from the sound source and the sum of the noise sources for all beam directions. Grey loudspeakers indicate the diffuse noise sources and dark green the signal source.

response when the direction of the beamformer coincides with the direction of a sound source. This is evident through the calculation of $\hat{G}^{+}$ for the diffuse field case with positive SNR values. For the SNRs of 20 and 10 dB in a single or multi sound source scenario, the $\hat{G}^{+}$ values towards the direction of the main sound source differ from the original level by 1–2 dB. It is also evident that in all cases there is low response towards any direction where there is no sound source, even in the case of diffuse noise only.

If speech signals are considered as sound sources, due to the sparsity and the varying nature of speech, the spectrum of the two speech signals when added can be approximated by the maximum of the two individual spectra in each time-frequency frame. It is then unlikely that two speech signals carry significant energy in the same time-frequency frame [38]. Hence, the $\hat{G}^{+}$ post-filter values will be calculated accurately for the steered direction, which motivates the use of the CroPaC algorithm in teleconferencing applications. In other words, for simultaneous talkers the resulting directivity of the CroPaC algorithm can be assumed to fall into case (a) in Fig. 4.

B. Suboptimal Conditions: A Real Microphone Array

An eight-microphone, rigid body, cylindrical array of radius 1.3 cm and height 16 cm is employed with sensors placed equidistantly in the horizontal plane every $45°$. The microphones are mounted on the perimeter at the

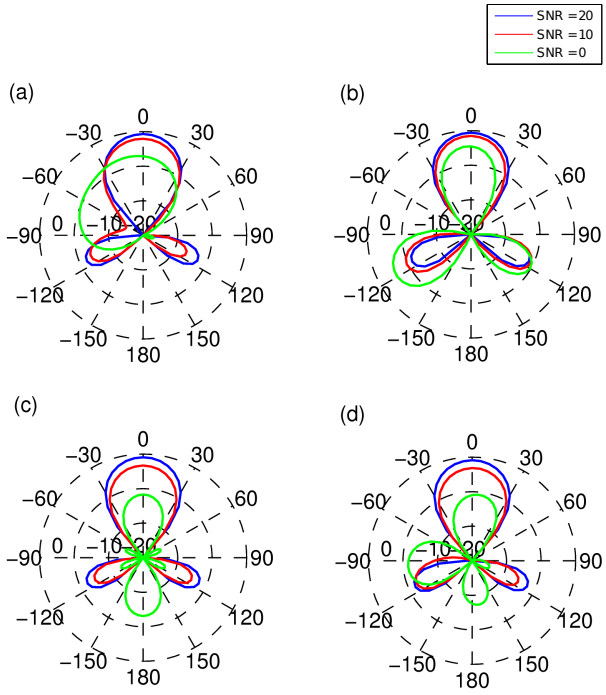

Fig. 4. Directional attenuation patterns of $\hat{G}^{+}$ in dB of the CroPaC algorithm with (a) a single sound source at 0° and an interfering source at -60°, (b) a sound source at 0° and an interfering source at -120°, (c) a sound source at 0° and an interfering source at 180°, and (d) a sound source at 0° and two interfering sources at -90° and 180°. The directional attenuation is calculated, under different SNRs for the sound source and the interfering sources for all beam directions with static sources.

half-height of the rigid cylinder. Although only five sensors are required in theory in a unified circular array to deliver microphone components of the 2^{nd} order, the additional sensors provide an increased aliasing frequency as compared to an array having the same radius with fewer sensors.

1) Directional Characteristics: Directivity measurements were performed in an anechoic environment to show the performance of the CroPaC algorithm utilizing the cylindrical microphone array with first- and second-order microphones. White noise of duration $2\,\mathrm{s}$ was used as a stimulus signal. The stimulus was fed to a single loudspeaker and the array was placed $1.5\,\mathrm{m}$ away from the loudspeaker, mounted on a turntable able to perform consecutive rotations of 5°. One measurement was performed for each angle. Each set of measurements was transformed into the time-frequency domain with an STFT, and the post-filter $\hat{G}^{+}$ values were calculated for each rotation angle with static sources. This way the directional characteristics were obtained in this setting. Figs. 5 and 6 show the resulting directional characteristics of the post-filter in the horizontal and vertical plane as a result of the combination of first- and second-order microphone inputs. The directional characteristics can be adjusted by choosing different combinations of directional microphones.

A consistent directivity is obtained in the horizontal plane where the $\hat{G}^{+}$ function is constant in the frequency

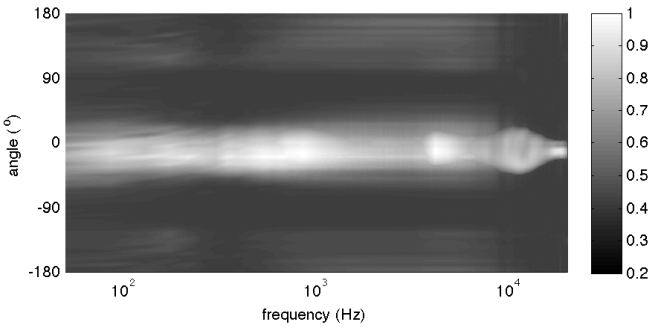

Fig. 5. Directional pattern of the horizontal beamformer in the horizontal plane.

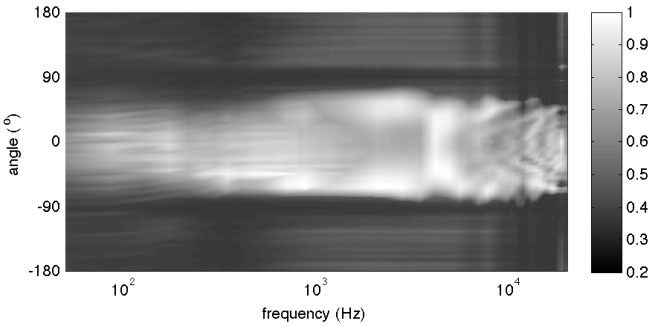

Fig. 6. Directional pattern of the horizontal beamformer in the vertical plane.

range from $50\,\text{Hz}$ to $14\,\text{kHz}$ which is approximately the spatial aliasing frequency for the cylindrical microphone array. The beamformer receives a constant $\hat{G}^+$ value in the horizontal plane in the look direction of $0°$ with an angle span of approximately $\pm 20°$. In the vertical plane, the CroPaC algorithm is capable of delivering valid $\hat{G}^+$ values for elevated sources that are not in the same plane as the microphone of the array. The maximum angle span where the beamformer provides high $\hat{G}^+$ values is $\pm 50°$ in elevation, in which a noticeable spectral coloration is visible for directions between $[20°, 50°]$ and $[-20°, -50°]$ due to the frequency dependent $\hat{G}^+$ values.

2) Attenuation Values: The CroPaC algorithm is now derived for the typical case of the cylindrical microphone array, from which the zeroth (S_0), first (S_1^1 and S_1^{-1}) and second order (S_2^1 and S_2^{-1}) microphones are encoded. The flow diagram is shown in Fig. 7.

The encoding equations to derive the directional microphones are calculated using (7). The temporal averaging coefficient α is frequency dependent and varies between 0.1 and 0.4. The lower values result in a higher average and are used for low frequencies and the higher values, which indicate less average, are used for the high frequencies. Example values for the frequency-dependent averaging coefficient are found in [39] for applause input signals and can be further optimized to suit the input signals. The spectral floor is set to $\lambda = 0.2$.

The array is positioned in the center of a room with a measured reverberation time of $\text{RT}_{60} = 500\,\text{ms}$, mounted

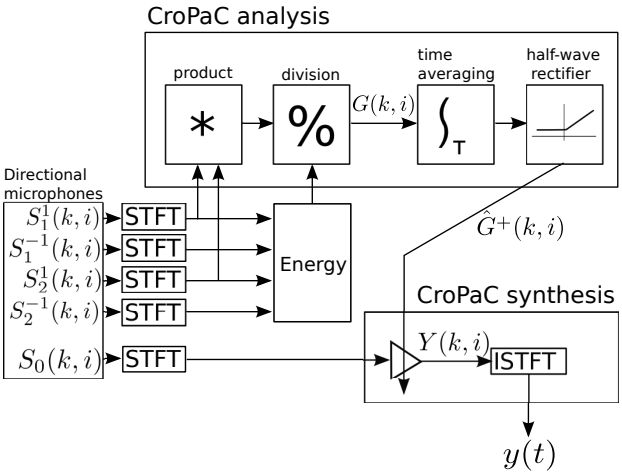

Fig. 7. Block diagram of the CroPaC algorithm implemented with zeroth (S_0), first (S_1^1, S_1^{-1}), and second (S_2^1, S_2^{-1}) order microphone signals.

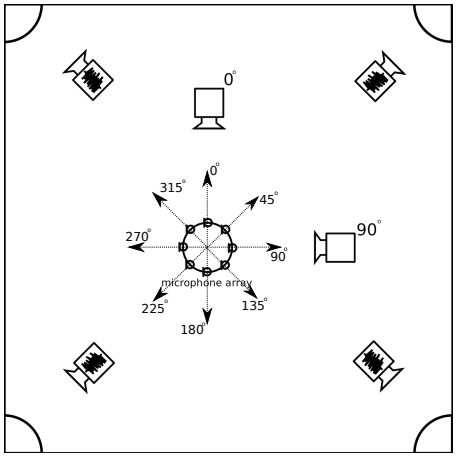

Fig. 8. Arrangement of the measurement system. The microphone array steers a full circle in 8 directions every $45°$ detecting sound from each direction. The active sources are two speakers at $0°$ and $90°$ and additional background noise.

on a tripod. The sound field scenario consisted of two loudspeaker placed at $0°$ and $90°$ in the azimuthal plane, 1.5 m away from the microphone array, transmitting speech signals simultaneously. The background noise in the room was mainly from a computer and air conditioning noise.

The attenuation values of CroPaC in this multi-speaker scenario are shown in Fig. 9. Eight different $\hat{G}^+$ values are calculated for different beam directions ($0°$, $45°$, $90°$, $135°$, $180°$, $225°$, $270°$ and $315°$). The CroPaC post-filter assigns attenuation values to each direction according to whether there is signal activity at that specific angle. This signal activity is indicated correctly at $0°$ and $90°$.

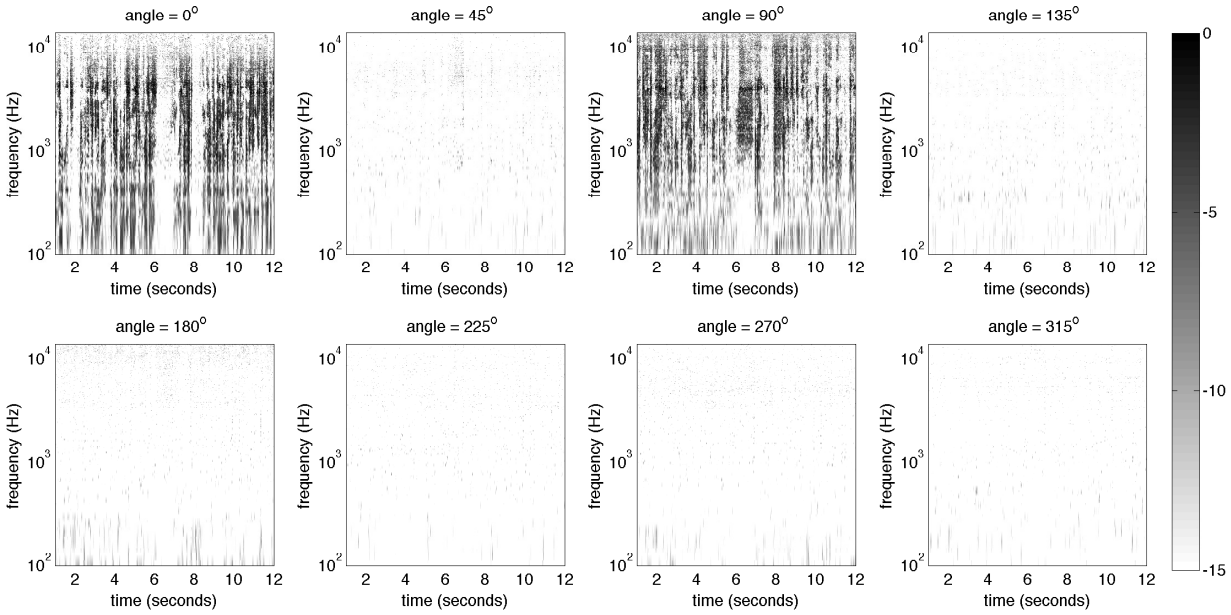

Fig. 9. $\hat{G}+$ post filter values in dB for 8 different directions every 45° in a real life multi-speaker senario with two active speakers and background noise.

3) Objective Quality Measures: The performance of the CroPaC algorithm is accessed in the aforementioned real acoustic conditions using objective quality measures. Additional background noise is generated with four loudspeakers placed at the corners of the moderately reverberant room facing towards diffusers to create a diffuse sound field. The levels of the additional background noise were of 10, 0 and -10 dB. The setup is shown in Fig. 8. The CroPaC post-filter is calculated for each time frequency frame (k, i) using first- and second-order microphones. The beamformer output and the McCowan post filter were also calculated and assessed in comparison with CroPaC. The Zelinski algorithm results are omitted as its performance has been found to be degraded when compared to the McCowan in a real acoustical scenario [5]. In Fig. 10, the waveforms of the different scenarios are shown for SNR of 0 dB. The same results are plotted as spectrograms in Fig. 11 by using a window size of 1024 samples and a hop size of 512 samples at a sampling frequency of 48 kHz with the frequency scaling set as logarithmic to highlight the differences in performance at low frequencies. The introduction of real higher-order microphones reduces the in-between correlation at low frequencies. This is evident in the spectrograms where the structure of the residual noise is apparent. The black background noise in the single microphone input and the beamformer is suppressed by the McCowan post-filter. However, the CroPaC output provides a greater suppression in the low-frequency region.

Two objective measures are employed to evaluate the performance of CroPaC and are compared to those of the McCowan post-filter: the frequency-weighted segmental SNR enhancement (segSNRE) and the Mel-frequency Cepstrum coefficients (MFCC) distance. The segSNRE is defined as the difference in segSNR between the enhanced

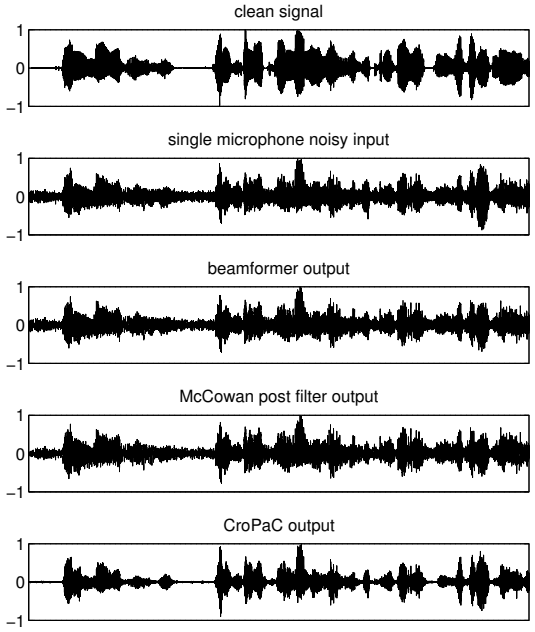

Fig. 10. Signal waveforms for unprocessed and processed signals.

output and the noisy input and utilizes frequency weighting derived from psychoacoustic properties of human hearing [40]. The signal cancellation aspect is evaluated by employing the concept of shadow filtering: the output signal is calculated by applying the same post filter to both the clean and mixed signal. This results in two available output signals: the processed clean and the enhanced output signal. The MFCC distance is then computed between these two signals. Lower values of MFCC indicate lower speech distortion [41].

In Table I the performance of CroPaC post-filter is shown for different spectral floor values and compared to the McCowan post-filter. When the McCowan post-filter shows a segSNRE of 8.3 and $6.6\,$dB for SNR of 10 and $0\,$dB, the CroPaC post-filter indicates a better segSNRE of 9.1 and $8.4\,$dB for the same SNR values and $\lambda = 0$. Higher spectral floor values provide a segSNRE up to 11.6 and $10.6\,$dB for SNR of 10 and $0\,$dB. Due to the relatively small size of the array of radius $1.3\,$cm, noise between microphones is highly correlated, which becomes evident in the performance of the McCowan post-filter. The CroPaC post-filter provides an improvement varying between 2.3 and $5.5\,$dB even at the very low SNR values of $-10\,$dB.

The results of the MFCC distance are shown in Table II. For $\lambda = 0$, the MFCC distance for the CroPaC is slightly higher than the McCowan post-filter due to the musical noise artifact present in this setting as discussed in Section III-D. For higher spectral floor values and all noise conditions, it is apparent that the CroPaC algorithm achieves lower MFCC distance values verifying the mitigation of the artifact as also discussed in Section III-D. A

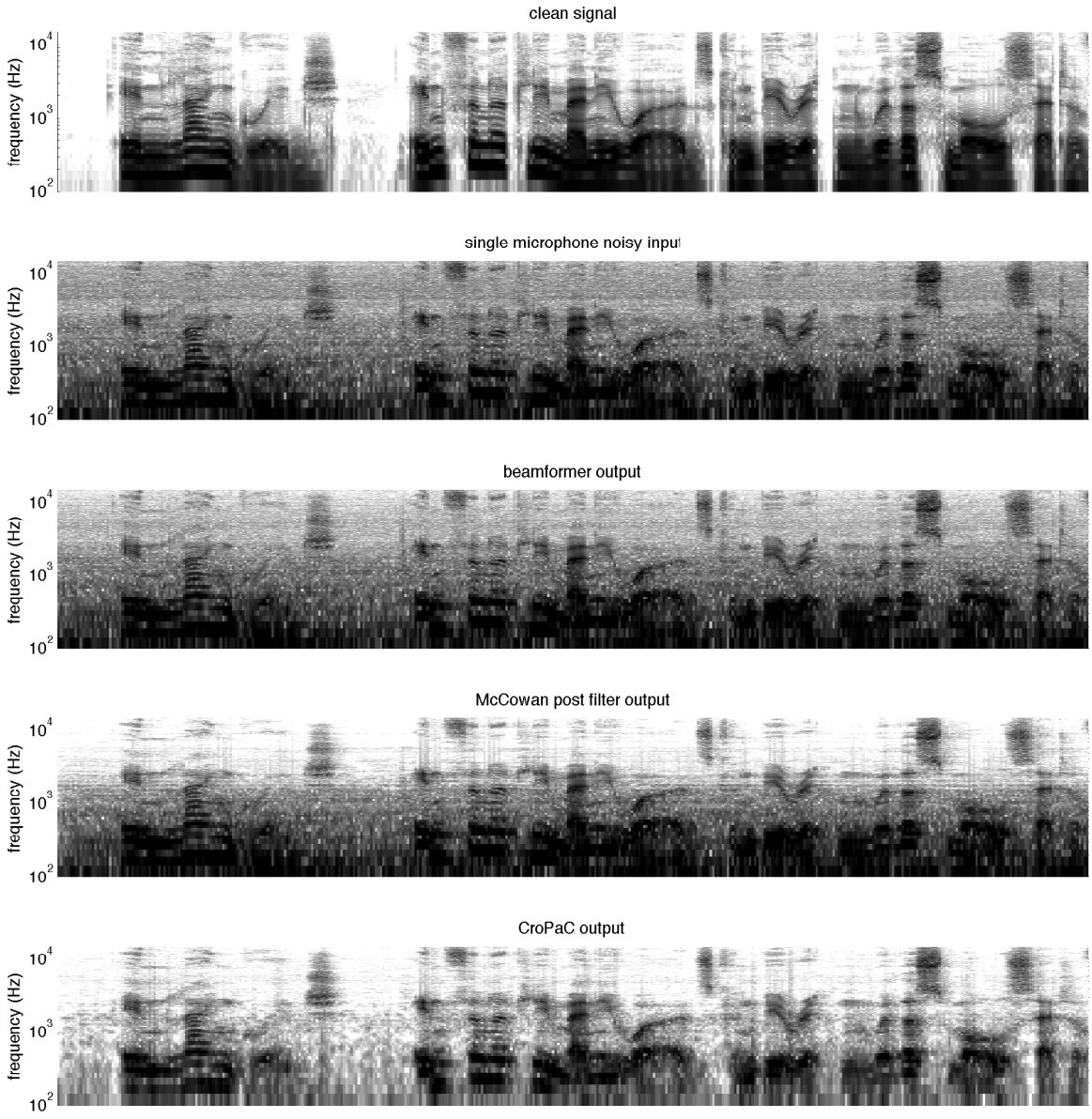

Fig. 11. Signal spectrograms for unprocessed and processed signals.

decrease in distance is evident when the spectral floor increases. The reason for this decrease is that higher spectral floor values result in a post-filter that attenuates interfering sources and noise less.

TABLE I
SEGSNRE RESULTS

	segSNRE		
Signal	10 dB	0 dB	−10 dB
McCowan	8.3	6.6	0.5
CroPaC ($\lambda = 0$)	9.1	8.4	2.3
CroPaC ($\lambda = 0.1$)	10	9.2	4.7
CroPaC ($\lambda = 0.2$)	11.4	10.3	5.3
CroPaC ($\lambda = 0.3$)	11.6	10.6	5.5

TABLE II
MFCC DISTANCE RESULTS

	MFCC distance		
Signal	10 dB	0 dB	−10 dB
McCowan	1.2	1.4	1.7
CroPaC ($\lambda = 0$)	2.1	2.3	2.8
CroPaC ($\lambda = 0.1$)	0.9	0.6	0.3
CroPaC ($\lambda = 0.2$)	0.4	0.3	0.02
CroPaC ($\lambda = 0.3$)	0.2	0.1	0.01

C. Listening Tests

The level of the spectral floor in the proposed post-filter is a trade-off between the effectiveness of the algorithm in terms of spatial filtering and the level of audible musical noise. The performance of the proposed post-filter was evaluated objectively in the previous sections for different spectral floor values. Additionally, listening tests were conducted to determine the level of the spectral floor that causes little or no annoyance due to artifacts in the output.

The listening tests were conducted in a listening room following the [ITS-R BS.1116.1 1997] recommendation and by using loudspeaker reproduction. During the test, each subject was positioned 2 m in front of a pair of loudspeakers in stereophonic arrangement, reproducing identical sound signals. Twelve volunteers, not including the authors, of ages between 25 and 35 years, all with earlier experience in listening tests and familiar with musical noise artifacts participated in the listening test.

The source signals used in the listening test were processed recordings with the eight real microphone cylindrical array in a reverberant room having a measured reverberation time of $RT_{60} = 500$ ms. Five different acoustical scenarios, consisting of a single or multiple speakers with different levels of background noise, were recorded. In particular, the single talker was positioned at 0°, the talkers in the dual-talker scenario were at 0° and 60°, 0° and 120° and 0° and 180°; and the positions in the three-talkers scenario were 0°, 90° and 180°. The background noise levels were 20, 10, 5, 0 and -10 dB. The recordings were processed with the CroPaC algorithm where the directional microphones were pointing at 0°. Five different values for the spectral floor were used: $\lambda_0 = 0$, $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, $\lambda_3 = 0.3$ and $\lambda_4 = 0.5$.

The subjective evaluation of the proposed algorithm was based on a multi-stimulus test with a hidden reference. The hidden reference was one of the unprocessed signals from the microphone array. The samples lasted from 15–20 seconds and were looped. The subject was asked to rate the level of audible artifacts in the recordings of 1–3 voices and a continuous noise source. The artifacts were explained in text to the subjects as "something audible in the reproduction which does not appear in a usual recording but is a result of a processing algorithm". A high rating (≤ 100) is given for the case of inaudible artifacts and a low one (≥ 0) for cases of audible artifacts.

A statistical analysis of the results was performed in SPSS [42], based on a repeated measures analysis of variance (RM-ANOVA) with the factors being the recording (single or multiple talkers), the SNR (different background levels) and the spectral floor λ. The assumptions of RM-ANOVA were tested with Mauchly's test and the results revealed that the assumption of sphericity was violated in the factors of recording $\chi^2(9) = 22.014$, $p < 0.05$, $\epsilon > 0.75$ and spectral floor $\chi^2(14) = 83.951$, $p < 0.05$, $\epsilon < 0.75$. Two types of corrections were used in further analysis: for $\epsilon < 0.75$, the Greenhouse-Geisser method was used and for $\epsilon > 0.75$ the Huynh-Feldt. The RM-ANOVA results are shown in Table III with all factors being significant with 95% confidence.

TABLE III
RM-ANOVA RESULTS

Source	F	p
recording	$F(3.173, 38.081) = 3.651$	0.019
SNR	$F(4, 48) = 55.757$	< 0.000
λ	$F(1.768, 21, 215) = 194.905$	< 0.000
recording*SNR	$F(16, 192) = 6.307$	< 0.000
recording*λ	$F(20, 240) = 8.247$	< 0.000
SNR*λ	$F(20, 240) = 20.074$	< 0.000
recording*SNR*λ	$F(80, 960) = 4.983$	< 0.000

The third-order interaction (recording*SNR*λ) is studied first. Further inspection of the specific interaction revealed that the cause of significance in this case was the high rating of 85 for the single talker recording with an SNR of 20 dB and $\lambda = 0$. The modulation caused in this case was not audible due to the high level of the SNR and the absence of interfering talkers. In all other cases where the SNR decreased, the artifacts became audible and the rating for single and multiple talker recordings varied between 20 and 35. The second-order interaction (recording*λ) was significant due to the high rating of the single talker recording (30) with λ_0 when compared to the rating of the other recordings (15–18). Similarly, the interaction (recording*SNR) was significant due to the high rating of the single talker recording for SNR $= 20$ dB. These results are omitted as they do not provide sufficient information on the effect of the spectral floor and SNR.

The analysis focused on the significant second-order interaction between the SNR and the spectral floor λ. Fig. 12 shows the marginal means with a 95% confidence interval between the two factors. The hidden reference was always perceived clearly with scores close to 100. Similar scores were given for spectral floor values of λ_3 and λ_4. The lowest spectral floor λ_0 was given the lowest scores. In particular, for the low SNR values of 5, 0 and

-10 dB, the scores for λ_0 were between 0–20 and for the higher SNRs of 10 and 20 dB between 20–40. For the spectral floor λ_1, the perception of artifacts varied significantly. For the high SNR value of 20 dB, a mean score of 87 was achieved, while when the SNR was 10 and 0 dB the mean scores were between 35–60, which indicates that artifacts were present. Low SNR values of -10 dB were given mean scores of 30. The scenario that is of most interest in the results of this listening test is the spectral floor λ_2. For SNR values of 20 and 10 dB scores above 90 were assigned, which indicates that there were only slight or no audible artifacts present. Low SNR values of 5 and 0 dB were given a mean score of 80 and for the lowest SNR of -10 dB a score of 70.

The interaction between the SNR and spectral floor revealed that only slight audible artifacts were present in the case of $\lambda_2 = 0.2$ and SNR≥ 5 dB. In addition, according to the objective results, the specific spectral floor value provides adequate spatial filtering performance and outperforms previous coherence-based algorithms such as the McCowan post-filter. However, the lower spectral floor values λ_1 were also given high scores for high SNR values.

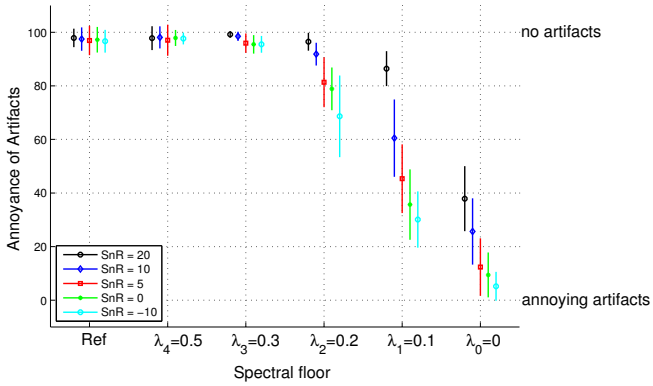

Fig. 12. Results for the listening tests for the interaction between spectral floor and SNR

V. Conclusions

In this study we propose the formulation of a post-filtering algorithm for directional microphones, derived from microphone arrays. By utilizing directional microphones, the correlation between microphones in multiple-source scenarios with an added diffuse noise is reduced, especially at low frequencies. The performance of the proposed algorithm indicates an improvement over the McCowan post-filter, especially in the low-frequency region. There are two main parameters that affect the performance of the proposed algorithm: the choice of the directional microphones and the level of the spectral floor. Whilst in the examples of this paper the microphones used to calculate the post-filter were of first and second order, other orders of directional microphones can be also used, depending on the number of available sensors in a microphone array. The level of artifacts caused by different values of the spectral floor was evaluated by conducting listening tests. For applications where the task is to recover

a sound signal corrupted by noise and quality is not an issue, the spectral floor value can be set close to zero. When sound quality is important, the spectral floor can be set to higher values, such as 0.2. The proposed algorithm can run in real-time with low latency and be applied to systems that use focusing or background noise suppression, such as teleconferencing, when the desired direction of arrival in defined. Moreover, although this method is rendered for monophonic reproduction, as the beam aims at one direction at a time, it can be extended to multichannel reproduction systems by having multiple beams towards each loudspeaker direction.

REFERENCES

[1] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds: Springer 2001, ch. 2, pp. 1938.

[2] S. L. Gay and J. Benesty, *Acoustic Signal Processing fo Telecommunications*, Eds. Kluwer Academic Publishers, 2000.

[3] K. U. Simmer, J. Bitzer and C. Marro "Post-Filtering Techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds: Springer 2001, ch. 2, pp. 40–60.

[4] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 2578–2581, 1988.

[5] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise eld coherence," *Speech and Audio Processing, IEEE Transactions on*, vol 34 no. 3, pp. 393–398, 2003.

[6] S. Fischer, K. D. Kammeyer and K. U. Simmer, "Adaptive microphone arrays for speech enhancement in coherent and incoherent noise fields," in *Proc 3rd joint meeting of the Acoustical Society of America and the Acoustical Society of Japan*, Honolulu, Hawaii, 1996.

[7] J. Bitzer, K. U. Simmer and K. D. Kammeyer, "Multichannel noise reduction algorithms and theoretical limits," in *Proc European Signal Processing Conference*, Rhodes, Greece, p. 105, 1998.

[8] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *Speech Audio Processing, IEEE Trans on*, vol 52 no. 7, pp. 709–716, 2004.

[9] E. Vincent, S. Araki, F.J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B.V. Gowreesunker, D. Lutter and N.Q.K. Duong, "The Signal Separation Evaluation Campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[10] E. Vincent, H. Sawada, P. Bofill, S. Makino and J.P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Proc. 7th Int. Conf. on Independent Component Analysis and Signal Separation*, pp. 552–559, 2007.

[11] E. Vincent, S. Araki and P. Bofill, "The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation," in *Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation*, pp 734–741, 2009.

[12] E. Vincent, R. Gribonval and C. Fevotte,"Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol.14, no.4, pp.1462,1469, July 2006.

[13] R. Mukai, H. Sawada, S. Araki and S. Makino, "Blind Source Separation for Moving Speech Signals using Blockwise ICA and Residual Crosstalk Subtraction," in *IEICE Trans. Fundamentals*, vol.E87-A, no. 8, pp.1941–1948, Aug. 2004.

[14] D. Wang, R. Vipperla, , N. Evans and T.F. Zheng, "Online Non-Negative Convolutive Pattern Learning for Speech Signals," *Signal Processing, IEEE Transactions on*, vol. 61, no. 1, pp. 44–56, 2013.

[15] S. Laurent and E. Vincent, "A general framework for online audio source separation," in *Proceedings of the 10th international conference on Latent Variable Analysis and Signal Separation*, Tel Aviv, Israel, 2012

[16] S. Moreau, J. Daniel and S. Bertet, "3D Sound Field Recording with Higher Order Ambisonics – Objective Measurements and Validation of Spherical Microphone," in *AES 120th Convention*, Paris, France, May 20–23, 2006.

[17] C. Faller, "Modifying the Directional Response of a Coincident Pair of Microphones by Postprocessing," *J. Audio Eng. Soc.*, vol 56, no. 10, Oct. 2008.

[18] C. Faller, "A Highly Directive 2-Capsule Based Microphone System," in *AES 123rd Convention*, New York, USA, October 5–8, 2007.

[19] V. Pulkki, "Spatial Sound Reproduction with Directional Audio Coding," *J. Audio Eng. Soc.*, vol 55, pp. 503–516, June 2007.

[20] M. Kallinger, H. Ochsenfeld, G. Del Caldo, F. Kuech, D. Mahne, R. Schultz-Amling and O. Thiergart, "A Spatial Filtering Technique for Directional Audio Coding," in *AES 126th Convention*, Munich, Germany, May 7–10, 2009.

[21] H. Teutsch and W. Kellermann, "Acoustic source detection and localization based on wave field decomposition using circular microphone arrays," *J. Audio Eng. Soc.*, vol 120, no. 5, November 2006.

[22] O. Kirkeby, P. A. Nelson, H. Hamada and F. Orduna-Bustamante, "Fast Deconvolution of Multichannel Systems Using Regularization," *Speech and Language Processing IEEE Trans Audio on*, vol. 6, no. 2, pp. 189-195, Mar. 1998.

[23] O. Kirkeby and P. A. Nelson, "Digital Filter Design for Inversion Problems in Sound Reproduction," *J. Audio Eng. Soc.*, vol. 47, no. 7/8, July/Aug. 1999.

[24] Earl G. Williams, "Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography," Academic Press, June 30, 1999.

[25] H. Teutsch, "Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition," Berlin Heidelberg: Springer-Verlag, 2007.

[26] B. Rafaely, "Analysis and Design of Spherical Microphone Arrays," *Speech and Language Processing IEEE Trans Audio on*, vol. 13, no. 1, pp 135–143, Jan. 2005.

[27] A. Farina, A. Capra, L. Chiesi and L Scopece, "A Spherical Microphone Array for Synthesizing Virtual Directive Microphones in Live Broadcasting and in Post Production," in *AES 40th International Conference*, Tokyo, Japan, October 8–10, 2010.

[28] L. Josefsson and P. Persson, "Conformal Array Antenna Theory and Design," John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

[29] S. Delikaris-Manias, C. Valagiannopoulos and V. Pulkki, "Optimal Directional Pattern Design Utilizing Arbitrary Microphone Arrays: A Continuous-Wave Approach," in *AES 134th Convention*, Rome, Italy, May 4–7, 2013.

[30] C.G. Clifford,"Coherence and time delay estimation," *Proceedings of the IEEE*, vol.75, no.2, pp.236–255, Feb. 1987.

[31] J. Makhoul, and M. Berouti,"High-frequency regeneration in speech coding systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing* , vol.4, no., pp.428–431, Apr. 1979.

[32] A.J. Manders, D.M. Simpson and S.L. Bell, "Objective Prediction of the Sound Quality of Music Processed by an Adaptive Feedback Canceller," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol.20, no.6, pp.1734–1745, Aug. 2012

[33] D. J. Freed and S. D. Soli, "An objective procedure for evaluation of adaptive antifeedback algorithms in hearing aids," *Ear Hear.*, vol. 27, no. 4, pp. 382–398, 2006.

[34] M. Kallinger, G. Del Galdo, F. Kuech, D. Mahne and R. Schultz-Amling, "Spatial filtering using directional audio coding parameters," in *IEEE International Conference on Acoustics, Speech and Signal Processing* , vol., no., pp.217–220, Apr. 2009.

[35] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing* , vol.4, no., pp.208–211, Apr. 1979.

[36] H. Cox, R. M. Zeskind and T. Kooij, "Practical Supergain", *Speech and Audio Processing, IEEE Transactions on*, vol 11 no. 6, pp. 709–16, 2003.

[37] V. Tourbabin, M. Agmon, B. Rafaely and J. Tabrikian, "Optimal Real-Weighted Beamforming With Application to Linear and Spherical Arrays," in *Audio, Speech, and Language Processing, IEEE Transactions on*, vol.20, no.9, pp.2575–2585, Nov. 2012.

[38] S. Roweis. "Factorial models and refiltering for speech separation and denoising," in *Proc. Eurospeech*, Sep. 2003.

[39] M.V. Laitinen, F. Kuech, S. Disch and V. Pulkki, "Reproducing Applause-Type Signals with Directional Audio Coding," *J. Audio Eng. Soc.*, vol 59, no 2, June 2011.

[40] J. Tribolet, P. Noll, B. McDermott and R. E. Crochiere, "A study of complexity and quality of speech waveform coders," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, pp. 586–590, 1978.

[41] S. R. Quackenbush, T. P. Barnwell and Clements MA. "Objective measures of speech quality," Englewood Cliffs, NJ, Prentice-Hall, Inc.; 1988

[42] IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Symeon Delikaris-Manias received his B.Sc. degree in mathematics from the University of Crete, Heraklion, Greece in 2006 and his M.Sc. degree in sound and vibration from the Institute of Sound and Vibration Research (ISVR), Southampton, UK, in 2008 with a thesis on inverse-filtering methods and cross-talk cancellation systems. He is currently pursuing his D.Sc degree in Electrical Engineering at Aalto University, Espoo, Finland.

Between 2008-2010, he was employed by PGacoustics, and was responsible for acoustic modeling and auralization. In 2010-2011 he was at the Center for Virtual Reality, Brest, France developing and evaluating sound-field recording and reproduction techniques. Since 2011, he is in the Department of Acoustics and Signals Processing, School of Electrical Engineering, Aalto University. His research interests are digital signal processing techniques for microphone arrays and multi-channel audio systems.

Mr. Delikaris-Manias is a member of the IEEE Signal Processing Society and the Audio Engineering Society.

Ville Pulkki received his M.Sc. and D.Sc. (Tech) degrees from Helsinki University of Technology in 1994 and 2001, respectively. He majored in acoustics, audio signal processing and information sciences. Between 94 and 97 he was a full time student at the Department of Musical Education in Sibelius Academy.

In his doctoral dissertation he developed Vector Base Amplitude Panning (VBAP), which is a method for positioning virtual sources to multichannel loudspeaker configurations. In addition, he studied the performance of VBAP with psychoacoustic listening tests and with modeling of auditory localization mechanisms. The VBAP method is now widely used in multi-channel virtual auditory environments, and in computer music installations. Later, he developed with his group a non-linear time-frequency-domain method for spatial sound reproduction and coding, Directional Audio Coding (DirAC). DirAC takes coincident first-order microphone signals as input, and processes output to arbitrary loudspeaker layouts or to headphones. He also researches computational functional model of the brain organs devoted to binaural hearing. He is leading a research group in Aalto University (earlier: Helsinki University of Technology, TKK or HUT), which consists of 15 researchers. The group conducts research also on head-related acoustics measurements, and conducts psychoacoustical experiments to better understand spatial sound perception.

Prof Pulkki enjoys being with his family (wife and two children), playing various musical instruments, building his summer place and dancing hip hop.