GENERATING VIRTUAL MICROPHONE SIGNALS IN NOISY ENVIRONMENTS

Konrad Kowalczyk[1,2]*, Alexandra Craciun*[1]*, and Emanuël A. P. Habets*[1,2]

[1]Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany
[2]International Audio Laboratories Erlangen[†], Erlangen, Germany

ABSTRACT

Spatial sound acquisition methods typically capture the sound scene with reference to the position of the recording device. Using a recently proposed *virtual microphone* (VM) technique, the position and characteristics of the recording device (such as the directivity response and orientation) can be modified. This technique relies on synthesizing a VM signal at an arbitrary position, which sounds perceptually similar to the signal that would be recorded with a physical microphone placed at the same location. In this paper, we present a method to generate a VM signal in the presence of noise. Noise reduction is accomplished using a parametric multichannel Wiener filter, where a trade-off parameter is applied in order to achieve a constant residual noise level in the generated VM signal, irrespective of the VM position. The simulated experiments show the applicability of the method for signal extraction in the presence of additive noise.

Index Terms— microphone arrays, parametric sound field model, spatial sound, Wiener filter, noise reduction

1. INTRODUCTION

The aim of spatial sound recording is to reproduce a captured sound field such that a listener perceives the reproduced sound as if he was present in the original sound scene. Standard techniques, such as stereophonic recording, make use of several directional microphones to capture the spatial information characterizing the sound field. Steerable directivity patterns can be attained by spatial filtering with microphone arrays [1]. Alternatively, one can apply methods based on parametric representations of the sound field, such as the Directional Audio Coding (DirAC) [2] method. Here the microphone arrays are used to determine one or more audio downmix signals, together with the associated spatial information describing the sound scene.

The main drawback of existing spatial sound acquisition approaches is that the spatial image is always recorded relative to the position of the microphone or microphone array. Therefore, in many applications it is not possible to listen to the sound scene from a desired position, e.g., close to the desired sound source. To this end, a technique has been proposed in [3] for computing the output signal of a *virtual microphone* (VM) placed at an arbitrary position in a room using two circular arrays. Based on the information gathered by distributed microphone arrays, a parametric sound field representation that is independent of the recording position can be computed. The far-end listener can then freely decide on the position and orientation of the spatial sound image. A similar goal can be obtained using sound field extrapolation methods [4]. Nevertheless, such methods commonly suffer from performance degradation for long distances [5].

In this paper, we employ linear arrays and extend the model described in [3] to account also for the additive background noise. Our goal is to generate a virtual microphone signal with a reduced noise power level, which can be achieved using a parametric multichannel Wiener filter (PMWF) [1, 6]. In particular, the trade-off parameter of the PMWF enables synthesizing a VM signal characterized by a constant residual noise level, independent of the position of the VM, as one expects from a physical microphone. Furthermore, in this paper, the solution for generating a VM signal using linear microphone arrays is presented, whereas in [3] circular arrays were used.

2. PROBLEM FORMULATION

For sound field analysis, the signals recorded by two microphone arrays are transformed to the time-frequency domain using a short-time Fourier transform (STFT). The time and frequency indices are denoted as n and k, respectively. However, in the following, we omit the time index for clarity. We assume that, for each time-frequency bin, the sound field captured by a microphone located at position $\boldsymbol{d} = [x\ y]^{\mathrm{T}}$ can be modeled as a sum of a direct sound component $P_{\mathrm{s}}(k, \boldsymbol{d})$ and a noise component $P_{\mathrm{v}}(k, \boldsymbol{d})$, i.e.,

$$P(k, \boldsymbol{d}) = P_{\mathrm{s}}(k, \boldsymbol{d}) + P_{\mathrm{v}}(k, \boldsymbol{d}), \qquad (1)$$

where additive noise is assumed to be uncorrelated with the source signal. Such a single-wave model is typically valid for speech sources under the assumption that the time-frequency overlap between different speakers is sufficiently small, i.e., the signals fulfill the so-called W-disjoint orthogonality condition [7]. The pressure of a spherical wave emitted by an

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2011-287760.

[†]A joint institution of the Friedrich-Alexander-University of Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

isotropic point-like source (IPLS) located at $\boldsymbol{d}_{\text{IPLS}}$ that is observed at position $\boldsymbol{d}$ can be written as

$$P_{\text{s}}(k, \boldsymbol{d}) = G_{\text{s}}(k, \boldsymbol{d}, \boldsymbol{d}_{\text{IPLS}})\, P_{\text{s}}(k, \boldsymbol{d}_{\text{IPLS}}), \tag{2}$$

where $G_{\text{s}}(k, \boldsymbol{d}, \boldsymbol{d}_{\text{IPLS}})$ is a complex propagation factor from $\boldsymbol{d}_{\text{IPLS}}$ to $\boldsymbol{d}$. In general, such a transfer function is given by

$$G_{\text{s}}(k, \boldsymbol{d}, \boldsymbol{d}_{\text{IPLS}}) = r^{-1}(k)\, e^{j\kappa r(k)}, \tag{3}$$

where κ denotes the wave number and $r(k) = \|\boldsymbol{d} - \boldsymbol{d}_{\text{IPLS}}(k)\|$ is the distance between these two positions.

In this paper, we aim to generate an artificial microphone signal $P(k, \boldsymbol{d}_{\text{VM}})$ at an arbitrary location $\boldsymbol{d}_{\text{VM}}$ in the sound scene with very similar acoustic characteristics as those of a real microphone actually placed at the same location. Given the direct sound $P_{\text{s}}(k, \boldsymbol{d})$, the VM signal can be generated by applying the weights of the virtual microphone $W(k, \boldsymbol{d}_{\text{VM}}, \varphi_{\text{VM}})$ for the desired VM look direction φ_{VM} (see Sec. 4 for details), which yields

$$P(k, \boldsymbol{d}_{\text{VM}}) = W(k, \boldsymbol{d}_{\text{VM}}, \varphi_{\text{VM}})\, P_{\text{s}}(k, \boldsymbol{d}). \tag{4}$$

In the presence of additive noise, $P_{\text{s}}(k, \boldsymbol{d})$ cannot be observed directly. The estimate of the direct sound pressure at the reference microphone can be obtained using a spatial linear filter

$$\hat{P}_{\text{s}}(k, \boldsymbol{d}) = \boldsymbol{h}^H(k)\, \boldsymbol{p}(k, \boldsymbol{d}), \tag{5}$$

where $\boldsymbol{p}(k, \boldsymbol{d}) = [P(k, \boldsymbol{d}_1), \ldots, P(k, \boldsymbol{d}_{\text{M}})]^T$ denotes the M-element microphone signal vector and $\boldsymbol{h}(k)$ denotes the filter weight vector. Let us denote a reference microphone position as $\boldsymbol{d} = \boldsymbol{d}_1$. Using (1), the estimated signal can be written as

$$\begin{aligned} \hat{P}_{\text{s}}(k, \boldsymbol{d}) &= \boldsymbol{h}^H(k)[\boldsymbol{p}_{\text{s}}(k, \boldsymbol{d}) + \boldsymbol{p}_{\text{v}}(k, \boldsymbol{d})] \\ &= P_{\text{fs}}(k, \boldsymbol{d}) + P_{\text{rn}}(k, \boldsymbol{d}), \end{aligned} \tag{6}$$

where $\boldsymbol{p}_{\text{s}}(k, \boldsymbol{d}) = [P_{\text{s}}(k, \boldsymbol{d}_1), \ldots, P_{\text{s}}(k, \boldsymbol{d}_{\text{M}})]^T$ and $\boldsymbol{p}_{\text{v}}(k, \boldsymbol{d}) = [P_{\text{v}}(k, \boldsymbol{d}_1), \ldots, P_{\text{v}}(k, \boldsymbol{d}_{\text{M}})]^T$ are the direct sound signal and noise signal vectors, respectively, $P_{\text{fs}}(k, \boldsymbol{d})$ denotes the filtered direct signal, and $P_{\text{rn}}(k, \boldsymbol{d})$ is the residual noise signal. The power of this residual noise is given by

$$E\{|P_{\text{rn}}(k, \boldsymbol{d})|^2\} = \boldsymbol{h}^H(k)\, \boldsymbol{\Phi}_{\text{v}}(k)\, \boldsymbol{h}(k), \tag{7}$$

where the noise power spectral density (PSD) matrix is given by $\boldsymbol{\Phi}_{\text{v}}(k) = E\{\boldsymbol{p}_{\text{v}}(k, \boldsymbol{d})\, \boldsymbol{p}_{\text{v}}^H(k, \boldsymbol{d})\}$. Using (4), the power of the VM signal can be written as

$$E\{|P(k, \boldsymbol{d}_{\text{VM}})|^2\} = W^2(k, \boldsymbol{d}_{\text{VM}}, \varphi_{\text{VM}})\, E\{|\hat{P}_{\text{s}}(k, \boldsymbol{d})|^2\} \tag{8}$$

and the noise power spectral density of the VM signal yields

$$\begin{aligned} \phi_{\text{v}}(k, \boldsymbol{d}_{\text{VM}}) &= E\{|P_{\text{rn}}(k, \boldsymbol{d}_{\text{VM}})|^2\} \\ &= W^2(k, \boldsymbol{d}_{\text{VM}}, \varphi_{\text{VM}})\, E\{|P_{\text{rn}}(k, \boldsymbol{d})|^2\}. \end{aligned} \tag{9}$$

As can be seen from (9), the power of the residual noise of the virtual microphone signal depends on the VM weights $W(k, \boldsymbol{d}_{\text{VM}}, \varphi_{\text{VM}})$, and thus it is dependent on the VM position. Since the additive noise component cannot be fully suppressed in practice, in order to generate a realistic virtual microphone signal, we aim to synthesize this signal such that the residual noise is independent of the selected VM position.

3. DIRECT SOUND PRESSURE ESTIMATION

3.1. Parametric multichannel Wiener Filter

In order to reduce noise at one of the arrays, a parametric multichannel Wiener filter (PMWF) [1, 8], denoted as $\boldsymbol{h}_{\text{PMWF}}(k)$, can be applied in (5). Assuming the noise STFT coefficients are Gaussian distributed, the PMWF can be written as [1, 8]

$$\boldsymbol{h}_{\text{PMWF}}(k) = \boldsymbol{h}_{\text{MVDR}}(k) H_{\text{PW}}(k, \beta), \tag{10}$$

where $\boldsymbol{h}_{\text{MVDR}}(k)$ is the array weight vector of a minimum variance distortionless response (MVDR) beamformer and $H_{\text{PW}}(k)$ denotes a parametric single-channel Wiener (PW) filter. The weights of the MVDR beamformer are given by

$$\boldsymbol{h}_{\text{MVDR}}(k) = \frac{\hat{\boldsymbol{\Phi}}_{\text{v}}^{-1}(k)\boldsymbol{g}(k, \hat{\theta})}{\boldsymbol{g}^H(k, \hat{\theta})\hat{\boldsymbol{\Phi}}_{\text{v}}^{-1}(k)\boldsymbol{g}(k, \hat{\theta})}, \tag{11}$$

where $\hat{\boldsymbol{\Phi}}_{\text{v}}(k)$ denotes the estimated noise PSD matrix. Assuming that the source is located in the far-field with respect to the microphone array, $\boldsymbol{g}(k, \hat{\theta})$ is the array steering vector for the estimated direction of arrival $\hat{\theta}$. The noise at the output of the MVDR beamformer is further reduced by a parametric single-channel Wiener filter, which is given by

$$H_{\text{PW}}(k, \beta) = \frac{\xi(k)}{\beta(k) + \xi(k)}, \tag{12}$$

where parameter $\beta(k)$ ($\beta(k) \geq 0$) enables the control of the trade-off between the residual noise suppression and the distortion of the direct sound signal. In this paper, the *a priori* SNR $\xi(k)$ is estimated using a decision-directed approach [9], which is given by (note that the time index n is written explicitly)

$$\begin{aligned} \xi(k, n) = \alpha\, H_{\text{PW}}^2(k, n-1)\, \gamma(k, n-1) \\ + (1-\alpha) \max[\xi_{\text{min}}, \gamma(k, n) - 1], \end{aligned} \tag{13}$$

where ξ_{min} is a small positive constant to avoid negative estimates of $\xi(k)$, α is a weighting factor, and the *a posteriori* SNR is estimated using the following relation:

$$\gamma(k) = \frac{|\boldsymbol{h}_{\text{MVDR}}^H(k)\boldsymbol{p}(k, \boldsymbol{d})|^2}{\boldsymbol{h}_{\text{MVDR}}^H(k)\hat{\boldsymbol{\Phi}}_{\text{v}}(k)\boldsymbol{h}_{\text{MVDR}}(k)}. \tag{14}$$

3.2. Residual noise control

Substituting (7) and (10) into (9), the residual noise of the virtual microphone signal can be written as

$$\begin{aligned} \phi_{\text{v}}(k, \boldsymbol{d}_{\text{VM}}) = W^2(k, \boldsymbol{d}_{\text{VM}}, \varphi_{\text{VM}}) H_{\text{PW}}^2(k, \beta) \\ \times \boldsymbol{h}_{\text{MVDR}}^H(k)\hat{\boldsymbol{\Phi}}_{\text{v}}(k)\boldsymbol{h}_{\text{MVDR}}(k). \end{aligned} \tag{15}$$

The noise reduction factor of the MVDR beamformer is given by $\psi_{\text{MVDR}}(k) = \phi_{\text{v}}(k, \boldsymbol{d})/[\boldsymbol{h}_{\text{MVDR}}^H(k)\hat{\boldsymbol{\Phi}}_{\text{v}}(k)\boldsymbol{h}_{\text{MVDR}}(k)]$, where $\phi_{\text{v}}(k, \boldsymbol{d})$ is the noise PSD at $\boldsymbol{d}$. Note that for an MVDR beamformer, the noise reduction factor is equivalent to the array gain. Using the above noise reduction definition, (15) can be written as $\phi_{\mathrm{v}}(k, \boldsymbol{d}_{\mathrm{VM}}) = W^2(k, \boldsymbol{d}_{\mathrm{VM}}, \varphi_{\mathrm{VM}}) H^2_{\mathrm{PW}}(k, \beta)$ $\phi_{\mathrm{v}}(k, \boldsymbol{d})/\psi_{\mathrm{MVDR}}(k) = c(k)\phi_{\mathrm{v}}(k, \boldsymbol{d})$. Thus the noise power reduction level $c(k)$ achieved by the PMWF can be written as

$$c(k) = \frac{W^2(k, \boldsymbol{d}_{\mathrm{VM}}, \varphi_{\mathrm{VM}})\, \xi^2(k)}{\psi_{\mathrm{MVDR}}(k)\, [\beta(k) + \xi(k)]^2}, \tag{16}$$

where $0 < c(k) < 1$ to ensure that the VM signal contains less noise than the noise received by the reference microphone. The value of the trade-off parameter $\beta(k)$, for which the desired noise reduction level $c(k)$ is achieved, can finally be determined by solving (16), which yields

$$\beta(k) = \xi(k) \left[\frac{W(k, \boldsymbol{d}_{\mathrm{VM}}, \varphi_{\mathrm{VM}})}{\sqrt{c(k)\, \psi_{\mathrm{MVDR}}(k)}} - 1 \right]. \tag{17}$$

Note that the allowed trade-off parameter values are restricted to the range $0 \leq \beta(k) \leq \beta_{\mathrm{max}}$. For $\beta(k) = 0$, the PMWF reduces to the MVDR beamformer, whereas for $\beta(k) = 1$, a multichannel Wiener filter is obtained.

4. VIRTUAL MICROPHONE SIGNAL SYNTHESIS

Given the direct sound component at the IPLS position, we can express the virtual microphone (VM) signal as

$$\begin{aligned} P(k, \boldsymbol{d}_{\mathrm{VM}}, \varphi_{\mathrm{VM}}) &= B(k, \varphi_{\mathrm{VM}}) \\ &\times G_{\mathrm{s}}(k, \boldsymbol{d}_{\mathrm{VM}}, \boldsymbol{d}_{\mathrm{IPLS}})\, P_{\mathrm{s}}(k, \boldsymbol{d}_{\mathrm{IPLS}}), \end{aligned} \tag{18}$$

where $G_{\mathrm{s}}(k, \boldsymbol{d}_{\mathrm{VM}}, \boldsymbol{d}_{\mathrm{IPLS}})$ is the propagation factor from each IPLS to the VM position and $B(k, \varphi_{\mathrm{VM}})$ denotes the weights corresponding to different characteristics of the virtual microphone, such as orientation and polar pattern. In order to calculate these weights, the DOAs of the IPLSs are computed for a VM position $\mathbf{d}_{\mathrm{VM}}$ with respect to the global coordinate system using the following relation:

$$\boldsymbol{\theta}_{\mathrm{VM}}(k, \boldsymbol{d}_{\mathrm{VM}}) = \frac{\mathbf{d}_{\mathrm{IPLS}} - \boldsymbol{d}_{\mathrm{VM}}}{\|\boldsymbol{d}_{\mathrm{IPLS}} - \boldsymbol{d}_{\mathrm{VM}}\|}. \tag{19}$$

In order to create the VM corresponding to a first-order directional microphone, the directivity pattern weights can be calculated as

$$B(k, \varphi_{\mathrm{VM}}) = \eta + (1 - \eta)\boldsymbol{\theta}^T_{\mathrm{VM}}(k, \boldsymbol{d}_{\mathrm{VM}})\boldsymbol{\varphi}_{\mathrm{VM}}, \tag{20}$$

where $\boldsymbol{\varphi}_{\mathrm{VM}} = [\cos(\varphi_{\mathrm{VM}}) \;\; \sin(\varphi_{\mathrm{VM}})]^T$ denotes the VM orientation vector for the desired look direction φ_{VM}, and factor η determines the directivity pattern, e.g., $\eta = 0.5$ yields a cardioid microphone, while $\eta = 1$ yields an omnidirectional microphone.

Applying (2), the direct sound pressure at the IPLS position can be computed by back-propagating the direct sound component from the reference microphone position to the IPLS according to

$$P_{\mathrm{s}}(k, \boldsymbol{d}_{\mathrm{IPLS}}) = G^{-1}_{\mathrm{s}}(k, \boldsymbol{d}, \boldsymbol{d}_{\mathrm{IPLS}})\, P_{\mathrm{s}}(k, \boldsymbol{d}). \tag{21}$$

Given the estimated direct sound pressure $\hat{P}_{\mathrm{s}}(k, \boldsymbol{d})$ obtained using the PMWF, the VM signal can finally be written as

$$P(k, \boldsymbol{d}_{\mathrm{VM}}, \varphi_{\mathrm{VM}}) = W(k, \boldsymbol{d}_{\mathrm{VM}}, \varphi_{\mathrm{VM}})\, \hat{P}_{\mathrm{s}}(k, \boldsymbol{d}), \tag{22}$$

where the overall weights that take into account the desired VM directivity pattern and the propagation compensation from each IPLS (observed at the reference microphone) to the VM position are given by

$$\begin{aligned} W(k, \boldsymbol{d}_{\mathrm{VM}}, \varphi_{\mathrm{VM}}) &= B(k, \varphi_{\mathrm{VM}}) \\ &\times G_{\mathrm{s}}(k, \boldsymbol{d}_{\mathrm{VM}}, \boldsymbol{d}_{\mathrm{IPLS}})\, G^{-1}_{\mathrm{s}}(k, \boldsymbol{d}, \boldsymbol{d}_{\mathrm{IPLS}}). \end{aligned} \tag{23}$$

To increase robustness and reduce artifacts, we aim to reconstruct only the magnitude of the VM signal, as proposed in [3]. Consequently, the propagation factor reduces to $G_{\mathrm{s}}(k, \cdot, \cdot) = r^{-1}(k)$.

5. IPLS POSITION ESTIMATION

In order to synthesize the VM signal, the IPLS positions need to be estimated for each time-frequency bin. This is achieved by means of triangulation, based on the positions of two uniform linear microphone arrays, which are assumed to be known, and the estimates of the directions of arrival (DOAs) of the IPLSs at each of them. In this work, the DOAs are estimated using the rotational invariance technique known as ESPRIT [10]. First, the eigenvalue decomposition of the microphone covariance matrix $\mathbf{R}(k, \boldsymbol{d}) = E\{\boldsymbol{p}(k, \boldsymbol{d})\boldsymbol{p}^H(k, \boldsymbol{d})\}$ is computed to separate the signal subspace from the noise subspace:

$$\mathbf{R}(k, \boldsymbol{d}) = \left\{ [\mathbf{u}_s \mathbf{U}_v] \begin{bmatrix} \lambda_s & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_v \end{bmatrix} [\mathbf{u}_s \mathbf{U}_v]^H \right\} (k, \boldsymbol{d}). \tag{24}$$

Assuming a single plane-wave model, the eigenvector $\mathbf{u}_s$ corresponding to the largest eigenvalue λ_s represents the signal subspace, while the remaining eigenvalues ($\boldsymbol{\Lambda}_v$) correspond to the noise subspace spanned by $\mathbf{U}_v$. Based on $\mathbf{u}_s$, two rotationally invariant subarrays $\mathbf{u}_1$ and $\mathbf{u}_2$ are created by taking the first and last $M - 1$ elements of $\mathbf{u}_s$, respectively. It can then be shown that the following relation is satisfied [10]:

$$\boldsymbol{u}_2(k, \boldsymbol{d}) = \boldsymbol{u}_1(k, \boldsymbol{d}) \cdot e^{j\mu(k)}, \tag{25}$$

where $\mu(k)$ denotes the spatial frequency, which can be estimated in the least-mean-square sense [10]. Finally, an estimate of the DOA is given by

$$\hat{\theta}(k) = \arcsin\left(\frac{\lambda \cdot \hat{\mu}(k)}{2\pi\Delta} \right), \tag{26}$$

where $\hat{\mu}(k)$ is the estimated spatial frequency and Δ is the microphone spacing.

Knowing the positions of both arrays and the estimated DOAs $\hat{\theta}_1$ and $\hat{\theta}_2$ at each of them, the line equation $y = m \cdot x + n$ can be used to calculate the IPLS position $\boldsymbol{d}_{\mathrm{IPLS}} = [x_{\mathrm{IPLS}}\, y_{\mathrm{IPLS}}]^T$ by solving the following relation for x_{IPLS}:

$$m_1 \cdot x_{\mathrm{IPLS}} + n_1 = m_2 \cdot x_{\mathrm{IPLS}} + n_2, \tag{27}$$

where $m_{\mathrm{i}} = \tan(\hat{\theta}_{\mathrm{i}})$, $n_{\mathrm{i}} = y_{\mathrm{i}} - m_{\mathrm{i}} \cdot x_{\mathrm{i}}$, and the array position is denoted as $[x_{\mathrm{i}}\; y_{\mathrm{i}}]^T$ for $i \in \{1, 2\}$. Finally, y_{IPLS} can be obtained from the line equation for one of the arrays, e.g., as

$$y_{\mathrm{IPLS}} = m_1 \cdot x_{\mathrm{IPLS}} + n_1. \tag{28}$$

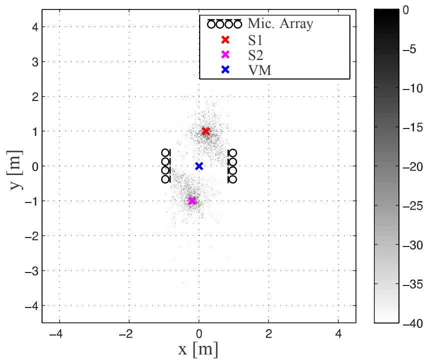

Fig. 1. Room setup and spatial power distribution, in dB, for a double-speaker scenario. True source 1 (S1), true source 2 (S2), and VM positions are denoted as red, purple, and blue crosses, respectively.

6. PERFORMANCE EVALUATION

6.1. Setup description

The presented method for VM signal generation in noisy environments is evaluated through a set of simulated experiments, in which the room impulse responses for a $9 \times 9 \times 3$ m room with a reverberation time of 250 ms were generated using an image-source method [11]. Two 4-element uniform linear arrays (with inter-microphone spacing of 0.02 m) were used to gather the 2D sound field information. The sample rate was set to 44.1 kHz and a 1024-point STFT with a sine window and 50% overlap was used for time-frequency analysis. For noise estimation, an ideal power-based voice activity detector was applied, and parameter α in (13) was set to 0.98.

Two scenarios were considered. In a double-talk scenario, illustrated in Fig. 1, the signals of two simultaneously active male speakers of 2 s duration were convolved with the simulated impulse responses, and a white Gaussian noise sequence with an averaged segmental signal-to-noise ratio (SNR) of 10 dB was added to the microphone signals. As a performance measure, the signal-to-interference-and-noise ratio (SINR) was used, where speaker 1 (S1) was selected as the desired speaker and speaker 2 (S2) as the interferer. In the second scenario, only speaker 1 was active and the same amount of noise was added. In both scenarios, the performance was evaluated using the fullband segmental SNR and SINR improvement measures, respectively denoted as ΔSNR and ΔSINR, which are defined as the difference between the S(I)NR at the VM output and the input S(I)NR values, in dB. The final ΔSNR and ΔSINR values were calculated by averaging the segmental results over a 2 s period.

6.2. Results and evaluation

The applicability of the VM method to extract the desired source signal in a multi-speaker environment is shown in a double-talk experiment. Fig. 1 illustrates the so-called spatial power density for the simulated environment, i.e., the IPLS positions with their corresponding powers averaged over time

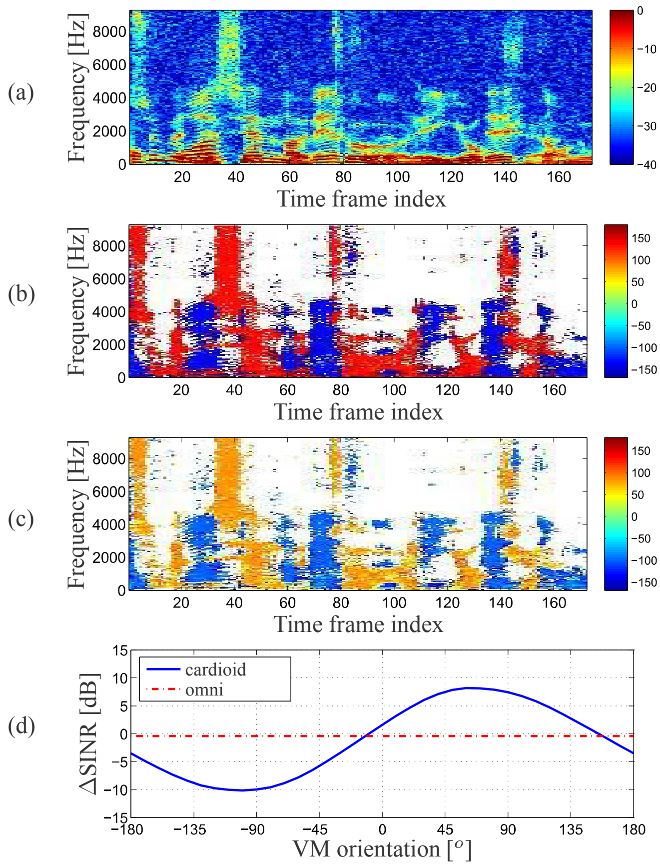

Fig. 2. Double-talk scenario: (a) spectrogram of the reference microphone signal at the right array, (b) DOA estimates at the right array during speech presence periods, (c) DOAs computed for the VM positioned at $(0, 0)$ for a 90^o look direction during speech presence periods, and (d) ΔSINR for a rotating cardioid VM and an omnidirectional VM, positioned at $(0, 0)$.

and frequency. As can be observed, the estimated source positions are centered around the actual source position, and the estimation variance was mainly due to the additive noise and spatial resolution of the array. The spectrogram of the reference microphone signal, the DOAs estimated at the right array positioned at $(0, 1)$, and the DOAs computed for the VM positioned at $(0, 0)$ are shown in Fig. 2(a-c). The estimated DOAs match well the DOAs corresponding to the true source positions (129^o and -141^o at the right array, and 79^o and -101^o at the VM, respectively). Fig. 2(d) depicts the ΔSINR for a cardioid VM ($\eta = 0.5$ in (20)) positioned at $(0, 0)$ and rotated from -180^o to 180^o; ΔSINR for an omnidirectional VM positioned at the same location is also shown for comparison. As can be observed, the trend shown matches the expected performance of a real cardioid microphone that would be placed in the same location and rotated analogously. An omnidirectional VM located between the two sources of almost equal energy also exhibits the expected behavior.

An experiment with a single talker was conducted to evaluate the performance of the VM signal synthesis with a constant residual noise level. Fig. 3 (top) shows the SNR im-

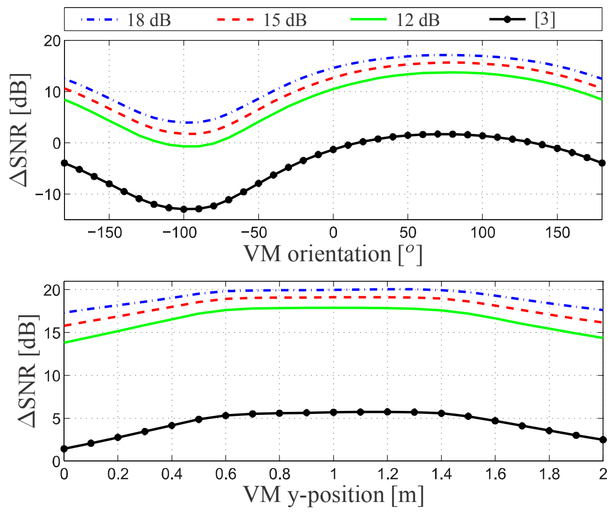

Fig. 3. ΔSNR for a rotating cardioid VM positioned at $(0, 0)$ (top) and ΔSNR for an omnidirectional VM moving long the y-axis, $x = 0.2$ (bottom), for the desired VM noise reduction of $12, 15, 18$ dB and the VM signal generated as in [3], in a single-source scenario.

provement of a cardioid microphone positioned at $(0, 0)$, for a full circle rotation of the VM and the desired noise reduction levels $c(k) = 12, 15, 18$ dB. The maximum ΔSNR is obtained when the cardioid microphone points in the direction of the source (79^o) and the minimum value is obtained for the opposite look direction. For comparison with the method presented in [3], the SNR improvement when the signal of the first microphone of the right array is used as a reference signal to generate the VM signal is also shown. The SNR improvement for an omnidirectional VM moving along a 2 m long straight line parallel to the y-axis ($x = 0.2$ is kept constant) is illustrated in Fig. 3 (bottom). As expected, the SNR improvement increases when the VM is moved closer to the source position, and it decreases for larger distances.

To show that the residual noise level is constant using the proposed technique, the ΔSNR of an omnidirectional VM is depicted in Fig. 4 for a changing VM position along an arc of a circle centered at the speaker position, for two different radii values ($r = 0.8$ m and $r = 1$ m). For comparison, the results for $\beta = 0.1$ are also provided. As can clearly be seen, the ΔSNR for the proposed method (where β is computed using (17)) is constant for a given radius, which shows that the noise level does not change since the source signal power remains constant. On the other hand, the ΔSNR values vary when the trade-off parameter β was constant.

7. CONCLUSIONS

In this paper, a method for synthesizing a virtual microphone signal in noisy acoustic environments is presented. The suppression of additive noise is achieved using a parametric multichannel Wiener filter, the trade-off parameter of which assures that the residual noise level remains constant for any arbitrarily selected VM position. The presented results show

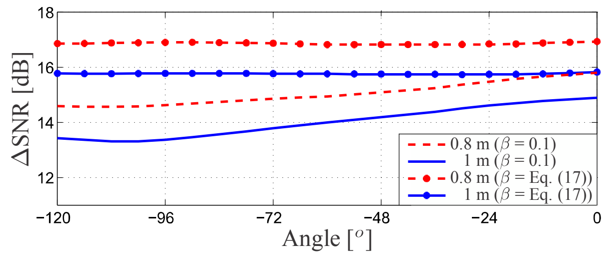

Fig. 4. ΔSNR for an omnidirectional VM moving on an arc of a circle between -120^o and 0^o and two different circle radii. Results are shown for a constant ($\beta = 0.1$) and varying β (calculated using (17) with $c(k) = 15$ dB).

the applicability of the proposed method for signal extraction in noisy environments.

8. REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag, Berlin, Germany, 2008.

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

[3] G. Del Galdo, O. Thiergart, T. Weller, and E. A. P. Habets, "Generating virtual microphone signals using geometrical information gathered by distributed arrays," in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Edinburgh, United Kingdom, May 2011.

[4] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, Academic Press, first edition, 1999.

[5] A. Kuntz and R. Rabenstein, "Limitations in the extrapolation of wave fields from circular measurements," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2007.

[6] A. Spriet, M. Moonen, and J. Wouters, "Spatially preprocessed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[7] S. Rickard and Z. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2002.

[8] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, 2010.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[10] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 984–995, 1989.

[11] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.