On Speech Sparsity for Computational Efficiency and Noise Reduction in Hearing Aids

Adrien Llave, Simon Leglaive

On Speech Sparsity for Computational Efficiency and Noise Reduction in Hearing Aids

Adrien Llave and Simon Leglaive
CentraleSupélec, IETR, Avenue de la Boulaie, 35510 Cesson-Sévigné, France

***Abstract*—Beamforming techniques are widely used in hearing aids to improve the signal-to-noise ratio. In a multi-speaker scenario, it is common to assume that the speech signals associated with each speaker do not overlap in the time-frequency domain. This so-called W-disjoint orthogonality assumption allows us to reduce the complexity of the beamforming algorithm. However, its validity decreases in presence of more than two speakers. In this study, we propose a beamforming algorithm relying on a less restrictive assumption regarding the sparsity of speech signals in the time-frequency domain. Its implications over the noise reduction performance and the computational complexity are discussed and compared with the Linearly Constrained Minimum Variance (LCMV) and the Minimum Variance Distortionless Response (MVDR) beamformers. We show that the proposed algorithm improves the noise reduction performance and reduces the computational cost compared to the LCMV beamformer without increasing the artifacts amount unlike the MVDR beamformer.**

***Index Terms*—Beamforming, speech sparsity, noise reduction, hearing aids**

I. Introduction

Noise reduction is a key feature in Hearing Aids (HA) and beamforming algorithms are the most efficient techniques in this context [14]. They are based on a constrained optimization problem [5] as the Linearly Constrained Minimum Variance (LCMV) beamformer. The aim is to minimize the power of the noise component at the beamformer output subject to the constraint of preserving the sources of interest. However, the noise reduction performance decreases with the number of speakers to preserve. Indeed, the preservation constraint for one given speaker removes one degree of freedom in the optimization problem, reducing the size of the sub-space over which the noise power minimization is achieved [19]. Moreover, adding a source of interest into the optimization problem increases the computational complexity of the resulting filter which is known to be a severe constraint of HA.

Some works addressed the LCMV computational efficiency. For instance, [7] assumed that the location of the speech sources does not change frequently, such that the filter can be updated from a time frame to the next one thanks to an iterative method with a low computational cost. Another work [12] proposed to consider that only a subset of speech sources move between two time frames. Then, they proposed a method to update the LCMV filter without recomputing it from scratch. However, those hypotheses are not suitable in the HA context, as the head is able to move quickly and often. Moreover, those methods are efficient for larger sensor arrays than the ones used in the HA and do not address the problem of the limited noise reduction performance of the LCMV when the number of speakers and microphones are close.

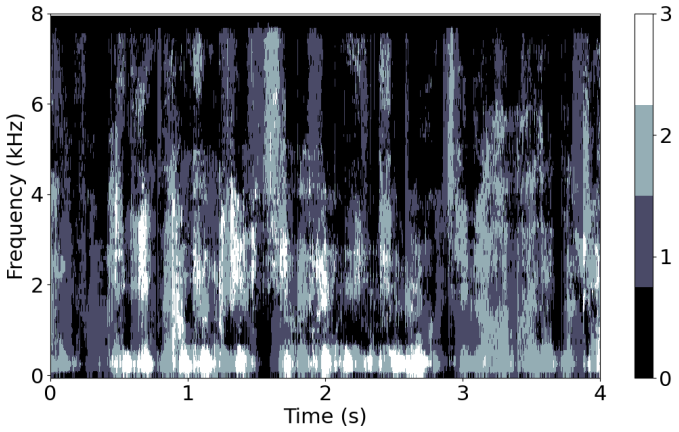

Figure 1. Example of the number of active sources in the STFT domain for a mixture of three sentences.

To solve this problem, we can consider an additional hypothesis, the W-disjoint orthogonality of speech sources in the Short-Time Fourier Transform (STFT) domain [17]. This consists in exploiting the sparsity of the speech signals in the STFT domain by assuming that they do not overlap so that only one source is active at a given time-frequency (T-F) point. This allows us to use a beamformer targeting only one source per T-F point, known as the Minimum Variance Distortionless Response (MVDR) beamformer. It exhibits good noise reduction performance in a two-speaker scenario [3], [23] and it has a reduced computational complexity.

However, for a number of sources greater than 2, this assumption becomes false on a non-negligible proportion of T-F points. For instance, in a three-speaker scenario, it is false for about 30 % of the T-F points[1]. We observe in Fig. 1 that the speech signals mostly overlap at low frequencies, which also correspond to the most energetic frequency area of speech. Therefore, although in most T-F points the hypothesis is valid, the remaining ones where it is not valid correspond to the most critical areas of the speech spectrum. In the example of Fig. 1, the T-F points subject to overlap are on average 20 dB louder than the average speech level.

In this study, we propose a beamforming method based on a milder hypothesis: the speech sources in the STFT domain

[1]obtained with the ideal time-frequency voice activity detector and the STFT parameters described in Section IV-A.

are allowed to overlap but most of the time they do not. To our knowledge, only one work studied a similar relaxed sparsity assumption for source separation problem (without diffuse noise) and with a Soundfield microphone array rather than hearing aids [9]. These differences in the application context lead to different implications. Furthermore, the computational efficiency was not a concern for the authors and has not been investigated. From this more flexible sparsity assumption, we derive a beamforming algorithm and assess its noise reduction performance and its computational complexity in the HA context. Interestingly, the proposed beamformer can be seen as a special case of the parametric multispeaker multichannel Wiener filter [13]. The performance is compared to the MVDR beamformer, result of the W-disjoint orthogonality assumption, and to the LCMV beamformer [19] for which all the sources are assumed to be present at each T-F point.

II. Signal model

We consider an auditory scene composed of Q point sources, denoted by $s_q(t)$. The transformation between the q^{th} source and the m^{th} microphone is modeled by a linear filtering whose impulse response is denoted by $h_{m,q}(t)$. It is also assumed that there is a different noise component per microphone, denoted by $n_m(t)$. Then, the signal received at the m^{th} microphone can be written as follows:

$$x_m(t) = \sum_{q=1}^{Q} h_{m,q}(t) \star s_q(t) + n_m(t), \tag{1}$$

where $\star$ is the convolution operator. This mixture model is usually expressed in the STFT domain. When the length of $h_{m,q}(t)$ is lower than the size of the STFT analysis window, convolution can be approximated by a simple product [1]:

$$x_m(k,\ell) = \sum_{q=1}^{Q} h_{m,q}(k)\ s_q(k,\ell) + n_m(k,\ell), \tag{2}$$

where k and ℓ are the frequency and time indices, respectively. This expression can be rewritten in matrix form by stacking the variables along the microphones and sources axes:

$$\mathcal{M}_1: \quad \mathbf{x}(k,\ell) = \mathbf{H}(k,\ell)\ \mathbf{s}(k,\ell) + \mathbf{n}(k,\ell), \tag{3}$$

with $\mathbf{H}(k,\ell) \in \mathbb{C}^{M\times Q}$ the mixing matrix containing the Acoustic Transfer Functions (ATFs), $\mathbf{n}(k,\ell) \in \mathbb{C}^{M}$ and $\mathbf{s}(k,\ell) \in \mathbb{C}^{Q}$.

Assuming that only one source is active at each T-F point (the so-called W-disjoint orthogonality assumption [17]), the previous expression can be written as follows:

$$\mathcal{M}_2: \quad \mathbf{x}(k,\ell) = \mathbf{h}_{q(k,\ell)}(k)\ s_{q(k,\ell)}(k,\ell) + \mathbf{n}(k,\ell), \tag{4}$$

where $q(k,\ell)$ is the index of the active speech source at the T-F point (k,ℓ).

The alternative hypothesis proposed in this study is to consider all intermediate configurations from $\kappa(k,\ell) = 0$ up to $\kappa(k,\ell) = Q$ active sources at T-F point (k,ℓ):

$$\mathcal{M}_3: \quad \mathbf{x}(k,\ell) = \tilde{\mathbf{H}}(k,\ell)\tilde{\mathbf{s}}(k,\ell) + \mathbf{n}(k,\ell), \tag{5}$$

where $\tilde{\mathbf{H}}(k,\ell) \in \mathbb{C}^{M\times\kappa(k,\ell)}$ and $\tilde{\mathbf{s}}(k,\ell) \in \mathbb{C}^{\kappa(k,\ell)}$. In the following, we refer to the models described in (3), (4) and (5) as $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$, respectively.

Furthermore, $s_q(k,\ell)$ and $\mathbf{n}(k,\ell)$ are modeled as random variables following a centered complex isotropic normal distribution with a variance $\phi_{s_q}(k,\ell)$ and covariance matrix $\mathbf{\Phi}_{\mathbf{n}}(k,\ell)$, respectively. The noise is assumed to be cylindrically spatially diffuse, allowing us to decompose its covariance matrix as the product of the time-invariant coherence matrix corresponding to a spatially cylindrical diffuse noise, denoted $\mathbf{\Gamma}_{\text{d}}(k)$, and a scaling factor, denoted $\phi_n(k,\ell)$ [6]. In practice, the matrix $\mathbf{\Gamma}_{\text{d}}(k)$ is estimated by averaging all the ATF of the horizontal plane [11]. Finally, the ATFs are assumed to be known for all the sources located in the horizontal plane. Moreover, similarly to [4], we assume the *oracle* knowledge of the directions of arrival of the sources present in the auditory scene.

III. Noise reduction methods

A. Algorithms derivation

The ideal beamformer output does not contain the noise component and is only composed of the sum of the Q speech sources filtered by the corresponding transfer function $g_q(k,\ell)$ containing, for instance, the desired localization cues [13]:

$$y(k,\ell) = \mathbf{g}^H(k,\ell)\mathbf{s}(k,\ell), \tag{6}$$

where $\mathbf{g}(k,\ell) = \left[g_1^*(k,\ell), ..., g_Q^*(k,\ell)\right]^T \in \mathbb{C}^Q$. They may be time dependent if the sources move for example.

The beamformer output, denoted by $\hat{y}(k,\ell)$, is built as a linear combination of the microphone signals mixed with the weights $\mathbf{w}(k,\ell) \in \mathbb{C}^M$:

$$\hat{y}(k,\ell) = \mathbf{w}^H(k,\ell)\mathbf{x}(k,\ell). \tag{7}$$

Determining $\mathbf{w}_{\mathcal{M}_1}(k,\ell)$, the weights of the beamformer for the $\mathcal{M}_1$ model, consists in minimizing the variance of the noise component at the output of the beamformer subject to the constraint of preserving the frequency response of the target sources:

$$\mathbf{w}_{\mathcal{M}_1}(k,\ell) = \underset{\mathbf{w}}{\text{argmin}} \left\{\phi_n(k,\ell)\mathbf{w}^H\mathbf{\Gamma}_{\text{d}}(k)\mathbf{w}\right\} \tag{8}$$

$$\text{s.t.} \quad \mathbf{w}^H\mathbf{H}(k,\ell) = \mathbf{g}^H(k,\ell). \tag{9}$$

Using the Lagrange multipliers, we obtain:

$$\mathbf{w}_{\mathcal{M}_1}(k,\ell) = \mathbf{\Gamma}_{\text{d}}^{-1}(k)\mathbf{H}(k,\ell)\left(\mathbf{H}^H(k,\ell)\mathbf{\Gamma}_{\text{d}}^{-1}(k)\mathbf{H}(k,\ell)\right)^{-1}\mathbf{g}(k,\ell). \tag{10}$$

This solution is called in the literature the LCMV beamformer [19].

The model $\mathcal{M}_2$ is the special case of $\mathcal{M}_1$ for which only one source $s_{q(k,\ell)}(k,\ell)$ is present at each T-F point (k,ℓ). We can then write the solution as:

$$\mathbf{w}_{\mathcal{M}_2}(k,\ell) = \frac{\mathbf{\Gamma}_{\text{d}}^{-1}(k)\mathbf{h}_{q(k,\ell)}(k)}{\mathbf{h}_{q(k,\ell)}^H(k)\mathbf{\Gamma}_{\text{d}}^{-1}(k)\mathbf{h}_{q(k,\ell)}(k)}. \tag{11}$$

This solution corresponds to the MVDR beamformer or more precisely to the beamformer maximizing the directivity

index [18] because we consider a noise coherence matrix corresponding to a spatially diffuse one.

Finally, the proposed $\mathcal{M}_3$ model leads, as the $\mathcal{M}_1$ model, to the LCMV beamformer by replacing $\mathbf{H}$ by $\tilde{\mathbf{H}}$ and $\mathbf{g}$ by $\tilde{\mathbf{g}}$:

$$\mathbf{w}_{\mathcal{M}_3} = \mathbf{\Gamma}_{\mathrm{d}}^{-1}\tilde{\mathbf{H}}\left(\tilde{\mathbf{H}}^H\mathbf{\Gamma}_{\mathrm{d}}^{-1}\tilde{\mathbf{H}}\right)^{-1}\tilde{\mathbf{g}}, \tag{12}$$

where the indices k and ℓ has been omitted for the sake of brevity. Let us recall that the dimensions of $\tilde{\mathbf{H}}(k,\ell) \in \mathbb{C}^{M\times\kappa(k,\ell)}$ and $\tilde{\mathbf{g}}(k,\ell) \in \mathbb{C}^{\kappa(k,\ell)}$ vary with the T-F indices (k,ℓ). By making the assumption that the speech source signals in the STFT domain can overlap but most of the time they do not, the average constraints number in the optimization problem is expected to be lower than for $\mathbf{w}_{\mathcal{M}_1}$, letting more degrees of freedom allocated to the noise reduction task. Furthermore, unlike the $\mathcal{M}_2$ for which it is assumed that one speech source is always present, $\mathcal{M}_3$ considers the case where no source is active, leading to $\mathbf{w}_3(k,\ell) = 0$.

B. Analysis of the solution

In this subsection, we analyze $\mathbf{w}_{\mathcal{M}_1}, \mathbf{w}_{\mathcal{M}_2}$ and $\mathbf{w}_{\mathcal{M}_3}$ as special cases of the Parametric Multispeaker Multichannel Wiener Filter (PMMWF) aiming at minimizing the noise power at the output of the beamformer as well as the distortion between the ideal speech sources and their estimates [13]. By removing the indices k and ℓ for brevity, we can write the determination of the filter, denoted $\mathbf{w}_{\text{PMMWF}}$ as the following optimization problem:

$$\mathbf{w}_{\text{PMMWF}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \begin{array}{c} \mathbf{w}^H\mathbf{\Phi}_{\mathbf{n}}\mathbf{w} \\ +\sum_{q=1}^{Q}\lambda_q\mathbb{E}\left[|g_q s_q - \mathbf{w}^H\mathbf{h}_q s_q|^2\right] \end{array} \right\}, \tag{13}$$

where λ_q, $q \in \{1,...,Q\}$ control the speech distortion amount. This problem accepts the following closed-form solution:

$$\mathbf{w}_{\text{PMMWF}} = \mathbf{\Gamma}_{\mathrm{d}}^{-1}\mathbf{H}\left(\phi_n\mathbf{\Lambda}^{-1}\mathbf{\Phi}_{\mathbf{s}}^{-1} + \mathbf{H}^H\mathbf{\Gamma}_{\mathrm{d}}^{-1}\mathbf{H}\right)^{-1}\mathbf{g}, \tag{14}$$

where $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1,...,\lambda_Q\}$ and $\mathbf{\Phi}_{\mathbf{s}} = \mathrm{diag}\{\phi_{s_1},...,\phi_{s_Q}\}$ is the speech sources covariance matrix. The optimal way to set $\mathbf{\Lambda}$ is not straightforward. Several strategies have been proposed, for example setting $\lambda_q = \mathrm{sig}\left(\phi_{s_q}/\phi_n\right)$ with $\mathrm{sig}\left(.\right)$ a sigmoid function [21], or setting λ_q as the posterior speech presence probabilities [2], or with $\lambda_q \rightarrow \infty\ \forall q$, reducing the PMMWF to the LCMV beamformer. The proposed beamformer $\mathbf{w}_{\mathcal{M}_3}$ can be interpreted as setting $\lambda_q \rightarrow \infty$ if source q is active, and $\lambda_q = 0$ otherwise.

C. Computational complexity analysis

In this subsection, we analyze the computational complexity of the beamformers presented previously, defined as the number of products required to compute the corresponding filter. To do so, we assume that it is not possible to pre-compute and store the filters. For instance, the number of possible $\mathbf{H}(k,\ell)$ is equal to the binomial coefficient $\binom{D}{Q}$ where D is the number of known directions. For a horizontal plane sampled with a step of $5°$, $D = 72$, leading to 59640 possible $\mathbf{H}$ per frequency with $Q = 3$. This makes it prohibitive to compute offline and store all possible filters $\mathbf{w}_{\mathcal{M}_1}$. In Tab. I, we provide the details for determining the number of products required to compute the LCMV beamformer $\mathbf{w}_{\mathcal{M}_1}$. The results for the three beamformers are presented in the Tab. II.

Operation	nb. of products	Operation	nb. of products
$\mathbf{N} = \mathbf{\Gamma}_{\mathrm{d}}^{-1}\mathbf{H}$	M^2Q	$\mathbf{z} = \mathbf{D}^{-1}\mathbf{g}$	$\frac{Q^3}{6} + Q^2$
$\mathbf{D} = \mathbf{H}^H\mathbf{N}$	MQ^2	$\mathbf{w}_{\mathcal{M}_1} = \mathbf{N}\mathbf{z}$	MQ

Table I
DETAIL OF THE NUMBER OF PRODUCTS REQUIRED TO COMPUTE THE LCMV BEAMFORMING FILTER. SOLVING $\mathbf{D}\mathbf{z} = \mathbf{g}$ REQUIRES $Q^3/6 + Q^2$ PRODUCTS, EXPLOITING THE FACT THAT $\mathbf{D}$ IS POSITIVE-DEFINITE [16].

Filter	Average number of products
$\mathbf{w}_{\mathcal{M}_1}$	$(MQ+Q^2)(M+1)+\frac{Q^3}{6}$
$\mathbf{w}_{\mathcal{M}_2}$	M^2+M
$\mathbf{w}_{\mathcal{M}_3}$	$\alpha_1(M^2+M)+\sum_{\kappa=2}^{Q}\alpha_\kappa\left((M\kappa+\kappa^2)(M+1)+\frac{\kappa^3}{6}\right)$

Table II
AVERAGE NUMBER OF PRODUCTS NEEDED TO COMPUTE THE BEAMFORMING FILTER. α_κ DENOTES THE PROPORTIONS OF T-F POINTS FOR WHICH κ SOURCES ARE ACTIVES.

It is worth noting that the number of products per time frame required by the computation of the proposed beamformer $\mathbf{w}_{\mathcal{M}_3}$ is no longer constant, as it depends on the number of active sources at each frequency. Its average depends on the proportions of T-F points $\alpha_\kappa \in [0;1]$ for which $\kappa \in \{0,...,Q\}$ sources are active ($\sum_\kappa \alpha_\kappa = 1$).

Finally, we have to mention that the computation of the LCMV filter in the 2-speaker case ($Q = 2$) can be accelerated with the efficient implementation proposed in [8]. Taking this improvement into account in our more general 3-speaker scenario is left for future work.

D. Voice activity detection (VAD)

In order to detect which speech source is present or not at each T-F point, we propose to use a voice activity detector (VAD) based on the thresholding of the SNR at the output of an MVDR beamformer steering to the q^{th} source, denoted by $\xi_{\text{MVDR},q}(k,\ell)$ [20] :

$$\text{VAD}_q(k,\ell) = \left\{ \begin{array}{ll} 1 & \text{if } \xi_{\text{MVDR},q}(k,\ell) > 10^{\frac{\tau}{10}} \\ 0 & \text{otherwise,} \end{array} \right. \tag{15}$$

where $\tau \in \mathbb{R}$ is the threshold. The estimation of the SNR at the output of the MVDR steering to the q^{th} source, denoted by $\hat{\xi}_{\text{MVDR},q}(k,\ell)$, is expressed as follows:

$$\hat{\xi}_{\text{MVDR},q}(k,\ell) = \frac{\hat{\phi}_{s_q}(k,\ell)}{\hat{\phi}_{n,q}(k,\ell)}\mathbf{h}_q^H(k)\mathbf{\Gamma}_{\mathrm{d}}^{-1}(k)\mathbf{h}_q(k), \tag{16}$$

where $\hat{\phi}_{s_q}(k,\ell)$ and $\hat{\phi}_{n,q}(k,\ell)$ are respectively the estimates of the q^{th} source and noise variances assuming that only the

q^{th} source is active [20]:

$$\hat{\phi}_{s_q}(k,\ell) = \mathbf{w}_{\text{MVDR},q}^H(k)(\mathbf{\Phi}_{\mathbf{x}}(k,\ell) - \hat{\phi}_{n,q}(k,\ell)\mathbf{\Gamma}_{\text{d}}(k))\mathbf{w}_{\text{MVDR},q}(k) \tag{17}$$

$$\hat{\phi}_{n,q}(k,\ell) = \frac{1}{M-1}\text{Tr}\left\{(\mathbf{I} - \mathbf{h}_q(k)\mathbf{w}_{\text{MVDR},q}^H(k))\mathbf{\Phi}_{\mathbf{x}}(k,\ell)\mathbf{\Gamma}_{\text{d}}^{-1}(k)\right\} \tag{18}$$

where $\mathbf{\Phi}_{\mathbf{x}}(k,\ell)$ is the microphone covariance matrix, estimated thanks to a recursive filter, and

$$\mathbf{w}_{\text{MVDR},q}(k) = \frac{\mathbf{\Gamma}_{\text{d}}^{-1}(k)\mathbf{h}_q(k)}{\mathbf{h}_q^H(k)\mathbf{\Gamma}_{\text{d}}^{-1}(k)\mathbf{h}_q(k)}. \tag{19}$$

IV. Experiments

In this section, we assess the three denoising algorithms in terms of noise reduction and algorithmic complexity. In the following, the LCMV, MVDR and the proposed beamformers refer to the filters $\mathbf{w}_{\mathcal{M}_1}$, $\mathbf{w}_{\mathcal{M}_2}$ and $\mathbf{w}_{\mathcal{M}_3}$, respectively.

A. Evaluation methods

The algorithms are tested by processing virtual auditory scenes composed of three speech sources of 4 s duration and a cafeteria noise played over two virtual speaker rings located at elevations $\pm 45°$ mixed at various SNR ranging from 0 to 10 dB with a 2.5 dB step. The speech signals are recorded from the France Culture radio station sampled at 16 kHz and spatialized on the horizontal plane at azimuths $\{-45°, 0°, 45°\}$. The Behind-the-Ears HA ATF ($M = 4$) used for the virtual auditory scene generation and the beamforming algorithms come from [15]. In total, 40 audio examples[2] are generated for each tested SNR. The algorithms are integrated into an overlap-add processing chain with a window size of 128 samples and an overlap of 50 %. Each frame is expressed in the frequency domain without zero padding. The MVDR and the proposed beamformers are tested using the VAD based on the ideal and the estimated SNR.

To assess the noise reduction, we consider the Signal-to-Artifact Ratio (SAR) and the improvements in terms of the Signal-to-Distortion Ratio (ΔSDR) and Signal-to-Interferer Ratio (ΔSIR) [22] which are defined respectively as the ratio between the power of the target signal, as defined in (6), and (i) the artifacts generated by the beamforming, (ii) the other components in the output signal, and (iii) the interfering noise component.

B. Results

First, let us compare the ΔSDR and the computational complexity in Fig. 2. We observe that the MVDR beamformer is more than 5 times less complex than the LCMV beamformer and that it improves the ΔSDR by about 1 dB over the latter. In the tested scenario, both algorithms have very poor distortion reduction performance. Using a VAD based on an oracle SNR, the proposed algorithm improves the performance of ΔSDR by 6.5 dB with an optimal detection threshold setting while only slightly increasing the algorithmic complexity compared to the MVDR beamformer ($+50\%$). It can be noted thanks to the ellipses representing the standard deviation of the data that the ΔSDR is negatively correlated with the computational complexity. This is because the more speech sources overlap, the greater the number of constraints in the optimization, thus increasing the computational cost and reducing the size of the subspace on which noise reduction can be performed, leading to a lower performance on average. When using the VAD based on the estimated SNR (and not oracle), the algorithmic complexity remains unchanged but the performance of the proposed method in terms of ΔSDR decreases significantly, even though it still outperforms the two other beamformers. This shows that the VAD has a great impact on the noise reduction performance of the proposed beamforming method.

Second, let us take a closer look at the denoising performance by studying the ΔSIR and the SAR as a function of the input SNR, as showed in Fig. 3. The MVDR beamformer is very efficient to reduce the noise (ΔSIR=13 dB at SNR=0 dB with the oracle VAD) compared to the LCMV beamformer

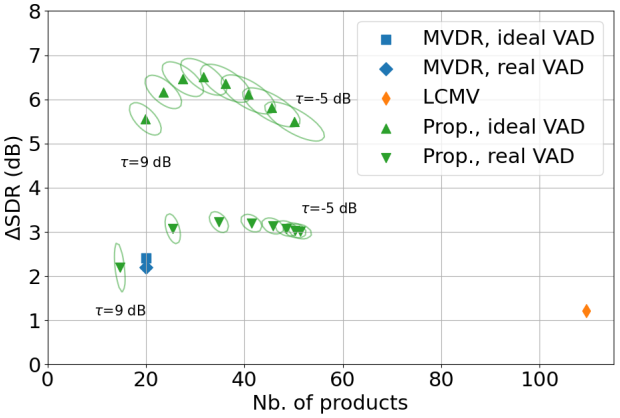

Figure 2. SDR improvement versus computational complexity averaged over the examples for an SNR of 0 dB. The ellipses show the standard deviation isovalue of a 2D gaussian distribution fitted over the results for the proposed algorithm with the VAD threshold τ ranging from -5 to 9 dB with a 2 dB step.

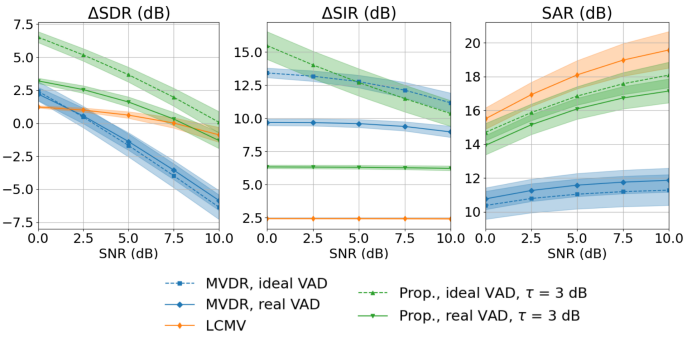

Figure 3. ΔSDR (left), ΔSIR (center) and SAR (right) for an input SNR varying from 0 to 10 dB. The line shows the average across the examples and the transparent surface the standard deviation. We only display the proposed algorithm performance using a VAD threshold τ=3 dB which maximizes the ΔSDR as shown in Fig. 2.

[2]Audio examples repository URL: https://a-llave.github.io/demo_apsipa2021

(ΔSIR=2.5 dB). It is expected as the first one uses only one degree of freedom of the optimization to address the preservation of speech sources. However, by preserving only one source per T-F point, this beamformer introduces more artifacts (SAR=10 dB at SNR=0 dB) compared to the LCMV beamformer (SAR=15 dB). For a high input SNR, the artifact amount introduced by the MVDR beamformer can become larger than the noise component level in the original mixture, resulting in a negative ΔSDR as can be seen in Fig. 3. The proposed method achieves SAR performance similar to that obtained with the LCMV beamformer, although slightly lower. Regarding the ΔSIR, it achieves 15 dB improvement (at SNR=0 dB) for a threshold setting maximizing the ΔSDR (τ=3 dB). However, this score decreases sharply when using the VAD based on the estimated SNR. Indeed, this VAD algorithm tends to make a lot of false positives, thus reducing the number of degrees of freedom for denoising. Nevertheless, as shown by the overall performance measure ΔSDR, the proposed method obtains similar or better results compared with the two other beamformers, while being efficient in terms of computational complexity as previously shown.

Finally, note that the proposed algorithm is more sensitive to the VAD estimation errors than the MVDR beamformer. Indeed, the latter needs to know only the most energetic source while the former needs to know precisely which source is active or not.

V. Conclusion

In this work, we proposed a new beamforming algorithm that exploits the sparsity of speech signals in the STFT domain, in a less restrictive manner compared with the popular W-disjoint orthogonality assumption. In a three-speaker scenario, experimental results show that the LCMV and MVDR beamformers exhibits two extreme behaviors: the first one preserves well the speech sources but does not achieve a good noise reduction, whereas the latter reduces dramatically the noise and the computational complexity but introduces a lot of artifacts. The proposed method achieves to be beneficial both in terms of noise reduction and speech distortion without increasing too much the computational cost. We limited the investigation to $Q = 3$ because we used an array of four microphones. Future work will have to investigate the performance of the proposed method for a larger microphone array [10] and to improve the VAD to get closer to the noise reduction performance upper bound.

References

[1] Yekutiel Avargel and Israel Cohen. On Multiplicative Transfer Function Approximation in the Short-Time Fourier Transform Domain. *IEEE Signal Processing Letters*, 14(5):337–340, May 2007.

[2] Saeed Bagheri and Daniele Giacobello. Exploiting Multi-Channel Speech Presence Probability in Parametric Multi-Channel Wiener Filter. In *Interspeech 2019*, pages 101–105. ISCA, September 2019.

[3] Sebastian Braun, Wei Zhou, and Emanuel A. P. Habets. Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, New Paltz, NY, USA, October 2015. IEEE.

[4] Ryan M. Corey and Andrew C. Singer. Dynamic range compression for noisy mixtures using source separation and beamforming. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 289–293, New Paltz, NY, October 2017. IEEE.

[5] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov. A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730, April 2017.

[6] Steven L. Gay and Jacob Benesty, editors. *Acoustic Signal Processing for Telecommunication*. Springer US, Boston, MA, 2000.

[7] Xiansheng Guo, Baogen Xu, Zhongchu Rao, Qun Wan, Zhengming Feng, and Yijiang Shen. Low-Complexity Iterative Adaptive Linearly Constrained Minimum Variance Beamformer. *Circuits, Systems, and Signal Processing*, 33(3):987–997, March 2014.

[8] Elior Hadad, Simon Doclo, and Sharon Gannot. The Binaural LCMV Beamformer and its Performance Analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):543–558, March 2016.

[9] Maoshen Jia, Jundai Sun, and Xiguang Zheng. Multiple Speech Source Separation Using Inter-Channel Correlation and Relaxed Sparsity. *Applied Sciences*, 8(123), January 2018.

[10] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier. Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses. *EURASIP Journal on Advances in Signal Processing*, 2009(1), December 2009.

[11] Thomas Lotter and Peter Vary. Dual-Channel Speech Enhancement by Superdirective Beamforming. *EURASIP Journal on Advances in Signal Processing*, 2006(1), December 2006.

[12] Shmulik Markovich-Golan, Sharon Gannot, and Israel Cohen. Low-Complexity Addition or Removal of Sensors/Constraints in LCMV Beamformers. *IEEE Transactions on Signal Processing*, 60(3):1205–1214, March 2012.

[13] Shmulik Markovich-Golan, Sharon Gannot, and Israel Cohen. A weighted multichannel Wiener filter for multiple sources scenarios. In *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*, pages 1–5, Eilat, Israel, November 2012. IEEE.

[14] Ryan W. McCreery, Rebecca A. Venediktov, Jaumeiko J. Coleman, and Hillary M. Leech. An Evidence-Based Systematic Review of Directional Microphones and Digital Noise Reduction Hearing Aids in School-Age Children With Hearing Loss. *American Journal of Audiology*, 21(2):295, December 2012.

[15] Chris Oreinos and Jörg M. Buchholz. Measurement of Full 3D Set of HRTFs for In-Ear and Hearing Aid Microphones on a Head and Torso Simulator. *Acta Acustica united with Acustica*, 99:836–844, 2013.

[16] William H. Press, Saul A. Teukolsky, William T. Vetterling, and rian P. Flannery. *Numerical Recipes*. Cambridge University Press, third edition, 2007.

[17] Scott Rickard and Özgür Yilmaz. On the Approximate W-Disjoint Orthogonality of Speech. In *ICASSP 2002 - 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 4, Orlando, USA, May 2002.

[18] Robert W. Stadler and William M. Rabinowitz. On the potential of fixed arrays for hearing aids. *The Journal of Acoustical Society of America*, 94(3):1332–1342, September 1993.

[19] Yôiti Suzuki, Shinji Tsukui, Futoshi Asano, Ryouichi Nishimura, and Toshio Sone. New Design Method of a Binaural Microphone Array Using Multiple Constraints. *IEICE Trans. Fundamentals*, 82(4):588–596, April 1999.

[20] Joachim Thiemann, Menno Müller, Daniel Marquardt, Simon Doclo, and Steven van de Par. Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene. *EURASIP Journal on Advances in Signal Processing*, 2016(1), December 2016.

[21] Oliver Thiergart and Emanuël A P Habets. An Informed Parametric Spatial Filter Based on Instantaneous Direction-of-Arrival Estimates. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):15, 2014.

[22] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing, Institute of Electrical and Electronics Engineers*, page 10, 2006.

[23] Mehdi Zohourian, Gerald Enzner, and Rainer Martin. Binaural Speaker Localization Integrated Into an Adaptive Beamformer for Hearing Aids. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3):515–528, March 2018.