A Multichannel Audio Denoising Formulation Based on Spectral Sparsity

İlker Bayram

***Abstract*—We consider the estimation of an audio source from multiple noisy observations, where the correlation between noise in the different observations is low. We propose a two-stage method for this estimation problem. The method does not require any information about noise and assumes that the signal of interest has a sparse time-frequency representation. The first stage uses this assumption to obtain the best linear combination of the observations. The second stage estimates the amount of remaining noise and applies a post-filter to further enhance the reconstruction. We discuss the optimality of this method under a specific model and demonstrate its usefulness on synthetic and real data.**

***Index Terms*—Multichannel audio denoising, sparsity, spectrogram, post-filter, beamforming, sufficient statistic, UMVU estimator.**

I. Introduction

Reconstructing an audio signal from multiple noisy observations requires us to overcome challenges that differ according to conditions. In this paper, we consider the case where the noise or interference terms in the different observations have low correlation. Such a case occurs if the observations are obtained from microphones scattered in an environment, or if each noise source is closer to one of the microphones (as in Fig. 14, discussed in Exp. 4). Another example is restoration of a sound recording given several degraded copies, where the degradation in each copy is assumed to affect different time-frequency regions. The proposed scheme in this paper consists of an adaptive weighting stage that linearly combines the multiple observations, followed by a post-filter. Such a two-stage approach is common in multichannel audio reconstruction (see e.g. [21, 35]). However, unlike previous work, our formulation does not require additional information about the second order statistics of noise and is based on the sparsity of the source signal's time-frequency representation.

Determination of adaptive weights is related to 'beamforming' formulations that aim to optimally combine multiple observations. Classical beamforming formulations employ models based on second order statistics of signals. In particular, in the minimum variance distortionless response (MVDR) beamformer (see [6, 21, 35] and the references therein), under a stationarity assumption, one restricts herself/himself to estimates obtained by linear time-invariant (LTI) filtering and summing of the observations. In such a setting, one seeks the filters that minimize the expected output power, where the sum of the filters are subject to certain constraints designed to avoid introducing any distortion to the signal of interest [10, 27]. More recently, subspace based multichannel estimation methods have been proposed [1, 16, 28], which make use of the fact that frames of speech/source signals typically reside in a lower-dimensional subspace than noise [18]. For these methods, closed form solutions for the filters can be derived but they involve correlation functions of the signals. Therefore, in order to apply such methods, if statistics of the signals are not available, they need to be estimated. We also refer to [36] for the description of a recent framework where noise statistics are estimated in a multiple speaker setup.

Post-filtering was originally proposed by Zelinski as an ad-hoc Wiener filter applied to the output of a delay-and-sum beamformer [39]. The idea was to treat the beamformer output as a 'noisy signal' and achieve a higher SNR through Wiener filtering. The statistics required for this single-channel Wiener filter were estimated by assuming that the signal of interest and noise terms are uncorrelated. Later, the relationship between this single-channel post-filter and multi-channel Wiener filtering was noted and modified post-filters were proposed [35]. Specifically, it was shown in [35] that a multichannel Wiener filter can alternatively be realized by applying a single channel Wiener filter to the output of an MVDR beamformer. In other words, the multi-channel Wiener filter can be factorized into an MVDR beamformer followed by a single-channel Wiener filter. For a relatively recent overview of post-filtering methods, we refer to [21]. An interesting scheme that is of interest for our formulation is [2] where the authors derive multichannel versions of the Ephraim-Malah estimator for speech. They show as a byproduct that a sufficient statistic of interest coincides with the beamformer output according to their model. This sufficient statistic requires knowledge of the statistical parameters of noise, which need to be estimated in practice.

The beamforming methods discussed above do not make explicit use of the sparsity of the source signal's time-frequency representation. Also, they typically need to estimate and keep track of second order statistics of the desired and unwanted sources and of the noise terms. In this paper, we formulate the selection of the adaptive weights as a convex minimization problem that looks for a reconstruction with a sparse time-frequency representation. The minimization problem does not require us to estimate statistics of the source or noise signals as a preprocessing step. The proposed formulation is different from the one in our previous work [3] and allows us to naturally develop a post-filter as well. In [3], the formulation was primarily in the time-domain and the (time-domain) noise samples were assumed to be independent with a time-varying

İ. Bayram is with the Dept. of Electronics and Communication Engineering, Istanbul Technical University, Istanbul, Turkey (e-mail : ibayram@itu.edu.tr).
This research is supported by TÜBİTAK (Project No :113E511).

variance. Here, we consider time *and* frequency varying noise. In order to handle noise with such characteristics, we work with time-frequency blocks.

For the post-filtering stage, we adopt an observation model similar to that used by [2]. Consequently, the output of the adaptive weighting stage acts as a complete sufficient statistic in our framework also. However, in our framework, the sufficient statistic is estimated by the adaptive weighting stage without any knowledge of the signal and noise statistics, only making use of the prior knowledge that the spectrogram of the reconstruction is (approximately) sparse. As in [2], adaptive weighting results in a noisy single-channel signal. To eliminate this remaining noise, we first estimate the variance of the remaining noise. We then use this estimate in a block-based denoiser similar in spirit to the denoising schemes proposed by Cai [9] and Yu et al. [38].

Outline

In Section II, we describe the formulations for the adaptive weighting and post-filtering stages. A short discussion on the optimality of the two-step reconstruction scheme is also included. An algorithm for obtaining the solutions of the formulation for determining the weights is given in Section III. Experiments demonstrating the utility of the formulations and comparisons with well-known algorithms can be found in Section IV. Finally, some concluding discussion is provided in Section V.

II. Problem Formulation

Observation Model

We assume that $x(n)$ is the signal of interest (in the time-domain) but we have M noisy observations $y_i(n)$ as

$$y_i(n) = x(n) + u_i(n), \text{ for } i = 1, 2, \ldots, M, \tag{1}$$

where $u_i(n)$ denotes the noise term affecting the i^{th} observation.

In this paper, we will work in the short-time Fourier transform (STFT) domain. The STFT of a time-domain signal $z(n)$ is denoted by $Z(k, s)$, where k and s denote the time and frequency parameters respectively. Given a window function $g(n)$ of length N, and a 'hop-size' K, the STFT of z is defined as,

$$Z(k, s) = \sum_n z(n)\, g(n - k\,K)\, \exp\left(-j\,\frac{2\pi}{N} s(n - k\,K)\right).$$

With this notation, we can express (1) in the STFT domain as,

$$Y_i(k, s) = X(k, s) + U_i(k, s), \text{ for } i = 1, 2, \ldots, M. \tag{2}$$

Noise Model

We model the STFT coefficients of the noise terms $U_i(k, s)$ as independent complex valued Gaussian random variables. That is, $U_i(k, s)$ and $U_{i'}(k', s')$ are independent if $(i, k, s) \neq (i', k', s')$. At a specific time-frequency point (k, s), the real and imaginary parts of $U_i(k, s)$ are also assumed to be independent zero-mean Gaussian random variables with the same variance $\sigma_i^2(k, s)$[1]. In other words, $U_i(k, s)$ is a circular normal random variable [30]. Note that such a model cannot be true if the STFT is overcomplete because independence is not preserved in that case. Nevertheless, we employ this assumption because it simplifies the problem significantly. Another assumption we will make about noise is that the variance field $\sigma_i^2(k, s)$ changes slowly with respect to (k, s).

Signal Model

We assume that $X(k, s)$, the STFT coefficients of the audio source, is a sparse, or compressible 2D signal – we note that this is a widely used assumption, thanks to the growing interest in sparse representations [25, 31]. In order to capture this property, we use the ℓ_1 norm (see [25] in this context) defined in this case as,

$$\|X\|_1 = \sum_{k,s} |X(k, s)|. \tag{3}$$

Before detailing them, in the following we briefly discuss why there are two stages in our formulation

Motivation for the Two-Stage Formulation

Given a prior distribution for the signal of interest, it is well-known that any optimal Bayesian estimate is a function of the sufficient statistic (see e.g. Sec. 4.2.1 in [4]). Unfortunately, lacking knowledge on the noise variance, we can only *estimate* the sufficient statistic in our scenario. This estimation comprises the 'adaptive weighting' stage of our method. It turns out that in our model, the sufficient statistic is an unbiased estimate of X. Therefore, by the Rao-Blackwell theorem, it is the uniformly minimum variance unbiased estimate (UMVUE) for X. Thus we can also regard the output obtained by adaptive weighting as an estimate of the signal itself. However, as will be clarified in Prop. 1 below, this estimate is noisy. The post-filter which acts on the sufficient statistic/UMVUE aims to eliminate the remaining noise and it may thus be interpreted as a 'denoising' operation.

In the following, we provide the details on the formulation of the two stages, namely the selection of the adaptive weights and post-filtering in Sec. II-A and Sec. II-B.

A. Selection of the Weights

In Sec. II-A1 and Sec. II-A2, we provide two related formulations for weight selection that have different complexities. We start by assuming that the noise variance field is constant for each observation, which allows to state a simple principle that relies on Prop. 1 below. This assumption is removed in the sequence.

[1]Notice therefore that $\mathbb{E}\big(|U_i(k, s)|^2\big) = 2\sigma_i^2(k, s)$.

A Complete Sufficient Statistic for $X(k,s)$: The following proposition motivates the formulation for determining the weights.

Proposition 1. Suppose we are given observations as in (2), where $U_i(k,s)$, and $U_{i'}(k',s')$ are independent if $(i,k,s) \neq (i',k',s')$ and for each i, $U_i(k,s)$ is a circular normal random variable whose real and imaginary parts have variance $\sigma_i^2(k,s)$. If we treat $X(k,s)$ as an unknown constant, a complete sufficient statistic for $X(k,s)$ is given by

$$\tilde{X}(k,s) = \sum_{i=1}^{M} \alpha_i(k,s)\, Y_i(k,s), \tag{4}$$

where

$$\alpha_i(k,s) = \sigma_i^{-2}(k,s)\, \sigma^2(k,s), \quad \text{for } \sigma^2(k,s) = \left(\sum_{i=1}^{M} \sigma_i^{-2}(k,s) \right)^{-1}. \tag{5}$$

$\tilde{X}(k,s)$ is a circularly symmetric random variable whose real and imaginary parts have variance $\sigma^2(k,s)$. Further, $\tilde{X}(k,s)$ is the UMVUE for $X(k,s)$.

Proof. See the appendix. □

Let us temporarily assume that $\sigma_i^2(k,s) = \sigma_i^2$. That is, for each observation, we assume that the noise variance is constant over the time-frequency plane.

Note now that in Prop. 1, α_i's satisfy

$$\sum_{i=1}^{M} \alpha_i = 1 \quad \text{and} \quad \alpha_i \geq 0,\ \forall i, \tag{6}$$

that is, the sufficient statistic/UMVUE is a convex combination of the observations. A converse statement is also valid: Any convex combination like $\sum_i \beta_i Y_i$ is an UMVUE for a specific set of σ_i's – for instance if $\sigma_i^2 = \beta_i^{-1}$. We remark that, in general, the weights of a beamformer in the STFT domain are complex-valued, so as to compensate for the time-delay in the different observations. However, here the weights are real-valued because the signals are time-aligned.

Observe that if we do not know σ_i's, we cannot form the sufficient statistic/UMVUE. However, we still know that the sufficient statistic lies somewhere in the convex hull of the observations. We also note that even if the signal of interest $X(k,s)$ is truly sparse, the sufficient statistic is not expected to be sparse, because it will contain Gaussian noise, as per Prop. 1. We propose to pick the element of the convex hull that is closest to being sparse, as the estimate of the sufficient statistic/UMVUE. In order to gauge proximity to the set of sparse signals, we employ the ℓ_1 norm. More precisely, we propose to form the estimate of the sufficient statistic/UMVUE as,

$$\hat{X}(k,s) = \sum_{i=1}^{M} \hat{\alpha}_i\, Y_i(k,s), \tag{7}$$

where

$$\hat{\alpha} = \arg\min_{\alpha} \left\| \sum_{i=1}^{M} \alpha_i Y_i(k,s) \right\|_1 \text{ s.t. } \begin{cases} \alpha_i \geq 0, \\ \sum_i \alpha_i = 1. \end{cases} \tag{8}$$

In the sequel, we modify this formulation by dropping the white noise assumption. But first we introduce some notation.

Notation for Time-Frequency Partitions: The proposed formulations below are based on partitions of the time-frequency coefficients with a special notation. Recall that $Y_i(k,s)$ are the STFT coefficients of the i^{th} observation. Suppose the time-frequency indices (k,s) take values in a set $\mathcal{T}$. We assume that $\mathcal{T}_l$, for $l = 1, \ldots, L$ forms a partition of the set $\mathcal{T}$ (see Fig. 1). Here, $\mathcal{T}_l$ consists of the indices for the 'l^{th} block'. In this paper, we use rectangular blocks on the time-frequency lattice as shown in Fig. 1. Finally, $Y_i^{(l)}$ denotes a vector that contains the time-frequency samples of $Y_i(k,s)$ for $(k,s) \in \mathcal{T}_l$.

1) Block-Based Formulation: If the noise variance field $\sigma_i^2(k,s)$ is not constant but slowly varying over the time-frequency lattice, the formulation in (8) is not suitable because it asks to employ a constant weight for each observation. To cope with this problem, we propose to partition the time-frequency lattice into blocks. Note that, when the noise variance field changes slowly, we can think of it as approximately constant on small neighborhoods or blocks. Then, on each block, we can use the formulation in (8) described above.

In the 'block-based formulation', each block is treated independently. For the l^{th} block, the optimal weights $\hat{\alpha}^{(l)} \in \mathbb{R}^M$ are chosen as,

$$\hat{\alpha}^{(l)} = \arg\min_{\alpha \in \mathbb{R}^M} \sum_{(k,s) \in \mathcal{T}_l} \left| \sum_{i=1}^{M} \alpha_i Y_i(k,s) \right| \text{ s.t. } \begin{cases} \alpha_i \geq 0, \\ \sum_i \alpha_i = 1. \end{cases} \tag{9}$$

After we determine $\hat{\alpha}^{(l)} = \left(\hat{\alpha}_1^{(l)}, \ldots, \hat{\alpha}_M^{(l)}\right)$, we form the estimate at the time-frequency point $(k,s) \in \mathcal{T}_l$ as,

$$\hat{X}(k,s) = \sum_{i=1}^{M} \hat{\alpha}_i^{(l)}\, Y_i(k,s). \tag{10}$$

Referring to Prop. 1 (and (5) specifically), we see that the 'best' α_i's are directly related to σ_i's. In that respect, using large blocks has the effect of regularizing the estimation of α_i's, since a large block allows to work with more samples influenced by the values of σ_i's. However, according to our model, σ_i's are not necessarily constant in each block. If the variation of σ_i's is high, a reliable estimate of α_i's is not easy to obtain. Therefore, one should ideally select the largest blocks such that the noise variation within the blocks is negligible for practical purposes.

The block-based formulation can be effective and it is computationally attractive because the blocks are treated independently (allowing for a parallel implementation). However, if computation time is not a major constraint, the formulation can be improved, as discussed next.

2) A Smoother Formulation: We would like to point to two deficiencies of the block-based formulation.

(i) The formulation relies on the assumption that, within a block, the noise variance field does not vary much. However, this assumption is dependent on the relative size of the blocks and the norm of the noise variance

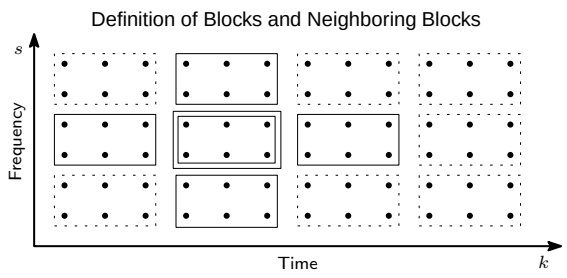

Fig. 1. The rectangles (solid or dashed) enclose time-frequency blocks used in the adaptive weighting stage. For the block framed with two rectangles, the neighboring blocks are shown with solid lines – note that we pick the closest blocks in the horizontal and vertical directions as neighbors. For blocks on the boundaries, we similarly take neighbors to be the closest blocks in the horizontal and vertical directions.

field's gradient with respect to the time and frequency indices k, s.

(ii) We would expect some interdependence between neighboring blocks, but this is not taken into account as the blocks are treated independently.

One could address the first point by taking smaller blocks. But working in blocks also provides some regularization for the problem and small blocks could lead to locally erratic selection of the weights due to a small sample size. To address these issues, we introduce a regularization term that penalizes swift changes in the weights of neighboring blocks. Note that such a modification also takes care of the second deficiency listed above.

More precisely, for the l^{th} block, let $\mathcal{N}(l)$ denote the labels of the neighboring blocks. For this paper, we define the neighbors of a given block as the closest vertical and horizontal blocks in the time-frequency plane (see Fig. 1). Also, let $\alpha = \left(\alpha^{(1)}, \ldots, \alpha^{(L)}\right)$ be the collection of weights for the whole set of blocks. Recall here that each $\alpha^{(l)}$ is a length-M vector. Under this setting, we define the regularization term as,

$$P(\alpha) = \sum_{l=1}^{L} \sum_{m \in \mathcal{N}(l)} \|\alpha^{(l)} - \alpha^{(m)}\|_2^2. \tag{11}$$

Using this P, we propose to select the optimal weights as,

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{l=1}^{L} \left(\sum_{(k,s)\in\mathcal{T}_l} \left| \sum_{i=1}^{M} \alpha_i^{(l)} Y_i(k,s) \right| \right) + \lambda\, P(\alpha), \quad \text{s.t.} \begin{cases} \sum_i \alpha_i^{(l)} = 1, \\ \alpha_i^{(l)} \geq 0, \ \forall i, \end{cases} \quad \text{for all } l, \tag{12}$$

where $\lambda \in \mathbb{R}_+$ is a regularization parameter.

The minimization problems in (9) and (12) are convex, thanks to the convexity of the cost function and the constraints. An algorithm that numerically solves these problems is provided in Section III.

B. Post-filtering

The estimate obtained by adaptive weighting ideally recovers the unbiased linear combination with the least amount of noise. But unless one of the observations is clean, this estimate will also be noisy, as implied by Prop. 1. In order to further suppress this noise, we employ a post-filter. But for that, we need to estimate the amount of remaining noise. We next discuss a simple estimator for the remaining noise variance.

1) Estimating the Noise After the Adaptive Weighting Stage: To estimate the remaining noise variance at a particular time-frequency point (k, s), we adopt an empirical approach. Specifically, we assume that the obtained adaptive weights are equal to the optimal weights. That is, we assume that, for the time-frequency point (k, s), the weights $\hat{\alpha}_i$ satisfy $\hat{\alpha}_i = \sigma_i^{-2}\, \sigma^2$ (see Prop. 1). Then, at that time-frequency point, an unbiased estimator for $\sigma^2(k, s)$ is (see the appendix),

$$\hat{\sigma}^2(k,s) = \frac{1}{2(M-1)} \sum_{i=1}^{M} \hat{\alpha}_i(k,s) \left| Y_i(k,s) - \hat{X}(k,s) \right|^2. \tag{13}$$

Note that we started with multiple observations affected by time-varying noise with *unknown* variance, but we now have at hand an unbiased estimate of the clean signal with an *estimated* noise variance field. The next step is to eliminate this remaining noise.

2) The Post-Filter: For denoising, one can employ any one of powerful denoising methods. We have experimented with two different methods that aim to achieve a sparse reconstruction in the STFT domain.

Soft-Thresholding: A simple approach is to apply a soft-threshold to each STFT coefficient. In that case, the post-filter takes the form

$$\bar{X}(k,s) = \left(1 - \frac{\tau(k,s)}{|\hat{X}(k,s)|}\right)_+ \hat{X}(k,s), \tag{14}$$

where $\tau(k, s)$ is a threshold value that depends on $\hat{\sigma}(k, s)$ and $(t)_+ = \max(t, 0)$. One choice of interest that performs well in practice is to take $\tau(k, s) = 2\, c\, \hat{\sigma}(k, s)$, where $1 \leq c \leq 3$.

Block-Thresholding: As an alternative, we consider a thresholding scheme based on blocks. This post-filter is derived from the block denoisers proposed by Cai [9] and Yu et al. [38].

Now suppose $\mathcal{C}_1, \ldots, \mathcal{C}_S$ form a partition of $\mathcal{T}$ – here we use different symbols $\mathcal{C}_s$ to emphasize that the blocks can be different than those used in the weight selection stage. Each block is treated independently. Also, we take the block sizes as $2^H \times 2^V$ (along the time×frequency axes) with $H \leq V$ for this paper.

Suppose now that we fix the block index so that the block of interest is denoted as B. In B, the audio signal might show transient or tonal behavior, or a combination [15]. Our thresholding scheme will depend on the characteristic of the signal in the block [38]. To determine the characteristic, we propose to perform simple tests. For a $v \in \{0, 1, \ldots, H\}$, we partition the block into subblocks of size $2^{H-v} \times 2^v$. Note that this gives a total of 2^V subblocks and the number of time-frequency samples in each subblock is constant with respect to v (see Fig. 2). We decide which subblock structure to use by minimizing a cost as a function of v. For a specific v, this cost function is computed by summing the ℓ_2 norms of the subblocks. We note that this procedure may be regarded as a

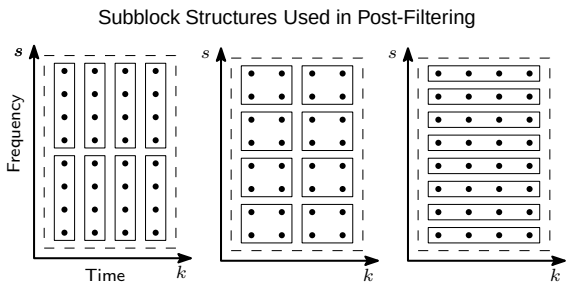

Fig. 2. Different subblock structures for $V = 3$, $H = 2$. From left to right, the subblock partitions for $v = 0, 1, 2$ are shown. We can think of the partitions as suitable for a 'transient' component when $v = 0$, a 'tonal component' when $v = 2$ and a 'combination' when $v = 1$.

Bayes test for a multiple hypothesis testing problem, where each choice of v is equally likely. For each v, the hypothesis H_v suggests that the large block B is subblock-sparse with subblock shape determined by the parameter v. The test is performed by comparing the mixed $\ell_{2,1}$ norms [26] for the large block B under different hypotheses.[2]

After choosing the subblock structure, we apply a separate threshold to each subblock. Suppose now that $\mathcal{I}$ holds the indices for a specific subblock. Our block estimator will be of the form

$$\bar{X}(k,s) = \eta \hat{X}(k,s) \quad \text{for all } (k,s) \in \mathcal{I}, \tag{15}$$

where $\eta \in \mathbb{R}$. Recall that since we assume that the optimal weights are chosen in the adaptive weighting stage, we have

$$\hat{X}(k,s) = X(k,s) + \sigma(k,s)\, Z(k,s), \tag{16}$$

where $Z(k,s)$ denote independent and identically distributed circularly symmetric random variables whose real and imaginary parts have unit variance. In this setting, the choice of η minimizing the sum of the mean squared error (MSE) distance to (the unknown constants) $X(k,s)$ is given by

$$\eta^* = \arg\min_{\eta} \sum_{(k,s)\in\mathcal{I}} \mathbb{E}\Big(\big|\eta \hat{X}(k,s) - X(k,s)\big|^2\Big) \tag{17}$$

$$= \frac{\sum\limits_{(k,s)\in\mathcal{I}} |X(k,s)|^2}{\sum\limits_{(k,s)\in\mathcal{I}} |\hat{X}(k,s)|^2}. \tag{18}$$

Lacking knowledge of $X(k,s)$, we cannot compute η^* in practice. But noting that

$$\mathbb{E}\left[\sum_{(k,s)\in\mathcal{I}} \big|\hat{X}(k,s)\big|^2\right] = \sum_{(k,s)\in\mathcal{I}} \big|X(k,s)\big|^2 + 2\sigma^2(k,s), \tag{19}$$

we can estimate η^* by employing $\hat{\sigma}^2(k,s)$. Noting also that $\eta^* \geq 0$, we use an estimate of η^* given as

$$\hat{\eta} = \left(\frac{\sum\limits_{(k,s)\in\mathcal{I}} \big|\hat{X}(k,s)\big|^2 - 2\,\hat{\sigma}^2(k,s)}{\sum\limits_{(k,s)\in\mathcal{I}} \big|\hat{X}(k,s)\big|^2}\right)_+ \tag{20}$$

$$= \left(1 - \frac{\sum\limits_{(k,s)\in\mathcal{I}} 2\hat{\sigma}^2(k,s)}{\sum\limits_{(k,s)\in\mathcal{I}} \big|\hat{X}(k,s)\big|^2}\right)_+ \tag{21}$$

III. Minimization Algorithm for Determining the Weigths

The only step that was left implicit in the description of the reconstruction method in Section II is the selection of the weights to be used in the adaptive weighting stage. The weights are obtained by solving a constrained convex minimization problem (see (9), (12)), with a non-differentiable cost function. Thanks to the growing interest in the signal processing literature on convex optimization, there are a number of alternative algorithms applicable for this problem, such as ADMM [8], PPXA [13, 14], or various saddle point algorithms [12, 19, 32]. Here, we describe an adaptation of a projected subgradient algorithm [5, 34] for the numerical solution of (9) and (12). We chose this algorithm because it is simple to implement, and its performance is comparable to other state-of-the-art algorithms (see [3] for an adaptation of a saddle point algorithm [12, 19, 32] for a related formulation).

A. The Projected Subgradient Algorithm

The projected subgradient algorithm may be regarded as an extension of the projected gradient algorithm, which is used in differentiable constrained minimization. Let us start with a definition from convex analysis (see [23] for a more detailed account).

Definition 1. If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, its *subdifferential* at x is denoted as $\partial f(x)$ and is defined to be the set of $v \in \mathbb{R}^n$ that satisfies

$$f(x) + \langle v, y - x\rangle \leq f(y), \text{ for all } y \in \mathbb{R}^n. \tag{22}$$

Any element $v \in \partial f(x)$ is said to be a *subgradient* of f at x.

Consider now a generic constrained minimization problem of the form,

$$\min_{x\in C} f(x), \tag{23}$$

where C is a closed convex set in $\mathbb{R}^n$ and $f(x)$ is a convex function. Also, let $P_C(\cdot)$ denote the projection operator onto C. For this problem, the projected subgradient algorithm [5, 34] constructs a sequence $x^i \in C$. In the description of the algorithm below, $\beta_i \in \mathbb{R}$ denotes a step-size for iteration i, selected according to a predetermined rule.

[2] In contrast, Yu et al. [38] check the quality of the reconstruction under the different hypotheses, using Stein's unbiased MSE estimate. Both approaches involve simple operations and give comparable results. We opt for tests because we think this approach is simpler to describe.

Algorithm 1 The Projected Subgradient Algorithm

$i \leftarrow 1$, initialize x^0
repeat
  Select $g^i \in \partial f(x^i)$
  $x^{i+1} \leftarrow P_C\big(x^i - \beta_i\, g^i\big)$
  $i \leftarrow i + 1$
until some convergence criterion is met

The algorithm is very similar to the projected gradient algorithm but convergence results for the projected subgradient algorithm are rather intricate. Specifically, if we use constant step sizes $\beta_i = \beta$, then convergence to a minimizer is not ensured, but instead the cost function values are guaranteed to visit infinitely often a neighborhood of the minimum value possible [5, 34]. More precisely,

Proposition 2. [5, 34] Suppose that

$$\sup_{x \in C} \sup_{v \in \partial f(x)} \|v\|_2 \leq c \tag{24}$$

for some constant c. In this case, if the stepsizes satisfy,

$$\lim_{i \to \infty} \beta_i = 0, \text{ and } \sum_i \beta_i \to \infty, \tag{25}$$

then,

$$\lim_{s \to \infty} \inf_{i \geq s} f(x^i) = \inf_{x \in C} f(x). \tag{26}$$

Note that this is not a convergence result. That is, x^i's do not necessarily converge to a minimizer. Also, the cost is not guaranteed to decrease with each iteration. We admit that these are discouraging remarks. Nevertheless, in practice, for our problem, in a reasonable number of iterations, the projected subgradient algorithm produces solutions that approximately satisfy the optimality conditions. We also note that, to be on the safe side, we can always track the cost function and keep the iterate that has produced the lowest cost upto iteration i. This gives a meta-algorithm that monotonely decreases the cost. Prop. 2 ensures that such an approach is guaranteed to yield an iterate with cost arbitrarily close to the minimum value possible.

B. Adaptation to the Problem

In order to apply the projected subgradient algorithm, we need expressions for the subdifferentials of the cost function and we need a procedure for projecting onto the unit simplex. We describe the subdifferentials for the two formulations in Section II below. The expressions are summarized here – the derivations can be found in an appendix.

1) Block-Based Formulation: Note that the cost is separable with respect to different blocks. We describe the subdifferential for the l^{th} block. Suppose we collect all $Y_i(k,s)$ with $(k,s) \in \mathcal{T}_l$ in a column vector Y_i. Then form the matrix Y by concatenating these column vectors as,

$$Y = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_M \end{bmatrix}. \tag{27}$$

Also, let α denote a length-M vector. With these definitions, the cost for the l^{th} block can be expressed as,

$$f(\alpha) = \|Y\,\alpha\|_1 = \sum_n |(Y\,\alpha)_n|, \tag{28}$$

The subdifferential of $f(\alpha)$ is,

$$\partial f(\alpha) = \text{real}\big(Y^H\, U\big), \tag{29}$$

where U denotes the set of vectors u which satisfy

$$u_i \in \begin{cases} \{(Y\,\alpha)_i / |(Y\,\alpha)_i|\}, & \text{if } (Y\,\alpha)_i \neq 0, \\ \{x \in \mathbb{C} : |x| \leq 1\}, & \text{if } (Y\,\alpha)_i = 0. \end{cases} \tag{30}$$

In order for Prop. 2 to apply, we need to show that the subgradient is bounded on the feasible set. But this follows directly from the expression for the subdifferential since U is a bounded set and $\text{real}(Y^H \cdot)$ is a linear mapping in a finite dimensional space.

2) Smoother Formulation: The subdifferential of a function of the form $f+g$ can be expressed as $\partial f + \partial g$ [23]. Therefore, since the data term is similar for both formulations, we just need to specify the subgradient for the penalty term $P(\alpha)$ in (12). Note that $P(\alpha)$ is actually a differentiable function. Its gradient with respect to the weight parameters of the l^{th} block namely $\alpha^{(l)}$ is given as,

$$\nabla_l\, P(\alpha) = 4 \sum_{m \in \mathcal{N}(l)} \left(\alpha^{(l)} - \alpha^{(m)}\right). \tag{31}$$

Note that since $\alpha^{(n)}$'s are required to lie on the unit simplex, and the unit simplex is included in the unit ℓ_2 ball, $\|\alpha^{(l)} - \alpha^{(m)}\|_2 \leq 2$ for the feasible set of α vectors. Since the number of neighbors is at most 4, we therefore obtain $\|\nabla_l\, P(\alpha)\|_2 \leq 32$ on the feasible set. Therefore, the subgradients of the cost function used in (12) are bounded and Prop. 2 applies in this case also.

3) Projection Onto the Unit Simplex: The final ingredient required for both of the formulations is the projection operator onto the unit simplex. For this purpose, we used the method described in [17]. We found this step rather time-consuming in the experiments. In cases where it is known that σ_i's are close to each other, one can argue that the optimal weights are similar. This allows to replace the unit simplex with a set that lies strictly inside the unit simplex, projections onto which are easier to realize. Another option might be to come up with algorithms that avoid explicit projections onto the unit simplex.

C. Optimality Conditions

Since the algorithm above is iterative, we need a criterion for terminating the algorithm. A simple approach is to limit the number of iterations and check if the solution satisfies the optimality conditions. Optimality conditions for constrained problems can be expressed easily in terms of 'normal cones' [23]. Recall that

Definition 2. [23] The *normal cone* of a convex set $C \subset \mathbb{R}^n$ at a point $x \in \mathbb{R}^n$, denoted by $N_C(x)$, is defined as the set of $v \in \mathbb{R}^n$ that satisfy,

$$\langle v - x, y - x \rangle \leq 0, \text{ for all } y \in C. \tag{32}$$

Alternatively, if D_x is the set of points whose projection onto C is x, then $N_C(x) = D_x - x$.

We can now state the optimality conditions. A point x^* is a solution of the convex minimization problem in (23) if and only if there exists a vector v such that '$v \in \partial f(x^*)$' and '$-v \in N_C(x^*)$' [23].

We already described the subdifferentials of the cost functions for the two formulations. Therefore it suffices to describe the normal cone of the unit simplex $S^M \subset \mathbb{R}^M$. At $\alpha \in S^M$, the normal cone is the set of vectors $v \in \mathbb{R}^M$ of the form

$$v = \begin{bmatrix} w_1 \\ \vdots \\ w_M \end{bmatrix} + t \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \tag{33}$$

where $t \in \mathbb{R}$ and w_i's satisfy

$$w_i \in \begin{cases} \{0\}, & \text{if } \alpha_i > 0, \\ \mathbb{R}_+, & \text{if } \alpha_i = 0. \end{cases} \tag{34}$$

In the following, we use these conditions to check if convergence has occured.

IV. Experiments

Experiment 1. In this experiment, we demonstrate how the characteristics of the desired signal affect the performance of the proposed beamforming formulation. We take a complex valued random signal $x(n)$ of length $N = 100$ with iid entries. More precisely,

$$x(n) = \begin{cases} z(n), & \text{if } v(n) \leq \tau, \\ 0, & \text{if } v(n) > \tau, \end{cases} \tag{35}$$

where $z(n)$ is an iid circularly symmetric complex valued Gaussian random vector and $v(n)$ is uniformly distributed on $[0, 1]$. Note that here τ controls the level of sparsity. The closer τ is to zero, the sparser the random vector is expected to be. Given x, we produce three observations y_1, y_2, y_3 of the form

$$y_i = x + \sigma_i \, u_i, \tag{36}$$

where u_i has the same pdf as z above, and σ_i is a constant. Note that according to Prop. 1, once the energy of x and σ_i's are given, the (expected) SNR of the UMVUE is given as

$$\text{SNR}_{\text{UMVUE}} = 10 \log_{10} \left(\frac{\epsilon_x^2}{2\sigma^2 \, N} \right), \tag{37}$$

where ϵ_x^2 is the energy of the clean signal, $\sigma^2 = \left(\sum_i \sigma_i^{-2} \right)^{-1}$. This allows to produce a graph of the SNR achieved by the formulation in (8) with respect to the UMVUE SNR. We conducted two experiments based on this setup.

In a first experiment, we took $\sigma_i = \theta$ for all i and varied θ to control the SNR. Note that the UMVUE is given by a simple average of the observations in this case. If we take $\tau = 0.1$ (leading to a sparse clean signal), we obtain the thick curve in Fig. 3a. Notice that the formulation does not make use of the knowledge of σ_i's but is able to achieve an SNR which is sometimes even higher than that of the UMVUE. On the other hand, when we take $\tau = 1$, the clean signal is actually a realization of a complex Gaussian vector and is not sparse. In this case, we end up with the thin curve in Fig. 3a – the SNR achieved is lower than that of the UMVUE as much as a few decibels especially for high SNRs. The reason for this discrepancy in the reconstruction performance is due to the nature of the clean signal and the assumptions leading to the formulation for selecting the weights. The formulation seeks a signal which is close to being sparse. When the clean signal is approximately sparse, the SNR therefore turns out to be as high as the UMVUE.

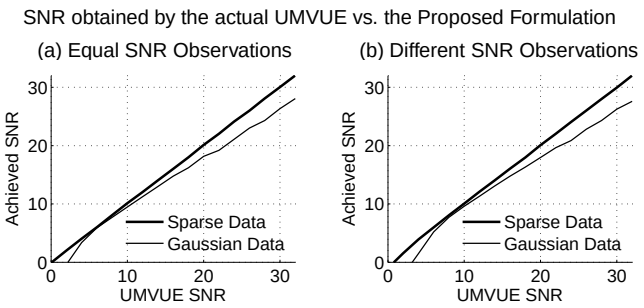

Fig. 3. Comparison of the SNR achieved by the proposed formulation and the UMVUE for observations that have (a) the same SNR, (b) different SNRs.

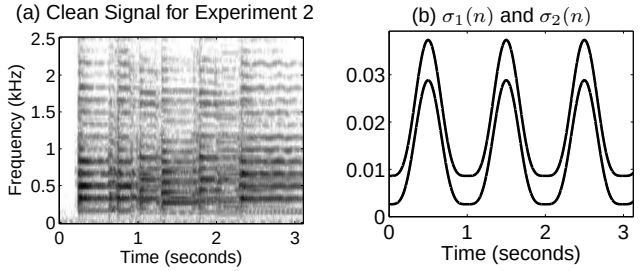

Fig. 4. (a) Spectrogram of the clean signal used in Experiment 2. (b) Time-varying noise standard deviations used for producing the two observations in Experiment 2.

In a second experiment, we repeat the first experiment but take $\sigma_1 = \sigma_2/\sqrt{2} = \sigma_3/\sqrt{3} = \theta$. In this case, the UMVUE is obtained by a convex combination of the observations with unequal weights. The resulting curves are shown in Fig. 3b. The behavior of the curves are similar to those in Fig. 3a. Therefore, we can argue that the proposed formulation can recover approximately the best convex combination of the observations.

Experiment 2. In order to evaluate the capability of estimating the amount of remaining noise after adaptive weighting, we consider a relatively simple noise distribution in this experiment. The spectrogram of the clean signal is shown in Fig. 4a. Using the clean signal $x(n)$, we produced two noisy observations of the form

$$y_i(n) = x(n) + \sigma_i(n) \, z_i(n), \tag{38}$$

where $z_i(n)$'s are iid standard normal random variables and $\sigma_i(n)$'s are deterministic sequences, unknown to the observer. The sequences $\sigma_i(n)$ used in this experiment are shown in Fig. 4b. Notice that the amount of noise peaks around the same regions for both observations. The SNRs of the two observations (not shown) are 6.00 and 9.00 dB.

We employ the formulation in Sec. II-A2 and we take

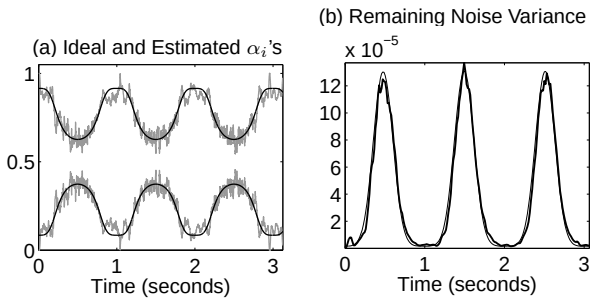

Fig. 5. (a) The ideal (black) and estimated (gray) α_i's for the adaptive weighting stage in Experiment 2. (b) Remaining noise variance after adaptive weighting in Experiment 2, estimated (thick black) and actual (thin gray).

each time-slice in the STFT domain as a block. The ideal $\alpha_i(k)$'s and those estimated by the beamforming formulation are shown in Fig. 5a. Observe that the estimated α_i's follow the ideal ones closely.

Using the estimated α_i's as weights, we obtain a weighted average signal whose spectrogram is shown in Fig. 6a (SNR = 10.85 dB). Observe that the there is a time-varying noise pattern on this signal. Specifically, noise energy peaks around $t = 0.5, 1.5, 2.5$ seconds, and dips around $t = 0, 1, 2, 3$ seconds. This can be expected because the observations are affected by noise whose variance follows a similar pattern (see Fig. 4b), and the signal is formed by a linear combination of the observations.

Given the estimated α_i's, we computed the remaining noise variance field. Averaging this variance field over frequencies (for each time instant), the variance of the actual remaining noise is shown in Fig. 5b (thin gray line). Note that, because different time instances are subject to different amounts of noise, ad-hoc thresholding with a fixed threshold is unlikely to perform well. The estimate of the amount of remaining noise per (13) is shown in Fig. 5b (thick black line). Note that the estimate is close to the actual amount of remaining noise. Using this estimate, we applied the block-based post-filter from Sec. II-B2 followed by an empirical Wiener filter [22]. We used 8×16 blocks for the post-filter, corresponding approximately to $\Delta t = 480$ ms, $\Delta\omega = 267$ Hz. The resulting output SNR is 20.67 dB. The spectrogram of this signal is shown in Fig. 6b. We note that it is the post-filtering stage rather than the adaptive weighting stage that contributed more to suppressing the noise. However, adaptive weighting was instrumental in determining the remaining noise pattern, which in turn rendered possible the application of effective denoising.

To demonstrate the convergence properties of the algorithm, we show in Fig. 7a the progress of the cost function with respect to iterations. We see that the cost settles to its final value after about 15 iterations. The cost seems to be monotonically decreasing although this is not guaranteed by the algorithm – see Sec. III-A. In Fig. 7b, the gradients at each time-frequency block are shown. Observe that by Fig. 5a, none of the variables assume the value zero. Therefore, by (33), (34), if the output of the algorithm is actually the solution, the gradients are expected to lie parallel to the vector $\begin{bmatrix}1 & 1\end{bmatrix}$. We see that this is indeed the case, ensuring that the cost has settled to the minimum value possible.

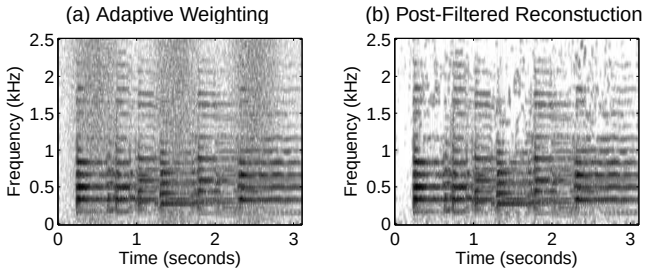

Fig. 6. Spectrograms of the resulting signals from Experiment 2. (a) Output of the adaptive weighting stage, SNR = 10.85 dB. (b) Reconstruction after applying the post-filter to the signal in (a), SNR = 20.67 dB.

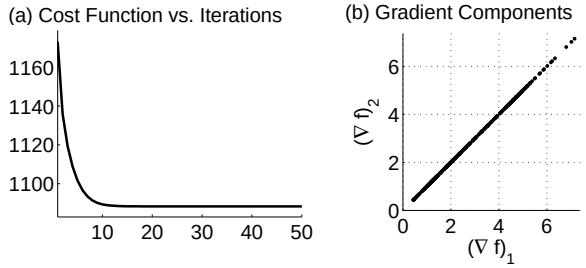

Fig. 7. Convergence of the Algorithm for Experiment 2. (a) Note that the algorithm monotonely decreases the cost. (b) The components of the gradient of the cost function lie in the direction of $\begin{bmatrix}1 & 1\end{bmatrix}^T$, as required by the optimality condition.

Experiment 3. In this experiment, we test the proposed scheme using more complicated noise patterns. The clean signal is now a speech signal (see Fig 8a). We use three different noise patterns. The first two noise signals have time-and-frequency varying characteristics, whereas the third is white noise. The energy of all of the noise signals are the same. Therefore the SNRs of the observations are the same

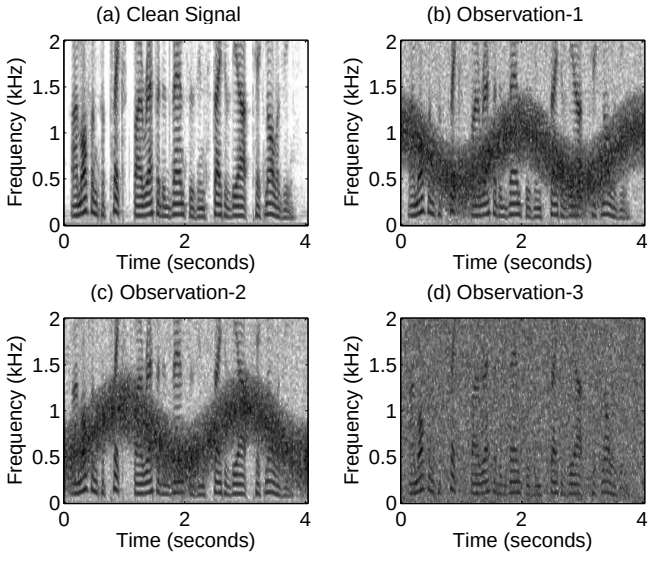

Fig. 8. Spectrogram of the clean signal and the observation signals from Experiment 3. The SNRs of all the observation signals are equal to -5 dB.

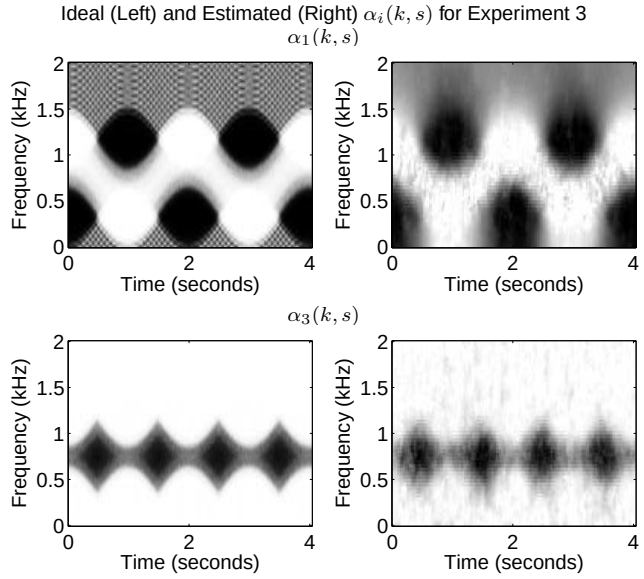

Fig. 9. Ideal (left column) and estimated (right column) α_i's used by the adaptive weighting stage in Experiment 3. The images are in linear scale – black and white indicate the values 1 and 0 respectively.

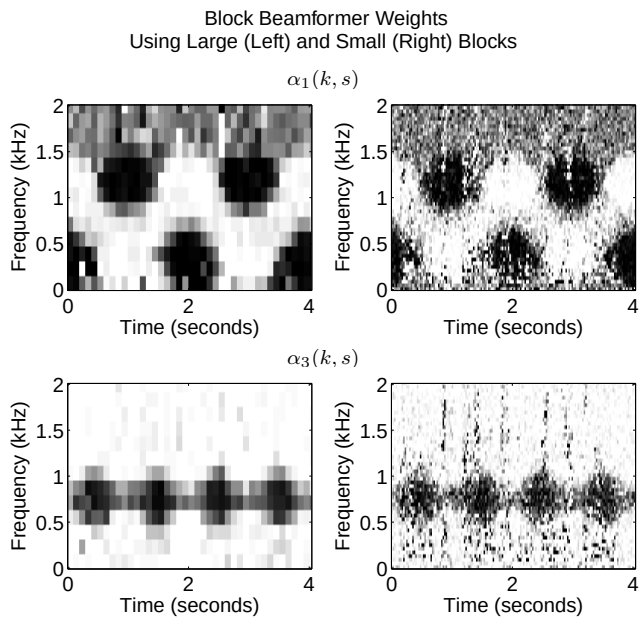

Fig. 10. Block beamformer weights for Experiment 3, using large (left) and small (right) blocks.

(equal to -5.00 dB). The spectrograms of the observations are shown in Fig. 8 b,c,d[3].

For the adaptive weighting stage, the ideal $\alpha_i(k,s)$'s, given by,

$$\alpha_i(k,s) = \sigma_i^{-2}(k,s) \left(\sum_{i=1}^{3} \sigma_i^2(k,s) \right)^{-1} \tag{39}$$

are shown in Fig. 9 on the left column. Observe for instance that, α_1 takes low values around the time-frequency regions where $\sigma_1^2(k,s)$ is greater than $\sigma_2^2(k,s)$ and $\sigma_3^2(k,s)$. Observe also that on the regions where the two curves (in the time-frequency plane) in Fig. 8 b,c intersect, α_3 receives a high value, because the third observation is less noisy in these regions.

The weights for the first and the third observation obtained by the block-based formulation are shown in Fig. 10 for two different block sizes. The weights on the left are obtained by using blocks of size 8×8 whereas the ones on the right are obtained with 2×2 blocks. As expected, large blocks lead to weights which have a coarse appearance in the time-frequency plane. On the other hand, weights obtained using small blocks vary a lot over the time-frequency plane. This is in line with our claim that large blocks help regularize the weight selection. The SNRs of the reconstructed signals which are 8.72 dB (large blocks) and 8.74 dB (small blocks) also support this claim. After post-filtering (the details are given below), the SNRs rise to 12.84 dB (large blocks) and 10.35 dB (small blocks). We think that the block-based formulation with 8×8 blocks performs quite well in this setting.

The weights obtained by the smoother formulation are shown on the right panel of Fig. 9. The reconstructed signal is shown in Fig. 11a. The actual and estimated time-frequency distribution of the remaining noise are shown in Fig. 11 b,c. Note that both the actual and the estimated remaining noise pattern implies that there is a region around 0.5-1 KHz (with a repeating diamond-like pattern) contaminated with noise. Using the estimate of the remaining noise, post-filtering with a soft-threshold applied in the STFT domain, followed by an empirical Wiener filter [22] yields an SNR of 12.52 dB. For the soft threshold, we set the parameter $c = 1$ (see Sec. II-B2). If we apply the block-based post-filter in Section II-B2, with blocks of size 8×16 ($\Delta k \times \Delta s$), the SNR increases to 13.21 dB. The spectrogram of the resulting reconstruction is shown in Fig. 11d.

In order to evaluate the performance of the algorithm against existing methods, we tried reconstructions with two different methods, namely the Frost beamformer [20, 21] and Zelinski's post-filtering method [39]. We note that both methods do not lead to the highest SNR in the literature [21, 29, 35] but we think they are both useful for benchmarking because they are well-known, simple and can achieve good performance.

Frost's beamformer [20] may be regarded as an adaptive realization of the MVDR beamformer, as discussed briefly in the introduction. We implemented the Frost beamformer in the STFT domain, as described in [21]. As noted in [21], the beamformer can be slow to adapt to the changes in the characteristics of noise. In order to reduce this effect, we adjusted the 'step-length' parameter (i.e., μ in [21], Table 47.1) so as to achieve a high SNR value. The resulting weights and the reconstruction obtained are shown in Fig 12. The weights show that Frost's beamformer is responsive to the changes in the time-frequency variations of noise, but the response is rather crude. Although about 3 dB better than simple averaging (Fig. 13a), this results in a reconstruction with much lower SNR than the proposed adaptive weighting reconstruction.

Zelinski's method is a simple scheme that performs quite well in practice [35]. The method first combines the observed

[3]The observed signals are provided with the MATLAB code available at '`http://web.itu.edu.tr/ibayram/SparseMC/`'.

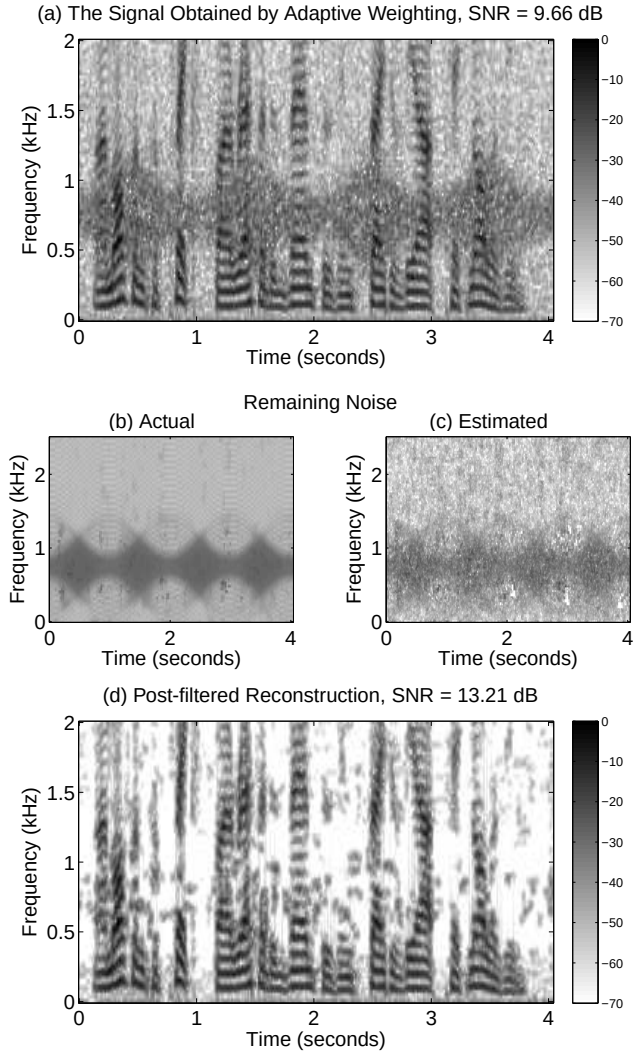

Fig. 11. The output signals for Experiment 3. (a) The signal obtained by adaptive weighting. (b,c) Remaining noise variances, actual (b) and estimated (c). (d) Final reconstruction obtained by post-filtering the signal in (a).

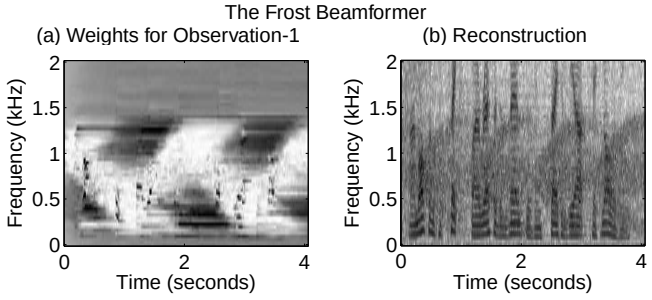

Fig. 12. (a) Weights used by the Frost Beamformer applied to the first observation in Experiment 3. The weights do follow the input noise patterns but, they are rather crude compared to those in Fig. 11. (b) Frost Beamformer reconstruction, SNR = 5.32 dB.

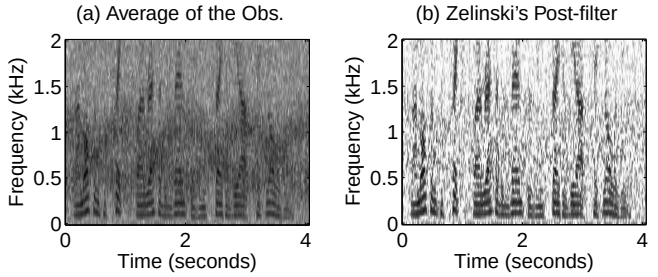

Fig. 13. Left : Average of the observations in Experiment 3, SNR = -0.20 dB. Right : Zelinski's post-filter applied to the average of the observations, SNR = 8.81 dB.

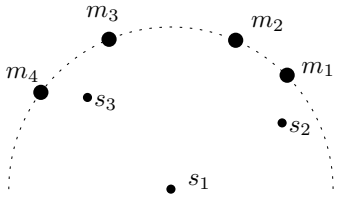

Fig. 14. The physical setup for Experiment 4. s_i's denote sources and m_i's denote microphones. The microphones are equidistant to s_1, but otherwise do not follow a specific pattern. We are interested in recovering s_1 while s_2 and s_3 are considered as noise.

signals by a simple average. If the noise variance of the observations are similar, this gives an estimate with an improved SNR. Then, the amount of remaining noise is estimated and a Wiener post-filter is applied to further suppress the noise [35, 39]. The spectrogram of the average of the observations is shown in Fig. 13a. We can clearly see the noise variance patterns of the individual noise terms. The spectrogram of the post-filtered signal (using the post-filter in [39]) is shown in Fig. 13b. Despite the disadvantages of simple averaging, post-filtering significantly improves the SNR. However, Zelinski's post filter leads to significant degradation perceptually and the quality is lower compared to the proposed reconstruction for this experiment.

Experiment 4. In this experiment, we study the performance of the proposed formulation in a more realistic scenario. The physical recording setup is depicted in Fig. 14. There are three sources and four microphones in the scene, where one of the sources, namely s_1, is equidistant to the microphones and produces the desired (clean) signal. The other two sources s_2 and s_3 produce the noise signals. Sources s_1 and s_2 are human speakers whereas s_3 is the ring-tone of a cell-phone. Recordings are made in a semi-anechoic chamber and the spectrograms of the observed signals are shown in Fig. 15. The SNRs of the observations are 0.78, 7.07, 8.19, -0.97 dB. We note that the second source s_2 (human speaker) is more dominant in the first observation whereas the third source s_3 (ring-tone) is more dominant in the fourth observation.

The weights obtained by the smoother formulation are shown in Fig. 16. Observe that the weights chosen by the proposed formulation have adapted well to the noise pattern. Specifically, we see that $\alpha_1(k,s)$ assumes low values in regions where s_2 (human speaker) contaminates the observations

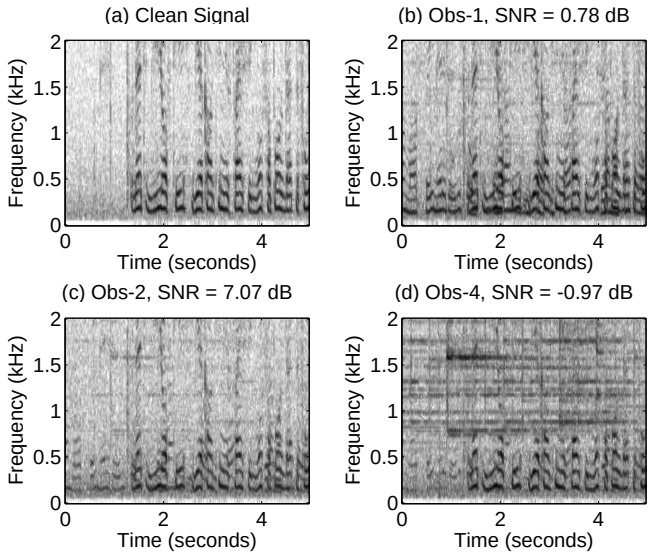

Fig. 15. Spectrogram of the clean signal and the observation signals in Experiment 4.

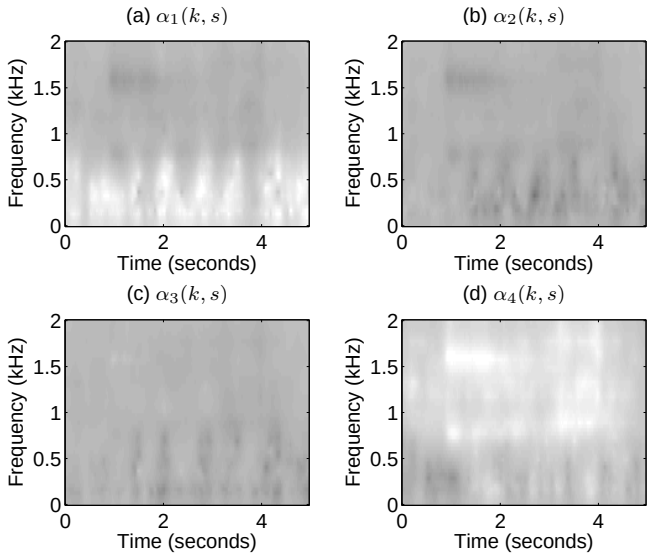

Fig. 16. Weights chosen by the proposed formulation in Experiment 4.

and $\alpha_4(k,s)$ takes low values around the harmonics of the ring-tone. The spectrogram of the reconstruction is shown in Fig. 17b. Compared to a simple average (see Fig. 17a), there is a significant improvement in SNR.

After post-filtering, despite the modest improvement in SNR, the artifacts are further removed. Compared to Zelinski's post-filter, there is more than 1 dB improvement in SNR. For a comparison, see Fig. 17c,d.

In our experiments with signals obtained in similar setups (including outdoors recordings), we found that adaptive weighting leads to a significantly better reconstruction than simple averaging. Post-filters such as Zelinski's post-filter or the ones proposed here can further suppress the unwanted sources but this is at the expense of degrading the original signal. Therefore, weighted averaging without any post-filtering

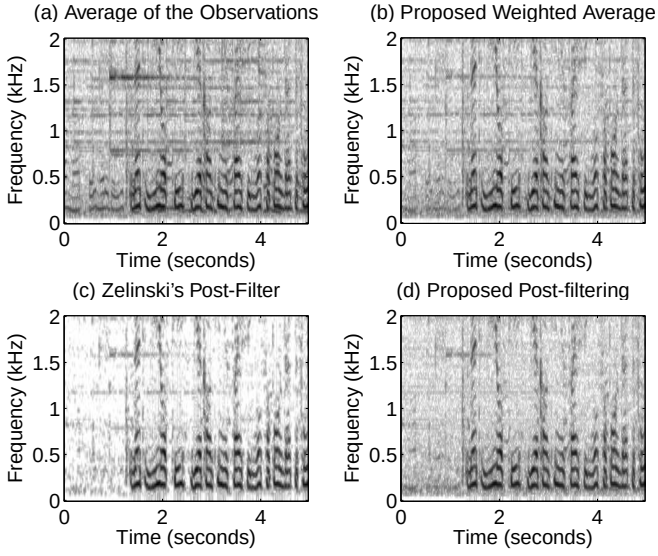

Fig. 17. Reconstructed signals for Experiment 4. (a) Regular averaging, SNR = 9.12 dB, (b) adaptive weighting via the proposed formulation SNR = 14.62 dB, (c) Zelinski's post-filter applied to the averaged signal, SNR = 13.57 dB, (d) proposed post-filter applied to the adaptive weighting output, SNR = 14.69 dB.

is also a viable option if a faithful reconstruction of the source is desired.

Experiment 5. In a final experiment, we provide a comparison of the proposed adaptive weighting stage against the Frost beamformer for varying observation SNRs. The setup is similar to that in Experiment 4. There are four microphones placed arbitrarily on a circle and the source of interest is in the center of the circle. But this time, for noise, we use recordings obtained outdoors and the noise signal contains speech as well as wind. The unwanted speech source is closer to the fourth microphone and farther from the first microphone. The spectrograms of the noise signals for the first and the fourth microphone are shown in Fig. 18a,b. Denoting these noise signals as $u_1(n),\ldots,u_4(n)$, and the source signal at the i^{th} microphone as $x_i(n)$, we produce the observation signals according to the following formula.

$$y_i(n) = x_i(n) + \gamma\, u_i(n), \text{ for } i = 1,\ldots,4. \tag{40}$$

Here, $\gamma \in \mathbb{R}_+$ is a parameter that allows us to control the SNR of the observations. For each value of γ, a quick reconstruction is obtained by averaging the observations. To obtain better estimates, we employ the Frost beamformer as well as the proposed weighted averaging. The gains in segmental SNR (SSNR) obtained with these methods (with respect to the first observation) are shown in Fig. 19. Observe that the proposed weighted averaging leads to a clearly higher gain in SSNR. The spectrogram of the proposed weighted average for the case with lowest SSNR is shown Fig. 18c. The spectrogram of the signal obtained with the Frost beamformer looks similar (not shown) but there are important differences. In order to better show these differences, we provide the absolute value of the ratio of the two spectrograms (Proposed/Frost) in Fig. 18d (in dB). In this image, regions where the Frost

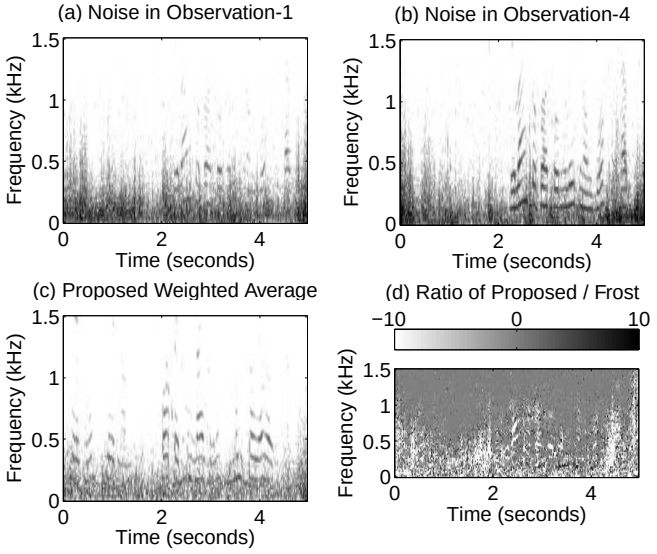

Fig. 18. (a,b) Noise signals used in Experiment 5. (c) Proposed weighted average for the lowest SSNR case. (d) Ratio of the spectrograms obtained with the proposed weighted averaging and the Frost beamformer.

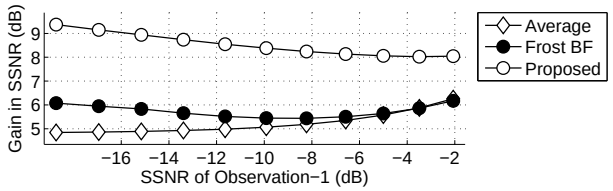

Fig. 19. The gains in SSNR with respect to the first observation's SSNR, obtained by a regular average, Frost's beamformer and the proposed weighted averaging for Experiment 5.

beamformer reconstruction has higher magnitude appear light and regions where the proposed reconstruction has higher magnitude appear dark. We observe that the ratio image is not very dark, which implies that the proposed weighted average does not contain components that are absent from the Frost beamformer reconstruction. On the other hand, we see some white regions with similar patterns as noise. In these time-frequency regions dominated by noise, the proposed weighted average assumes lower values than the Frost beamformer reconstruction, which in turn helps obtain a reconstruction with a higher SSNR.

V. Conclusion

We proposed a framework for the reconstruction of an audio signal from its time-aligned noisy observations. The proposed method consists of an adaptive weighting stage, followed by a post-filter. The adaptive weighting stage does not require us to estimate the noise statistics. Instead, the formulation makes explicit use of the sparsity of the desired signal's time-frequency image. The proposed formulation differs from previous work on the subject in this regard. We demonstrated in Experiments 4, 5 that this approach is useful in a scenario where multiple audio sources are present in a scene and the main source of interest is monitored with equidistant microphones.

In the two-stage framework, we regard the adaptive weighting stage as a device to estimate the UMVU estimate (UMVUE), which turns out to be a complete sufficient statistic. At this point, it is reasonable to ask if it is not possible to obtain a reliable approximation of the UMVUE with a simpler scheme, say by a simple average of the observations as in the Zelinski's post-filtering method. In fact, [7, 37] imply that simple averaging can be a good estimate if the noise variance in the different recordings are approximately the same (also see Ex. 7.42, 7.43 in [11]). However, if the noise variances differ significantly, it can be shown that simple averaging leads to a poor estimate. In such a case, since the optimal reconstruction has to be a function of the sufficient statistic, post-filtering is unlikely to yield a good estimate, as demonstrated in Experiment 4.

In certain scenarios, real-time operation can be critical. However, the 'smoother formulation' requires the whole signal and is not directly usable for real-time processing. One approach for adapting the formulation to real-time operation might be to take into account only the past. An alternative is to employ independent processing of blocks as in the 'block-based formulation' and derive real-time algorithms with a small delay. The block-based formulation might also be of interest as it allows parallel processing. Although the proposed algorithm for this formulation is also iterative, it can be made to work in real time by limiting the number of iterations and warm-starting the iterations.

The proposed framework assumes that the observations are merely noisy and the source of interest is not otherwise distorted. This is indeed the case if the signals are recorded outdoors or in a room with negligible reverberation. However, if the recordings are made in a reverberant room, the observation model ceases to be valid. For such cases, one possibility is to use the relative transfer functions as in [28] to transform the observations and reconstruct a relatively clean but reverberated version of the source signal. Another possibility is to extend the current formulation so that it can handle arbitrary reverberation in the different recordings. In such an extension, it would also be of interest to make the formulation robust to the inexact knowledge of the room impulse responses, since those are usually not known precisely (and change rapidly with respect to position) in practice. We think that it would be interesting to extend the formulation in this direction and we hope to investigate this problem in future work.

Acknowledgements

The author thanks the anonymous reviewers and Damien J. Duff, Istanbul Technical University, for their comments and suggestions which helped improve the paper. The author also thanks Aziz Koçanaoğulları, Istanbul Technical University, for his help in acquiring the data for Experiments 4, 5.

Appendix
Proof of Proposition 1

Thanks to the independence assumption, we can drop the time-frequency indices (k, s), in order to simplify the notation.

Let X be the unknown complex constant where X^r and X^j denote the real and imaginary parts of X. Since the real and imaginary parts of U_i are independent, the joint pdf of the observations $Y_1, \ldots, Y_M$, namely $f(y_1, \ldots, y_M)$, is given as

$$f = \overbrace{\left(\prod_{i=1}^{M} \frac{1}{2\pi\sigma_i^2}\right)}^{c} \exp\left(-\sum_{i=1}^{M} \frac{1}{2\sigma_i^2}|y_i - X|^2\right) \tag{41}$$

$$= c\, \exp\left(-\sum_{i=1}^{M} \frac{1}{2\sigma_i^2}|y_i|^2\right) \exp\left(X^r \sum_{i=1}^{M} \frac{1}{\sigma_i^2} y_i^r\right)$$
$$\exp\left(X^j \sum_{i=1}^{M} \frac{1}{\sigma_i^2} y_i^j\right) \exp\left(-|X|^2 \sum_{i=1}^{M} \frac{1}{2\sigma_i^2}\right), \tag{42}$$

where y_i^r and y_i^j denote the real and imaginary parts of y_i. It follows by the factorization theorem [24, 33] that $S^r = \sum_{i=1}^{M} \sigma_i^{-2} Y_i^r$ is a sufficient statistic for X^r and $S^j = \sum_{i=1}^{M} \sigma_i^{-2} Y_i^j$ is a sufficient statistic for X^j. Thus

$$S = S^r + j\, S^j = \sum_{i=1}^{M} \sigma_i^{-2}\, Y_i \tag{43}$$

is a sufficient statistic for X. Further, since Y_i's are independent and circularly normal [30], S is also circularly normal (i.e., S^r and S^j are independent). Also, $\mathbb{E}(S) = \sigma^{-2}\, X$, where

$$\sigma^2 = \left(\sum_{i=1}^{M} \sigma_i^{-2}\right)^{-1}, \tag{44}$$

and

$$\mathrm{var}(S^r) = \mathrm{var}(S^j) = \sum_{i=1}^{M} \sigma_i^{-4}\, \sigma_i^2 = \sum_{i=1}^{M} \sigma_i^{-2} = \sigma^{-2}. \tag{45}$$

It follows that S is complete with respect to X because if $\mathbb{E}(g(S)) = 0$ for all X, then

$$\iint g(t^r, t^j) \times$$
$$\exp\left(-\frac{\sigma^2}{2}\left[(t^r - X^r\sigma^{-2})^2 + (t^j - X^j\sigma^{-2})^2\right]\right) dt^r\, dt^j$$
$$= 0, \tag{46}$$

for all (X^r, X^j), which implies that $g(t^r, t^j) = 0$ for all t. It then follows by the Rao-Blackwell theorem [24, 33] that the UMVUE is given by an unbiased function of S. Since $\tilde{X} = \sigma^2\, S$ is unbiased, it must therefore be the UMVUE. Note in this case that $\tilde{X}$ is circularly normal and the variances of the real and imaginary parts of $\tilde{X}$ are given by $\sigma^4\ \mathrm{var}(S^r) = \sigma^2$.

Appendix
Expected Value of the Estimator in (13)

In order to show that the estimator in (13) is unbiased, let us simplify the notation and drop the time-frequency indices (k, s). Specifically, let Y_i's for $i = 1, 2, \ldots, M$ denote complex valued observations of a constant X in the form $Y_i = X + \sigma_i\, Z_i$, where Z_i's denote iid complex valued noise terms. We assume that the real and imaginary parts of Z_i's are independent standard Gaussian random variables (i.e., Z_i is circularly normal [30]), so that $\mathbb{E}(|Z_i|^2) = 2$. In this setting, let

$$\sigma^2 = \left(\sum_{i=1}^{M} \sigma_i^{-2}\right)^{-1}, \quad \alpha_i = \sigma^2\, \sigma_i^{-2}, \quad \hat{X} = \sum_{i=1}^{M} \alpha_i Y_i. \tag{47}$$

Now let,

$$\hat{\sigma}^2 = \frac{1}{2(M-1)} \sum_{i=1}^{M} \alpha_i \left|Y_i - \hat{X}\right|^2 \tag{48}$$

Since α_i's add to unity, we can write

$$\hat{X} = X + \sum_{i=1}^{M} \alpha_i \sigma_i Z_i. \tag{49}$$

Using this, let us compute the expected value of $s = 2(M-1)\hat{\sigma}^2$.

$$\mathbb{E}(s) = \sum_{i=1}^{M} \alpha_i\, \mathbb{E}\Big(\left|Y_i - \hat{X}\right|^2\Big) \tag{50}$$

$$= \sum_{i=1}^{M} \Bigg[\alpha_i\, \mathbb{E}\Big(\left|(1-\alpha_i)\sigma_i Z_i\right|^2 + \sum_{m\neq i} |\alpha_m \sigma_m Z_m|^2\Big)\Bigg] \tag{51}$$

$$= 2\sigma^2 \sum_{i=1}^{M} \left[(1-\alpha_i)^2 + \alpha_i\,(1-\alpha_i)\right] \tag{52}$$

$$= 2(M-1)\,\sigma^2. \tag{53}$$

Thus, $\hat{\sigma}^2$ in (48) is an unbiased estimator of σ^2.

Appendix
Derivations of the Expressions in Sec. III-B1

We think of a complex vector $z \in \mathbb{C}^n$ as n real number pairs (z_k^r, z_k^i) for $k = 1, \ldots, n$ (essentially interpreting $\mathbb{C}^n$ as $\mathbb{R}^{2n}$). Now let B_∞ denote the unit ball of the ℓ_∞ norm in $\mathbb{C}^n$. In this appendix, let us also define an inner product of two complex vectors as,

$$\langle u, z\rangle = \sum_{k=1}^{n} \Big(u_k^r\, z_k^r + u_k^i\, z_k^i\Big). \tag{54}$$

Note that this is just the real part of the regular complex valued inner product and is a valid inner product itself. We can now write,

$$\|z\|_1 = \sup_{v\in B_\infty} \langle v, z\rangle. \tag{55}$$

Therefore (see e.g. Chp.C,D in [23]), $\partial(\|z\|_1)$ is the set of vectors v that satisfy,

$$v_k \in \begin{cases} \{w \in \mathbb{C} : |w| \le 1\}, & \text{if } z_k = 0, \\ \{z_k/|z_k|\}, & \text{if } z_k \neq 0. \end{cases} \tag{56}$$

Finally, note that if $\alpha \in \mathbb{R}^M$, and Y is a complex $n \times M$ matrix, multiplying α on the left by Y may be regarded as a linear mapping from $\mathbb{R}^M$ to $\mathbb{R}^{2n}$. If we similarly denote the real and the imaginary parts of Y as Y^r and Y^i, the transpose of this operation applied on a complex vector z

is given as $(Y^r)^T z^r + (Y^i)^T z^i$. From the calculus rules of subdifferentials (see Thm.D.4.2.1 [23]) we finally have that,

$$\partial f(\alpha) = Y^r U^r + Y^i U^i, \tag{57}$$

where U is the set of vectors u as described in (30). This is equivalent to (29).

REFERENCES

[1] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura. Speech enhancement based on the subspace method. *IEEE Trans. Speech and Audio Processing*, 8(5):497 – 507, September 2000.

[2] R. Balan and J. Rosca. Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase. In *Proc. IEEE Sensor Array and Multichannel Signal Proc. Workshop*, pages 209 – 213, Rosslyn, VA USA, 2002.

[3] İ. Bayram. Combining multiple observations of audio signals. In *Proc. SPIE 8858, Wavelets and Sparsity XV*, San Diego, CA USA, 2013.

[4] J. O. Berger. *Statistical Decision theory and Bayesian Analysis*. Springer, 2nd edition, 1993.

[5] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[6] J. Bitzer and K. U. Simmer. Superdirective microphone arrays. In M. Brandstein and D. Ward, editors, *Microphone Arrays*. Springer, 2001.

[7] D. A. Bloch and L. E. Moses. Nonoptimally weighted least squares. *The American Statistician*, 42(1):50–53, February 1988.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, July 2011.

[9] T. T. Cai. Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *The Annals of Statistics*, 27(3):898–924, June 1999.

[10] J. Capon. High resolution frequency-wavenumber spectrum analysis. *Proc. IEEE*, 57:1408–1418, August 1969.

[11] G. Casella and R. L. Berger. *Statistical Inference*. Cengage Learning, 2nd edition, 2001.

[12] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, May 2011.

[13] P. L. Combettes and J.-C. Pesquet. A proximal decomposition method for solving convex variational inverse problems. *Inverse Problems*, 24(6), December 2008.

[14] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, New York, 2011.

[15] L. Daudet and B. Torrésani. Hybrid representations for audiophonic signal encoding. *Signal Processing*, 82(11):1595–1617, November 2002.

[16] S. Doclo and M. Moonen. GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Trans. Signal Processing*, 50(9):2230– 2244, September 2002.

[17] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1 ball for learning in high dimensions. In *Proc. 25^{th} Int. Conf. on Machine Learning*, pages 272–279, New York, NY USA, 2008.

[18] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech and Audio Processing*, 3(4):251–266, July 1995.

[19] E. Esser, X. Zhang, and T. F. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Sciences*, 3(4):1015–1046, November 2010.

[20] O. L. Frost III. An algorithm for linearly constrained adaptive array processing. *Proc. IEEE*, 60(8):926–935, August 1972.

[21] S. Gannot and I. Cohen. Adaptive beamforming and postfiltering. In *Handbook of Speech Processing*. Springer, 2008.

[22] S. P. Ghael, A. M. Sayeed, and R. G. Baraniuk. Improved wavelet denoising via empirical Wiener filtering. In *Proc. SPIE 3169 Wavelet Applications in Signal and Image Proc.*, San Diego, CA USA, 1997.

[23] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2004.

[24] S. Kay. *Fundamentals of Statistical Signal Processing, Vol I : Estimation Theory*. Prentice Hall, 1993.

[25] C. Kereliuk and P. Depalle. Sparse atomic modelling of audio : A review. In *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, pages 89–92, Paris, France, 2011.

[26] M. Kowalski. Sparse regression using mixed norms. *J. of Appl. and Comp. Harm. Analysis*, 27(3):303–324, November 2009.

[27] R. T. Lacoss. Adaptive combining of wideband array data for optimal reception. *IEEE Trans. Geoscience Electronics*, 6(2):78–86, May 1968.

[28] S. Markovich, S. Gannot, and I. Cohen. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *IEEE Trans. Audio, Speech and Language Processing*, 17(6):1071 – 1086, August 2009.

[29] C. Marro, Y. Mahieux, and K. U. Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Trans. Speech and Audio Processing*, 6(3):240–259, May 1998.

[30] B. Picinbino. On circularity. *IEEE Trans. Signal Processing*, 42(12):3473–3482, December 1994.

[31] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies. Sparse representations in audio and music : From coding to source separation. *Proc. IEEE*, 98(6):995–1005, June 2010.

[32] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the Mumford-Shah functional. In *Proc. IEEE Int. Conf. on Computer Vision*, Bombay, India, 2009.

[33] H. V. Poor. *An Introduction to Signal Detection and*

Estimation. Springer, 2nd edition, 1998.

[34] N. Z. Shor. *Minimization Methods for Non-Differentiable Functions*. Springer, 1985.

[35] K. U. Simmer, J. Bitzer, and C. Marro. Post-filtering techniques. In M. Brandstein and D. Ward, editors, *Microphone Arrays*. Springer, 2001.

[36] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada. A multichannel MMSE-based framework for speech source separation and noise reduction. *IEEE Trans. Audio, Speech and Language Processing*, 21(9):1913–1928, September 2013.

[37] J. W. Tukey. Approximate weights. *Annals of Mathematical Statistics*, 19(1):91–92, March 1948.

[38] G. Yu, S. Mallat, and E. Bacry. Audio denoising by time-frequency block thresholding. *IEEE Trans. Signal Processing*, 56(5):1830–1839, May 2008.

[39] R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pages 2578 – 2581, New York, NY USA, 1988.

İlker Bayram received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from Middle East Technical University (METU), Ankara, Turkey, in 2002 and 2004 respectively. He received the Ph.D. degree in electrical engineering from Polytechnic Institute of New York University, Brooklyn, NY, USA in 2009. Following that, for a year, he worked as a post-doctoral researcher in the Biomedical Imaging Group at École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. In 2010, he joined the Department of Electronics and Telecommunications Engineering, Istanbul Technical University, Istanbul, Turkey, where he is currently an associate professor. His research interests are in applications of time-frequency representations and sparse signal processing.