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

We consider the estimation of an audio source $x(n)$, given multiple noisy observations $y_1(n), \ldots, y_M(n)$. Specifically, we suppose

$$y_i(n) = x(n) + u_i(n), \tag{1}$$

where $u_i(n)$ denotes the noise signal affecting the i^{th} observation. We will assume that the noise terms affecting different observations (that is u_i and u_j with $i \neq j$) are not correlated, or have low correlation. Our motivation in adopting such a model is the reconstruction problem when multiple unreliable recordings are available. Such a problem arises if we have multiple recordings of a source obtained with microphones scattered in an environment such that different noise sources affect different microphones. Note that in this scenario, we are interested only in a specific source and the noise sources near a certain microphone may not be affecting the recording obtained with some other microphone.

In this paper, we propose a method for estimating $x(n)$ that does not require prior information about noise and requires solving a convex minimization problem. The whole scheme consists of two main stages, depicted with the 'adaptive weighting' and 'post-filter' boxes in Fig. 1.

Put briefly,

(i) The adaptive weighting stage operates without any information about noise, and is based solely on a sparsity based model of the audio signal of interest. This stage determines the near optimal weights to linearly combine the data by solving a convex minimization problem.

(ii) Then, based on a simple estimate of the remaining noise standard deviation, the post-filter is applied to further enhance the signal obtained by adaptive weighting. The post-filter relies on a sparsity based model for the desired signal.

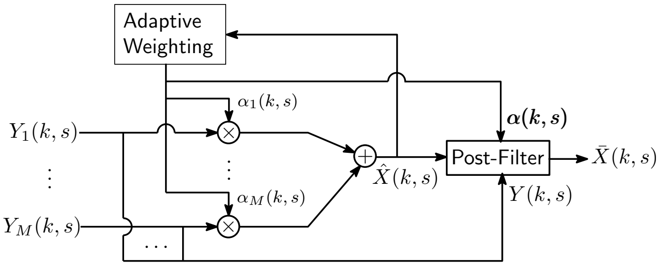


Fig. 1: The general structure employed for estimating the source consists of a adaptive weighting stage followed by a post-filter.

A. Adaptive Weighting

The determiation of the adaptive weights is related to 'beamforming' formulations that aim to optimally combine multiple observations. Before we describe our approach, let us briefly review the key ideas that are of interest in multi-channel audio processing.

Related Prior Work: Classical beamforming formulations employ models based on second order statistics of the signals. In particular, in the minimum variance distortionless response (MVDR) beamformer [6, 20, 38], under a stationarity assumption, one restricts herself to estimates obtained by LTI filtering the observations and summing them. 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 so as to avoid introducing any distortion to the signal of interest [10, 30]. Closed form solutions for the filters can be derived but they involve correlation functions of the signals [6, 20, 38]. Therefore, in order to apply such methods, if statistics of the signals are not available, they need to be estimated. However, considering the instantaneous output power as an approximation of the expected output power, one can derive adaptive algorithms that work in real-time, and that do not make use of such prior knowledge about the statistics of the source or the noise signals [19, 22].

Proposed Approach: Here, we assume that whole recordings are available and seek an offline algorithm. One attempt in this direction might be to lift the stationarity assumption in the MVDR beamformer formulation and use time-varying filters. Then, one could attempt to minimize the total power, subject to a similar 'no-distortion' constraint. However, we noted in [4] that such a formulation leads to trivial (zero) solutions. In this paper, we propose a convex minimization problem derived using models based on the spectral sparsity of the reconstruction. The framework is different than the one in our previous work [4] and allows to naturally develop a post-filter as well. In [4], the formulation was primarily in the time-domain and the (time-domain) noise samples were assumed to be independent with a time-varying variance. Here, we consider time *and* frequency varying noise. In order

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).

to handle noise terms with such characteristics, we work with time-frequency blocks. Moreover, we propose an estimator for determining the amount of remaining noise in the signal obtained by adaptive weighting. This allows to employ a post-filter so as to further suppress noise, improving the SNR significantly.

One aspect of the formulation we would like to underline is its use of sparsity. The weight selection problem for the adaptive weighting stage is formulated such that the reconstructed signal is close to being sparse in the STFT domain. Sparsity of the reconstruction in the STFT domain was used as a driving criterion in blind source separation (BSS) previously [1, 42]. However, BSS assumes that we have at hand a mixture of signals, and each source signal is interesting in its own right. In contrast, we assume that we have several noisy recordings of a single source and in these recordings, the noise terms may not show statistically interesting behavior (i.e. they may not be necessarily non-Gaussian as typically assumed in BSS or ICA [25]).

B. *Post-filtering*

Related Prior Work: Post-filtering was originally proposed by Zelinski as an ad-hoc Wiener filter applied to the output of a delay-and-sum beamformer [41]. 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 the noise terms are uncorrelated. Later, the relation of this single-channel post-filter with multi-channel Wiener filtering was noted and modified post-filters were proposed [38]. Specifically, it was shown in [38] that a multichannel Wiener filter can equivalently 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.

An interesting scheme that implies the optimality of this two-step procedure is introduced in [2], where the authors derive multi-channel versions of the Ephraim-Malah estimator for speech. They show as a byproduct that, a sufficient statistic of interest coincides with the beamformer according to their model. This sufficient statistic requires a knowledge of the statistical parameters of noise, which need to be estimated in practice.

Proposed Approach: In this paper, we adopt a similar observation model as [2]. Consequently, the output of the adaptive weighting stage acts as a complete sufficient statistic in our framework too. However, in our scheme, the sufficient statistic is estimated by the adaptive weighting stage without any knowledge of the signal and noise statistics, solely making use of the spectral sparsity of the audio signal of interest. 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 to devise a block-based denoiser similar in spirit to the denoising schemes of Cai [9] and Yu et al. in [40].

Contribution

Consider an abstract estimation problem as follows. Suppose X is a 'sparse' vector of interest and we observe

$$Y_i = X + \sigma_i\, U_i, \tag{2}$$

where U_i denote random vectors with independent and identically distributed standard normal components and σ_i are unknown constants. The formulations in [10, 30] are viable for this problem when X is also Gaussian. However, 'sparsity' calls for a different approach. Our problem is similar to the abstract problem described above but it also presents a challenge as the statistics of the noise terms vary over the time-frequency plane. Therefore, the contribution of this paper is two-fold. We propose a two-stage estimation scheme and we adapt the scheme for audio signals where the noise components are time-frequency varying. We are not aware of a similar formulation in the literature that addresses such a problem.

Notation

We denote signals in the time-domain using small case letters, as in $z(n)$. The STFT of $z(n)$ is denoted as $Z(k,s)$, 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, we define the STFT of z as,

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

We assume throughout the paper that the number of observed signals is M.

The propoped formulations are based on partitions of the time-frequency coefficients with a special notation given as follows. Let $Y_i(k,s)$ denote the STFT coefficients of the i^{th} observation. Suppose the time-frequency parameters (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 e.g., Fig. 2). Also, $Y_i^{(l)}$ denotes a vector that contains the time-frequency samples of $Y_i(k,s)$ for $(k,s) \in \mathcal{T}_l$. Finally, we denote the components of a vector $\alpha \in \mathbb{R}^M$ using subscripts, as $\alpha = \begin{pmatrix}\alpha_1 & \alpha_2 & \ldots & \alpha_M\end{pmatrix}$.

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

Recall that $x(n)$ is the signal of interest but we have observations $y_i(n)$ which contain additive noise $u_i(n)$. Let $X(k,s)$, $Y_i(k,s)$, $U_i(k,s)$ denote the STFTs of these signals so that

$$Y_i(k,s) = X(k,s) + U_i(k,s), \tag{3}$$

for $i = 1, 2, \ldots, M$.

Noise Model: We model the time-frequency 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 $\sigma_i^2(k,s)$. In other words, $U_i(k,s)$ is a circular normal random variable [32]. Note that such a model cannot be true if the STFT is overcomplete because the noise terms cannot be independent 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 [27, 33]. In order to capture this property, we use the ℓ_1 norm (see [27] in this context) defined in this case as,

$$\|X\|_1 = \sum_{(k,s)\in\mathcal{T}} |X(k,s)|. \tag{4}$$

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. But before that 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 [5]). 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. Therefore we can 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. Thus, the post-filter which acts on the sufficient statistic/UMVUE may be interpreted as a 'denoising' operation as well.

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

We will provide two related formulations of varying complexity in Sec. II-A1 and Sec. II-A2. But in order to set the stage, we start by assuming that the variance field is constant for each observation. This allows to state a simple formulation that relies on Prop. 1 below. We will remove this assumption in Sec. II-A1 and Sec. II-A2.

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

Proposition 1. For a fixed (k,s) pair, suppose we are given observations as in (3), with the described noise model. If we treat $X(k,s)$ as an unknown constant for each (k,s) pair, a complete sufficient statistic for $X(k,s)$ is given by

$$\tilde{X}(k,s) = \sum_{i=1}^{M} \alpha_i(k,s)\, Y_i(k,s), \tag{5}$$

where

$$\alpha_i(k,s) = \sigma_i^{-2}(k,s)\,\sigma^2(k,s), \qquad \text{for } \sigma^2(k,s) = \left(\sum_{i=1}^{M} \sigma_i^{-2}(k,s)\right)^{-1}. \tag{6}$$

$\tilde{X}(k,s)$ 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. We can therefore drop the indices (k,s) under this assumption.

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

$$\sum_{i=1}^{M} \alpha_i = 1 \quad \text{and} \quad \alpha_i \geq 0,\ \forall i. \tag{7}$$

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 $\sigma_i^2 = \beta_i^{-1}$.

The foregoing discussion implies 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, 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,

$$\hat{X}(k,s) = \sum_{i=1}^{M} \hat{\alpha}_i\, Y_i(k,s), \tag{8}$$

where

$$\hat{\alpha} = \arg\min_{\alpha} \left\| \sum_{i=1}^{M} \alpha_i Y_i(k,s) \right\|_1 \quad \text{s.t.} \begin{cases} \alpha_i \geq 0, \\ \sum_i \alpha_i = 1. \end{cases} \tag{9}$$

In the sequel, we modify this formulation by dropping the white noise assumption.

1) Block-Based Formulation: If the variance field $\sigma_i^2(k,s)$ is not constant but slowly varying over the time-frequency lattice, the formulation in (9) 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 described above.

Recall from the 'Notation' in the Introduction that $\mathcal{T}$ denotes the set of time-frequency indices (k,s). Also, $\mathcal{T}_l$ for $l = 1, \ldots, L$ form a partition of $\mathcal{T}$. 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. 2.

In the 'block-based formulation', each block is treated independently of the other blocks. For the l^{th} block, the optimal weights $\hat{\alpha}^{(l)} = \big(\hat{\alpha}_1^{(l)}, \ldots, \hat{\alpha}_M^{(l)}\big)$ are chosen as,

$$\hat{\alpha}^{(l)} = \arg\min_{\alpha \in \mathbb{R}^M} \sum_{(k,s)\in\mathcal{T}_l} \left| \sum_{i=1}^{M} \alpha_i Y_i(k,s) \right| \quad \text{s.t.} \begin{cases} \alpha_i \geq 0, \\ \sum_i \alpha_i = 1. \end{cases} \tag{10}$$

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

$$\hat{X}(k,s) = \sum_{i=1}^{M} \hat{\alpha}_i^{(l)} \, Y_i(k,s). \tag{11}$$

Referring to Prop. 1 (and (6) 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 in 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 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 shortcomings 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 size of the noise variance 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.

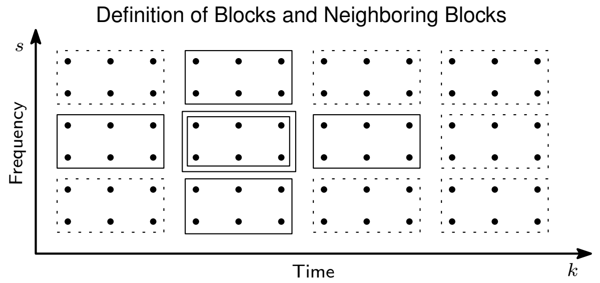


Fig. 2: The rectangles (solid or dashed) enclose time-frequency blocks used in the adaptive weighting stage. For the block enclosed in a double rectangle, the neighboring blocks are shown with solid lines – note that we pick the closest blocks in the horizontal and vertical directions as neighbors. For blocks on the boundaries, we similarly take neighbors to be the closest blocks in the horizontal and vertical directions.

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 overcome such behavior, we introduce a regularization term that penalizes swift changes in the weights of neighboring blocks. Note that such a modification also adresses the second shortcoming listed above.

More precisely, for the l^{th} block, let $\mathcal{N}(l)$ denote the indices 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. 2). Also, let $\alpha = \big(\alpha^{(1)}, \ldots, \alpha^{(L)}\big)$ 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,

$$P(\alpha) = \sum_{l=1}^{L} \sum_{m\in\mathcal{N}(l)} \|\alpha^{(l)} - \alpha^{(m)}\|_2^2. \tag{12}$$

Given the regularization function above, we propose to select the optimal weights as,

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{l=1}^{L} \left(\sum_{(k,s)\in\mathcal{T}_l} \left| \sum_{i=1}^{M} \alpha_i^{(l)} Y_i(k,s) \right| \right) + \lambda\, P(\alpha), \quad \text{s.t.} \begin{cases} \sum_i \alpha_i^{(l)} = 1, \\ \alpha_i^{(l)} \geq 0, \ \forall i, \end{cases} \text{for all } l, \tag{13}$$

where λ is a regularization parameter.

The minimization problems in (10) and (13) 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 adaptive weights obtained by the first stage are equal to the optimal weights.

Suppose now 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[1] (see the appendix),

$$\hat{\sigma}^2(k,s) = \frac{1}{2(M-1)} \sum_{i=1}^{M} \hat{\alpha}_i(k,s)\,\big|Y_i(k,s) - \hat{X}(k,s)\big|^2. \quad (14)$$

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

$$\bar{X}(k,s) = \left(1 - \frac{\tau(k,s)}{|\hat{X}(k,s)|}\right)_+ \hat{X}(k,s), \quad (15)$$

where $\tau(k, s)$ is a threshold value that possibly 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 $1 \le c \le 3$.

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. [40].

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 $H \le V$ 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 [40]. 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. 3). 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 norm of each subblock. We note that this procedure may be regarded as a 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 [28] for the large block B under different hypotheses.[2]

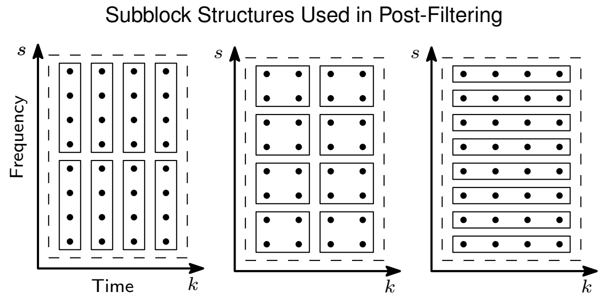


Fig. 3: Different subblock structures for $V = 3$, $H = 2$. From left to right, the subblock partitions for $v = 0, 1, 2$ are shown. We can think of the partitions as suitable for a 'transient' component when $v = 0$, a 'tonal component' when $v = 2$ and a 'combination' when $v = 1$.

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

$$\bar{X}(k,s) = \eta \hat{X}(k,s) \quad \text{for all } (k,s) \in \mathcal{I}, \quad (16)$$

where $\eta \in \mathbb{R}$. Notice now that if $X(k, s)$ denotes the unknown time-frequency samples to be estimated, then the choice of η minimizing the mean squared error (MSE) is given by,

$$\eta^* = \frac{\sum_{(k,s)\in\mathcal{I}} X^2(k,s)}{\sum_{(k,s)\in\mathcal{I}} \hat{X}^2(k,s)}. \quad (17)$$

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

$$\mathbb{E}\left[\sum_{(k,s)\in\mathcal{I}} \hat{X}^2(k,s)\right] = \sum_{(k,s)\in\mathcal{I}} X^2(k,s) + \sum_{(k,s)\in\mathcal{I}} \sigma^2(k,s), \quad (18)$$

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

$$\eta = \left(\frac{\sum_{(k,s)\in\mathcal{I}} \hat{X}^2(k,s) - \sum_{(k,s)\in\mathcal{I}} \hat{\sigma}^2(k,s)}{\sum_{(k,s)\in\mathcal{I}} \hat{X}^2(k,s)}\right)_+ \quad (19a)$$

$$= \left(1 - \frac{\sum_{(k,s)\in\mathcal{I}} \hat{\sigma}^2(k,s)}{\sum_{(k,s)\in\mathcal{I}} \hat{X}^2(k,s)}\right)_+ \quad (19b)$$

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

[1] Recall from Prop. 1 that the variance of the remaining noise is $2\sigma^2(k, s)$.

[2] In contrast, Yu et al. [40] 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.

obtained by solving a constrained convex minimization problem (see (10), (13)), 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, 18, 34]. Here, we describe an adaptation of a projected subgradient algorithm [16, 36] for the numerical solution of (10) and (13). We chose this algorithm because it is simple to implement, and its performance is comparable to other state-of-the-art algorithms (see [4] for an adaptation of a saddle point algorithm [12, 18, 34] 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

$$f(x) + \langle v, y - x \rangle \leq f(y), \quad \text{for all } y \in \mathbb{R}^n. \tag{20}$$

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,

$$\min_{x \in C} f(x), \tag{21}$$

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 [16, 36] 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.

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 [16, 36]. More precisely,

Proposition 2. [16, 36] Suppose that

$$\sup_{x \in C} \sup_{v \in \partial f(x)} \|v\|_2 \leq c \tag{22}$$

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

$$\lim_{i \to \infty} \beta_i = 0, \text{ and } \sum_i \beta_i \to \infty, \tag{23}$$

then,

$$\lim_{s \to \infty} \inf_{i \geq s} f(x^i) = \inf_{x \in C} f(x). \tag{24}$$

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,

$$Y = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_M \end{bmatrix}. \tag{25}$$

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

$$f(\alpha) = \|Y\,\alpha\|_1 = \sum_i |(Y\,\alpha)_i|, \tag{26}$$

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

$$\partial f(\alpha) = \text{real}\big(Y^H\, U\big), \tag{27}$$

where U denotes the set of vectors u which satisfy

$$u_i \in \begin{cases} \{(Y\,\alpha)_i / |(Y\,\alpha)_i|\}, & \text{if } (Y\,\alpha)_i \neq 0, \\ \{x \in \mathbb{C} : |x| \leq 1\}, & \text{if } (Y\,\alpha)_i = 0. \end{cases} \tag{28}$$

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 (13). 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,

$$\nabla_l P(\alpha) = 4 \sum_{m \in \mathcal{N}(l)} \alpha^{(l)} - \alpha^{(m)}. \tag{29}$$

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 (13) 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. We have used the method described in [17] for this task. 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,

$$\langle v - x, y - x \rangle \leq 0, \text{ for all } y \in C. \tag{30}$$

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 (21) 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

$$v = \begin{bmatrix} w_1 \\ \vdots \\ w_M \end{bmatrix} + t \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \tag{31}$$

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

$$w_i \in \begin{cases} \{0\}, & \text{if } \alpha_i > 0, \\ \mathbb{R}_+, & \text{if } \alpha_i = 0. \end{cases} \tag{32}$$

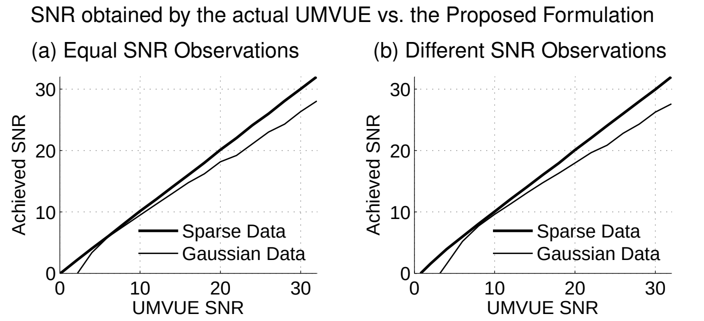


Fig. 4: Comparison of the SNR achieved by the proposed formulation and the UMVUE.

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, given as

$$x(n) = \begin{cases} z(n), & \text{if } v(n) \leq \tau, \\ 0, & \text{if } v(n) > \tau, \end{cases} \tag{33}$$

where $z(n)$ is a complex valued Gaussian random vector with iid real and imaginary parts 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

$$y_i = x + \sigma_i \, u_i, \tag{34}$$

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

$$\text{SNR}_{\text{UMVUE}} = 10 \log_{10} \left(\frac{\epsilon_x^2}{2\sigma^2 N} \right), \tag{35}$$

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 (9) 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. 4a. 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. 4a – the SNR achieved can be 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

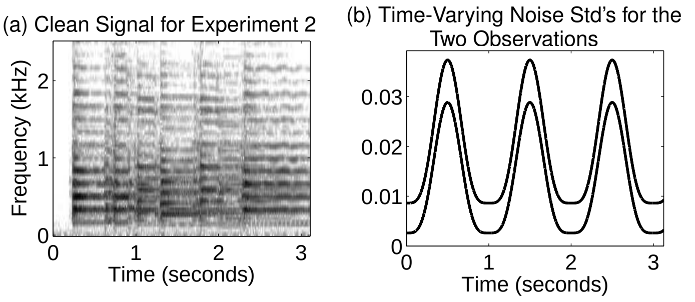


Fig. 5: (a) Spectrogram of the clean signal used in Experiment 2. (b) Time-varying noise standard deviations used for producing the two observations in Experiment 2.

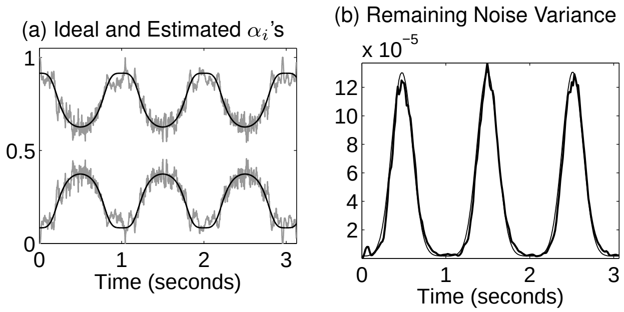


Fig. 6: (a) The ideal (black) and estimated (gray) α_i's for the adaptive weighting stage in Experiment 2. (b) Remaining noise variance after adaptive weighting in Experiment 2, estimated (solid) and actual (gray).

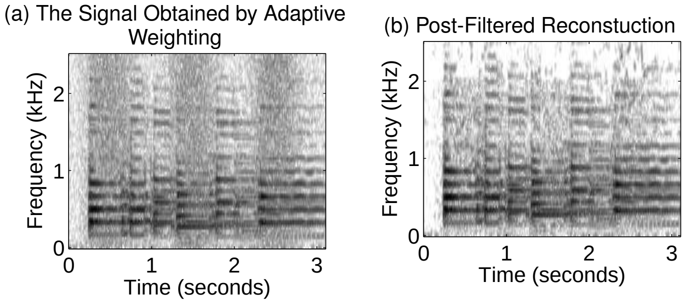


Fig. 7: Spectrograms of the resulting signals from Experiment 2. (a) Output of the adaptive weighting stage, SNR = 10.87 dB. (b) Reconstruction after applying the post-filter to the signal in (a), SNR = 20.91 dB.

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.

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. 4b. The behavior of the curves are similar to those in Fig. 4a. 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. 5a. Using the clean signal $x(n)$, we produced two noisy observations of the form

$$y_i(n) = x(n) + \sigma_i(n)\, z_i(n), \tag{36}$$

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. 5b. 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 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. 6a. 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. 7a (SNR = 10.87 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. 5b), 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. 6b (black 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 (14) is shown in Fig. 6b (gray 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 [21]. 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. 7b. 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. 8a 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 expected in general (and is not guaranteed by the algorithm – see Sec. III-A). In Fig. 8b, the gradients at each time-frequency block are shown. Observe that by Fig. 6a,

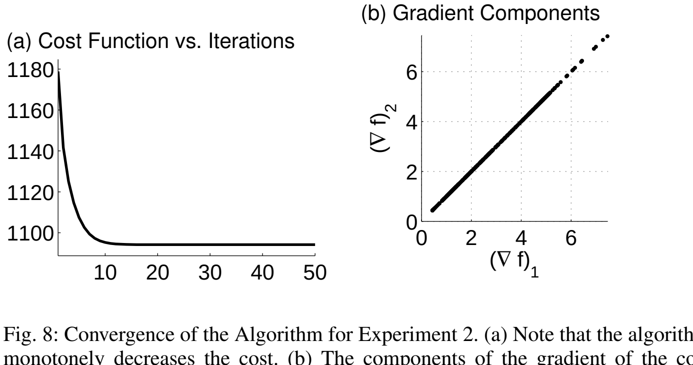


Fig. 8: Convergence of the Algorithm for Experiment 2. (a) Note that the algorithm monotonely decreases the cost. (b) The components of the gradient of the cost function lie in the direction of $[1\,1]^T$, as required by the optimality condition.

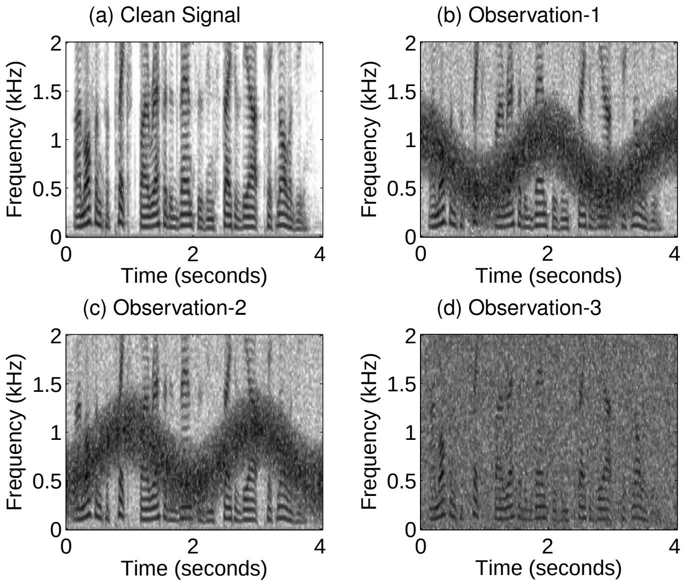


Fig. 9: Spectrogram of the clean signal and the observation signals from Experiment 3. The SNRs of all the observation signals are equal to -5 dB.

none of the variables assume the value zero. Therefore, by (31), (32), 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.

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 9a). 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 (equal to -5.00 dB). The spectrograms of the observations are shown in Fig. 9 b,c,d[3].

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

$$\alpha_i(k,s) = \sigma_i^{-2}(k,s) \left(\sum_{i=1}^{3} \sigma_i^2(k,s) \right)^{-1} \tag{37}$$

are shown in Fig. 10 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. 9 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. 11 for two different block sizes. The weigths 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.

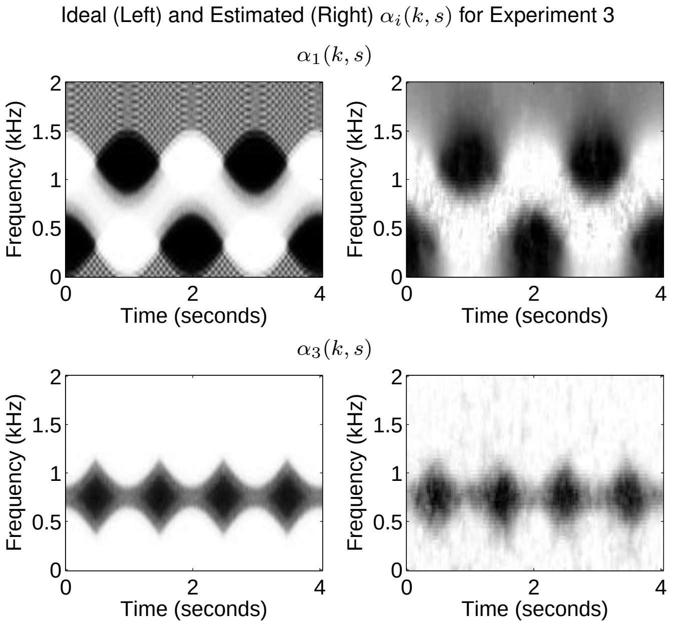


Fig. 10: Ideal (left column) and estimated (right column) α_i's used by the adaptive weighting stage in Experiment 3. The images are in linear scale – black and white indicate the values 1 and 0 respectively.

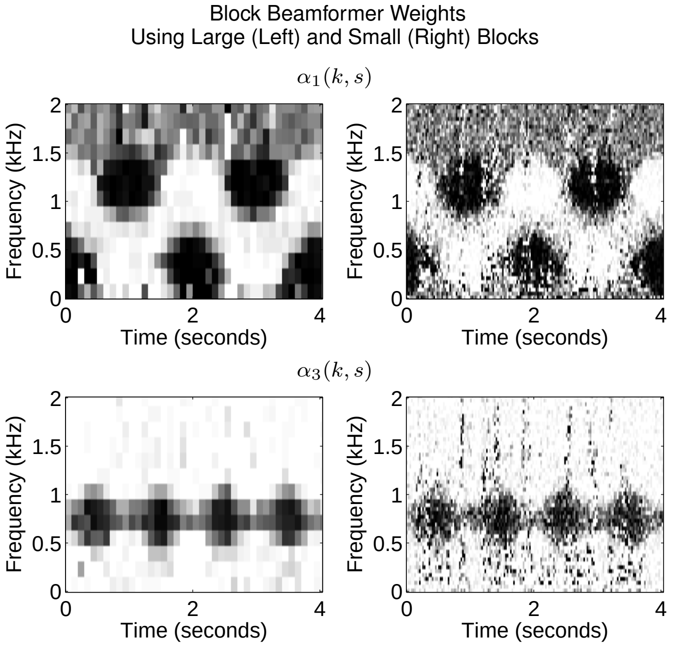


Fig. 11: Block beamformer weights for Experiment 3, using large (left) and small (right) blocks.

The weights obtained by the smoother formulation are shown on the right panel of Fig. 10. The reconstructed signal is shown in Fig. 12a. The actual and estimated time-frequency distribution of the remaining noise are shown in Fig. 12 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-

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

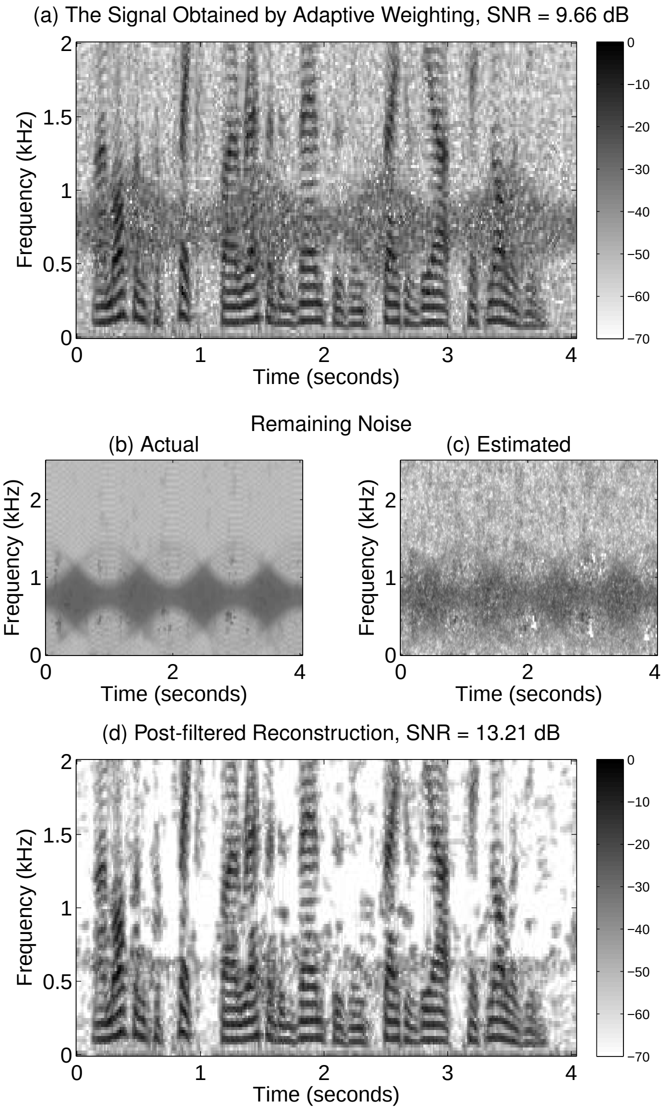


Fig. 12: The output signals for Experiment 3. (a) The signal obtained by adaptive weighting. (b,c) Remaining noise variances, actual (b) and estimated (c). (d) Final reconstruction obtained by post-filtering the signal in (a).

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 [21] 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. 12d.

In order to evaluate the performance of the algorithm with existing methods, we tried reconstructions with two different methods, namely the Frost beamformer [19, 20] and Zelinski's post-filtering method [41]. We note that both methods do not lead to the highest SNR in the literature [20, 31, 38] but we think they are both useful for benchmarking because they are well-known, simple and can achieve good performance.

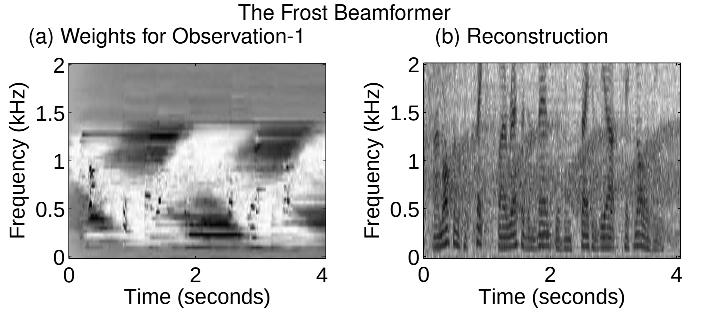


Fig. 13: (a) Weights used by the Frost Beamformer applied to the first observation in Experiment 3. The weights do follow the input noise patterns but, they are rather crude compared to those in Fig. 12. (b) Frost Beamformer reconstruction, SNR = 5.32 dB.

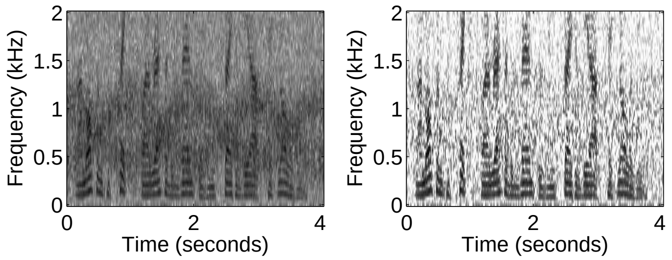


Fig. 14: Left : Average of the observations in Experiment 3, SNR = -0.20 dB. Right : Zelinski's post-filter applied to the average of the observations, SNR = 8.81 dB.

Frost's beamformer [19] 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 [20]. As noted in [20], 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 [20], Table 47.1) so as to achieve a high SNR value. The resulting weights and the reconstruction obtained are shown in Fig 13. 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. 14a), this results in a reconstruction with much lower performance than the proposed adaptive weighting reconstruction.

In a nutshell, Zelinski's method first combines the observed 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 [38, 41]. Although it is simple, Zelinski's method performs quite well in practice [38]. The spectrogram of the average of the observations is shown in Fig. 14a. 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 [41]) is shown in Fig. 14b. 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. 15. There are three sources and four microphones in the scene, where one of the sources is equidistant to the microphones. Sources s_1 and s_2 are human speakers whereas s_3 is a cell-phone (a ring-tune). Recordings are made in a semi-anechoic chamber and the observed signals'

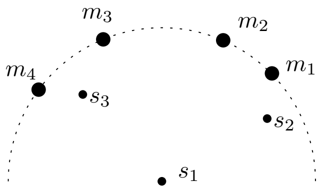
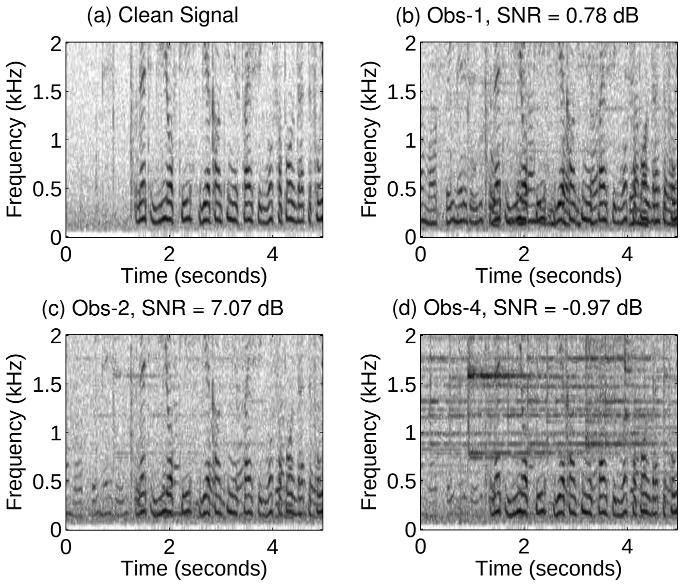


Fig. 15: The physical setup for Experiment 4. s_i's denote sources and m_i's denote microphones. The microphones are equidistant to s_1, but otherwise do not follow a specific pattern.


(a) Clean Signal

(b) Obs-1, SNR = 0.78 dB

(c) Obs-2, SNR = 7.07 dB

(d) Obs-4, SNR = -0.97 dB

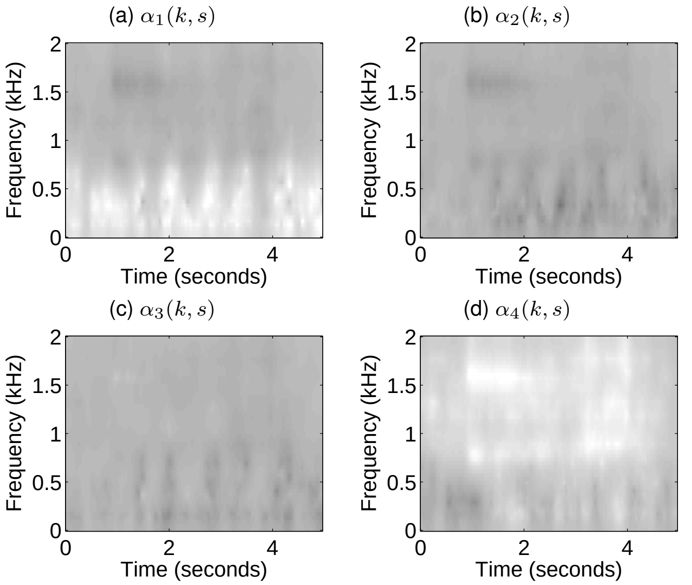
Fig. 16: Spectrogram of the clean signal and the observation signals from Experiment 4.

spectrograms are shown in Fig. 16. 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 is more dominant in the fourth observation.

The weights obtained by the smoother formulation are shown in Fig. 17. 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 and $\alpha_4(k, s)$ takes low values around the harmonics of the ring-tone. The spectrogram of the reconstruction is shown in Fig. 18b. Compared to a simple average (see Fig. 15), 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. 18c,d.

In our experiments with signals obtained in similar setups (including recordings in open air), 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 is also a viable option if a faithful reconstruction of the source is desired.

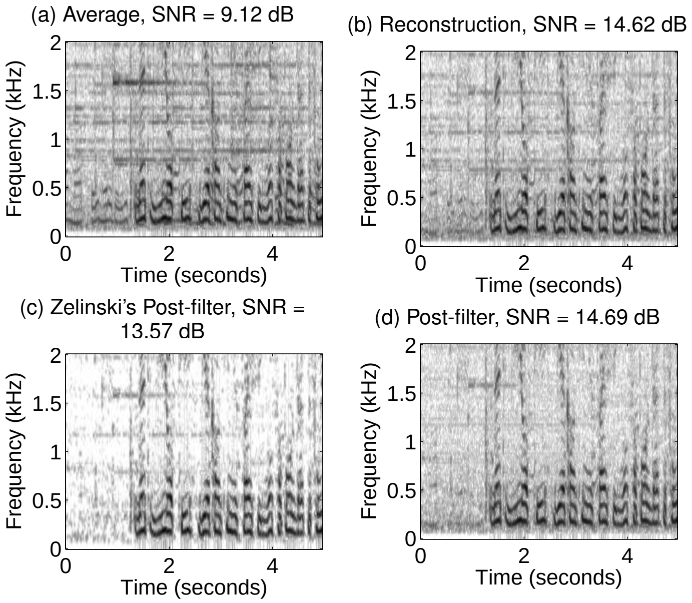

(a) $\alpha_1(k, s)$

(b) $\alpha_2(k, s)$

(c) $\alpha_3(k, s)$

(d) $\alpha_4(k, s)$

Fig. 17: Weights chosen by the proposed formulation for Experiment 4.


(a) Average, SNR = 9.12 dB

(b) Reconstruction, SNR = 14.62 dB

(c) Zelinski's Post-filter, SNR = 13.57 dB

(d) Post-filter, SNR = 14.69 dB

Fig. 18: Reconstructed signals for Experiment 4.

Experiment 5. The physical setup of this experiment is similar to Experiment 4, but this time there are only two speech sources. The source of interest is again equidistant to the microphones. The clean signal, the noise in the first microphone and the observed signals are shown in Fig. 19.

The ideal weights (if the noise signals were known) for adaptive weighting are shown in Fig. 20. The weights obtained by solving the proposed formulation are shown in Fig.21. Observe that the ideal weights depend highly on the signal of interest and the noise signals and some harmonics of the noise signal are visible. Such rapid change in the time-frequency plane is not easy to achieve and thus the weights obtained by the formulation are coarser. However, we can still see some broad correlation between the ideal and the obtained weights.

The average of the observations and the reconstruction obtained via adaptive weighting is shown in Fig. 22a,b. Adaptive weighting significantly improves the reconstruction – removing the unwanted source to a great extent. Zelinski's post-filter (applied to the aver-

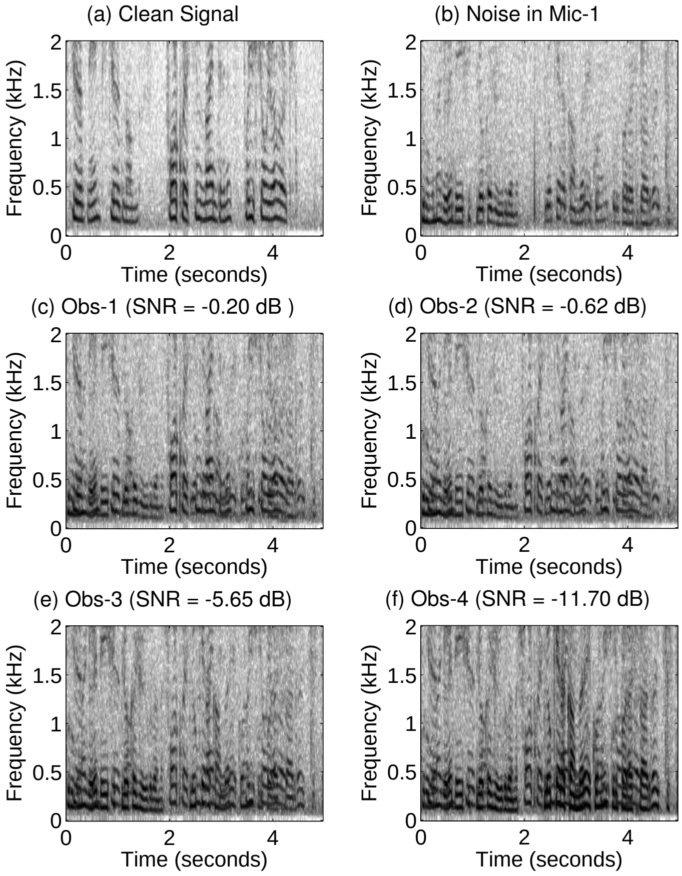


Fig. 19: Input signals for Experiment 5.

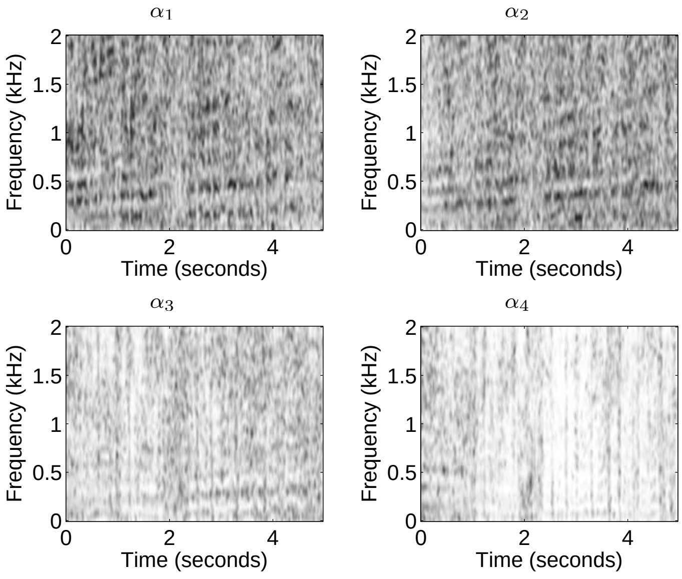


Fig. 20: Ideal Weights for Experiment 5.

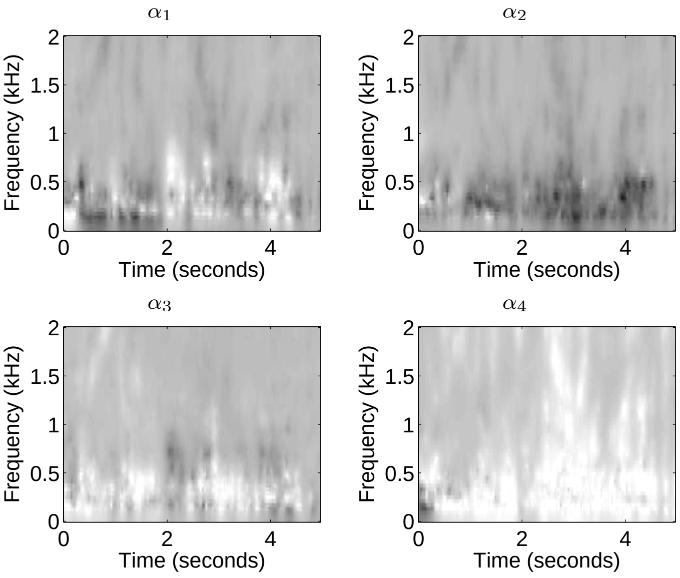


Fig. 21: Weights obtained by the proposed formulation for Experiment 5.

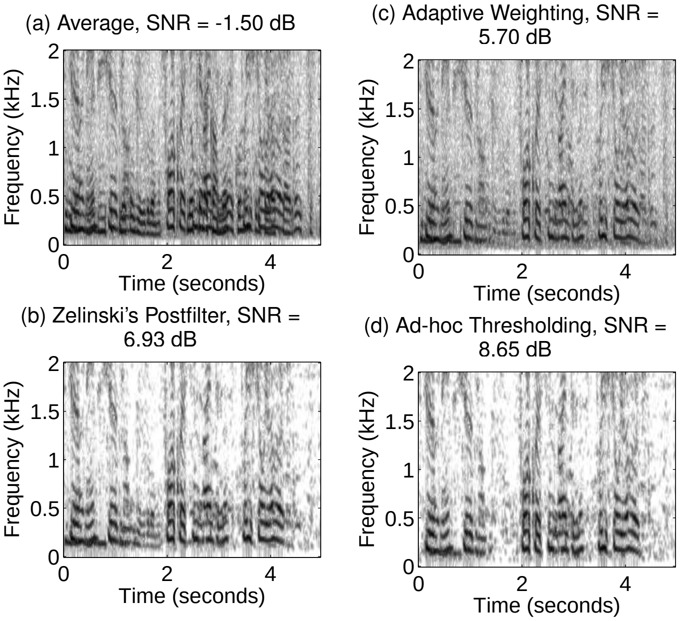


Fig. 22: Reconstructed signals for Experiment 5.

age of the observations) significantly suppresses the contribution from the unwanted source. However, because the average of the observations is a poor estimate, the source signal is also audibly degraded. In fact, despite the lower SNR, the adaptive weighting reconstruction may be preferable because the signal distortion is much lower [24]. In fact, when we apply an ad-hoc post-filter similar to Zelinski's post-filter to the adaptive weighting reconstruction, we obtain a higher SNR with a better suppression of the unwanted source, as shown in Fig. 22d.

Experiment 6. In a last experiment, we consider data recorded in open air. The setup is similar to that in the last two experiments. There are two speakers in the scene. One of them is equidistant to the microphones, whereas the other is closer to the first microphone. Also, there is significant wind noise in the recordings. The spectrogram of the four observations are shown in Fig. 23. Because we do not have the ground truth, we cannot report the

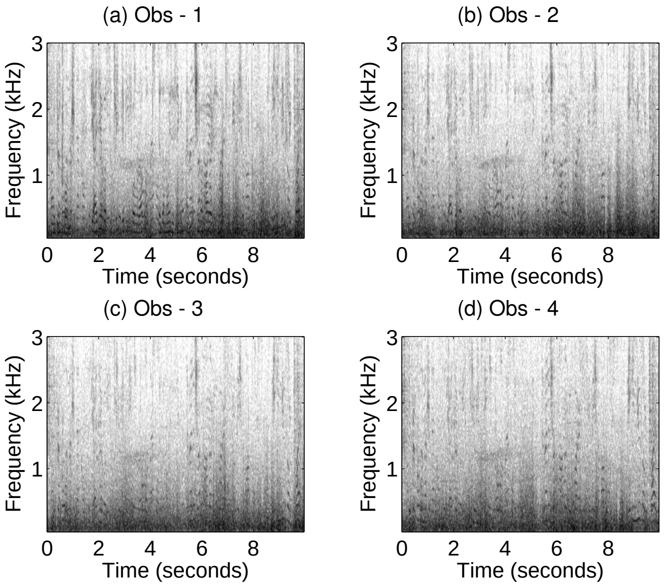


Fig. 23: Spectrograms of the observation signals from Experiment 6.

SNRs for this experiment. However, the audio signals are made available online.

The reconsruction obtained by averaging the observations are shown in Fig. 24a. While it is not easy to recognize from the spectrogram, the unwanted source is clearly audible in the averaged signal. If we use weighted averaging as proposed in this paper, the reconstruction is as shown in Fig. 24b. The energy of the unwanted source is significantly suppressed in this reconstruction, while the source of interest is retained with minimal distortion. The post-filtered reconstructions are shown in Fig. 24c,d. Zelinski's post-filter significantly improves the simple average. However, since the average signal is far from ideal to begin with, the source of interest is also degraded in this reconstruction. In contrast, the proposed block-based post-filter distorts the signal less but it is less effective in removing the noise. Perceptually, we think that the reconstruction obtained by weighted averaging (without post-filtering) is a better choice for this experiment.

V. Discussion and Conclusion

We proposed a framework for the reconstruction of an audio signal given its time-aligned noisy observations. The proposed formulation does not require information about noise but uses prior information about the source, namely sparsity of its spectrogram, in order to combine the observations via adaptive weighting. We noted that this stage essentially forms an estimate of a complete sufficient statistic for the underlying signal of interest. Thanks to this property, any Bayesian 'optimal' estimate, that takes as input the given multichannel data, can be realized as a single stage filter applied to the output of the adaptive weighting stage (see also [2] in this context). For this, we presented results based on soft thresholding and a block-based thresholding method. We would like to note that other denoising methods that take into account the time-frequency characteristics of the signal of interest [3, 29, 37, 40] could also be adapted for use as a post-filter.

We regard the adaptive weighting stage as a device to estimate the UMVU estimate, which turns out to be a complete sufficient

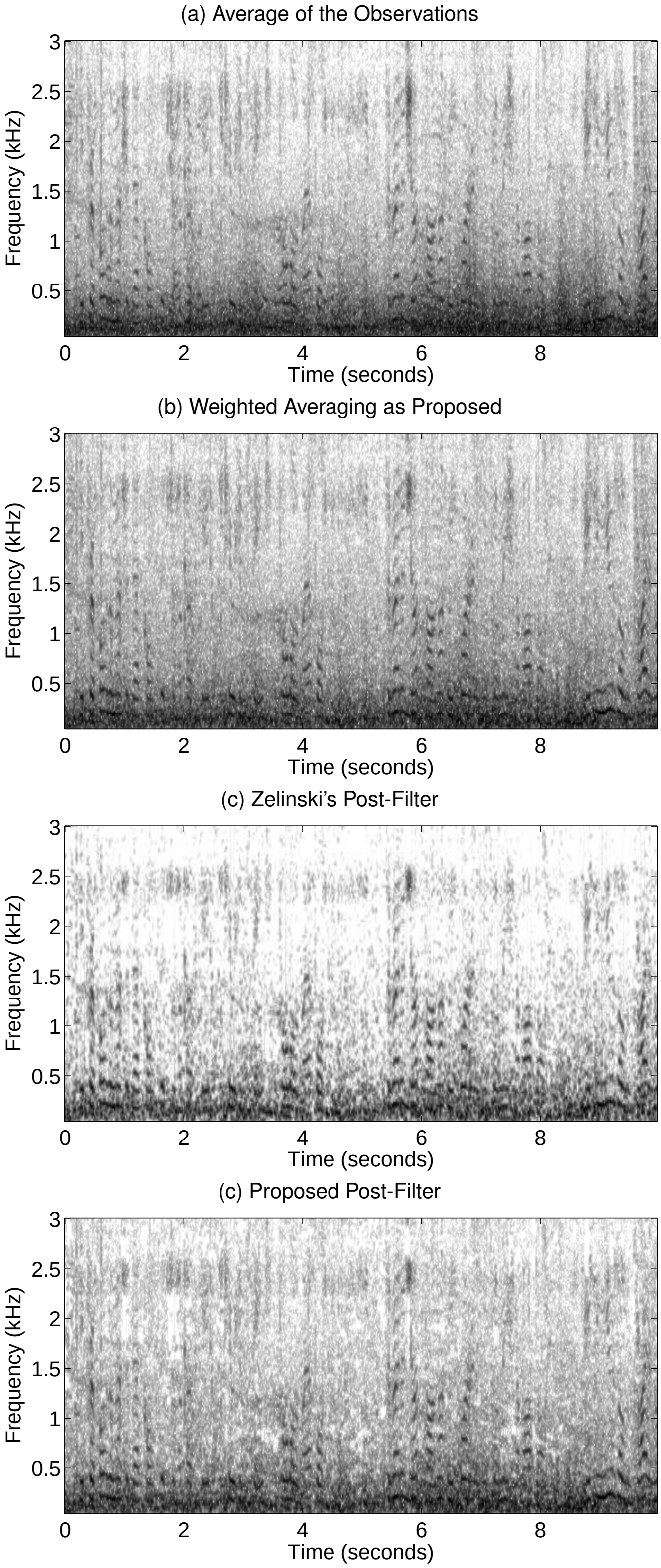


Fig. 24: Spectrograms of the reconstructed signals from Experiment 6.

statistic. One question of interest is, can we obtain a reliable estimate with a simpler scheme, say by a simple average of the observations as in the Zelinski's post-filtering method? In fact, simple averaging can be argued to be a good estimate if the noise variance in the different recordings are approximately the same [7, 39] (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, because 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 3.

In general, determination of the adaptive weights is more effective when time-frequency blocks are treated jointly. However, independent processing of blocks can also produce reconstructions with acceptable performance. Independent processing might also be of interest because it allows to derive real-time algorithms with a small delay (as well as parallel processing).

Appendix
Proof of Proposition 1

In order to simplify the notation, let us drop the time-frequency indices (k, s). 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 $Y_1, \ldots, Y_M$ are given as

$$\begin{aligned} f(y_1, \ldots, y_M) &= \overbrace{\left(\prod_{i=1}^{M} \frac{1}{2\pi\sigma_i^2} \right)}^{c} \exp\left(-\sum_{i=1}^{M} \frac{1}{2\sigma_i^2} |y_i - X|^2 \right) \\ &= c \exp\left(-\sum_{i=1}^{M} \frac{1}{2\sigma_i^2} |y_i|^2 \right) \exp\left(X^r \sum_{i=1}^{M} \frac{1}{\sigma_i^2} y_i^r \right) \\ &\qquad \exp\left(X^j \sum_{i=1}^{M} \frac{1}{\sigma_i^2} y_i^j \right) \exp\left(-|X|^2 \sum_{i=1}^{M} \frac{1}{2\sigma_i^2} \right), \end{aligned}$$

where y_i^r and y_i^j denote the real and imaginary parts of y_i. It follows by the factorization theorem [26, 35] 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

$$S = S^r + j\, S^j = \sum_{i=1}^{M} \sigma_i^{-2}\, Y_i \tag{38}$$

is a sufficient statistic for X. Further, since Y_i's are independent and circularly normal [32], S is also circularly normal (i.e., S^r and S^j are independent). Also, $\mathbb{E}(S) = \sigma^{-2} X$, where

$$\sigma^2 = \left(\sum_{i=1}^{M} \sigma_i^{-2} \right)^{-1}, \tag{39}$$

and

$$\operatorname{var}(S^r) = \operatorname{var}(S^j) = \sum_{i=1}^{M} \sigma_i^{-4}\, \sigma_i^2 = \sum_{i=1}^{M} \sigma_i^{-2} = \sigma^{-2}. \tag{40}$$

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

$$\begin{aligned} &\int\!\!\int g(t^r, t^j) \times \\ &\exp\left(-\frac{\sigma^2}{2} \left[(t^r - X^r \sigma^{-2})^2 + (t^j - X^j \sigma^{-2})^2 \right] \right) dt^r\, dt^j = 0, \end{aligned}$$

for all (X^r, X^j), which implies that $g(t^r, t^j) = 0$ for all t. It follows by the Rao-Blackwell theorem [26, 35] 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 $\sigma^4 \operatorname{var}(S^r) = \sigma^2$.

Appendix
Expected Value of the Estimator in (14)

In order to show that the estimator in (14) 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 [32]), so that $\mathbb{E}(|Z_i|^2) = 2$. In this setting, let

$$\sigma^2 = \left(\sum_{i=1}^{M} \sigma_i^{-2} \right)^{-1}, \quad \alpha_i = \sigma^2\, \sigma_i^{-2}, \quad \hat{X} = \sum_{i=1}^{M} \alpha_i Y_i. \tag{41}$$

Now let,

$$\hat{\sigma}^2 = \frac{1}{2(M-1)} \sum_{i=1}^{M} \alpha_i \left| Y_i - \hat{X} \right|^2 \tag{42}$$

Since α_i's add to unity, we can write

$$\hat{X} = X + \sum_{i=1}^{M} \alpha_i Z_i. \tag{43}$$

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

$$\begin{aligned} \mathbb{E}(s) &= \sum_{i=1}^{M} \alpha_i\, \mathbb{E}\Big(\left| Y_i - \hat{X} \right|^2 \Big) && \text{(44a)} \\ &= \sum_{i=1}^{M} \left[\alpha_i\, \mathbb{E}\Big(\left| (1-\alpha_i) Z_i \right|^2 + \sum_{m \neq i} |\alpha_m\, Z_m|^2 \Big) \right] && \text{(44b)} \\ &= 2\sigma^2 \sum_{i=1}^{M} \left[(1-\alpha_i)^2 + \alpha_i\, (1-\alpha_i) \right] && \text{(44c)} \\ &= 2(M-1)\, \sigma^2. && \text{(44d)} \end{aligned}$$

Thus, $\hat{\sigma}^2$ in (42) 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,

$$\langle u, z\rangle = \sum_{k=1}^{n} u_k^r\, z_k^r + u_k^i\, z_k^i. \tag{45}$$

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,

$$\|z\|_1 = \sup_{v \in B_\infty} \langle v, z\rangle. \tag{46}$$

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

$$v_k \in \begin{cases} \{w \in \mathbb{C} : |w| \leq 1\}, & \text{if } z_k = 0, \\ \{z_k/|z_k|\}, & \text{if } z_k \neq 0. \end{cases} \tag{47}$$

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,

$$\partial f(\alpha) = Y^r\, U^r + Y^i\, U^i, \tag{48}$$

where U is the set of vectors u as described in (28). This is equivalent to (27).

References

[1] R. Aichner, H. Buchner, and W. Kellerman. Convolutive blind source separation for noisy mixtures. In E. Hänsler and G. Schmidt, editors, *Speech and Audio Processing in Adverse Environments*. Springer, 2008.

[2] R. Balan and J. Rosca. Microphone array speech enhencement by Bayesian estimation of spectral amplitude and phase. In *Proc. IEEE Sensor Array and Multichannel Signal Proc. Workshop*, 2002.

[3] İ. Bayram. Mixed-norms with overlapping groups as signal priors. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2011.

[4] İ. Bayram. Combining multiple observations of audio signals. In *Proc. SPIE Conf. on Wavelets and Sparsity*, 2013.

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

[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, 2011.

[9] T. T. Cai. Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *The Annals of Statistics*, 27:898–924, 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] D.Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[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. 25th Int. Conf. on Machine Learning*, 2008.

[18] 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.

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

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

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

[22] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas and Propagation*, 30(1):27–34, January 1982.

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

[24] Y. Hu and P. C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio, Speech and Language Processing*, 16(1):229–238, January 2008.

[25] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

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

[27] C. Kereliuk and P. Depalle. Sparse atomic modelling of audio : A review. In *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2011.

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

[29] M. Kowalski, K. Siedenburg, and M. Dörfler. Social sparsity!

Neighborhood systems enrich structured shrinkage operators. *IEEE Trans. Signal Processing*, 61(10):2498–2511, May 2013.

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

[31] 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.

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

[33] 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.

[34] 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*, 2009.

[35] H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer, second edition, 1998.

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

[37] K. Siedenburg and M. Dörfler. Structured sparsity for audio signals. In *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2011.

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

[39] J. W. Tukey. Approximate weights. *Annals of Mathematical Statistics*, 19:91–92, 1948.

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

[41] 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)*, 1988.

[42] M. Zibulevsky, B.A. Pearlmutter, P. Bofill, and P. Kisilev. Blind source separation by sparse decomposition. In S. J. Roberts and R. M. Everson, editors, *Independent Component Analysis: Principles and Practice*. Cambridge, 2001.