Combining Multiple Observations of Audio Signals

İlker Bayram

Istanbul Technical University, Istanbul, Turkey

ABSTRACT

We consider the problem of reconstructing an audio signal from multiple observations, each of which is contaminated with time-varying noise. Assuming that the time-variation is different for each observation, we propose an estimation formulation that can adapt to these changes. Specifically, we postulate a parametric reconstruction and choose the parameters so that the reconstruction minimizes a cost function. The cost function is selected so that audio signals are penalized less compared to arbitrary signals with the same energy. As cost functions, we experiment with a recently proposed prior as well as mixed norms placed on the short time Fourier coefficients.

Keywords: Beamforming, audio fusion, mixed norm, group sparsity, variational formulation, time-varying noise

1. INTRODUCTION

Suppose we have noisy observations of an audio signal x, as

$$y_1(k) = x(k) + n_1(k),$$
$$y_2(k) = x(k) + n_2(k),$$

where $k \in \mathbb{Z}$ denotes time and n_i's denote noise terms with possibly different and unknown time-varying characteristics. In this paper, we consider the problem of combining the information from such multiple observations to estimate the original signal.

As an example, spectrograms of two noisy observations are shown in Fig.1. The noise terms affect the signal of interest in different instances in time. Note that, because the noise variance varies with time, it is not desired to 'denoise' the observations with an off-the-shelf denoising algorithm that might have stationarity assumptions. Rather than resorting to such preprocessing, we formulate the reconstruction problem as a minimization problem.

One approach to construct a cost function might be to employ sums of

(i) data terms, penalizing deviations from the observations,
(ii) regularization terms, that penalize the deviation from a prior model.

In such a setting, it is desirable to employ time-varying weights for data terms of the different observations, because the validity of the observations depend on the time-variation of the noise terms. Normally, one could weight the distances from the observations based on the noise level. But the noise characteristics are unknown, therefore it is not obvious how to choose the weights. To avoid this difficulty, we do not include such data terms in our formulation. Instead, we postulate a parametric reconstruction and impose constraints on the parameters in order to make use of the properties that the reconstruction is expected to have. Specifically, the reconstruction we propose for the problem outlined above is

$$\hat{x}(k) = \hat{\alpha}_1(k)\, y_1(k) + \hat{\alpha}_2(k)\, y_2(k), \tag{1}$$

where,

$$(\hat{\alpha}_1, \hat{\alpha}_2) = \underset{\alpha_1,\alpha_2}{\operatorname{argmin}}\; g\Big(y_1(k)\,\alpha_1(k) + y_2(k)\,\alpha_2(k)\Big) + \lambda\left(\mathrm{TV}(\alpha_1) + \mathrm{TV}(\alpha_2)\right), \text{ subject to } \begin{cases}\alpha_1(k) + \alpha_2(k) = 1,\\ \alpha_i(k) \geq 0.\end{cases} \tag{2}$$

E-mail: ibayram@itu.edu.tr

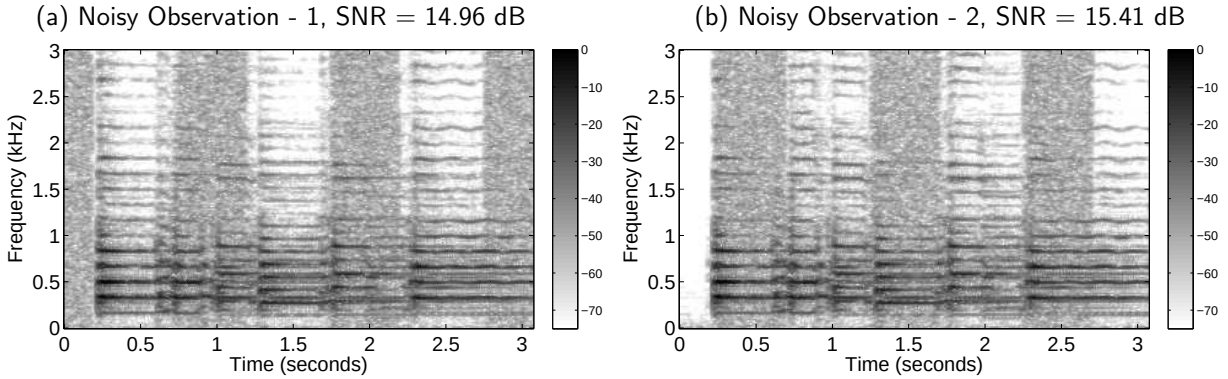

Figure 1. Spectrograms of (a) First Observation, (b) Second Observation. Notice that the noise terms in the two observations contaminate different portions of the desired signal. However, the time-varying noise variance information is not provided. Our goal is to reconstruct the original signal from these observations.

Here, the function $g(\cdot)$ penalizes deviations from spectral sparsity of its argument, and the TV term denotes the total variation of its arguments. We experimented with different $g(\cdot)$ functions as well as different variants of total variation. Note that the constraints on $\alpha_i(k)$, for each k, ask that we take a convex combination of the observations. If $g(\cdot)$ is a functional that has a minimum at zero, then the best linear (or, even affine) combination will be zero (as discussed further in the following). Employing a convex combination instead of a linear or affine combination, we avoid the zero solution. But this is not the sole reason behind restricting α_i's. In Section 2.1, we provide further motivation for this constraint through the discussion of a simpler problem.

Related Work

The formulation described above may be regarded as an instance of a beamforming formulation (see Refs. 1,2,3 for an overview). Beamforming typically involves multiple sensors recording data from one or multiple sources. When the sensor positions as well as the source direction is known, the challenge is to combine the multiple observations to form a close estimate of the source. An interesting formulation for this problem, which was employed in Refs. 4, 5, can be described as follows. Put briefly, the formulation applies an adaptive filter to each sensor observation and then sums the filtered observations to produce an output. The filter coefficients are chosen so as to minimize the expected output power subject to side constraints. The side constraints are imposed in order to ensure that the signal of interest is preserved. By employing stochastic approximations to the terms that appear in the solution, Frost proposed a beamformer in Ref. 5. Griffiths and Jim showed in Ref. 6 that the beamformer proposed in Ref. 5 can be rearranged as the sum of two beamformers that operate in parallel on the input. One of these beamformers is fixed and supplies the signal from the direction of interest, whereas the other beamformer helps reduce the total power. The formulations discussed above usually assume that the source/sensor positions are known. Cox et al.[7] discuss modifications to the formulations so that slight deviations from such prior information do not lead to significant degradation of the beamformer output. This is achieved by incorporating terms that penalize the sensitivity of the beamformer output to such prior location information. More recently, Parra and Alvino[8] combine ideas from beamforming and blind source separation. Specifically, for a problem that involves multiple sources and multiple sensors, they aim to separate the sources. For this, they consider cost functions proposed originally for blind source separation, subject to constraints derived from classical beamforming methods. We note that the problem discussed here does not fit well to this last scenario, since we do not assume that the noise components behave as coherent sources that act on the sensors.

Relation with the Proposed Formulation

We now briefly describe the approach in Ref. 5 and relate it to the proposed formulation. In the setting above, which assumes given two noisy observations, the idea in Ref. 5 is to apply filters $h_i(k)$ to the observations $y_i(k)$ and sum the filtered outputs to obtain the estimate as,

$$\hat{x}(k) = y_1(k) * h_1(k) + y_2(k) * h_2(k). \tag{3}$$

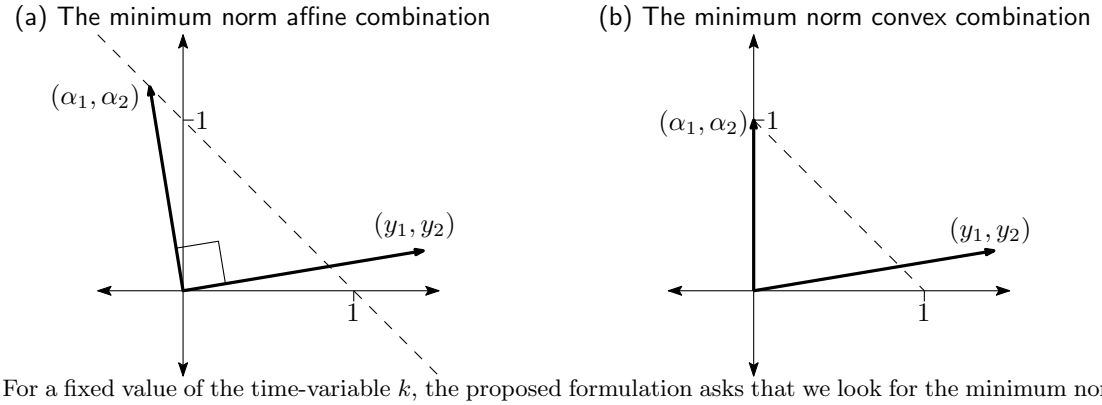

Figure 2. For a fixed value of the time-variable k, the proposed formulation asks that we look for the minimum norm linear combination of $y_1(k)$, $y_2(k)$ subject to constraints. This is achieved by asking that the output be given by $\alpha_1(k)\, y_1(k) + \alpha_2(k)\, y_2(k)$, and restricting $\alpha_i(k)$'s to lie in certain sets. (a) For fixed k, if we ask that α_i's sum to unity, this is equivalent to asking that the vector (α_1, α_2) lie on the dashed line. With this constraint, the minimum norm linear combination y_i's will be zero, unless $y_1(k) = y_2(k)$. (b) If we further ask that $\alpha_i \geq 0$, this reduces the infinite dashed line in (a) to a segment in the first orthant (shown as a dashed segment). Under this additional constraint, the minimum size linear combination will be zero only if $\operatorname{sgn}\big(x_1(k)\big) \neq \operatorname{sgn}\big(x_2(k)\big)$.

The formulation models the signal of interest and the noise components as uncorrelated stochastic processes, where the noise terms are also assumed to be zero mean. The sum of the filters $h_i(k)$ are constrained to be equal to some fixed filter $h(k)$. This constraint allows to express the output as,

$$\hat{x}(k) = x(k) * h(k) + \Big[n_1(k) * h_1(k) + n_2(k) * h_2(k) \Big]. \tag{4}$$

Now assume that the source $x(k)$ and the noise terms $n_i(k)$ are independent stationary processes. Under these assumptions, we can express the expected power as,

$$\mathbb{E}\big(\hat{x}^2(k)\big) = \left(\sum_{n,l} h(n)\, R_x(n-l)\, h(l) \right) + \left(\sum_{i=1}^{2} \sum_{n,l} h_i(n)\, R_i(n-l)\, h_i(l) \right),$$

where R_x and R_i denote the autocorrelation functions of x and n_i respectively. In this expression, the first term inside the parentheses is independent of the choice of the filters h_i, since h is fixed. But the second term is a non-negative term that is due to noise and is a function of h_i's. Therefore, minimizing the expected output power by a proper choice of h_i is expected to reduce the power of the noise terms and not the source. Based on this observation, the beamformer in Ref. 5 aims to adaptively minimize the total output power by varying the coefficients of h_i's slowly in time.

One attempt to adapt this scheme to a deterministic setting might be to try to minimize the total output power and allow the coefficients of h_i to vary with time. However, this approach has a shortcoming. Specifically, suppose that each h_i is a single tap filter, so that the output is formed as in (1), along with the constraint $\alpha_1(k) + \alpha_2(k) = 1$, for all k (as also required by Lacoss[4]). Now if zero is also a minimum of the function $g(\cdot)$ (which it is, for the functions used in this paper), the best choice of $\alpha_i(k)$ will set $\hat{x}(k)$ to zero, provided $y_1(k) \neq y_2(k)$. This is demonstrated in Fig. 2a. If there are no further constraints on $\alpha_i(k)$, this in turn leads to a zero reconstruction for all k. Such a behavior is avoided by the formulation in (1). Instead of an *affine* combination, the formulation seeks the best *convex* combination, by further constraining α_i's to be non-negative (see Fig. 2b). We note however that this is not the main motivation behind employing the convex combination. A more detailed treatment is provided in the sequel.

2. PROBLEM FORMULATIONS

We noted in the Introduction that the formulation is based on a parametric reconstruction. In order to motivate this parametric reconstruction, we start by considering a simple problem.

2.1 Estimating a Constant From Noisy Observations

Suppose we have K observations of a constant Θ as

$$y_i = \Theta + \sigma_i \, n_i, \quad \text{for } i = 1, 2, \ldots, K. \tag{5}$$

where n_i's denote unit variance, zero-mean, independent random variables. Given the observations, we form the estimate of Θ as a linear combination of y_i's as,

$$\hat{\Theta} = \underbrace{\begin{bmatrix} y_1 & \ldots & y_K \end{bmatrix}}_{y^T} \underbrace{\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix}}_{\alpha} \tag{6}$$

We are interested in the choice of α that minimizes $\mathbb{E}\big(|\Theta - \hat{\Theta}|^2\big)$. Note that,

$$\mathbb{E}\big(|\Theta - \hat{\Theta}|^2\big) = \Theta^2 - 2\,\Theta^2 \sum_{i=1}^{K} \alpha_i + \alpha^T \, R \, \alpha, \tag{7}$$

where

$$R = \operatorname{diag}\big(\sigma_1^2, \, \sigma_2^2, \, \ldots, \sigma_K^2\big) + \Theta^2 \begin{bmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{bmatrix}. \tag{8}$$

Setting the gradient with respect to α to zero, we see that the best choice $\hat{\alpha}$ is the solution of,

$$R \, \hat{\alpha} = \Theta^2 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{9}$$

It can be checked that

$$\hat{\alpha} = \frac{1}{\Theta^{-2} + \sum_{i=1}^{K} \sigma_i^{-2}} \begin{bmatrix} \sigma_1^{-2} \\ \sigma_2^{-2} \\ \vdots \\ \sigma_K^{-2} \end{bmatrix} \tag{10}$$

satisfies (9). Observe that with increasing Θ^2, the best linear combination tends to a convex combination of the observations, where the weights are inversely proportional to the noise variance of each observation. In the following, we will use this observation as a guideline to formulate the reconstruction problem.

2.2 Adapting to the Current Problem

We now consider the general version of the problem introduced in the beginning of the paper. Instead of two, let us have K observations,

$$y_i(k) = x(k) + n_i(k), \quad \text{for } i = 1, 2, \ldots, K. \tag{11}$$

We will propose two formulations. The first one will not include any regularization term applied on α_i's, and therefore no assumption on their behavior. The second one will assume that the variation of the noise characteristics will be slow and employ this knowledge by including terms that penalize the time-variation of α_i's.

Formulation without Regularization Terms

In view of the discussion in Sec. 2.1, we seek an estimate of $x(k)$ in the set of convex combinations of $y_i(k)$'s. For this, define,

$$x_\alpha(k) = \sum_{i=1}^{K} \alpha_i(k)\, y_i(k).$$

where $\alpha_i(k)$'s are weights to be determined. Observe that the reconstruction $x_\alpha(k)$ is a function of $\alpha_i(k)$'s. In order to choose the 'best' $\alpha_i(k)$'s, we need a criterion or cost function. Now, let $g(\cdot)$ be a non-negative functional that favors an audio signal over noise. By this, we mean that if y and z are two signals with equal energy, where y is an audio signal and z is some noise signal, then $g(y) \leq g(z)$. In a broad sense, one could think of $g(\cdot)$ as defining a prior distribution $p(\cdot)$ of the form $p(\cdot) \propto e^{-g(\cdot)}$. For signals with fixed energy, this prior distribution is expected to assume higher values for audio signals. Given such a function, we choose $\alpha_i(k)$ as

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}}\ g\big(x_\alpha(k)\big), \quad \text{subject to } \begin{cases} \sum_{i=1}^{K} \alpha_i(k) = 1, \\ \alpha_i(k) \geq 0, \text{ for } i = 1, 2, \ldots, K. \end{cases} \tag{12}$$

We experimented with different $g(\cdot)$ functions. Specifically, we used the prior introduced in Ref.9, as well as mixed norms[10] placed on the short time Fourier transform coefficients. The rationale behind these choices for a cost function is the expectation that the spectrogram of the audio signal will be roughly sparse, or sparse with a structured appearance – see Refs. 11, 12, 13, 14, for recent, related works.

The functionals mentioned above are convex. Thanks also to the convexity of the constraints in (12), this implies that the reconstruction is obtained by solving a convex minimization problem. The special forms of the functionals also allow to express the problem in (12) as a saddle point problem, for which convergent algorithms are available. The details are provided in Sec. 3,4.

Formulation with Regularization Terms on α_i's

We noted that we do not assume knowledge about the noise characteristics of the different observations. If, however, it is known that the noise characteristics change slowly over time, this knowledge can be incorporated into the formulation. In that case, it can be argued that the best $\alpha_i(k)$'s should vary slowly in time as well. This suggests an update of the formulation for estimating the best $\alpha_i(k)$'s as,

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}}\ g\big(x_\alpha(k)\big) + \lambda \sum_{i=1}^{K} \big\|D\alpha_i\big\|, \quad \text{subject to } \begin{cases} \sum_{i=1}^{K} \alpha_i(k) = 1, \\ \alpha_i(k) \geq 0, \text{ for } i = 1, 2, \ldots, K. \end{cases} \tag{13}$$

where, 'D' is a difference operator that maps $\alpha_i(k)$ to $d_i(k) = \alpha_i(k) - \alpha_i(k-1)$. The norm $\|\cdot\|$ used in the formulation penalizes the deviation of $d_i(k)$ from zero. When it is taken as the ℓ_1 norm, we get TV regularization. We have experimented with TV as well as the ℓ_2 norm for $\|\cdot\|$, but we will present results based on the latter choice. We demonstrate in Sec. 5 that including this regularization term improves the performance.

3. RELATION WITH A SADDLE POINT PROBLEM

3.1 A Generic Saddle Point Problem

In the following, we will rewrite the problem formulation in (13) as a saddle point problem. Specifically, we will see that it can be expressed as an instance of the generic saddle point problem given as,

$$\min_{\alpha \in A} \max_{z \in Z} \langle L\,\alpha, z \rangle, \tag{14}$$

where A, Z are closed convex sets and L is a linear operator. There are numerous algorithms that are proposed for such problems, see e.g. Refs.15, 16, 17, 18, 19. In particular, an algorithm presented in Refs. 17, 18, 19 takes the form,

Algorithm 1 A Saddle Point Algorithm[17,18,19]

Initialize $\alpha^0, \bar{\alpha}^0, z^0 \in D$. Set $\gamma \leq 1/\sigma(L)$. ▷ Initialization
for $n \geq 0$ **do**
$z^{n+1} \leftarrow P_Z\big(z^n + \gamma\, L\, \bar{\alpha}^n\big)$
$\alpha^{n+1} \leftarrow P_A\big(\alpha^n - \gamma\, L^T\, z^{n+1}\big)$
$\bar{\alpha}^{n+1} \leftarrow 2\alpha^{n+1} - \alpha^n$
end for

Here, $P_A(\cdot)$, $P_Z(\cdot)$, denote projection operators onto A and Z respectively. Also, $\sigma(L)$ denotes the spectral norm of L. In Section 4, we will adapt this algorithm to our problem. But first, we need to put (13) into the form (14).

3.2 Rewriting the Original Problem

In order to express (13) as a saddle point problem, we need a few definitions. In the following, we will take $g(\cdot)$ to be a mixed norm[10] placed on the STFT coefficients of its input (for another alternative, see Ref. 9).

STFT Operator

Let S denote an STFT operator. Specifically, we think of S as an operator that maps a one-dimensional signal $x(n)$, to a two-dimensional array $c(s,k)$, through the relation,

$$c(s,k) = \sum_n x(n)\, w(n-kN)\, e^{j\omega_s(n-kN)}, \tag{15}$$

where $w(\cdot)$ denotes a real-valued time-localized window function and N is the 'hop-size' that determines the redundancy of the transform. Notice that here $c(s,k)$ actually represents a time-frequency image of the signal. The first parameter s is related to frequency and k is associated with time. In general, for a fixed (s,k) pair, $c(s,k)$ is a complex number. However, we can also (and will do, in the rest of the paper) think of it as a real vector with two components (the real and imaginary parts). Consequently, the transpose of S is obtained as follows. For $x' = S^T c$, we have

$$x'(n) = \sum_s \sum_k \Big[\mathrm{Re}\{c(s,k)\}\, \cos\big(\omega_s(n-kN)\big) + \mathrm{Im}\{c(s,k)\}\, \sin\big(\omega_s(n-kN)\big)\Big]\, w(n-kN) \tag{16}$$

$$= \mathrm{Re}\left\{\sum_s \sum_k c(s,k)\, w(n-kN)\, e^{-j\omega_s(n-kN)}\right\}. \tag{17}$$

Mixed Norm

Given a two-dimensional (vector) array $c(s,k)$ as described above, the mixed norm definition we will use in this paper is,

$$\|c\|_{2,1} = \sum_s \sum_m \sqrt{\sum_{l=0}^{M-1} \big|c(s, mM+l)\big|^2}. \tag{18}$$

Here, M may be regarded as a parameter of the mixed norm. If c is the time-frequency map of an audio signal and e is an arbitrary two-dimensional array with the same energy, $\|c\|_{2,1}$ is expected to be smaller than $\|e\|_{2,1}$. For further discussions, as well as a more general investigation of mixed norms in this context, we refer to Ref.10. We note that if we take the group size M as one, the norm above becomes the ℓ_1 norm of the coefficients.

Notice that the mixed norm defined above actually is the sum of the ℓ_2 norms of *groups of coefficients*, where M denotes the group size. From this observation, it follows that*

$$\|c\|_{2,1} = \max_{t \in T} \langle c, t \rangle \tag{19}$$

*More generally, the mixed norm as defined here is sublinear, from which the same conclusion follows – see Ref. 20, Chp.C.

for a special choice of the set T – see e.g. Ref. 21. This allows us to express $g(x)$ as,

$$g(x) = \max_{t \in T} \langle S\, x, t \rangle.$$

Regularization Term

The regularization term applies on the unknowns of the formulation, namely $\alpha_i(k)$. The particular regularization term we consider in this paper penalizes the differences '$\alpha_i(k) - \alpha_i(k-1)$'. To ease notation, let us define the time-varying vector

$$\alpha(k) = \begin{bmatrix} \alpha_1(k) \\ \vdots \\ \alpha_K(k) \end{bmatrix}. \tag{20}$$

Also, let D be an operator defined so that

$$d(k) = D\,\alpha(k) = \begin{bmatrix} \alpha_1(k) - \alpha_1(k-1) \\ \vdots \\ \alpha_K(k) - \alpha_K(k-1) \end{bmatrix}. \tag{21}$$

Note that D takes the time-difference of each component. Now let us similarly define

$$p(k) = \begin{bmatrix} p_1(k) \\ \vdots \\ p_K(k) \end{bmatrix}. \tag{22}$$

By constraining $p_i(k)$ to lie in different sets, the inner product of d and p, that is,

$$\big\langle d, p \big\rangle = \sum_{i=1}^{K} \sum_{k} d_i(k)\, p_i(k), \tag{23}$$

can be used to obtain different regularizers. Specifically, let P_T be the set of all time-valued vectors such that if $p \in P_T$, then $|p_i(k)| \leq 1$ for all i and k. Then,

$$\max_{p \in \lambda\, P_T} \big\langle d, p \big\rangle = \lambda \sum_{i=1}^{K} \mathrm{TV}(\alpha_i). \tag{24}$$

Another regularization term we will use is the sum of the ℓ_2 norms of $d(k)$. For this, let P_{ℓ_2} be the set defined as,

$$P_{\ell_2} = \left\{ p : \|p_i(k)\|_2 = \sqrt{\sum_k |p_i(k)|^2} \leq 1, \text{ for } i = 1, 2, \ldots, K \right\}. \tag{25}$$

Then, we define $\|\cdot\|$ so that

$$\max_{p \in \lambda\, P_{\ell_2}} \big\langle d, p \big\rangle = \lambda\, \|D\,\alpha(k)\|. \tag{26}$$

In the following, we will consider P_{ℓ_2} in the development. We note that the other case, where P_T is employed instead, can be similarly developed.

Combining the Terms

Now let, Y denote the operator that maps α_i's to $\sum_{i=1}^{K} y_i\, \alpha_i$. Using Y, the STFT operator S and the difference operator D, define the operator

$$L = \begin{bmatrix} S\,Y \\ D \end{bmatrix}. \tag{27}$$

Also, let us define the sets

$$A = \left\{ \alpha : \sum_{i=1}^{K} \alpha_i(k) = 1 \ \ \forall k, \ \ \alpha_i(k) \geq 0 \ \ \forall i, k \right\}, \tag{28}$$

$$Z = T \times \lambda P_{\ell_2} \tag{29}$$

and the vector

$$z = \begin{bmatrix} t \\ p \end{bmatrix} \tag{30}$$

Under these definitions, the formulation (13) becomes equivalent to

$$\min_{\alpha \in A} \max_{z \in Z} \langle L\,\alpha, z \rangle, \tag{31}$$

which is (14).

Recall that the algorithm requires the spectral norm of L in order to determine a convergent step size. Now note that, $\sigma(Y) = \max_{i,k} |y_i(k)|$ and $\sigma(D) = 2$. Assuming that the STFT is self-inverting (or that $S^T\,S = I$), we have $\sigma(L) \leq \sqrt{\sigma(Y^T\,Y) + \sigma(D^T\,D)}$. Therefore, it suffices to set

$$\gamma = \frac{1}{\sqrt{\left(\max_{i,k} |y_i(k)|\right)^2 + 4}} \tag{32}$$

to ensure that the algorithm converges.

4. MINIMIZATION ALGORITHM

Based on the definitions given so far, Algorihm 1 takes the following form.

Algorithm 2 The Algorithm for (13)

Initialize $\alpha^0, \bar{\alpha}^0, t^0, p^0$. Set $\gamma \leq 1/\sigma(L)$. ▷ Initialization
for $n \geq 0$ **do**
  for all k **do**
    $r(k) \leftarrow \sum_{i=1}^{N} y_i(k)\,\bar{\alpha}_i(k)$ ▷ $r \leftarrow Y\,\bar{\alpha}$
    $d(k) \leftarrow \bar{\alpha}(k) - \bar{\alpha}(k-1)$ ▷ $d \leftarrow D\,\bar{\alpha}$
  end for
  $t^{n+1} \leftarrow P_T(t^n + \gamma\,S\,r)$
  $p^{n+1} \leftarrow P_{\lambda\,P_{\ell_2}}(p^n + \gamma\,d)$
  $s \leftarrow S^T\,t$
  for all k **do**
    $v(k) \leftarrow p^{n+1}(k+1) - p^{n+1}(k)$ ▷ $v \leftarrow D^T\,p$
    for $i = 1 : K$ **do**
      $u_i(k) \leftarrow s(k)\,y_i(k)$ ▷ $u \leftarrow Y^T\,S^T\,t^n$
    end for
  end for
  $\alpha^{n+1} \leftarrow P_A\big(\alpha^n - \gamma\,s - \gamma\,u\big)$
  $\bar{\alpha}^{n+1} \leftarrow 2\alpha^{n+1} - \alpha^n$
end for

REMARK 4.1. See Ref.21 for a discussion on how to realize the projection operator $P_T(\cdot)$.

REMARK 4.2. The projection operator $P_{P_{\ell_2}}(t)$ applies on a time-varying vector, $t(k)$, with K entries $t_1(k), \ldots, t_K(k)$ and it projects each $t_i(k)$ to the nearest time-varying vector in the unit ℓ_2 ball.

REMARK 4.3. Projection onto the set A (defined in (28)) requires to project onto the unit K-dimensional simplex (where K is the number of observations) – for a fast method, see Ref. 22.

Convergence of this algorithm is ensured since it is a special case of a formally convergent algorithm. Nevertheless, it might be of interest to check whether the algorithm has reached a saddle point. We provide some conditions below.

4.1 Conditions of Convergence

Recall the generic saddle point problem

$$\min_{\alpha\in A}\max_{z\in Z}\langle L\,\alpha, z\rangle. \tag{33}$$

A point (α^*, z^*) such that $\alpha^* \in A$, $z^* \in Z$ is a solution of this problem if and only if

$$z^* \in \operatorname*{argmax}_{z\in Z}\langle L\,\alpha^*, z\rangle, \tag{34}$$

$$\alpha^* \in \operatorname*{argmin}_{\alpha\in A}\langle \alpha, L^T\,z^*\rangle. \tag{35}$$

It follows by the definitions so far and the properties of norms that (34) is equivalent to

$$t^* \in T, \quad \langle t^*, S\,Y\,\alpha^*\rangle = \|S\,Y\,\alpha^*\|_{2,1}, \tag{36a}$$

$$p^* \in \lambda\,P_{\ell_2}, \quad \langle p^*, D\,\alpha^*\rangle = \lambda\,\|D\,\alpha^*\|. \tag{36b}$$

Another alternative for (34) is

$$t^* = P_T\big(t^* + S\,Y\,\alpha^*\big), \tag{37a}$$

$$p^* = P_{\lambda\,P_{\ell_2}}\big(p^* + D\,\alpha^*\big). \tag{37b}$$

Let us now give an alternative expression for (35). For this, let $f = Y^T\,S^T\,t^* + D^T\,p^*$. Note that f is a time-varying vector where $f(k) = \begin{bmatrix} f_1(k) & \dots & f_K(k)\end{bmatrix}^T$. In this setting, (35) is equivalent to

$$\alpha^* = P_A\big(\alpha^* - f\big). \tag{38}$$

5. EXPERIMENTS

EXPERIMENT 5.1. Our first experiment is based on the noisy observations shown in Fig. 1. The underlying clean signal is an excerpt from a stringed instrument recording. The sampling frequency is 32 kHz and the signal duration is 3.125 sec (10^5 samples). The SNRs of the observations are 14.96 dB and 15.41 dB. The noise terms affect the signals in different instances in time so that the original signal lies in the set of (time-varying) convex combinations as described in the Introduction. In fact, the clean signal can be obtained easily by choosing $\alpha_1(k) = 1$, $\alpha_2(k) = 0$ when the noise term affects the second observation and $\alpha_1(k) = 0$, $\alpha_2(k) = 1$ when the noise term affects the first observation. Note that there are also regions where there is no noise affecting the signal. For those samples, the choise of $\alpha_i(k)$ does not matter – the sample value obtained by any convex combination is equal to the original value.

For the cost function $g(\cdot)$, we used mixed norms with a choice of group size as $M = 5$ (see Sec. 3.2). Recall that the cost function also involves an STFT operator. For that, we used a smooth window of length 1024 samples with a Hop-size of 256 samples. The window is chosen so that the resulting STFT is actually a Parseval frame. We note that this choice of the cost function (including the STFT) is employed in all of the experiments. Notice however that the sampling frequency for the last two experiments is different than the sampling frequency for the first two experiments.

With no regularization (as in (12), or (13) with $\lambda = 0$), we obtain a reconstruction with an SNR of 38.33 dB (see Fig. 3a). The samples of $\alpha_1(k)$ are shown in Fig. 3c. Note that due to the constraints of the formulation, we have $\alpha_2(k) = 1 - \alpha_1(k)$ for this example. Observe that the scheme fairly chose the correct observation most of the time. For instance, around $t = 1$, Observation-1 is noisy (see Fig. 1) and the scheme correctly chose small

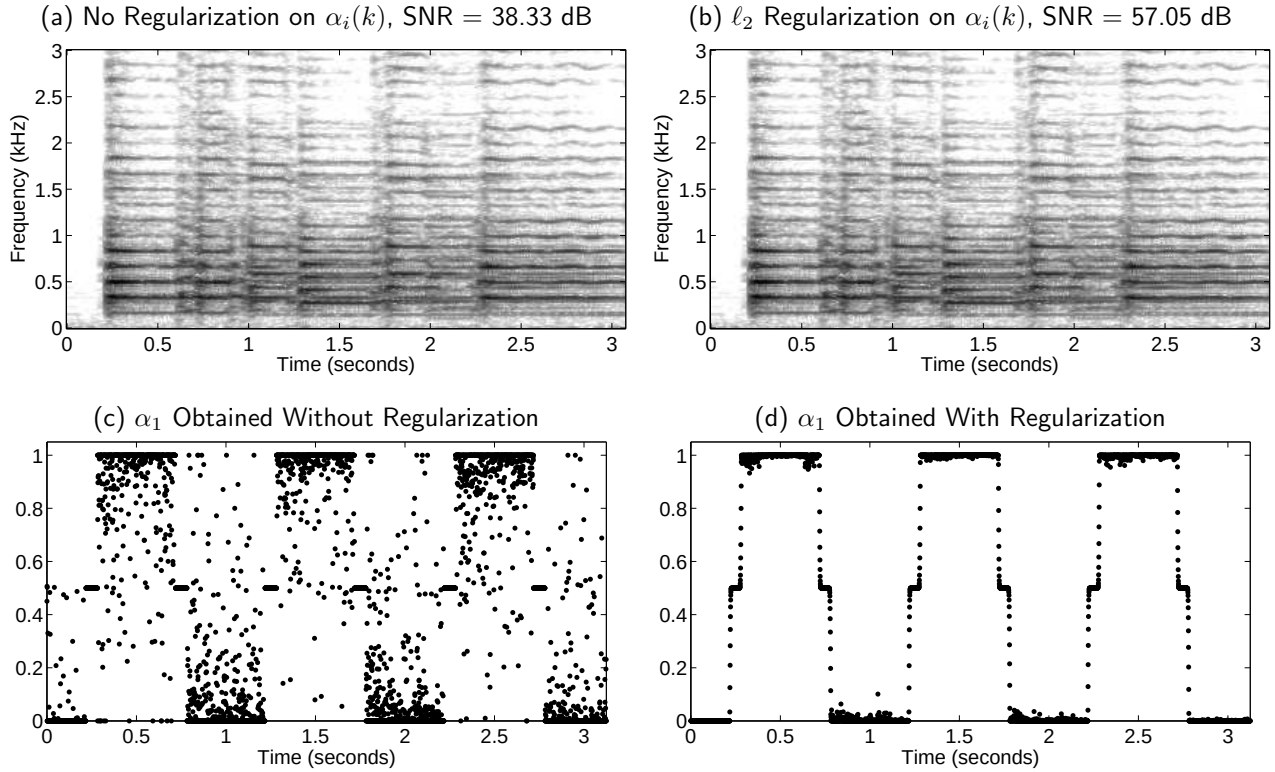

Figure 3. Left Column : Reconstruction with no regularization applied on $\alpha_i(k)$. Right column : Reconstruction where an ℓ_2 penalty term on α_i's is added for regularization. Top Row : Spectrograms of the reconstructions. Bottom Row : $\alpha_1(k)$. Note that $\alpha_2(k) = 1 - \alpha_1(k)$ for this example. From the bottom row, we see that even with no regularization, the scheme is fairly successful in choosing the correct observation. When the regularization term is added, the correct observation is almost always selected by the formulation, without further knowledge about the signal.

values for $\alpha_1(k)$, suppressing the weight of this noisy observation. However, we also observe that $\alpha_1(k)$ varies a lot as a function of k.

The formulation in (13), that also includes a regularization term, gives a reconstruction much closer to the original signal (SNR = 57.05 dB – see Fig. 3b). The samples of $\alpha_1(k)$ are shown in Fig. 3d. The scheme, for almost all samples, succesfully chose the clean observation. Notice that the only prior information supplied to the formulation is pertaining to the slow time-variation of the desired $\alpha_i(k)$'s; no specific information about the clean signal is provided. The prior knowledge about the signal of interest is implicitly encoded in the choice of the cost function $g(\cdot)$.

EXPERIMENT 5.2. We use the same clean signal as in Exp. 5.1. However, this time, we produce three observations, with a more complicated noise variance pattern. The noise standard deviations for the three observations are shown in Fig 4, along with the spectrograms of the observations. Notice that the observations have different SNRs.

The spectrogram of the reconstruction obtained by solving the regularized formulation is shown in Fig. 5a. The reconstruction achieves an SNR of 26.48 dB. Also, Fig. 6 compares the obtained $\alpha_i(k)$'s with the 'ideal' $\alpha_i(k)$'s defined as,

$$\alpha_i(k) = \frac{\sigma_i(k)^{-2}}{\sigma_1(k)^{-2} + \sigma_2(k)^{-2} + \sigma_3(k)^{-2}}, \quad \text{for } i = 1, 2, 3. \tag{39}$$

Note that this expression is obtained by setting Θ^{-2} to zero in (10). We think that the proposed formulation has been successful in handling this case, although no prior noise information was provided. In fact, the proximity of the ideal and obtained $\alpha_i(k)$'s suggest that we can work (10) backwards; that is, we can make use of the

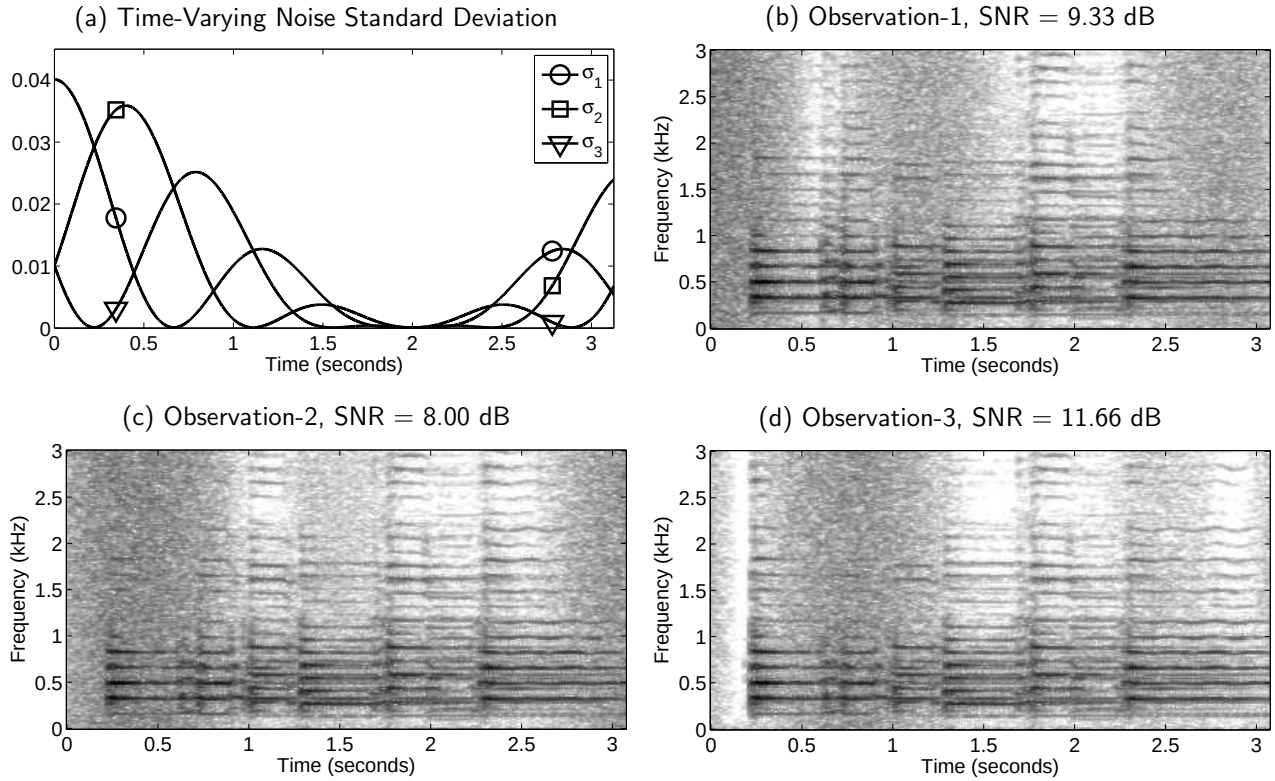

Figure 4. Experiment 5.2 involves three observations contaminated with different time-varying noise terms. Top Left: Noise standard deviation for the three observations. The rest of the figures show the spectrograms of the observations.

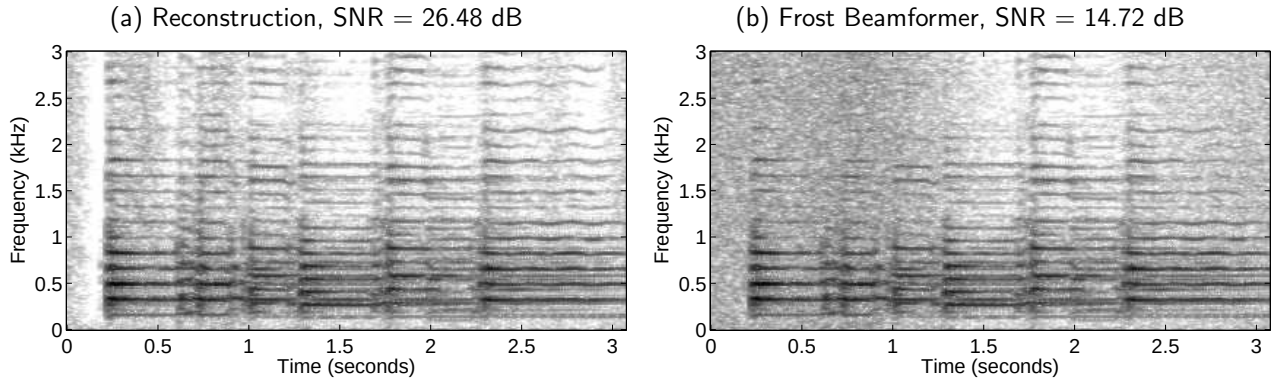

Figure 5. Reconstruction results for Experiment 5.2.(a) Spectrogram of the reconstruction achieved by the formulation using the observations shown in Fig. 4. (b) The reconstruction obtained by the Frost beamformer.

relation in α_i's and σ_i's in (10) to obtain estimates of the relative noise level. Exp. 5.4 provides a more controlled experiment on this point.

In order to compare the formulation with a well-known method, we employed the Frost beamformer[5] – see the Introduction for a brief explanation of this beamformer. In the Frost beamformer, we let the adaptive filters for each observation have four taps. The spectrogram of the reconstruction is shown in Fig. 5b. For this example, the Frost beamformer performed slightly better than a simple average. The proposed formulation leads to a reconstruction that is better than that provided by the Frost beamformer, both in terms of SNR and perceptual quality. We note that the proposed formulation, like the Frost beamformer, does not require intricate prior information regarding the signal. However, in its current form, the proposed algorithm runs offline whereas the Frost beamformer can be implemented as an online adaptive algorithm.

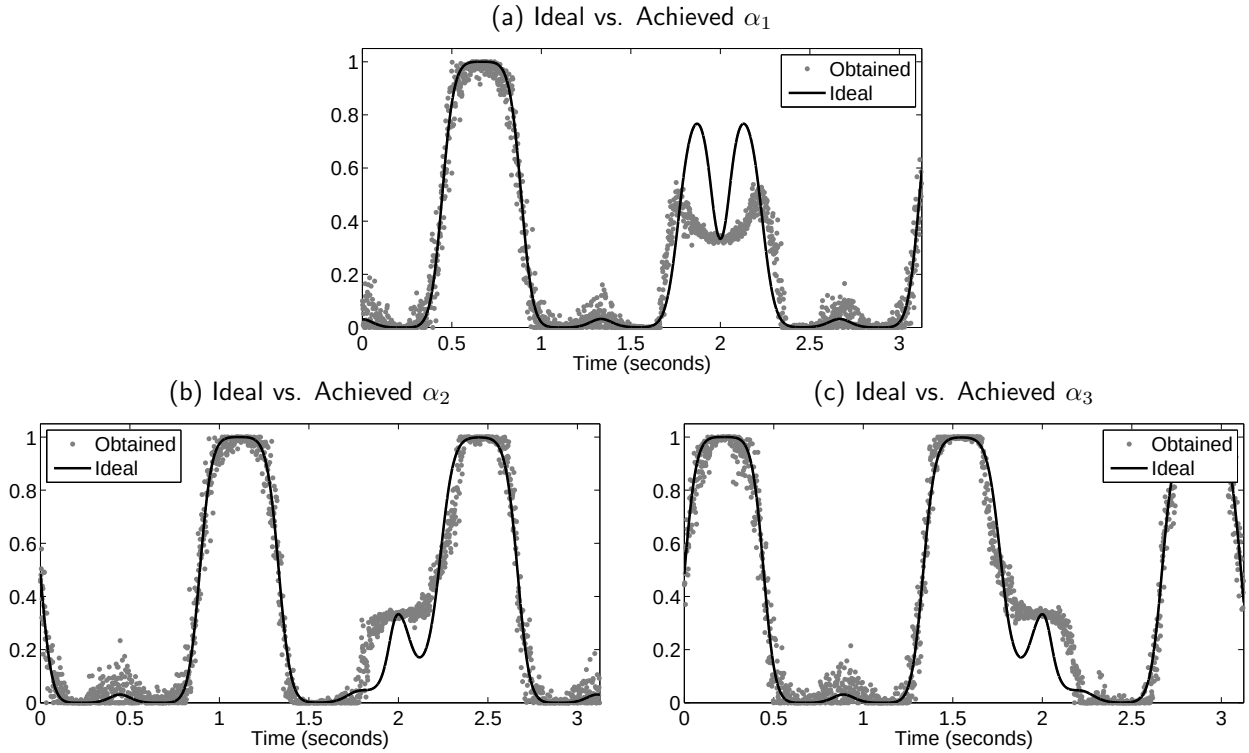

Figure 6. The 'ideal' α_i's obtained by setting Θ^{-2} to zero in (10) vs. α_i's obtained by the proposed algorithm. Note that to obtain the 'ideal' curves, we need to know the noise variances of each observation. This information is not supplied to the proposed reconstruction algorithm.

EXPERIMENT 5.3. In this experiment, we use a speech signal to test the capabilities of the formulation for a more complicated input. The noise variance patterns are similar to those used in Experiment 5.1. The spectrograms of the two noisy observations are shown in the left panel of Fig. 7.

The reconstruction obtained using the proposed formulation with and without regularization are also shown in Fig. 7. For reconstruction, we used the same parameters as in Exp. 5.1. Note that the difference between including or not including a regularization term on $\alpha_i(k)$'s is similar to that in Exp. 5.1. However, the improvement is less pronounced. In particular, observe that the variation of $\alpha_i(k)$'s in Fig. 8b is higher than those in Fig. 3d. We think this might be due to the nature of the input signal. The stringed instrument has a cleaner spectrum than speech and better fits the model implied by the mixed norm. Nevertheless, if we increase the weight of the regularization parameter, the variations of $\alpha_i(k)$'s decrease.

EXPERIMENT 5.4. In this experiment, we use the same clean signal as in Experiment 5.3. But this time, we use stationary noise processes. Our goal is to see whether the formulation is able to consistently recover the 'best' convex combination in this setting.

The (constant) noise variance for Observations-1,2,3 are $\sigma_1^2 = 10^{-4}$, $\sigma_2^2 = \sigma_1^2/2$, $\sigma_3^2 = \sigma_1^2/3$ respectively. The SNRs of the observations are 14.47 dB, 17.48 dB, 19.33 dB. Note that if we set

$$\hat{\alpha}_i(k) = \frac{\sigma_i^{-2}}{\sigma_1^{-2} + \sigma_2^{-2} + \sigma_3^{-2}}, \quad \text{for } i = 1, 2, 3, \tag{40}$$

then,

$$\hat{\alpha}_1(k) = \frac{1}{6} \approx 0.167, \quad \hat{\alpha}_2(k) = \frac{2}{6} \approx 0.333, \quad \hat{\alpha}_3(k) = \frac{3}{6}. \tag{41}$$

If we use the same weight for the regularization parameter λ as in Exp. 5.3, the reconstruction has an SNR of 22.83 dB. $\alpha_i(k)$'s used for this reconstruction are shown in Fig. 9a,b,c. Ideally, we would like to see

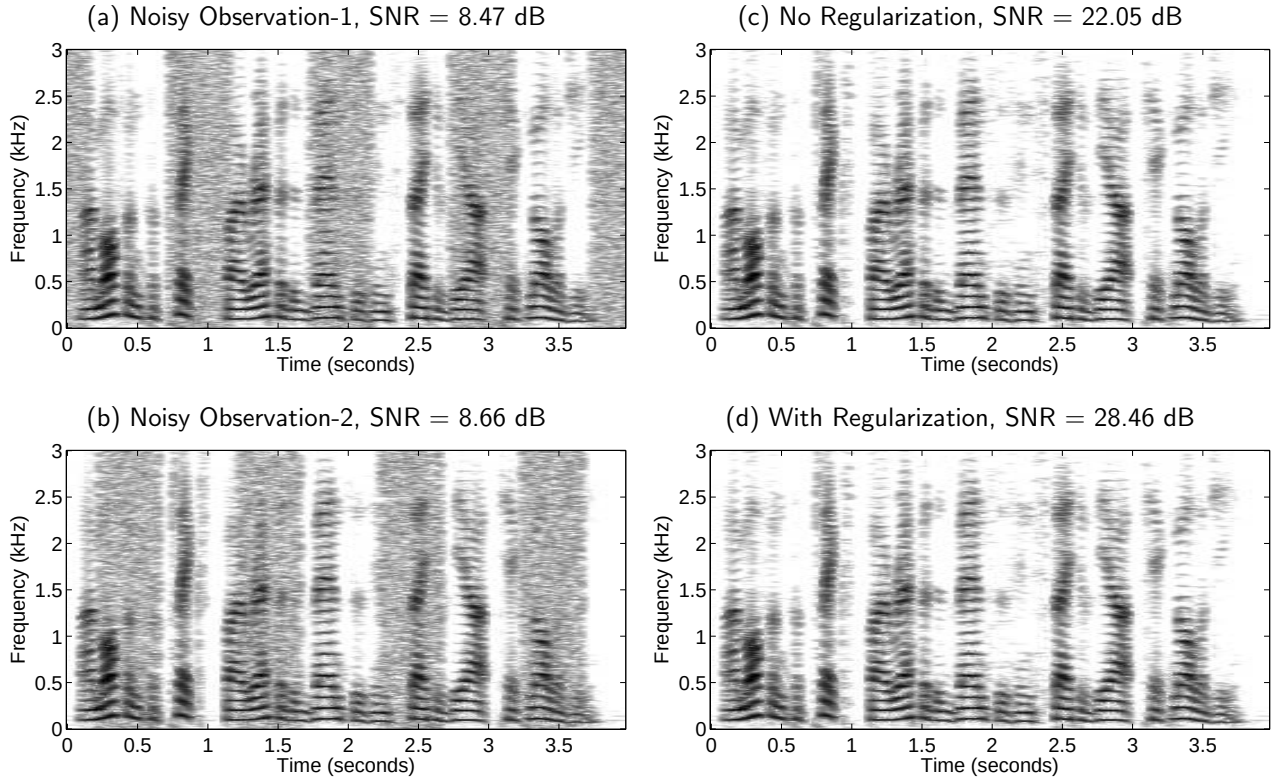

Figure 7. The spectrograms of the signals from Experiment 5.3. The clean signal is a speech signal. Similar noise patterns as in Experiment 5.1 are used to produce the two observations – at each instant, at least one of the observations contains no noise. The clean signal and the observations are shown on the left panel. The right panel shows (c) the reconstruction obtained without a regularization term, (d) with an ℓ_2 regularization term. (f) The reconstruction obtained with the Frost beamformer. Note that both regularized and non-regularized formulations proposed in this paper achieve a better reconstruction than the Frost beamformer.

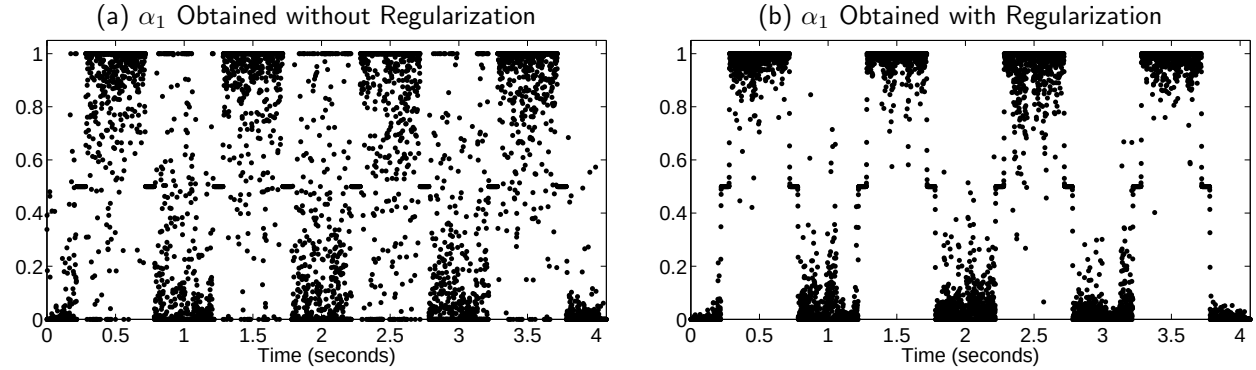

Figure 8. $\alpha_1(k)$ for Experiment 5.3. Note that $\alpha_2(k) = 1 - \alpha_1(k)$. (a) $\alpha_1(k)$ obtained where no regularization term is included in the problem formulation. Note that the main tendency is correct. (b) $\alpha_1(k)$ obtained with an added ℓ_2 regularization term. The same weight as in Experiment 5.1 is used for weighting the regularization term. Observe that adding the regularization term helped improve the selection of α_i's.

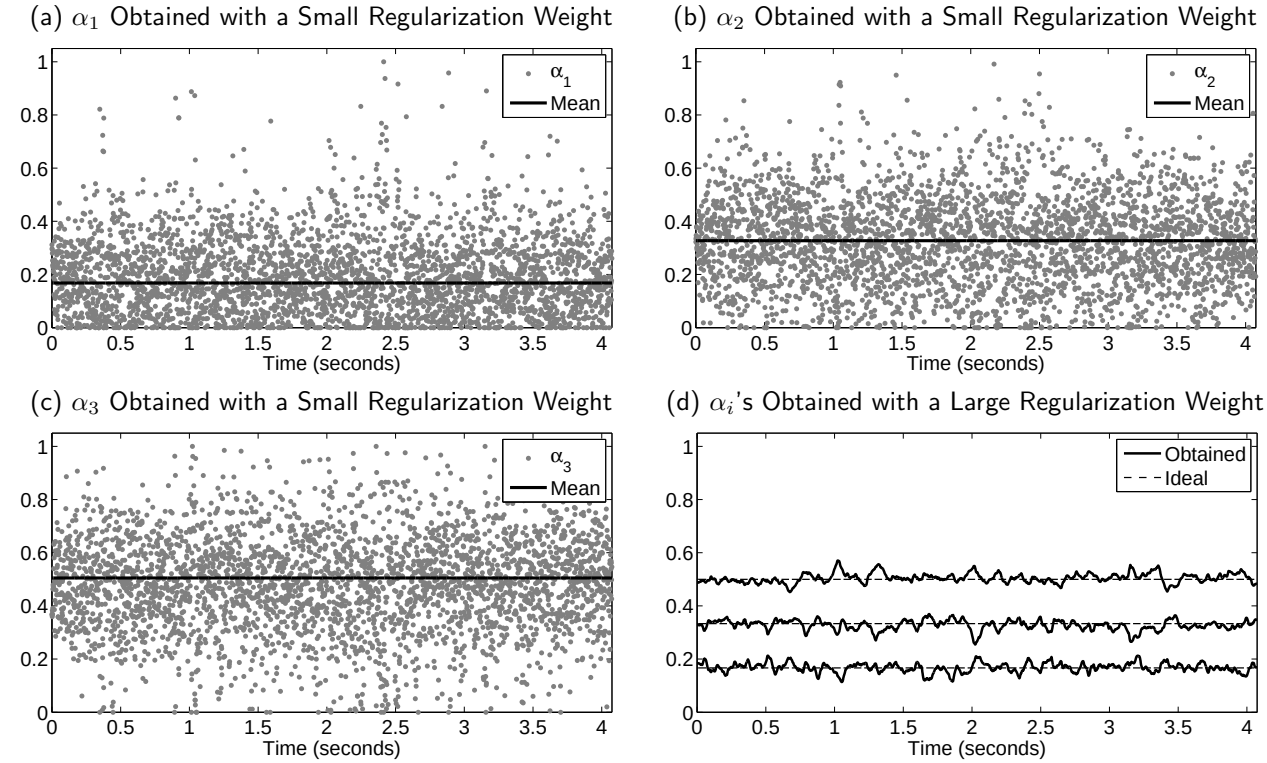

Figure 9. $\alpha_i(k)$'s for Experiment 5.4. (a,b,c) Each figure shows three different $\alpha_i(k)$'s, obtained with no regularization and their average values – see the text for the values. (d) If we increase the weight of the regularization term to $10 \times \lambda$, the formulation chooses $\alpha_i(k)$'s that are closer to being constant, and also to $\hat{\alpha}_i(k)$'s (indicated by dashed lines).

constant $\alpha_i(k)$'s since noise is stationary for each observation. Although this is not the case, we observe that the reconstruction puts more weight to the less noisy observations on average. In fact, when we look at the averages of $\alpha_i(k)$'s, we find that $\text{mean}\big(\alpha_1(k)\big) = 0.191$, $\text{mean}\big(\alpha_2(k)\big) = 0.325$, $\text{mean}\big(\alpha_3(k)\big) = 0.484$. Notice that these values are close those in (41). These average values are depicted in Fig. 9a,b,c with solid lines.

If we increase the weight of the regularization term to $10 \times \lambda$, we obtain $\alpha_i(k)$'s as shown in Fig. 9d. Observe that these $\alpha_i(k)$'s have a much lower variation. Moerover, we find that the means of $\alpha_i(k)$ are closer to $\hat{\alpha}_i(k)$'s. Specifically, $\text{mean}\big(\alpha_1(k)\big) = 0.168$, $\text{mean}\big(\alpha_2(k)\big) = 0.327$, $\text{mean}\big(\alpha_3(k)\big) = 0.505$. The SNR of the reconstruction is 22.32 dB. Although the SNR has not improved, the distribution of $\alpha_i(k)$'s are closer to our expectations.

6. CONCLUSION

We considered the problem of combining audio signals obtained using multiple sensors, when time-varying noise terms with different behaviors contaminate the observations. In such a setting, if the characteristics of the noise terms are known, it is feasible to employ this knowledge by including 'data terms' in a cost function, that penalize deviations from the observations according to the noise level. However, lacking knowledge on noise characteristics, we showed that a formulation of the problem as a convex minimization problem (without data terms) also leads to good performance.

The cost function employed in the formulation requires the selection of an STFT operator and a 'group size' for mixed norms.[10] Once the cost function is selected, the only parameter of the algorithm is the weight of the regularization term. Even with a weight of zero, we demonstrated that the algorithm can achieve a reasonable reconstruction. However, including such regularization, especially if it is known that noise terms have slowly-varying characteristics, usually gives a better reconstruction.

The formulation assumes that recorded observations are available and is meant to reconstruct the source offline. However, the scheme may also be used as a base to derive real-time algorithms. One issue that might

arise in practice is that the data from the sensors might not be aligned, i.e. the observations might involve delayed versions of the source, with different delays. In such a case, the algorithm requires the application of a prior time-synchronization step (as in Ref. 23).

REFERENCES

1. Van Veen, B. and Buckley, K., "Beamforming techniques for spatial filtering," in [*Digital Signal Processing Handbook*], CRC, Boca Raton (1997).
2. Gannot, S. and Cohen, I., "Adaptive beamforming and postfiltering," in [*Handbook of Speech Processing*], Springer (2008).
3. Johnson, D. H. and Dudgeon, D. E., [*Array signal processing*], Prentice Hall (1993).
4. Lacoss, R. T., "Adaptive combining of wideband array data for optimal reception," *IEEE Trans. Geoscience Electronics* **6**, 78–86 (May 1968).
5. Frost III, O. L., "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE* **60**, 926–935 (Aug. 1972).
6. Griffiths, L. J. and Jim, C. W., "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propagation* **30**, 27–34 (Jan. 1982).
7. Cox, H., Zeskind, R. M., and Owen, M. M., "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, and Signal Proc.* **35**, 1365–1376 (Oct. 1987).
8. Parra, L. C. and Alvino, C. V., "Geometric source separation : Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech and Audio Processing* **10**, 352–362 (Sept. 2002).
9. Bayram, I. and Kamasak, M. E., "A simple prior for audio signals," *IEEE Trans. Audio, Speech and Language Processing* **21**, 1190–1200 (June 2013).
10. Kowalski, M., "Sparse regression using mixed norms," *J. of Appl. and Comp. Harm. Analysis* **27**, 303–324 (Nov. 2009).
11. Kowalski, M. and Torrésani, B., "Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients," *Signal, Image and Video Processing* **3**(3), 251–264 (2009).
12. Siedenburg, K. and Dörfler, M., "Structured sparsity for audio signals," in [*Proc. Int. Conf. on Digital Audio Effects (DAFx)*], (2011).
13. Christensen, M. G. and Sturm, B. L., "A perceptually reweighted mixed-norm method for sparse approximation of audio signals," in [*Proc. Asilomar Conf. on Signals, Systems and Computers*], (2011).
14. Bayram, I., "Mixed-norms with overlapping groups as signal priors," in [*Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*], (2011).
15. Popov, L. D., "A modification of the Arrow-Hurwicz method for search of saddle points," *Matematicheskie Zametki* **28**, 777–784 (Nov 1980).
16. Korpelevich, G., "The extragradient method for finding saddle points and other problems," *Ekonomika i Matematcheskie Metody* **12**, 747–756 (1976).
17. Pock, T., Cremers, D., Bischof, H., and Chambolle, A., "An algorithm for minimizing the Mumford-Shah functional," in [*Proc. IEEE Int. Conf. on Computer Vision*], (2009).
18. Chambolle, A. and Pock, T., "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision* **40**, 120–145 (May 2011).
19. Esser, E., Zhang, X., and Chan, T. F., "A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science," *SIAM J. Imaging Sciences* **3**, 1015–1046 (Nov. 2010).
20. Hiriart-Urruty, J.-B. and Lemaréchal, C., [*Fundamentals of Convex Analysis*], Springer (2004).
21. Bayram, I. and Akyildiz, O. D., "Primal-dual algorithms for audio decomposition using mixed norms," *Signal Image and Video Processing* (2013).
22. Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T., "Efficient projections onto the ℓ_1 ball for learning in high dimensions," in [*Proc.* 25^{th} *Int. Conf. on Machine Learning*], (2008).
23. Valin, J.-M., Rouat, J., and Michaud, F., "Enhanced robot audition based on microphone array source separation with post-filter," in [*Proc. IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*], (2004).