A Convex Model and l_1 Minimization for Musical Noise Reduction in Blind Source Separation

Wenye Ma, Meng Yu, Jack Xin, and Stanley Osher

Abstract

Musical noise often arises in the outputs of time-frequency domain binary mask based blind source separation methods. Post-processing is desired to enhance the separation quality. An efficient musical noise reduction method is presented, based on a convex model of time-domain sparse filters. The sparse filters are computed by l_1 regularization and the split Bregman method. The proposed musical noise reduction method is used as a post-processing tool for binary mask or non-binary mask based blind source separation methods. Evaluations by both synthetic and room recorded speech and music data show that the method outperforms existing musical noise reduction methods in terms of objective and subjective measures.

Keywords: blind source separation, time-frequency domain, musical noise, convex model, time-domain sparse filters, l_1 minimization, split Bregman method.

AMS Subject Classification: 90C25, 65K10, 68T05.

The authors M. Yu and W. Ma contributed equally to this work. M. Yu and J. Xin were partially supported by NSF DMS-0911277 and DMS-0712881; W. Ma and S. Osher were partially supported by NSF DMS-0914561, NIH G54RR021813 and the Department of Defense.

M. Yu and J. Xin are with the Department of Mathematics, University of California, Irvine, CA, 92697 USA (e-mail: myu3@uci.edu; jxin@math.uci.edu).

W. Ma and S. Osher are with Department of Mathematics, University of California, Los Angeles, CA, 90095, USA (e-mail: mawenye@math.ucla.edu; sjo@math.ucla.edu).

I. Introduction

Sound signals in daily auditory scenes often appear as mixtures when multiple speakers or sound sources are active. It is of both fundamental and practical interest to recover the sound source signals from the received mixtures with minimal information of the environment, mimicing what human ears can do by paying attention to a selected speaker. Blind source separation (BSS) methods aim to achieve this goal, based on some a-priori knowledge of the source signal properties. Following the physics of sound mixing, let us consider N sources $s_n(t)$, $n = 1, \cdots, N$, to be convolutively mixed. At J sensors, the recorded mixture signals $x_j(t)$, $j = 1, \cdots, J$, are :

$$x_j(t) = \sum_{n=1}^{N} \sum_{d=0}^{D} h_{jn}(d)\, s_n(t-d), \tag{1.1}$$

where D is the delay length on the order of 10^3–10^4 in a standard room, $h_{jn}(d)$ is the discrete Green's function of the room, also known as the room impulse response (RIR), from source n to receiver j. The mathematical problem is to recover both $h_{jn}(d)$ and $s_n(t)$ from $x_j(t)$ which is severely ill-posed.

A major branch of BSS is the so called independent component analysis (ICA) which assumes that the source signals are orthogonal to (or independent of) each other [7]. ICA is a more general methodology than recovering sound signals. The time domain ICA attempts to estimate h_{jn}'s directly and has to deal with a high dimensional noncovex optimization problem ([7], [12]). Frequency domain ICA solves an instantaneous ($D = 0$) version of (1.1) in each frequency bin after applying the discrete Fourier transform (DFT) to (1.1) frame by frame:

$$X_j(f,\tau) \approx \sum_{n=1}^{N} H_{jn}(f)\, S_n(f,\tau), \tag{1.2}$$

where (X_j, H_{jn}, S_n) are T-point DFT of (x_j, h_{jn}, s_n) respectively, τ is the frame number. The larger T/D is, the better the approximation. Due to the absence of periodicity in d of h_{jn} and s_n, DFT does not transform convolution to local product exactly. The frequency domain approach is limited to using a long DFT, in addition to computations to sort out scaling and permutation ambiguities when synthesizing multi-frequency estimation of $S_n(f,\tau)$ back to a time domain output ([7], [11]). Imperfections and errors in scaling and permutation in the frequency domain may lead to artifacts in the time domain signals at the final output.

The time-frequency (TF) approaches have been developed ([18], [1] among others) more recently. It is based on the working assumption that $S_n(f,\tau)$ and $S_{n'}(f,\tau)$ ($n \neq n'$) are relatively sparse or have almost no overlap in (f,τ) domain. The non-overlap assumption is satisfied quite well by clean speech signals, though is found to deteriorate in reverberant room (a regular room with reflecting surfaces) conditions [4]. It follows from (1.2) and the non-overlap assumption that:

$$X_j(f,\tau) \approx H_{jk}(f)\, S_k(f,\tau), \tag{1.3}$$

where $k \in [1, N]$ is such that S_k is the dominant source at (f,τ). The source signals can be classified by clustering on TF features. In the two receiver case (similar to two ears), a common feature vector is:

$$\Theta(f,\tau) = \left[\frac{|X_2(f,\tau)|}{|X_1(f,\tau)|}, \frac{1}{2\pi f} \text{angle}(X_2(f,\tau)/X_1(f,\tau)) \right], \tag{1.4}$$

which are amplitude ratio and normalized phase difference (phase delay) at each point (f,τ). The angle ranges in $(-\pi, \pi]$. In view of (1.3), $X_2(f,\tau)/X_1(f,\tau) \approx H_{2k}(f)/H_{1k}(f)$, so the feature vector Θ reflects the Fourier transform of RIRs from the dominant source k. The success of the method relies on the formation of clusters in the histogram of the feature vector. The number of clusters is the number of identified source signals, see Fig. 1 for an illustration of two peaks in the Θ histogram with input data

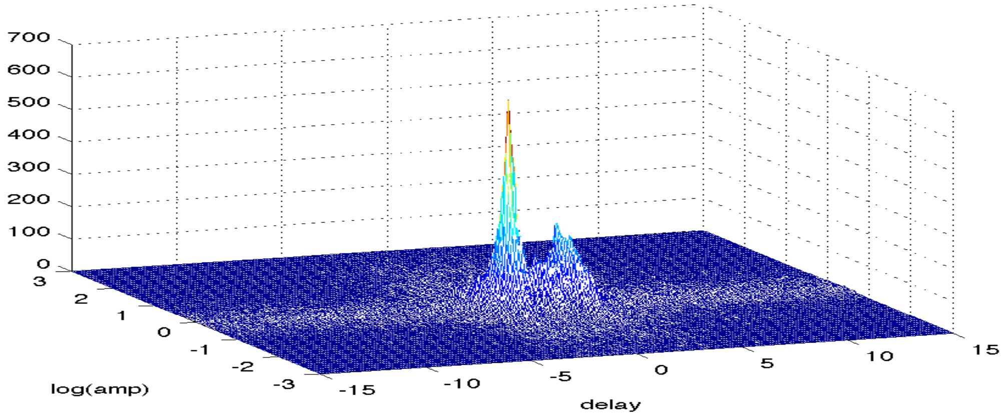

Fig. 1. *Histogram of Θ feature points (amplitude ratio and phase delay) of 2 mixtures of 2 speech signals, showing 2 distinct peaks.*

being a mixture of two speech signals. Each TF point (f, t) whose Θ belongs to cluster C_k (by comparing distances from $\Theta(f,\tau)$ to the cluster centroids) is regarded as occupied by the Fourier spectrum of the k-th source signal. One can then define a binary mask (BM) function:

$$M_k(f,\tau) = \begin{cases} 1 & \Theta(f,\tau) \in C_k \\ 0 & \text{otherwise.} \end{cases} \tag{1.5}$$

An estimation of the k-th source in TF domain is:

$$\tilde{S}_k(f,\tau) = M_k(f,\tau) X_1(f,\tau), \tag{1.6}$$

where X_1 may be replaced by X_2 as another choice. Finally, taking inverse DFT (iDFT) yields the estimate of $s_k(t)$. The method is robust in the sense that more than two source signals may be recovered from two receivers.

However some remarks are in order. First, the phase of the estimated signal in (1.6) is same as that of the mixture signal. While the amplitude of the dominant k-th source is a good approximation of the mixture signal at those points in C_k, it is not clear that the phase of the k-th signal is close to that of the mixture signal. Phase errors exist in (1.6). Second, the angle function in (1.4) can cause aliasing errors if the phase of $H_{2k}(f)/H_{1k}(f)$ goes out of $(-\pi,\pi]$. For example if $H_{2k}(f)/H_{1k}(f) = \exp\{i\,\phi_k\, f\}$, with $|\phi_k\ f| > \pi$, then the angle part of Θ is equal to the remainder of $\phi_k\ f$ modulo π, missing the true value $\phi_k\ f$ and causing artifacts in clustering and classification. Here d_k represents a typical delay of the dominant source in the model (1.1). This restriction translates into an upper limit of a few centimeters on the interdistance of the two receivers, and is recently relaxed [17] by a technique of oversampled Fourier transform and modular arithmetics. Third, the binary mask function M_k makes a zero or one (winner-take-all) decision in the TF domain, which easily leads to nonlinear nonlocal distortions perceived as ringing sounds (musical noise [3]) in the time domain. Fourth, the non-overlap working assumption is violated to various degrees when music signals are in the sources or when the number of source signals increases.

A few methods were proposed recently ([2], [3]) to suppress musical noise. The main ingredients of these methods are: (1) employing the overlap-add method for reconstructing the waveform outputs from estimated spectra of source signals; (2) using a finer shift of window function while taking short time Fourier transform (STFT); (3) adopting non-binary masks. One choice for (3) is based on the so called sigmoid function, namely M_k is defined by $\mathcal{M}_k(f,\tau) = 1/[1+\exp(g(d_k(f,\tau)-\theta_k))]$, here θ_k and g are shape parameters, $d_k(f,\tau)$ is the distance between cluster members and their centroids. The other choice for (3) comes from Bayesian inference. The mask function is a conditional probability function $\mathcal{M}_k(f,\tau) = P(C_k|\boldsymbol{X}(f,\tau))$ where C_k is the k-th cluster and $\boldsymbol{X}(f,\tau)$ the mixture spectrogram (absolute value of the DFT vector as a function of frequency and frame number). In short, the above noise reduction methods relied on either a gradual change of the Fourier spectra or non-binary masks to increase smoothness of processing in the TF domain.

In this paper, we introduce a simple and efficient *time domain method* to suppress musical noise like artifacts in the output of binary mask based TF domain BSS. Our method can be also used as a postprocessing tool for removing artifacts in any other frequency domain based processing. The idea is to formulate a convex optimization problem for seeking sparse filters to cancel the interference and re-estimate the source signals in the time domain. As a result, we effectively reduced errors in phase aliasing and the discontinuous masking operations of the initial TF mask based method. The sparse filters are computed by l_1 norm regularization and the split Bregman method for which fast convergence was recently studied [10]. The paper is organized as follows. In section II, we propose a way to modify the mask function to reduce fuzzy points in the feature space that lie in almost equal distances to two cluster centroids. This treatment reduced clustering errors and extended the TF binary mask based BSS [1] in the regime where the microphone spacing exceeds the effective range of [18] and phase aliasing errors occur. In section III, a convex musical noise suppression model is introduced based on a convex optimization problem with l_1 norm regularization. In section IV, the computational framework by the split Bregman method is shown. In section V, evaluations of the proposed method demonstrate its merits in comparison with existing methods. Even in the case of large and unknown microphone spacing, the proposed masking and musical noise reduction method enhances the recovered speech and music signals significantly. The concluding remarks are in section VI.

II. Initial Source Estimation

The initial sound separation is carried out by the TF domain binary mask method [18] as described in the introduction with K-means algorithm for clustering. We shall however propose some improvements towards the accurate estimation of the feature parameters with less restriction on the receiver interdistances. Because the single source dominance assumption at each TF point may not be valid with the increase of source number N or reverberation time (convolution length D in model (1.1)), we introduce a stricter criterion below for clustering accuracy. At each TF point (f,τ), the confidence coefficient of $(f,\tau) \in C_k$ is defined by $CC(f,\tau) = \frac{d_k}{\min_{j\neq k} d_j}$, where d_j is the distance between $\Theta(f,\tau)$ and the centroid of the j-th peak. The new mask function is defined for some $\rho > 0$ as

$$\mathcal{M}_k(f,\tau) = \begin{cases} 1 & (f,\tau) \in C_k \ \text{ and } \ CC(f,\tau) \leq \rho \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

The motivation for the refined mask (2.1) is to reduce the number of fuzzy feature points which have nearly equal distances to at least two cluster centers. The refined mask function (2.1) applies to the situation where the unknown receiver spacing is not small enough and phase aliasing errors are present [18], [1]. Similar to the TF binary mask BSS method of [1], we adopt the amplitude only feature $\mathbf{\Theta}(f,\tau) = \left[\frac{|X_1(f,\tau)|}{|X(f,\tau)|}, ..., \frac{|X_M(f,\tau)|}{|X(f,\tau)|}\right]$, where $|X(f,\tau)|$ is a normalization factor and M is the number of receivers (sensors). Such phase free feature vector, though robust to receiver inter-distances and free of phase aliasing errors, is found to less discriminative and produce lower quality separations [1]. Our modified mask function (2.1) helps to compensate for this loss of separation quality, and sets a better stage for the subsequent time-domain noise reduction and quality enhancement of the recovered source signals.

Besides the above binary mask based BSS methods, non-binary masks such as sigmoid function based mask and Bayesian inference based mask discussed in section I were also implemented as initial step of separation for our proposed musical noise reduction method. Further reduction of musical noise is observed for these methods as well after our post-processing.

III. Musical Noise Reduction Model

Let us first consider the determined case of mixing model with 2 sensors and 2 sources ($N = J = 2$). The output signals of the TF domain mask based BSS are denoted as $y_k(t)$, $k = 1, 2$. The mixing model (1.1) can be abbreviated as $x_j(t) = \sum_{k=1}^{2} h_{jk} * s_k$, where the star denotes linear convolution (the inner sum of (1.1)). The following algebraic identities hold:

$$\begin{aligned} h_{22} * x_1(t) - h_{12} * x_2(t) &= (h_{22} * h_{11} - h_{12} * h_{21}) * s_1(t), \\ h_{21} * x_1(t) - h_{11} * x_2(t) &= (h_{21} * h_{12} - h_{11} * h_{22}) * s_2(t). \end{aligned} \tag{3.1}$$

The identities are also known as cross-channel cancellation for blind channel identification in communication theory [15]. Now the modeling idea is to replace the convolutions of source signals on the right hand side of (3.1) by the initial separations y_1 and y_2 respectively. We then seek a pair of filters u_{jk}, $j, k = 1, 2$, such that

$$u_{1k} * x_1 - u_{2k} * x_2 \approx y_k. \tag{3.2}$$

In general, y_1 or y_2 may differ from $(h_{22} * h_{11} - h_{12} * h_{21}) * s_1(t)$ or $(h_{21} * h_{12} - h_{11} * h_{22}) * s_2(t)$ by a convolution [12]. Identities (3.1) imply a family of solutions to (3.2) of the form: $u_{11} = g_1 * h_{22}$ ($u_{12} = g_2 * h_{21}$) and $u_{21} = g_1 * h_{12}$ ($u_{22} = g_2 * h_{11}$), where g_1 and g_2 are a pair of unknown filters. In other words, the solutions u_{jk} may differ from the room impulse responses (RIRs, or the h_{jk}'s) by a convolution g_k. The optimal choice of g_1 (g_2) is the so called de-reverberation filter which minimizes the length or support of $g_1 * h_{12}$ ($g_2 * h_{11}$) and $g_1 * h_{22}$ ($g_2 * h_{21}$). Without knowledge of RIRs (h_{jk}'s) however, we shall use l_1 norm regularization of u_{1k} and u_{2k} to achieve this goal indirectly as follows.

Let us consider a duration D of $y_k(t)$, and seek a pair of sparse filters u_{jk}, $j, k = 1, 2$, to minimize the energy (l_2 norm) of $u_{1k} * x_1 - u_{2k} * x_2 - y_k$ subject to l_1-norm regularization. The l_2 norm comes from the Gaussian fit of the unknown noise (mismatch) distribution. The resulting convex optimization problem for $t \in D$ is:

$$(u_{1k}^*, u_{2k}^*) = \arg \min_{(u_{1k}, u_{2k})} \frac{1}{2} ||u_{1k} * x_1 - u_{2k} * x_2 - y_k||_2^2 + \mu(||u_{1k}||_1 + ||u_{2k}||_1). \tag{3.3}$$

Let us denote the length of signal in D as L_D and the length of filter solution as L. With u_{jk}'s l_1-regularized, we aim to recover minimal-length (minimal-support) solutions of (3.3). The l_1 regularization helps to fully resolve the major spikes in u_{jk}'s, or the early arrival part of RIR's, so that the filter solutions are stable and robust up to leading peaks under reverberant and noisy conditions. In matrix form, the convex objective (3.3) becomes:

$$u_k^* = \arg \min_{u_k} \frac{1}{2} ||Au_k - y_k||_2^2 + \mu||u_k||_1 \tag{3.4}$$

where u_k is formed by stacking up u_{1k} and u_{2k}, and $L_D \times 2L$ matrix A is (T is transpose):

$$A = \begin{pmatrix} x_1(1) & x_1(2) & ... & ... & x_1(L_D-1) & x_1(L_D) \\ & x_1(1) & ... & ... & x_1(L_D-2) & x_1(L_D-1) \\ & & \ddots & & & \vdots \\ & & & x_1(1) & ... & x_1(L_D-L+1) \\ -x_2(1) & -x_2(2) & ... & ... & -x_2(L_D-1) & -x_2(L_D) \\ & -x_2(1) & ... & ... & -x_2(L_D-2) & -x_2(L_D-1) \\ & & \ddots & & & \vdots \\ & & & -x_2(1) & ... & -x_2(L_D-L+1) \end{pmatrix}^T$$

Once u_{1k}^* and u_{2k}^* are found, we compute $u_{1k}^* * x_1 - u_{2k}^* * x_2$ for a better approximation of s_k with muscial noise reduced. If the acoustic environment does not change much, the estimation during $t \in D$ still applies when $t \notin D$. Otherwise, an adaptive estimation can be repeated at a later time interval. The objective (3.3) takes the same form as that in image denoising [10].

The above derivation generalizes to M sensors and N sources ($M \geq 3$ and $N = M$). We approximate y_k by a linear combination of the mixtures x_j, $j = 1, 2, ..., M$. When $t \in D$, for a proper value of $\mu > 0$, we minimize:

$$(u^*_{jk}) = \arg\min_{u_{jk}} \frac{1}{2}||\sum_{j=1}^{M} u_{jk} * x_j - y_k||_2^2 + \mu \sum_{j=1}^{M} ||u_{jk}||_1 \tag{3.5}$$

and estimate s_k by $\hat{s}_k = \sum_{j=1}^{M} u^*_{jk} * x_j$. Though two sensors are enough for mask based BSS methods, the remaining $M - 2$ sensors are also used here for reducing the musical noise.

IV. Minimization by Bregman Method

In this section, we adapt the split Bregman method and apply it to the musical noise reduction model (3.4) in reverberant conditions such as in a normal room with acoustic reflections. The split Bregman method was introduced by Goldstein and Osher [10] for solving l_1, total variation, and related regularized problems. It has connections to Lagrangian-based alternating direction methods in convex optimization [8]. The split Bregman method aims to solve the unconstrained problem:

$$\min_u J(\Phi u) + H(u), \tag{4.1}$$

where J is convex but not necessarily differentiable, H is convex and differentiable, and Φ is a linear operator. The general split Bregman iteration with initial values $d^0 = 0$, $u^0 = 0$, $b^0 = 0$, is:

$$d^{k+1} = \arg\min_d \frac{1}{\lambda} J(d) - \langle b^k, d - d^k \rangle + \frac{1}{2}||d - \Phi u^k||_2^2 \tag{4.2}$$

$$u^{k+1} = \arg\min_u \frac{1}{\lambda} H(u) + \langle b^k, \Phi(u - u^k) \rangle + \frac{1}{2}||d^{k+1} - \Phi u||_2^2 \tag{4.3}$$

$$b^{k+1} = b^k - (d^{k+1} - \Phi u^{k+1}) \tag{4.4}$$

where λ is a positive constant, and $\langle \cdot, \cdot \rangle$ is regular inner product.

If J is the l_1 norm, the subproblem (4.2) has explicit solutions. The subproblem (4.3) is also easy to solve since the objective is differentiable. Convergence of the split Bregman method for the case of $J(u) = \mu||u||_1$ was analyzed [6], and the result is:

Theorem IV.1. *Assume that there exists at least one solution u^* of (4.1). Then we have the following properties for the split Bregman iterations (4.2),(4.3), and (4.4):*

$$\lim_{k \to \infty} \mu||\Phi u^k||_1 + H(u^k) = \mu||\Phi u^*||_1 + H(u^*)$$

Furthermore,

$$\lim_{k \to \infty} ||u^k - u^*||_2 = 0$$

if u^* is the unique solution.

Now we implement the split Bregman method on our proposed musical noise reduction model. Let $J(u) = \mu||u||_1$, $\Phi = I$, and $H(u) = \frac{1}{2}||Au - f||_2^2$. Setting $d^0 = 0$, $u^0 = 0$, and $b^0 = 0$, we have the iterations:

$$d^{k+1} = \arg\min_d \frac{\mu}{\lambda}||d||_1 - \langle b^k, d - d^k \rangle + \frac{1}{2}||d - u^k||_2^2 \tag{4.5}$$

$$u^{k+1} = \arg\min_u \frac{1}{2\lambda}||Au - f||_2^2 + \langle b^k, u - u^k \rangle + \frac{1}{2}||d^{k+1} - u||_2^2 \tag{4.6}$$

$$b^{k+1} = b^k - (d^{k+1} - u^{k+1}) \tag{4.7}$$

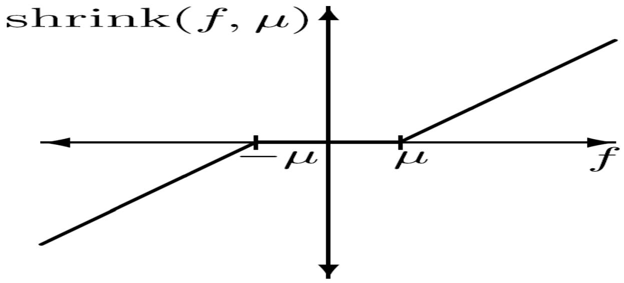

Fig. 2. *Illustration of the one dimensional shrink operator in the* ***while*** *loop of the algorithm in section IV.*

Explicitly solving (4.5) and (4.6) gives the simple algorithm

```
Initialize $u^0 = 0, d^0 = 0, b^0 = 0$
While $||u^{k+1} - u^k||_2/||u^{k+1}||_2 > \epsilon$
  (1) $d^{k+1} = \text{shrink}(u^k + b^k, \frac{\mu}{\lambda})$
  (2) $u^{k+1} = (\lambda I + A^T A)^{-1}(A^T f + \lambda(d^{k+1} - b^k))$
  (3) $b^{k+1} = b^k - d^{k+1} + u^{k+1}$
end While
```

Here shrink is the soft threshold function defined by $\text{shrink}(v, t) = (\tau_t(v_1), \tau_t(v_2), \cdots, \tau_t(v_n))$ for $v = (v_1, v_2, \cdots, v_n) \in \mathbb{R}^n$ and $t > 0$, where $\tau_t(x) = \text{sign}(x)\max\{|x| - t, 0\}$ see Fig. 2 for a one dimensional plot. Noting that the matrix A is fixed, we can precalculate $(\lambda I + A^T A)^{-1}$, then the iterations only involve matrix multiplication.

Unlike previous applications of Bregman methods to under-determined problems in compressed sensing, here A is an m by n matrix with $m \gg n$ (over-determined). The complexity of calculating $(\lambda I + A^T A)^{-1}$ is $O(mn^2) + O(n^3) = O(mn^2)$. The complexity of each iteration is $O(n^2)$. The Forward-Backward splitting method [8] is also a candidate for this problem. It does not involve matrix inversion. But the complexity of each iteration is $O(mn)$. We can accelerate it by precalculating $A^T A$ and $A^T f$ to reduce the complexity in each iteration to $O(n^2)$, where $A^T A$ has complexity $O(mn^2)$. However, the Forward-Backward splitting method usually needs more iterations to converge than the split Bregman method. We tested various cases and found that the convergence time of the split Bregman method is less than that of the Forward-Backward splitting method by about 40%. Our entire algorithm is summarized as follows:

Input: Observed mixture signals, x_j, $j = 1, ..., M \geq 2$.
Output: Estimated sources with musical noise suppressed, $\hat{s}_k$, $k = 1, ..., N$ $(N = M)$.
Initial separation: Extract signals y_k, $k = 1, ..., N$ by TF mask approaches with a proper ρ.
Filter estimation: Apply the **split Bregman** method to obtain the filters u^*_{jk}, $j = 1, ..., M$ for each source k, according to (3.5).
Musical Noise Suppression: $\hat{s}_k = \sum_{j=1}^{M} u^*_{jk} * x_j$.

V. Evaluation and Comparison

The parameters for the proposed method are chosen as $\mu = \epsilon = 10^{-3}$, $\eta = 1$, $\lambda = 2\mu$, $L_D = 30000$, and $L = 1000$. So matrix A is 30000×2000, and $A^T A$ is 2000×2000. As suggested in [3], the STFT frame size is 512 and frame shift is $512/8$. For simplicity, we denote by BM1 the so called DUET method of

[18], and by BM2 the extended binary mask BSS method of [1] with the modified feature $\Theta(f,\tau)$ in section II.

The performance measure [16] is calculated in two steps, provided that the true source signals and sensor noises are known. The first step is to decompose by orthogonal projection an estimate $\hat{s}(t)$ of a source $s(t)$ into a sum of four terms:

$$\hat{s}(t) = s_{target}(t) + e_{interf}(t) + e_{noise}(t) + e_{artif}(t) \tag{5.1}$$

where $s_{target}(t)$ is an allowed deformation of the target source $s(t)$, $e_{interf}(t)$ is an allowed deformation of the interferring (unwanted) sources, $e_{noise}(t)$ for sensor noises, and $e_{artif}(t)$ for artifacts of the separation algorithms such as musical noise or other forbidden distortions of the sources. The second step is to compute performance criteria on the decibel (dB) scale as follows ([16], [9]).

- The Signal to Distortion Ratio (SDR)

$$SDR \triangleq 10\log_{10}\frac{||s_{target}||_2^2}{||e_{interf} + e_{noise} + e_{artif}||_2^2} \tag{5.2}$$

- The Signal to Interferences Ratio (SIR)

$$SIR \triangleq 10\log_{10}\frac{||s_{target}||_2^2}{||e_{interf}||_2^2} \tag{5.3}$$

- The Signal to Artifacts Ratio (SAR)

$$SAR \triangleq 10\log_{10}\frac{||s_{target} + e_{interf} + e_{noise}||_2^2}{||e_{artif}||_2^2} \tag{5.4}$$

Besides these objective measures, the average Perceptual Evaluation of Speech Quality (PESQ,[14]) score was computed as a measure of performance. This measure was designed to estimate the subjective quality of speech. The output is an estimate of the Mean Opinion Score (MOS), a number between 1 and 5. The meanings of the scores in relation to speech quality are: 1-Bad, 2-Poor, 3-Fair, 4-Good and 5-Excellent.

To test the musical noise reduction portion of our method, synthetic mixture data are used to recover a source signal where energy loss due to binary mask is simulated. The masked signal plays the role of BSS output y_k in section III. Measured binaural RIRs ($h_{jk},\ j,k = 1,2$) are used to generate mixtures x_1 and x_2. For the spectrogram of $h_{11} * s_1$ (or absolute value of $S_{11} = STFT(h_{11} * s_1)$), a mask $\mathcal{M}$ of the same size as S_{11} is defined. The mask is multiplied entry by entry to S_{11} to produce a distorted waveform signal $s_d = iSTFT(S_{11} \circ \mathcal{M})$, where $\circ$ is entrywise product. We recover $h_{11} * s_1$ from the two mixture signals x_1, x_2 and s_d (in place of y_1) with the Bregman iterations in section IV. Fig. 3 shows the spectrogram of the source signal $h_{11} * s_1$ (left), the spectrogram of the distorted signal s_d (middle) and the spectrogram of the recovered signal (right) in some time frames. The test is repeated under different reverberation times: anechoic, 150 milliseconds (ms), and 580 ms. Though a little interference from s_2 is introduced, i.e. a little decrease of SIR, the gain in SDR [9] is found to be significant in the low input SDR regime (Fig. 4). This phenomenon is observed in processing room recorded data as well.

Comparison of several musical noise suppression methods is carried out on room recorded data. The set-up is shown in Fig. 5. In case of 2 sources, their locations are at S_1 and S_2 in Fig. 5, and the sensors Mic_1 and Mic_2 provide data for separation and noise suppression. In case of 4 sources, all the loudspeakers and microphones contribute to the musical noise reduction but only Mic_2 and Mic_3 are used for separation. Table I and II list results of different musical noise suppression methods discussed in section I. Compared with BM1, sigmoid mask and Bayesian mask methods, our method leads in the overall quality PESQ [14], and with a significant margin in SDR and SAR. The SIR improvement is however not uniformly better. In the case of 4 sources, SIR improvement lags the other methods. When the number of sources increases, ρ in the mask (2.1) should increase accordingly to control the growth of zero-paddings.

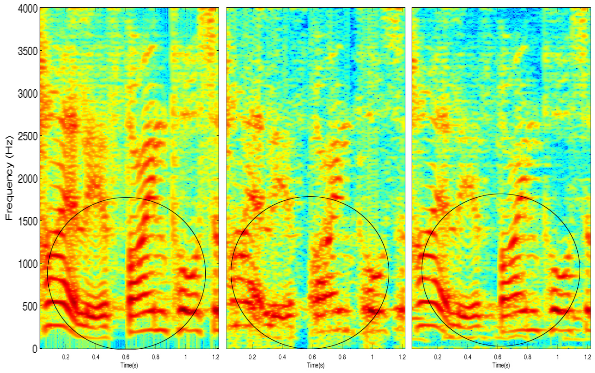

Fig. 3. *Spectrograms of a source signal (left), the distorted source signal s_d(middle) and the recovered source signal (right). From the left to the middle spectrogram,* 80% *of the energy is masked out. The reverberation time is 150 ms and the input SIR* ≈ -5.9 *dB (decibel). Patterns inside the circles illustrate the improvement by the proposed method.*

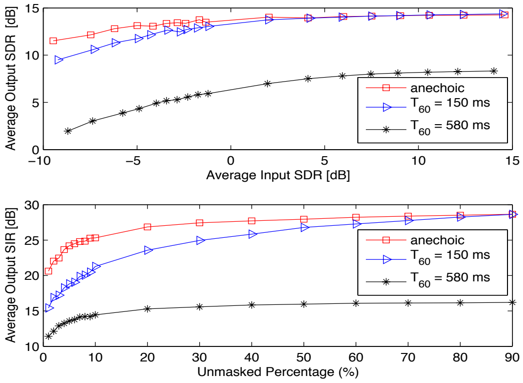

Fig. 4. *Upper panel: average signal to distortion (SDR) ratios of input and output signals in synthetic test. Lower panel: signal to interference ratio (SIR) vs. unmasked percentage (percentage of 1's) in the mask. The data points in the upper panel have the same unmasked percentage as those in the lower panel.*

Next we remove Mic_1 and Mic_4 from the set-up of Fig. 5, so only 2 microphones Mic_2 and Mic_3 are active. The unknown microphone spacing between Mic_2 and Mic_3 is reset to $[15, 20]$ cm outside the effective range of binary mask BSS methods ([18], [1]). The azimuths of the two loudspeakers (emitting speech and music signals sampled at 8000Hz and of 5 second duration) are changed to $0°$ and $60°$. We continue to use the refined mask (2.1) with a nearly optimal value of $\rho = 0.5$ based BM2, sigmoid mask and Bayesian mask respectively as the initial separation for our method. As discussed in section II, since BM2 may not work well, eliminating fuzzy feature points by a proper value of ρ helps to gain a good SIR but sacrifice the signal quality. However, as seen in Table III, the overall quality is improved significantly by both the Bayesian mask and our method without losing SIR.

Furthermore, we conduct a subjective test on ten listeners with normal hearing to evaluate the reduction of musical noise. The paired comparison test requires each listener to rank the four methods according to the performance of musical noise reduction in the groups of experiments conducted in Tables I, II and III. The percentage of our method's preference over three other methods in musical noise reduction is shown in Table IV. Since the initially estimated music sources contain more musical noise, the contrasts between these methods on the music channel are more pronounced.

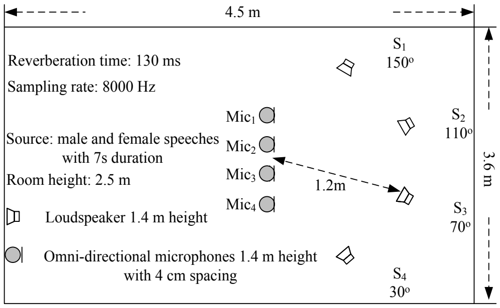
4.5 m
Reverberation time: 130 ms
Sampling rate: 8000 Hz
Source: male and female speeches with 7s duration
Room height: 2.5 m
Loudspeaker 1.4 m height
Omni-directional microphones 1.4 m height with 4 cm spacing
Mic1
Mic2
Mic3
Mic4
1.2m
S1 150°
S2 110°
S3 70°
S4 30°
3.6 m

Fig. 5. *Configuration and parameters of the room recording.*

TABLE I

Comparison of musical noise reduction methods on room recorded speech data. Average evaluation results are shown for 2 sources case. BM1 with conventional mask (1.5); SM (Sigmoid mask); BYM (Bayesian mask). The initial separation for our method employs BM1 with refined mask (2.1) (where $\rho = 0.50, 0.25, 0.10, 0.05$*), SM and BYM respectively (where* $\rho = 1$*).*

Method	PESQ	SIR	SDR	SAR
Input	1.37	0.04	0.02	46.48
BM1[18]	2.24	13.24	6.44	9.37
SM[2]	2.17	11.38	6.52	9.14
BYM[2]	2.33	13.30	7.20	10.20
BM1+Ours-0.50	2.18	9.47	8.58	17.74
BM1+Ours-0.25	2.21	10.07	9.26	17.97
BM1+Ours-0.10	2.22	10.18	9.51	18.94
BM1+Ours-0.05	**2.40**	**13.41**	**12.18**	**19.05**
SM+Ours	2.35	10.34	9.35	17.97
BYM+Ours	2.34	11.25	9.86	18.04

VI. Conclusions

We proposed and evaluated an efficient time domain method for reducing musical noise in the output of TF mask based BSS methods. By a more selective TF mask, we reduced percentage of fuzzy points on TF domain to improve separation quality. We employed fast Bregman iterations to minimize a convex l_1 norm regularized objective to compute sparse time-domain filters for musical noise reduction. The time domain filters effectively reduced musical noise and enhanced the overall quality of the recovered music and speech signals in terms of both objective and subjective measures.

VII. Acknowledgments

The authors would like to thank Dr. Ernie Esser for many valuable discussions on convex optimization methods, and the anonymous referees for their helpful and constructive comments.

TABLE II

Comparison of musical noise reduction methods on room recorded speech data. Average evaluation results are shown for 4 sources case. BM1 with conventional mask (1.5); SM (Sigmoid mask); BYM (Bayesian mask). The initial separation for our method employs BM1 with refined mask (2.1) (where $\rho = 0.50, 0.25, 0.10, 0.05$), SM and BYM respectively (where $\rho = 1$).

Method	PESQ	SIR	SDR	SAR
Input	1.10	-4.49	-4.51	26.54
BM1[18]	1.89	9.39	3.79	6.57
SM[2]	1.71	8.22	2.21	5.40
BYM[2]	1.83	8.21	3.34	6.65
BM1+Ours-0.50	1.90	6.30	5.85	19.06
BM1+Ours-0.25	**1.91**	**6.35**	**5.89**	**18.93**
BM1+Ours-0.10	1.84	5.63	5.23	18.76
BM1+Ours-0.05	1.75	5.36	4.79	16.86
SM+Ours	1.82	4.94	4.31	15.38
BYM+Ours	1.82	4.76	4.36	17.10

TABLE III

Comparison of musical noise reduction methods on speech/music mixtures with unknown large microphone spacing. Refined mask (2.1) with $\rho = 0.5$, sigmoid mask and Bayesian mask are employed respectively.

Method	PESQ	SIR	SDR	SAR
Input	1.50	1.90	1.85	33.16
BM2	1.63	16.87	3.58	4.01
SM [2]	2.07	22.10	8.86	9.10
BYM[2]	2.52	16.54	11.66	14.50
BM2+Ours	2.45	16.52	12.81	16.32
SM+Ours	2.33	16.66	11.48	17.76
BYM+Ours	2.60	15.42	11.73	20.32

References

[1] S. Araki, H. Sawada, R. Mukai and S. Makino,"Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors." Signal Processing, 87, 1833-1847, 2007.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Blind sparse source separation with spatially smoothed time-frequency masking", in Proc. Int. Workshop on Acoustic Echo and Noise Control, 2006.

[3] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask", in Proc. ICASSP, vol. III, pp. 81-84, 2005.

[4] S. Araki, H. Sawada, S. Makino, "K-means Based Underdetermined Blind Speech Separation", Chapter 9 in "Blind Speech Separation", S. Makino, T-W. Lee, H. Sawada, eds, Springer 2007.

[5] L. Bregman, "The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming", *USSR Comput Math and Math. Phys.*, v7: 200-217, 1967.

[6] J. Cai, S. Osher and Z. Shen, "Split Bregman Methods and Frame Based Image Restoration", *Multiscale Model. Simul.* 8(2):337–369, 2009.

[7] S. Choi, A. Cichocki, H. Park, and S. Lee, *Blind source separation and independent component analysis: A review*, Neural Inform. Process. Lett. Rev., 6 (2005), pp. 1–57.

[8] E. Esser, "Applications of Lagrangian-Based Alternating Direction Methods and Connections to Split Bregman", CAM report, 09-31, UCLA, 2009.

[9] C. Fevotte, R. Gribonval, and E. Vincent, "BSS_EVAL toolbox user guide – Revision 2.0", Tech. Rep. 1706, IRISA, Rennes, France, April 2005.

[10] T. Goldstein and S. Osher, "The split Bregman algorithm for l_1 regularized problems", SIAM J. Imaging Sci. 2:323-343, 2009.

[11] J. Liu, J. Xin, Y. Qi, "A Dynamic Algorithm for Blind Separation of Convolutive Sound Mixtures", Neurocomputing, 72(2008), pp 521-532.

[12] J. Liu, J. Xin, Y. Qi, F-G Zeng, "A time domain algorithm for blind separation of convolutive sound mixtures and l_1 constrained minimization of cross correlations", Comm Math Sciences, vol. 7, No.1, 109-128, 2009.

[13] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, "An iterative regularization method for total variation based image restoration", *SIAM Multiscale Model. and Simu.*, 4:460-489, 2005.

[14] PESQ, ITU-T Rec. P. 862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs", International Telecommunication Union, Geneva, 2001.

TABLE IV

Subjective evaluation on musical noise reduction. Notation $>$ ($<$) means the output of our method is perceived with less (more) musical noise, while $\approx$ means "hard to distinguish". Binary Mask is BM1 (BM2) for Table I and II (III).

Method	Test Category		$>$	$\approx$	$<$
Ours vs Binary mask	Table I	2 sources	55%	35%	10%
	Table II	4 sources	55%	35%	10%
	Table III	Speech	94%	4%	2%
	Table III	Music	99%	-	1%
Ours vs Sigmoid mask[2]	Table I	2 sources	55%	40%	5%
	Table II	4 sources	78%	15%	7%
	Table III	Speech	25%	64%	11%
	Table III	Music	76%	3%	21%
Ours vs Bayesian mask[2]	Table I	2 sources	35%	40%	25%
	Table II	4 sources	50%	43%	7%
	Table III	Speech	13%	66%	21%
	Table III	Music	74%	3%	23%

[15] L. Tong, G. Xu, T. Kailath, “Blind identification and equalization based on second order statistics: A time domain approach”, IEEE Information Theory, 40(2):340-349, 1994.

[16] E. Vincent, R. Gribonval, C. Fvotte, “Performance measurement in blind audio source separation”, IEEE Transactions on Audio, Speech, Language Processing 14(4): 1462-1469, 2006.

[17] Y. Wang, O. Yilmaz, Z. Zhou, “A Novel Phase Aliasing Correction Method for Robust Blind Source Separation Using DUET”, preprint, 2010.

[18] O. Yilmaz and S.Rickard, “Blind separation of speech mixtures via time-frequency masking”, IEEE Trans. Signal Processing, vol. 52, no. 7, 1830-1847, 2004.

[19] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. ”Bregman iterative algorithms for compressed sensing and related problems”, SIAM J. Imaging Sciences 1(1):143-168, 2008.