A Convex Speech Extraction Model and Fast Computation by the Split Bregman Method

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

***Abstract*—A fast speech extraction (FSE) method is presented using convex optimization made possible by pause detection of the speech sources. Sparse unmixing filters are sought by l_1 norm regularization and the split Bregman method. A subdivided split Bregman method is developed for efficiently estimating long reverberations in real room recordings. The speech pause detection is based on a binary mask source separation method. The FSE method is evaluated objectively and subjectively, and found to outperform existing blind speech separation approaches on both synthetic and room recorded data in terms of the overall computational speed and separation quality.**

***Index Terms*—convexity, sparse filters, split Bregman method, fast blind speech extraction.**

I. Introduction

BLIND speech separation (BSS) aims to recover source signals from their mixtures without detailed knowledge of the mixing process [1]. However, it remains a challenge to retrieve sound sources recorded in real-world environment such as in cluttered rooms. The physical reason is that sound reflections (reverberations) in enclosed rooms cause signal mixing at current time to depend on source signals and their long delays (history dependent). Mathematically, the mixing process is convolutive in time and the unknowns are high dimensional. Various efforts have been made to separate convolutive mixtures. Three major approaches are: time-domain BSS, frequency domain BSS, and time-frequency (TF) domain BSS.

Time domain BSS is based on optimizing a cost function (measuring entropy or non-Gaussianity) for time domain signals, for example independent component analysis (ICA). The approach is theoretically reasonable, and achieves a good separation if the optimization can be done accurately as is the case for mixtures with minimal time delay (low reverberation). However at the fundamental level, all time domain methods based on ICA attempt to optimize **non-convex** objectives, for which no global convergence is mathematically guaranteed. This weakness poses a difficult problem for actual convergence and robustness of approximation in real-world settings where high dimensional (on the order of thousands) optimization under measurement noise is encountered. The lack of robustness under perturbations may be explained by potentially many local minima of a non-convex objective where approximating sequences can get stuck in. Even if local convergence of optimizing sequence occurs, it may be computationally expensive. For example, the time domain scaled natural gradient method [1] is typically time consuming in the regime of long reverberations. Moreover, small divisors and divergence may occur in silent durations of mixture signals, in other words, the method is not stable in the presence of a small piece of silent duration. Though a nonlocally weighted soft constrained natural gradient method [2] resolves such issues and renders the method asymptotically consistent, convergence is still rather slow.

In frequency domain BSS, the observed time domain signals are converted into frequency domain time series signals by the short-time (windowed) Fourier transform (STFT). The convolutive mixture can be approximated with multiple instantaneous mixtures (no time delay), each of which is defined in a frequency bin. The approximation is however only valid if the window size is much larger than the length of delay. Under such condition, one can then employ any instantaneous BSS algorithm to separate the mixtures bin by bin. However, the permutation and scaling ambiguity of a BSS solution turns into a serious problem in reconstructing time domain output. The order of the output in each frequency bin must be determined correctly so that the separated frequency components that originate from the same source are grouped together before taking inverse STFT. Large window size, permutation and scaling issues limit the effectiveness of frequency domain BSS in reverberant conditions. In contrast, the time-frequency domain methods ([3], [4]) by spectral data clustering are both simple and efficient partly because they do not resolve room impulse responses. The basic working assumption is that at most one source signal is dominant at each time-frequency point of the mixture spectrogram. In other words, the Fourier spectra of the source signals rarely overlap in time. Such a non-overlapping property in the TF domain deteriorates however in reverberant conditions ([3], [4]), causing clustering errors and musical noise in the output.

In this paper, a novel fast time domain speech extraction (FSE) method is proposed based on the assumption that intelligible speech signals contain pauses. Pause detection is a problem of independent interest, which we handle here by processing the output of a modified TF domain clustering method. Because we only detect silence durations from the initial separation, tolerance of artifacts in TF domain clustering is higher. During silent durations of the target speech signal, information of the interference (background) is collected and

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).

allows us to formulate a **convex** optimization problem for finding part of the impulse response functions which suffice to estimate the target speech. A sparse solution is then computed by l_1 norm regularization and the split Bregman method for which fast convergence was recently studied [5]. The proposed time domain approach is free from the permutation problem in frequency domain BSS and relaxes the TF domain non-overlap hypothesis. It also does not rely on speech data statistics, and so enjoys high efficiency in data usage and computation.

This paper is organized as follows. In section II, the convex optimization problem for FSE is introduced. In section III, computational framework by l_1 norm regularization is shown. The Bregman method and the split Bregman method are explained with algorithmic schemes and convergence proofs. In subsections III-C and III-D, algorithms for moderately and highly reverberant acoustic environments are illustrated. The subdivided split Bregman method is proposed for FSE with long reverberations and large number of sources. In section IV, an onset-offset detection method of speech is outlined. In section V, the length of selected silent speech duration is studied, and the comparison between the split Bregman method and the subdivided split Bregman method is illustrated under different lengths of the filters. Evaluations of FSE show its merits in both speed and separation quality in comparison with existing methods. Discussion and conclusions are in section VI. Our method also applies to non-speech signal extraction from convolutive mixtures as long as pause detection of target signal is possible.

II. Fast Speech Extraction Model

Let us consider two sensors and two sound sources which can be either two speech signals or one speech signal and one non-speech background interference (music or other ambient noises). FSE method shall sequentially extract speech signals if there are more than one speech sources. Let us denote one of the two sources as the target speech signal s_T, and the other one as background interference s_B. The mixing model is

$$x_i(t) = h_{i1} * s_B(t) + h_{i2} * s_T(t) \tag{1}$$

where t is time; $i = 1, 2$; and $*$ is linear convolution. Instead of finding an unmixing filter W such that $W * (x_1, x_2)$ recovers (s_T, s_B), we extract speech signal s_T by eliminating (not recovering) interference s_B. Suppose that the target speech contains pauses. Then there is a union D of disjoint time intervals where $s_T \approx 0$, while interference s_B is active. It follows from (1) that $h_{21} * x_1(t) - h_{11} * x_2(t) \approx 0$ for $t \in D$. The elimination by cross multiplication was known in blind channel identification [6] and background suppression [7]. Inside D, we seek a pair of sparse filters u_i $(i = 1, 2)$ to minimize the energy of $u_2 * x_1 - u_1 * x_2$ in the region D. Ideally, $u_1 \approx h_{11}$ and $u_2 \approx h_{21}$, that is the solutions are expected to be a pair of sparse acoustic room impulse response (RIR). The sparse RIR model is theoretically sound [8], and has been shown useful for estimating RIRs in real acoustic environments [9]. Filter sparseness is achieved by l_1-norm regularization. The resulting convex optimization problem for $t \in D$ is:

$$(u_1^*, u_2^*) = \arg\min_{(u_1,u_2)} \frac{1}{2}||u_2 * x_1 - u_1 * x_2||_2^2 + \frac{\eta^2}{2}(\sum_{i=1}^{2} u_i(1) - 1)^2 + \mu(||u_1||_1 + ||u_2||_1) \tag{2}$$

where the second term is to fix scaling and prevent zero (trivial) solution. Denote the length of D by L_D and that of u_i by L. D can be as short as even 0.25 s' duration, which makes FSE method efficient on the data usage and different from other BSS methods that are based on the high order statistics of data. In matrix form, convex objective (2) becomes:

$$u^* = \arg\min_u \frac{1}{2}||Au - f||_2^2 + \mu||u||_1 \tag{3}$$

where u is formed by stacking up u_1 and u_2; vector $f = (0, 0, \cdots, 0, \eta)^T$ with length $L_D + 1$; and $(L_D + 1) \times 2L$ matrix A (T is transpose) is:

$$A = \begin{pmatrix} x_1(1) & x_1(2) & \dots & \dots & x_1(L_D-1) & x_1(L_D) & \eta \\ & x_1(1) & \dots & \dots & x_1(L_D-2) & x_1(L_D-1) & 0 \\ & & \ddots & & & \vdots & \vdots \\ & & & x_1(1) & \dots & x_1(L_D-L+1) & 0 \\ -x_2(1) & -x_2(2) & \dots & \dots & -x_2(L_D-1) & -x_2(L_D) & \eta \\ & -x_2(1) & \dots & \dots & -x_2(L_D-2) & -x_2(L_D-1) & 0 \\ & & \ddots & & & \vdots & \vdots \\ & & & -x_2(1) & \dots & -x_2(L_D-L+1) & 0 \end{pmatrix}^T$$

When $t \notin D$, cross multiplication of (1) shows that $\hat{s}_T = u_2^* * x_1 - u_1^* * x_2 \approx h_{21} * x_1 - h_{11} * x_2 = (h_{21} * h_{12} - h_{11} * h_{22}) * s_T$. Interference s_B is eliminated and $\hat{s}_T$ sounds same as s_T to human ear. Here we assumed that the acoustic environment does not change much so that estimates of h_{11} and h_{21} during D still apply when $t \notin D$. For a convex objective with non-negativity filter constraints for sparsity, see [7].

Extraction of a speech source from $M \geq 3$ mixtures of N sources ($N = M$) is similar. Let a source s_n $(1 \leq n \leq N)$ be silent in $t \in D$, for proper value of $(\eta, \mu) > 0$, we minimize:

$$\frac{1}{2}||\sum_{j=1}^{M} u_{jn} * x_j||_2^2 + \frac{\eta^2}{2}(\sum_{j=1}^{M} u_{jn}(1) - 1)^2 + \mu(\sum_{j=1}^{M} ||u_{jn}||_1),$$

and estimate s_n by $\hat{s}_n = \sum_{j=1}^{M} u_{jn} * x_j$.

III. Minimization by Bregman Method

In this section, we introduce Bregman distance [10], Bregman and split Bregman methods of non-smooth convex optimization. We show that the split Bregman method boils down to simple operations such as shrinkage, matrix multiplication, and one-time matrix inversion. Then we adapt the split Bregman method and apply it to the convex speech extraction model (3) in reverberant conditions.

A. Bregman iterative regularization

The Bregman method was first applied [11] to the image denoising model of Rudin-Osher-Fatemi [12] with the non-smooth total variation (TV) regularization:

$$u = \arg\min_u \mu \int |\nabla u| + \frac{1}{2}||u - f||_2^2 \tag{4}$$

where f is the observed noisy data and μ is a positive parameter related to signal to noise ratio. The Bregman distance is

$$D_J^p(u,v) = J(u) - J(v) - \langle p, u-v\rangle$$

where $J(u) = \mu\int|\nabla u|$ and $p \in \partial J$ is a subgradient of J at the point v. The Bregman distance is not a distance in the usual sense because $D_J^p(u,v) \neq D_J^p(v,u)$ in general. However, it measures the closeness of two points since $D_J^p(u,v) \geq 0$ for all u and v, and $D_J^p(u,v) \geq D_J^p(w,v)$ for all w on the line segment connecting u and v. The Bregman iterative regularization procedure [11] is to solve a sequence of unconstrained subproblems

$$u^{k+1} = \arg\min_u D_J^{p^k}(u,u^k) + \frac{1}{2}||u-f||_2^2 \tag{5}$$

for $k = 0, 1, \ldots$, starting with $u^0 = 0$ and $p^0 = 0$. Since $J(u) = \mu\int|\nabla u|$ is not differentiable everywhere, the subgradient of J may not be unique. However, it follows from the optimality of u^{k+1} in (5) that the inclusion $0 \in \partial J(u^{k+1}) - p^k + u^{k+1} - f$ holds or:

$$p^{k+1} = p^k + f - u^{k+1}. \tag{6}$$

Bregman iteration refers to the mapping from $(u^k, p^k) \rightarrow (u^{k+1}, p^{k+1})$.

Now consider a more general constrained minimization problem:

$$\min_u J(u), \quad \text{s.t.} \quad H(u) = 0 \tag{7}$$

where J is convex but not necessarily differentiable, such as the l_1 norm or TV norm, and H is convex and differentiable with zero as its minimum value. Traditionally, this problem may be solved by continuation methods. One solves a sequence of unconstrained problems

$$\min_u J(u) + \lambda_k H(u). \tag{8}$$

By choosing a sequence of positive numbers λ_k with $\lambda_k \rightarrow \infty$, one gets the solution of the constrained problem (7). Instead of solving (8), one solves a sequence of subproblems using the iterative regularization procedure as above:

$$u^{k+1} = \arg\min_u D_J^{p^k}(u,u^k) + H(u) \tag{9}$$

$$p^{k+1} = p^k - \nabla H(u^{k+1}). \tag{10}$$

with $u^0 = 0$ and $p^0 = 0$.

In [11], the authors analyzed the convergence of Bregman iterative scheme (9)-(10) and showed that under fairly weak assumptions on J and H, $H(u^k) \rightarrow 0$ as $k \rightarrow \infty$. For some cases, it is shown later [13] that this procedure solves the original problem (7). Here we restate two particular convergence results in [11].

Theorem III.1. *Assume that J and H are convex functionals and H is differentiable, and that the solutions to the subproblem in (9) exist. Let u^* be a minimizer of H, we then have*

(1) *monotonic decrease in H: $H(u^{k+1}) \leq H(u^k)$,*
(2) *convergence to a minimizer of H: $H(u^k) \leq H(u^*) + J(u^*)/k$.*

Theorem III.1 shows that $H(u^k)$ converges to $H(u^*)$. In particular, if H has minimal value 0, then u^k gets arbitrarily close to the solution of the constraint (7). If $H(u) = \frac{1}{2}||Au - f||_2^2$ and $Au = f$ has a solution, then $H(u^k)$ converges to 0 in finitely many steps [13]. The advantage of Bregman method is that it transforms a constrained problem into a sequence of unconstrained subproblems. It is different from the continuation methods since the parameter $\lambda_k = 1$ (uniform) for all subproblems. These subproblems are solvable in closed form when J is l_1 norm, as we show below in the context of the so called split Bregman method.

B. Split Bregman method

The split Bregman method was introduced by Goldstein and Osher [5] for solving l_1, TV, and related regularized problems in imaging. It has connections to Lagrangian-based alternating direction methods in convex optimization [14]. Consider the unconstrained problem:

$$\min_u J(\Phi u) + H(u),$$

where J and H are as in (7), and Φ is linear operator. In case of (3), $J(u) = \mu||u||_1$, $H(u) = \frac{1}{2}||Au-f||_2^2$, and $\Phi = I$. The key idea is to introduce an auxiliary variable $d = \Phi u$, and solve the constrained problem

$$\min_{d,u} J(d) + H(u), \ \text{s.t.} \ \ d = \Phi u \tag{11}$$

or

$$\min_{d,u} E(d,u), \ \text{s.t.} \ \ \frac{\lambda}{2}||d - \Phi u||_2^2 = 0$$

where $E(d,u) = J(d) + H(u)$ and λ is a positive constant. Then we can Bregmanize the problem as in (9). We replace $E(d,u)$ by its associated Bregman distance and update the subgradients p_d^k and p_u^k respectively. Given that $u^0 = 0$, $d^0 = 0$, $p_d^0 = 0$, and $p_u^0 = 0$, we have the iterations:

$$\begin{aligned}(u^{k+1}, d^{k+1}) =& \arg\min_{u,d} J(d) + H(u) - \langle p_d^k, d - d^k\rangle \\ & - \langle p_u^k, u - u^k\rangle + \frac{\lambda}{2}||d - \Phi u||_2^2 \\ p_d^{k+1} =& p_d^k - \lambda(d^{k+1} - \Phi u^{k+1}) \\ p_u^{k+1} =& p_u^k - \lambda\Phi^T(\Phi u^{k+1} - d^{k+1})\end{aligned}$$

For simplicity, we introduce a new variable $b^k = p_d^k/\lambda$. And we notice that $p_d^k = \lambda b^k$ and $p_u^k = -\lambda\Phi^T b^k$, and thus the iterations become:

$$\begin{aligned}(u^{k+1}, d^{k+1}) =& \arg\min_{u,d} J(d) + H(u) - \lambda\langle b^k, d - d^k\rangle \\ & + \lambda\langle b^k, \Phi(u - u^k)\rangle + \frac{\lambda}{2}||d - \Phi u||_2^2 \\ b^{k+1} =& b^k - d^{k+1} + \Phi u^{k+1}\end{aligned}$$

with $u^0 = 0$, $d^0 = 0$ and $b^0 = 0$. The iterates d^{k+1} and u^{k+1} can be updated alternatively. We first fix u^k to update d^{k+1} and then fix d^{k+1} to update u^{k+1}. The general split Bregman

iteration with initial values $d^0 = 0$, $u^0 = 0$, $b^0 = 0$, is:

$$d^{k+1} = \arg\min_d \frac{1}{\lambda} J(d) - \langle b^k, d - d^k \rangle + \frac{1}{2}||d - \Phi u^k||_2^2 \tag{12}$$

$$u^{k+1} = \arg\min_u \frac{1}{\lambda} H(u) + \langle b^k, \Phi(u - u^k) \rangle + \frac{1}{2}||d^{k+1} - \Phi u||_2^2 \tag{13}$$

$$b^{k+1} = b^k - (d^{k+1} - \Phi u^{k+1}) \tag{14}$$

If J is the l_1 norm, the subproblem (12) has explicit solutions. The subproblem (13) is also easy to solve since the objective is differentiable. Convergence of the split Bregman method for the case of $J(u) = \mu||u||_1$ is analyzed [15], and the result is:

Theorem III.2. *Assume that there exists at least one solution u^* of (11). Then we have the following properties for the split Bregman iterations (12),(13), and (14):*

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

Furthermore,

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

if u^* is the unique solution.

C. Implementation of FSE for moderate reverberations

In this subsection, we implement our proposed FSE method for the moderate reverberation case. Let $J(u) = \mu||u||_1$, $\Phi = I$, and $H(u) = \frac{1}{2}||Au - f||_2^2$.

Applying the split Bregman method and setting $d^0 = 0$, $u^0 = 0$, and $b^0 = 0$, we have the iterations:

$$d^{k+1} = \arg\min_d \frac{\mu}{\lambda}||d||_1 - \langle b^k, d - d^k \rangle + \frac{1}{2}||d - u^k||_2^2 \tag{15}$$

$$u^{k+1} = \arg\min_u \frac{1}{2\lambda}||Au - f||_2^2 + \langle b^k, u - u^k \rangle + \frac{1}{2}||d^{k+1} - u||_2^2 \tag{16}$$

$$b^{k+1} = b^k - (d^{k+1} - u^{k+1}) \tag{17}$$

Explicitly solving (15) and (16) 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 $\text{shrink}(v, t) = (\tau_t(v_1), \tau_t(v_2), \cdots, \tau_t(v_n))$ with $\tau_t(x) = \text{sign}(x)\max\{|x| - t, 0\}$, see Fig. 1. Noting that the matrix A is fixed, we can precalculate $(\lambda I + A^T A)^{-1}$, then the iterations only involve matrix multiplication and are extremely fast as a result. For moderate reverberation, the length of room impulse response (RIR) is not too long. The size of matrix $\lambda I + A^T A$ is $NL \times NL$, N being the number of sources. The computational cost for matrix inversion is not high. The above algorithm runs fast for the purpose of FSE.

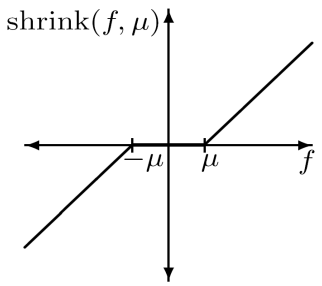


Fig. 1. *Demonstration of shrink operator in subsection III-C*

D. Subdivided Split Bregman for Long Reverberations

In the strong reverberation regime, RIR length is on the order of thousands. In order to have a more accurate solution, the length of u should be large accordingly. The length of u also goes up when $N \geq 3$. To reduce cost of matrix inversion when u is high dimensional, we subdivide u into r parts: $u = (u_1, u_2, \cdots, u_r)^T$ with $u_i \in \mathbb{R}^{\frac{NL}{r}}$. Correspondingly $A = [A_1, A_2, \cdots, A_r]$. The minimization problem is:

$$u = \arg\min_u \frac{1}{2}||\sum_{i=1}^r A_i u_i - f||_2^2 + \mu \sum_{i=1}^r ||u_i||_1.$$

The split Bregman method is applied to update each subdivided part of u sequentially (update u_i by fixing the other $r-1$ u_j's).

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) **For** i from 1 to r

$$u_i^{k+1} = (\lambda I + A_i^T A_i)^{-1}(A_i^T(f - \sum_{j\neq i} A_j u_j) + \lambda(d_i^{k+1} - b_i^k))$$

end For

(3) $b^{k+1} = b^k - d^{k+1} + u^{k+1}$

end While

where d_i and b_i are the subdivided parts of d and b. We precalculate inverse matrices $(\lambda I + A_i^T A_i)^{-1}$, each $\frac{NL}{r}$ dimensional. With proper choice of the number r, the computation speed can be improved significantly, as shown in section V.

IV. Source Activity Detection

The necessary preparation for FSE is silence detection of the speech sources. To maintain the overall speed of the proposed method, silence detection is based on the binary mask (BM) separation method DUET, the Degenerate Unmixing Estimation Technique [3], a fast method of blind speech separation without resolving RIRs. Though musical noise may occur due to binary operation in TF domain, DUET appears reliable for identifying silence periods of a target speech from a mixture (a robust speech feature). A brief review of DUET algorithm is given here. The standard mixing model for two

receivers and multiple sources is $x_j(t) = \sum_{k=1}^{N} h_{jk} * s_k$, where $j = 1, 2$, $*$ is the convolution and h_{jk} represents the impulse response from source s_k to sensor j. The time-domain signals $x_j(t)$, $j = 1, 2$, sampled at frequency f_s are first converted into frequency-domain time-series signals $X_j(f, \tau)$ with STFT. To group TF points into N clusters such that the points within each cluster are dominated by a single source signal, the feature parameters associated with each TF point are defined as $a(f, \tau) = |r(f, \tau)|$ and $\delta(f, \tau) = \frac{-1}{f} \angle r(f, \tau)$, where $r(f, \tau) = \frac{X_2(f,\tau)}{X_1(f,\tau)}$, $|\cdot|$ denotes the magnitude and $\angle\cdot$ denotes the phase angle of a complex number. Sufficient values of $a(f, \tau)$ and $\delta(f, \tau)$ generate a smooth two dimensional histogram. The K-means clustering algorithm finds the N most prominent peaks in the histogram. Each peak corresponds to one source in the mixture and the value for $a(f, \tau)$ and $\delta(f, \tau)$ at that peak are the feature parameters for that source. Once the feature parameters for each source have been estimated, DUET assigns the energy in each TF point to the source whose peak lies closest to that point in the feature space of a and δ. The individual separated signal spectrogram $Y_n(f, \tau)$ is estimated based on the clustering result. The TF binary mask for the n-th source signal is:

$$\mathcal{M}_n(f, \tau) = \begin{cases} 1 & (f, \tau) \in \text{cluster } C_k \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

Then $Y_n(f, \tau) = \mathcal{M}_n(f, \tau) X_J(f, \tau)$, where $n = 1, ..., N$ and J is a selected sensor index. Finally, inverse STFT (iSTFT) is applied to $Y_n(f, \tau)$ with overlap-add method [16] to recover the waveform $y_n(t)$.

The ratio $R_n(\tau) = \frac{||Y_n(\cdot,\tau)||_2^2}{||Y_B(\cdot,\tau)||_2^2}$ is used for detecting the silence part of source n, where Y_B is the sum of background sources. Though the separation quality may degrade if reverberation is long, the onset-offset feature is robust and detectable if we delete certain "fuzzy points" and reduce binary masking errors. Specifically, at each TF point (f, τ), the confidence coefficient of $(f, \tau) \in C_n$ is defined by

$$CC(f, \tau) = \frac{d_n}{\min_{j \neq n} d_j},$$

where d_j is the distance between the value of a and δ at (f, τ) and that at j-th peak. The mask is redefined for some $\rho > 0$ as

$$\mathcal{M}_n(f, \tau) = \begin{cases} 1 & (f, \tau) \in C_n \ \& \ CC(f, \tau) \leq \rho \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

The ρ is usually set to be $1/2$ to alleviate clustering error. We check the mean and variance of the ratio R_n frame by frame with proper frame size and overlapping. The time intervals with small mean and variance values are selected as the region where source n is almost silent. The entire FSE algorithm is:

V. Evaluation and Comparison

The implementation is in Matlab 2009b and the evaluation is done in the Windows 7 Home Premium operation system with Intel Core i5-M520 2.40 GHz CPU and 3.00 GB memory. The parameters for FSE are chosen as $\mu = \epsilon = 10^{-3}$, $\eta = 1$, and $\lambda = 2\mu$ throughout the evaluation.

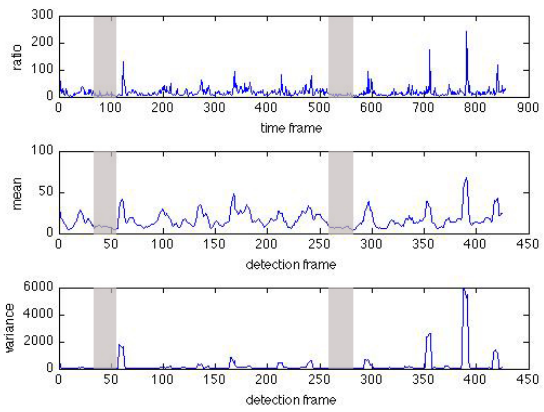


Fig. 2. *Source activity detection (mixture of speech and music). Top: ratio $R(\tau)$; middle: mean of $R(\tau)$; bottom: variance of $R(\tau)$. Detection frame size is 10 with shift as 2. The range of detection frame is half of time frame. Segments marked by the shadows are selected regions for D where the target speech signal is weak.*

Algorithm 1: FSE Overall Scheme

Input: Acoustic mixing signals, $x_j, j = 1, ..., M$ $(M \geq 2)$
Output: Extracted speech source $\hat{s}_n$, $n \in [1, N]$.
Activity Detection: Find durations of total length L_D where speech source n is either weak or silent
if *Room reverberation and number of sources are low* **then**
  Apply **split Bregman** method directly to obtain filters u_{jn}, $j = 1, ..., M$
else
  Apply **subdivided split Bregman** method to obtain filters u_{jn}, $j = 1, ..., M$
Speech Extraction: Calculate $\hat{s}_n = \sum_{j=1}^{M} u_{jn} * x_j$.

We first evaluate the proposed FSE method, study the relation between the length of selected silent speech duration and the extraction quality, and compare the split Bregman algorithm with subdivided split Bregman algorithm using synthetically mixed data (two sensors and two sources).

[Setup 1]: The room size is $5 \times 9 \times 3.5$ m, and the impulse responses are measured by two omni-directional microphones (middle of the room and 1.5 m above the floor) with the spacing 15 cm. The sources are 1 m away from the sensors with the azimuth 30° and 90°, and the same height as sensors. The reverberation times of impulse responses are from 0 s (anechoic) to 1.0 s. In order to illustrate the separation quality and speed of our proposed method, we simplify the detection step by knowing roughly about 0.5 s' silent duration D (e.g. 2.3 s - 2.8 s for the speech source in the up-left panel of Fig. 4) of target speech source ahead of time. The other source (e.g. lower-left panel in Fig. 4) is either speech or background music. The duration of the sources is 5 s and the sampling rate is 16000 Hz. Two mixtures (e.g. two right panels in Fig.

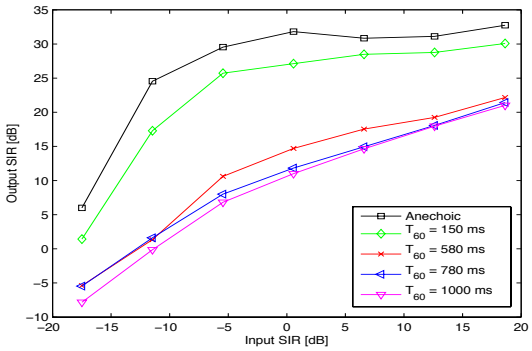


Fig. 3. *Output SIR vs. input SIR for the proposed FSE method with different reverberation times.*

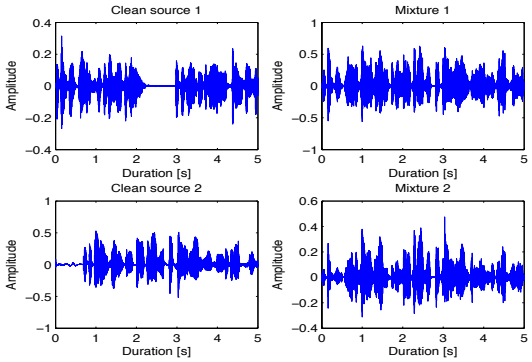


Fig. 4. *Clean speech source (up-left panel), background interference (lower-left panel) and two corresponding synthetical mixtures at* $T_{60} = 150$ *ms (up-right and lower-right pannels).*

4) are synthesized by measured RIRs according to (1). As the reverberation time goes up, the length of solution u (e.g. sparse solution u with 400 taps in Fig. 6) increases accordingly from 400 taps to 2000 taps. Shown in Fig. 3 are the average output signal to interference ratios (SIRs) achieved by FSE for the various reverberation times and input SIRs. Extracted speech sources in two channels are shown in Fig. 5, corresponding to the two sources in Fig. 4.

With different lengths of selected silent speech durations, FSE achieves various separation qualities, seen in Fig. 7. Basically, the separation effect is consistently improved with the increasing size of the silence region D (0.15 s, 0.30 s, 0.45 s, 0.60 s and 0.75 s). The separation reaches a plateau at 0.5 s. Length of 0.5 s total silence is an idea choice, which balances the computational speed and separation quality.

Table I illustrates the average iterations, computation time [s] and SIR improvement (SIRI [dB]) of the split Bregman algorithm and the subdivided split Bregman algorithm by different lengths of unmixing filters. The data are synthetic mixtures of two sources same as in **[Setup 1]** with however the reverberation time $T_{60} = 780\ ms$ and the input SIR $\approx -5.9\ dB$. The comparison indicates that the subdivided split Bregman ($r = 2$ here) performs better than the split Bregman if the length of unmixing filters is larger than 800 taps. When the length L is above 2000, the split Bregman runs out of memory. There is a trade-off between improved separation and

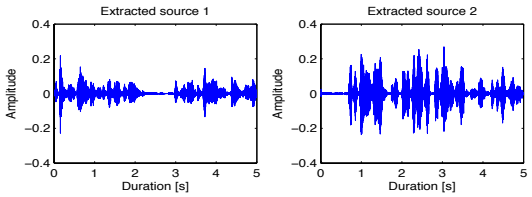


Fig. 5. *Extracted two speech sources from the two mixtures in Fig. 4 by FSE*

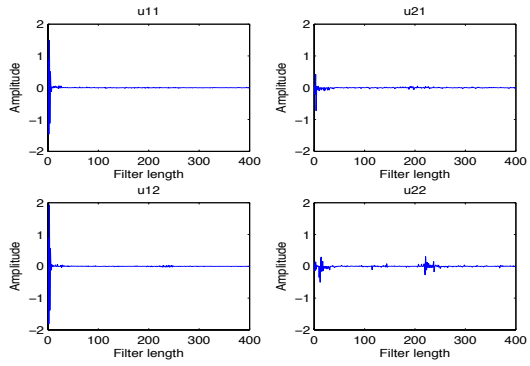


Fig. 6. *Sparse filters u's with 400 taps, u_{11} and u_{21} (u_{12} and u_{22}) are used to estimate source 1 (source 2) in Fig. 5.*

computation costs. From Table I, $L = 800$ already achieves a good separation.

TABLE I
Comparison of the (divided) split Bregman algorithms

	Split Bregman			Subdivided Split Bregman		
L	Iteration	Time	SIRI	Iteration	Time	SIRI
50	**57**	**0.028**	6.386	57	0.332	**6.392**
100	**50**	**0.058**	6.214	50	0.531	**6.221**
200	**42**	**0.209**	6.766	43	0.796	**6.776**
400	44	**0.780**	8.069	**43**	1.565	**8.111**
800	62	4.386	9.107	**50**	**4.064**	**9.195**
1200	63	10.994	10.364	**41**	**7.019**	**10.401**
1600	71	21.684	**11.379**	**66**	**14.820**	11.265
2000	103	38.161	**12.306**	**77**	**23.132**	12.159
2800	-	-	-	**104**	**48.245**	**12.984**
3600	-	-	-	**123**	**83.295**	**13.466**

The comparison of a list of existing BSS methods is shown in Table II in terms of computation time, SIR, signal to distortion ratio (SDR) and signal to artifact ratio (SAR). The data are synthetic mixtures of two speech sources as in **[Setup 1]** with reverberation time $T_{60} = 150\ ms$ and input SIR $\approx -5.9\ dB$. To compare the computation time of the algorithms directly, the proposed FSE method extracts two speech sources sequentially with the silent unions for the two speech sources known ahead of time. Table II indicates that the proposed FSE achieves the best separation quality in objective measures at almost the speed of FastICA.

Room recorded mixture data are used to evaluate and compare the above BSS methods by the Perceptual Evaluation

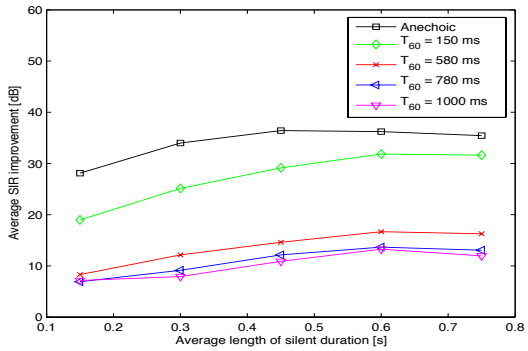


Fig. 7. *The relationship between average separation effect — SIR improvement (SIRI) and the length of selected silent speech duration with different reverberation times. The input SIR is about -5.9 dB.*

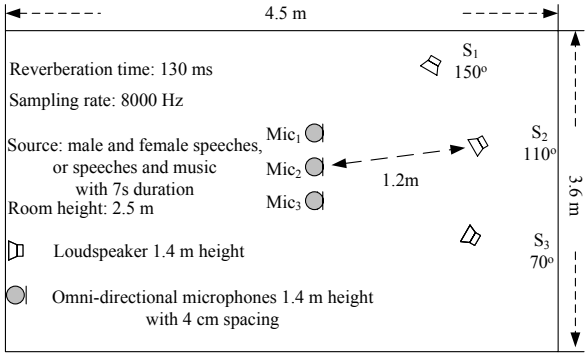


Fig. 8. *Configuration and parameters of the room recording.*

TABLE II
Comparison of BSS methods on synthetic mixture data

	Time [s]	SIR [dB]	SDR [dB]	SAR [dB]
Parra[17]	7.16	5.55	1.62	5.34
IVA[18]	42.72	14.59	7.21	9.52
SNGTD[1]	122.35	11.28	4.67	7.21
FastICA[1]	1.32	9.31	4.12	7.05
FSE	1.56	26.60	15.35	16.39

of Speech Quality (PESQ) [19]. **[Setup 2]**: The room size is 4.4 $\times$ 3.6 $\times$ 2.5 m with reverberation time $T_{60} = 130\ ms$. The loudspeakers and omni-directional microphones are 1.4 m high from the floor. The sensors are set in the middle of the room with 4 cm spacing linearly arranged. For the two sensors and two sources case, sources come from speaker S_1 and S_2, and Mic_2 and Mic_3 are turned on, see Fig. 8. For the case of three sensors and three sources, all the speakers and microphones in Fig. 8 are included. The mixture data are male and female speeches with the duration about 7 s and sampling rate 8000 Hz. Now with the source activity detection added, the separation quality of the proposed FSE exceeds those of the known methods, as seen from Table III. The speech sources activity detection is done within 2 to 3 seconds, and does not affect the efficiency of the FSE method. DUET BSS method [20] is included in Table III as the microphone spacing is small enough so that there is no phase-wrap ambiguity to degrade its performance.

TABLE III
Average PESQ of BSS methods on real recording mixture data. PRE PESQ is the average PESQ of the mixture data. Time for FSE is shown as detection time + speech extraction time.

	2 sources (time[s])	3 sources (time[s])
PRE PESQ	1.37	1.00
Parra	1.57 (7.9)	1.44 (16.0)
FastICA	1.90 (2.1)	1.70 (3.3)
SNGTD	2.07 (120)	1.88 (265)
IVA	2.35 (49.0)	2.02 (52.2)
DUET	2.36 (2.2)	2.00 (4.3)
FSE	2.58 (1.9+2.4)	2.15 (2.3+3.8)

TABLE IV
Subjective evaluation on blind speech separation. Here $>$ ($<$) means the output of our method is perceived better (worse) than the other method in terms of separation quality and voice clearness respectively, while $\approx$ means "hard to distinguish".

Method	Test Category	$>$	$\approx$	$<$
FSE vs IVA	Separation	71.5%	4.8%	23.7%
	Clearness	53.3%	5.5%	41.2%
FSE vs DUET	Separation	65.3%	5.8%	28.9%
	Clearness	45.5%	12.4%	42.1%

In the above objective evaluations, IVA, DUET and FSE lead other approaches. For further study, we evaluate these three approaches by subjective test. Mixture data are collected in the same environment as **[Setup 2]**, which contains both two sources and three sources cases. At lease one source is speech. Extracted speech sources by three different methods are evaluated by 10 human subjects with normal hearing. We utilized the paired comparison (PC) test, which requires each listener to rank the three methods according to the performance of separation quality and sound clearness. The preference percentages of our method to the other two methods is shown in Table IV, and they are calculated as

$$PC_{>} = \frac{\#\ of\ pairs\ where\ FSE\ is\ better}{\#\ of\ all\ pairs\ in\ the\ test} \tag{20}$$

$$PC_{<} = \frac{\#\ of\ pairs\ where\ FSE\ is\ worse}{\#\ of\ all\ pairs\ in\ the\ test} \tag{21}$$

$$PC_{\approx} = \frac{\#\ of\ pairs\ where\ difference\ is\ not\ significant}{\#\ of\ all\ pairs\ in\ the\ test} \tag{22}$$

Human perception test confirms that the proposed FSE method outperforms the other BSS methods in terms of speech separation quality and clarity.

VI. Discussion and Conclusion

We proposed and evaluated a fast and efficient blind speech extraction method as long as target speeches make pauses. A convex optimization problem is formulated and solved by the split Bregman method to yield sparse unmixing filters. Binary mask blind speech separation method is modified to detect the speech source onset-offset activity. Experimental results

indicate that the proposed method outperforms conventional blind speech separation methods in terms of the overall computation speed and separation quality. The limitation of the proposed method is that it relies on a robust silence detection in a long reverberation multi-talker environment which will be studied further in future work.

ACKNOWLEDGMENT

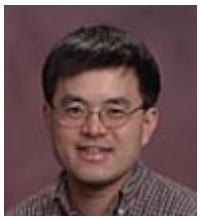
The authors would like to thank Yang Wang for helpful discussions.

REFERENCES

[1] S. Makino *et al. (eds.)*, Blind Speech Separation, Springer 2007.
[2] J. Xin, M. Yu, Y. Qi, H. Yang, and F-G Zeng, "A nonlocally weighted soft-constrained natural gradient algorithm for blind source separation of reverberant speech", IEEE Workshop on Application of Signal Processing to Audio and Acoustics, 81-84, Oct. 2009, New Paltz, NY, USA.
[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking", IEEE Trans. Signal Processing, vol. 52, no. 7, pp. 1830-1847, July 2004.
[4] 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.
[5] T. Goldstein and S. Osher, "The split Bregman algorithm for L1 regularized problems", *SIAM J. Imaging Sci.*, 2(2), 323-343, 2009.
[6] 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.
[7] Y. Wang, Z. Zhou, "Background suppression in audio through learning", in preparation, 2010.
[8] J. Allen, D. Berkley. "Image method for efficiently simulating small-room acoustics". J. Acoustical Society America, 65:943-950, 1979.
[9] D. Duttweiler. "Proportionate normalized least-mean-squares adaptation in echo cancelers". IEEE Trans. Speech Audio Processing, 8:508-518, 2000.
[10] 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.
[11] 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.
[12] L. Rudin, S. Osher, E. Fatemi, "Nonlinear total variation based noise removal algorithms", *Physica* D, 60, 259-268, 1992.
[13] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterativealgorithms for l_1-minimization with application to compressed sensing", *SIAM J. Imaging Sci.*, 1(1):143-168, 2008.
[14] E. Esser, "Applications of Lagrangian-Based Alternating Direction Methods and Connections to Split Bregman", CAM report, 09-31, UCLA, 2009.
[15] J. Cai, S. Osher and Z. Shen, "Split Bregman Methods and Frame Based Image Restoration", *Multiscale Model. Simul.* 8(2):337–369, 2009.
[16] 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. ICASSP2005, Mar. 2005, vol. III, pp. 81-84.
[17] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources", IEEE Trans. Speech Audio Processing, vol. 8, no. 3, 320-327, May 2000.
[18] T. Kim, H. Attias, S-Y Lee, and T-W Lee, "Blind source separation exploiting higher-order frequency dependencies", IEEE Trans. Audio, Speech Language Processing, vol. 15, no. 1, pp 70-79, 2007.
[19] 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.
[20] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures", in Proc. ICASSP 2000, vol. 12, 2985-2988, 2000.

Meng Yu received his B.S in scientific & engineering computing at Peking University in 2007 and M.S in computational and applied mathematics at University of California, Irvine in 2009. He is a Ph.D. candidate in acoustic speech and voice signal processing.

Wenye Ma received the B.S. and M.S. degree in mathematics from University of Science and Technology of China, in 2004 and 2007, and the M.A. in mathematics from University of California, Los Angeles in 2009. He is currently working towards the Ph.D. degree in mathematics at Unversity of California, Los Angeles. His research interests include optimization and its applications to image and signal analysis.



Jack Xin received his B.S in computational mathematics at Peking University in 1985, M.S and Ph.D in applied mathematics at New York University in 1988 and 1990. He was a postdoctoral fellow at Berkeley and Princeton in 1991 and 1992. He was assistant and associate professor of mathematics at the University of Arizona from 1991 to 1999. He was a professor of mathematics from 1999 to 2005 at the University of Texas at Austin. He has been a professor of mathematics in the Department of Mathematics, Center for Hearing Research, Institute for Mathematical Behavioral Sciences, and Center for Mathematical and Computational Biology at UC Irvine since 2005. He is a fellow of the John S. Guggenheim Foundation. His research interests include applied analysis and computation in nonlinear and multiscale problems, mathematical modeling in speech and hearing sciences, and sound signal processing.

Stanley Osher received his Phd degree in 1966 from New York University's Courant Institute of Mathematical Sciences. He is a Professor of Mathematics, Computer Science and Electrical Engineering at UCLA. He is also an Associate Director of the NSF funded Institute for Pure and Applied Mathematics. He is a member of the National Academy of Sciences, the American Academy of Arts and Sciences and is one of the top 25 most highly cited researchers in mathematics and computer sciences. He has received numerous academic honors and has co-founded three successful companies, each based largely on his own (joint) research. His current interests mainly involve image science.