Multi-Channel l_1 Regularized Convex Speech Enhancement Model and Fast Computation by the Split Bregman Method

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

***Abstract*—A convex speech enhancement (CSE) method is presented based on convex optimization and pause detection of the speech sources. Channel spatial difference is identified for enhancing each speech source individually while suppressing other interfering sources. Sparse unmixing filters indicating channel spatial differences are sought by l_1 norm regularization and the split Bregman method. A subdivided split Bregman method is developed for efficiently solving the problem in severely reverberant environments. The speech pause detection is based on a binary mask source separation method. The CSE method is evaluated objectively and subjectively, and found to outperform a list of existing blind speech separation approaches on both synthetic and room recorded speech mixtures in terms of the overall computational speed and separation quality.**

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

I. Introduction

BLIND speech separation (BSS) aims to recover source signals from their mixtures without detailed knowledge of the mixing process. However, it remains a challenge to retrieve sound sources recorded in real-world environments 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. Two major approaches are: time-domain BSS and frequency domain BSS [1].

Frequency domain approaches approximately reduce the convolutive unmixing problem into instantaneous BSS problem in each frequency bin [2], [3]. However, permutation ambiguity and frame length issue of short time Fourier transform (STFT) limit the performance [4]. As discussed in [5], a frequency-domain BSS which works well in low (100-200 ms) reverberation has degraded performance in medium (200-500 ms) and high ($>$ 500 ms) reverberations. Although much progress has been made for the permutation problem in recent years [6], [7], [8], few methods have been proposed with good separation results in a highly reverberant environment [5]. Thus, our proposed speech enhancement method treats the noisy signals in the time domain to avoid permutation and scaling ambiguity problems.

Time domain BSS approaches consider the impulse responses of a room as FIR filters, and directly estimate filter coefficients all together. Most of them are based on optimizing a statistical cost function measuring entropy, mutual information, and non-Gaussianity [9], [10], [11], [12], [13]. Those approaches may achieve a good separation if the optimization can be done accurately. However at the fundamental level, most of time domain methods 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 [9], [10], [13], as observed in [5]. Another disadvantage is that performance depends strongly on the initial value [14].

One more major challenge for both time and frequency domain BSS methods is that they do not work well for source enhancement if the desired source signals are very weak in comparison to the unwanted sources in the mixtures. Motivated by blind channel identification (BCI) [15] and blind sparse channel identification (BSCI) [16], a novel fast time domain convex speech enhancement (CSE) method is proposed in this paper based on the assumption that intelligible speech signals contain pauses. The proposed CSE requires a duration (or sum of durations) of around 100 ms [1] where one speech source is either silent or has extremely weak energy, as shown in the section V-C. In fact, 100 ms is not a short duration for detection in speech. As observed by [17], conversational speech rarely has a higher than 50% "on" time. So the "off" time of 100 ms almost surely exists in conversational speech signals of a few seconds. Pause detection is a problem of

The authors M. Yu and W. Ma contributed equally to this work. M. Yu and J. Xin were partially supported by NSF DMS-0911277, DMS-0948247 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. Phone: (949)-824-5309. Fax: (949)-824-7993.

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. Phone: (310)-825-1758. Fax: (310)-206-6673.

[1] Under moderate reverberant conditions.

independent interest, which we handle here by processing the output of a modified time-frequency (TF) domain clustering method. Because we only detect silence durations from the initial separation, tolerance of artifacts in TF domain clustering is allowed to be higher. In the example of two sources, during silent durations of the target speech signal, information of the interference (background) is collected and allows us to formulate an l_1 norm regularized **convex** optimization problem on a pair of sparse filters for cross channel cancellation. The cancellation suffices to identify channel spatial differences and estimate the target speech. A sparse solution is computed by the split Bregman method for which fast convergence was recently studied [18]. Unlike speech separation methods in TF domain [19], the proposed time-domain optimization does not rely on the sparsity hypothesis of speech spectrogram. Moreover, convexity guarantees global and fast convergence of the algorithm. Since our method aims to exploit the spatial difference which *does not assume the statistics of speech data* (e.g. independence or non-Gaussian assumptions of the source signals), filter estimation is more economical in data usage and robust under non-stationary speech data [20]. The l_1 regularization renders the estimation less sensitive to reverberant and noisy conditions, as discussed in section II. Because our method aims at cross channel cancellation instead of dereverberation [16], it *does not attempt to invert impulse responses* as [16]. Though the proposed method and the method in [16] share similar framework and are both based on spatial difference (between channels of the active source to microphones), our model ends up with knowledge of spatial difference for the cancellation of the interferences, while [16] ends up with an estimate of real impulse responses. Therefore, the sparseness regularization in [16] reduces the dimension of feasible space for the solution and may not be sufficient to fully represent the real impulse responses under highly reverberant conditions. In addition, with this regularization the proposed method does not require an assumption on the absence of common zeros of channel functions (z-transforms of impulse responses) [11], [21], [22], as discussed in section II.

The idea of adapting an interference canceller when the target signal is absent is well known. Such a strategy was proposed [17] in 1990 in the context of adaptive delay and sum beamformer. The input signal is a noisy speech. The noise cancellation filter is adapted when speech (target signal) is not present. The phase delays in the delay and sum beamformer are extracted from the cross correlation function of the two receivers, which requires a positive signal to noise ratio (SNR) of the input. Energy threshold method for speech silence detection deteriorates if input SNR falls under 5 dB. Our proposed source activity detection method based on initial separation by clustering allows speech silence detection to succeed for lower input SNR values. Moreover, the proposed CSE method works under more reverberant conditions, see section V.

The paper is organized as follows. In section II, the convex optimization problem for CSE is introduced. In section III, the computational framework by l_1 norm regularization is shown. In subsections III-B and III-C, the algorithms for moderately and highly reverberant acoustic environments are illustrated. The subdivided split Bregman method is proposed for CSE with long reverberations and large number of sources. In section IV, an onset-offset detection method of speech is outlined. In section V, the proposed CSE method is studied in terms of model and parameter choices, and the comparison among the split Bregman algorithm, the subdivided split Bregman algorithm and one recently popular l_1 optimization algorithm is investigated. Silent speech segment detection is evaluated under various reverberant conditions. Evaluations of CSE show its merits in both speed and separation quality compared with existing methods. Discussion and conclusions are in section VI. Our method also applies to non-speech signal enhancement from convolutive mixtures as long as pause detection of target signal is possible.

II. Convex Speech Enhancement 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). CSE method shall sequentially enhance 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

$$x_j(t) = h_{j1} * s_B(t) + h_{j2} * s_T(t) \tag{1}$$

where t is time; $j = 1, 2$; and $*$ is linear convolution. Instead of finding an unmixing filter W such that $W * (x_1, x_2)$ recovers (s_T, s_B), we enhance 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

$$h_{21} * x_1(t) - h_{11} * x_2(t) \approx 0 \quad for\ t \in D. \tag{2}$$

The elimination by cross-channel cancellation was known in blind channel identification [15] and background suppression [5]. It forms an l_2 norm optimization problem

$$(u_1^*, u_2^*) = \arg\min_{u_1, u_2} \frac{1}{2} ||u_2 * x_1 - u_1 * x_2||_2^2 \tag{3}$$

where u_1 and u_2 are subject to certain nontrivial constraints, e.g., $||u_1||_2^2 + ||u_2||_2^2 = 1$ to avoid trivial zero solution. Solved by eigenvalue decomposition as shown in [15], it suffers from limitations that a) the system should be noiseless and b) the two filters have no common zeros. As pointed in [16], by replacing the constraint $||u_1||_2^2 + ||u_2||_2^2 = 1$ with a convex singleton linear constraint $u_1(l) = 1$, where $u_1(l)$ is the lth element of filter u_1, the new optimization becomes

$$(u_1^*, u_2^*) = \arg\min_{u_1(l)=1} \frac{1}{2} ||u_2 * x_1 - u_1 * x_2||_2^2 \tag{4}$$

This least squares (LS) formulation is more robust to ambient noise than the eigenvalue decomposition approach in (3) since the singleton linear constraint in (4) has much less coupling in filter energy allocation [16]. It removes two degrees of freedom in filter estimates: a constant time delay (by fixing l) and a constant scalar factor (by fixing $u_1(l) = 1$). However,

it requires that the direct path in filter u_2 is no more than l samples earlier than the one in filter u_1. Instead of solving the constrained problem, we formulate an unconstrained problem by placing $\frac{\eta^2}{2}(\sum_{j=1}^{2} u_j(1)-1)^2$ in the objective,

$$(u_1^*, u_2^*) = \arg\min_{(u_1,u_2)} \frac{1}{2}||u_2 * x_1 - u_1 * x_2||_2^2 + \frac{\eta^2}{2}(\sum_{j=1}^{2} u_j(l)-1)^2 \quad (5)$$

where the second term $\frac{\eta^2}{2}(\sum_{j=1}^{2} u_j(l)-1)^2$ is to fix scaling and prevent zero (trivial) solution, η is a trade-off parameter to balance this term with the error of cross-channel cancellation. The choice of l is arbitrary as long as it is smaller than the length of u_j. Then the relative amplitude between u_1 and u_2 at tap l will be determined automatically with the convergence of algorithm iteration. Since the tap l does not represent the direct channel in this model, the limitation of the direct path in filter u_2 is removed.

As mentioned in [15] and [23], a common zero of the two channels would make (3), (4) and (5) unable to identify channel impulse responses. An l_1 regularization of the solution prevents this by shrinking the dimension of feasible space for solution u_j^* $(j = 1, 2)$, so that the optimal solutions are not necessarily the real room impulse responses h_{11} and h_{21}. Inside D, we seek a pair of sparse filters u_j $(j = 1, 2)$ to minimize the energy of $u_2 * x_1 - u_1 * x_2$ in the region D. Ideally, $u_1 \approx h_{11}$ and $u_2 \approx h_{21}$, however, the solutions are expected to be a pair of sparse acoustic room impulse responses (RIR). The sparse RIR model is theoretically sound [24], and has been shown useful for echo cancellation in real acoustic environments [25]. Filter sparseness is achieved by l_1-norm regularization which improves the robustness of the method. The resulting convex optimization problem for $t \in D$ is:

$$(u_1^*, u_2^*) = \arg\min_{(u_1,u_2)} \frac{1}{2}||u_2 * x_1 - u_1 * x_2||_2^2 + \frac{\eta^2}{2}(\sum_{j=1}^{2} u_j(l)-1)^2 + \mu(||u_1||_1 + ||u_2||_1), \quad (6)$$

where the parameter μ is used to balance the sparsity of solution with other terms.

Denote the length of D by L_D and that of u_j by L. L_D can be as short as even 100 msec duration.[2] As a result, this spatial difference based method carries out the CSE problem efficiently in terms of the data usage and is different from other BSS methods that rely on the high order statistics of data. Since the solution u_j is l_1 regularized, i.e., the optimal solution is sparse, the surplus length of it would be 0 while solving (6), which overcomes the over-fitting problem. In other words, if the given order of the filter u_j $j = 1, 2$, is larger than an expected value to represent the spatial difference between channels, sparse regularization would automatically suppress the spurs for the surplus part. In addition, sparseness allows the solution u_j to resolve the major spikes of the channel impulse response filters which comprise the relative time delay. In this sense, the regularization helps estimation of spatial difference to be insensitive and robust under reverberant conditions.

In section V, (5) is solved by a least-squares (LS) algorithm and compared with (6) in terms of speech enhancement quality. Since the spatial difference between the two channels is indicated by the peak delay between u_1 and u_2, the performance of direction of arrival (DOA) estimation through the relative time delay between u_1 and u_2 is evaluated as a reference for the accuracy of the solution u_1 and u_2, as shown in section V.

In matrix form, convex objective (6) becomes:

$$u^* = \arg\min_u \frac{1}{2}||Au - f||_2^2 + \mu||u||_1 \quad (7)$$

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 $x_j(t)$, $j = 1, 2$ in (1) shows that

$$\begin{aligned} \hat{s}_T &= u_2^* * x_1 - u_1^* * x_2 \\ &= h_{21} * x_1 - h_{11} * x_2 + e \\ &= (h_{21} * h_{12} - h_{11} * h_{22}) * s_T + e, \end{aligned} \quad (8)$$

where e is the cross-channel cancellation error. The error e is expected to be low energy and insignificant to hearing if leading peaks of RIRs ((h_{21}, h_{11})) are captured accurately in (u_1^*, u_2^*). If the length of filters u_j, $j = 1, 2$, is greater or equal to the length of real impulse response filters h_{jk}, $j, k = 1, 2$, the surplus part of u_j will be forced to 0's due to the sparsity regularization, such that the estimated cancellation filters u_j, $j = 1, 2$, approximate the real impulse response filters. However, due to computational complexity, the cancellation filter length L is less than the length of real impulse response filters in reverberant conditions. An underestimation problem arises. Fortunately, instead of resolving the whole impulse response filters, the sparsity regularization encourages the solution to resolve the spatial difference by focusing on the direct path and early reverberation parts. Therefore, the filter length L is less than the length of room impulse responses, yet it suffices to capture the leading peaks in RIRs. In (8), the interference s_B is eliminated up to an error e. Though doubly-convolved, $\hat{s}_T$ is expected to be a signal with the same speech contents as spoken by the target speaker, so $\hat{s}_T$ approximates s_T well for a human ear in our later experiements. Here we assumed that the acoustic environment does not change much so that estimates of h_{11} and h_{21} during

[2]For moderate reverberant conditions, L is significantly shorter than L_D.

D still apply when $t \notin D$. For a convex objective with non-negativity filter constraints for sparsity, see [26].

In order to extend the model to a more general case, let us consider the enhancement for 3 sources and 3 mixtures, where each mixture is the sum of sources coming from different channels as

$$\begin{aligned} x_1 &= h_{11} * s_1 + h_{12} * s_2 + h_{13} * s_3, \\ x_2 &= h_{21} * s_1 + h_{22} * s_2 + h_{23} * s_3, \\ x_3 &= h_{31} * s_1 + h_{32} * s_2 + h_{33} * s_3. \end{aligned} \tag{9}$$

With the same assumption as 2 sources case, the signal s_3 to be enhanced is silent in the duration D. Let u_j, $j = 1, 2, 3$ to be the cancellation filters. If the filters satisfy

$$u_1 * h_{11} + u_2 * h_{21} + u_3 * h_{31} = 0 \tag{10}$$

$$u_1 * h_{12} + u_2 * h_{22} + u_3 * h_{32} = 0, \tag{11}$$

then

$$u_1 * x_1 + u_2 * x_2 + u_3 * x_3 = (u_1 * h_{13} + u_2 * h_{23} + u_3 * h_{33}) * s_3 \tag{12}$$

However, the existence of the solution u_j (j = 1,2,3) satisfying (10) and (11) is to be proved.

As studied in [26], we take the Fourier transform to show the existence of u_j, $j = 1, 2, 3$, such that

$$\hat{u}_1(\xi)\hat{h}_{11}(\xi) + \hat{u}_2(\xi)\hat{h}_{21}(\xi) + \hat{u}_3(\xi)\hat{h}_{31}(\xi) = 0 \tag{13}$$

$$\hat{u}_1(\xi)\hat{h}_{12}(\xi) + \hat{u}_2(\xi)\hat{h}_{22}(\xi) + \hat{u}_3(\xi)\hat{h}_{32}(\xi) = 0 \tag{14}$$

where ξ is the frequency index in frequency domain. Let $\mathcal{F}$ be the field of all real trigonometric rational functions, i.e. functions of the form f/g where both f and g are trigonometric polynomials with real coefficients. Set $H = [\hat{h}_{jk}]$ with $j = 1, 2, 3$ and $k = 1, 2$. Since the number of columns is larger than the number of rows, there exists a $V = [v_1, v_2, v_3]^T$ in $\mathcal{F}^3$ such that $HV = 0$. Now let $F(\xi)$ be a trigonometric polynomial that is the common denominator of all trigonometric rational functions v_j, $j = 1, 2, 3$, and $U_j(\xi) = F(\xi)v_j(\xi)$. Each U_j is a trigonometric polynomial with real coefficients. Let u_j, $j = 1, 2, 3$ be filters in time domain such that $\hat{u}_j = U_j$. Then u_j's satisfy (10) and (11).

To find the solution u_j, $j = 1, 2, 3$, we take the mixtures x_j, $j = 1, 2, 3$ in the duration D as the training data. Since $s_3 = 0$ in D, it follows from (12) that the filters satisfying (10) and (11) result in $u_1 * x_1 + u_2 * x_3 + u_3 * x_3 = 0$ in D. Thus we use the data in D to learn the interfering sources and estimate the cancellation filters by minimizing

$$\begin{aligned} (u_1^*, u_2^*, u_3^*) = \arg\min_{u_1,u_2,u_3} & \frac{1}{2}||u_1 * x_1 + u_2 * x_2 + u_3 * x_3||_2^2 \\ & + \frac{\eta^2}{2}(\sum_{j=1}^{3} u_j(l) - 1)^2 \end{aligned} \tag{15}$$

Unfortunately, the numerical experiments have shown that without further constraints the cancellation filters obtained by (15) based on the training data in D do not work well on the whole utterance in the time domain. This results from a serious problem, which is over-fitting. The energy of $(u_1 * x_1 + u_2 * x_2 + u_3 * x_3)$ in D is minimized by exploiting the underlying channel information, i.e., both (10) and (11) are satisfied. However, the cancellation model might memorize the training data rather than learning to generalize from the model, i.e. $(u_1 * h_{11} + u_2 * h_{21} + u_3 * h_{31}) * s_1 + (u_1 * h_{12} + u_2 * h_{22} + u_3 * h_{32}) * s_2 = 0$ in D, while neither $(u_1 * h_{11} + u_2 * h_{21} + u_3 * h_{31})$ nor $(u_1 * h_{12} + u_2 * h_{22} + u_3 * h_{32})$ is equal to 0.

There are a few ways to avoid over-fitting, including cross-validation, regularization or Baysian priors on the parameters, and so on. By cross-validation, we have to chop the training data in D into a few chunks and use cross-validation to prevent the model memorizing the training data. Another efficient approach as we take here is regularization. With the same regularization illustrated in 2 sources case, we adjust the optimization (15) as

$$\begin{aligned} (u_1^*, u_2^*, u_3^*) = \arg\min_{u_1,u_2,u_3} & \frac{1}{2}||u_1 * x_1 + u_2 * x_2 + u_3 * x_3||_2^2 \\ & + \frac{\eta^2}{2}(\sum_{j=1}^{3} u_j(l) - 1)^2 \\ & + \mu(||u_1||_1 + ||u_2||_1 + ||u_3||_1). \end{aligned} \tag{16}$$

The optimal solutions u_j^*, $j = 1, 2, 3$ are used by (12) to enhance the source s_3 from the mixtures.

Enhancement 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:

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

for which the matrix form is the same as (7) except that the matrix A and vector f change accordingly with the increased number of sources. We estimate s_n by

$$\hat{s}_n = \sum_{j=1}^{M} u_{jn} * x_j. \tag{17}$$

III. Minimization by Bregman Method

Because of the presence of the l_1-regularization term, the optimization problem (7) is difficult to solve. Traditional methods such as gradient decent methods can not be applied since l_1 norm is not differentiable. The split Bregman method was introduced by Goldstein and Osher [18] for solving l_1, total variation, and related regularized problems, which is one of the most effective method for such non-smooth convex optimization problems. In this section, we introduce the split Bregman method and show that it 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 enhancement model (7). In the strong reverberation conditions, length of u_j should be large accordingly. We then propose the subdivided split Bregman method for the regime of long reverberations and large number of sources.

A. Split Bregman method

The goal of the original Bregman method [27], [28] is to solve the general constrained minimization problem:

$$\min_x E(x) \text{ s.t. } G(x) = 0 \tag{18}$$

where E is a convex function and H is convex and differentiable with zero as its minimum value. The original Bregman method is based on the concept of Bregman distance for a convex function E, for single variable case, which is given as:

$$D_E^p(x, y) = E(x) - E(y) - \langle p, x - y\rangle \tag{19}$$

where $p \in \partial E$ is a subgradient of E at the point y. Using (19), the problem (18) can be solved by Bregman iterations:

$$x^{k+1} = \arg\min_u D_E^{p^k}(x, x^k) + G(x) \tag{20}$$

$$p^{k+1} = p^k - \nabla G(x^{k+1}) \tag{21}$$

The advantage of Bregman iteration is to transform a constrained problem into a sequence of unconstrained subproblems. In [28], [29], the authors analyzed the convergence of Bregman iterative scheme (20)-(21) and showed that under fairly weak assumptions, this procedure solves the original problem (18). Here we restate a particular convergence result in [29].

Theorem III.1. *Assume that E is a convex function and $G(x) = h(Mx - f)$ where $h(\cdot)$ is a convex and differentiable function that only vanishes at $\mathbf{0}$, M is a matrix, f is a vector, and $Mx = f$ has at least one solution. If the solutions to the subproblem in (20) and (21) exist, then*

(1) *$Mx^k = f$ in finitely many steps,*
(2) *Such a x^k is a solution to the original problem (18).*

The split Bregman method aims to solve the unconstrained problem:

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

where J is a convex function but not necessarily differentiable, H is a convex differentiable function, and Φ is linear operator. In case of (7), $J(u) = \mu||u||_1$, $H(u) = \frac{1}{2}||Au - f||_2^2$, and $\Phi = I$. The solution of this l_1 regularized problem can be explicitly expressed only if A has some special structure (e.g. orthogonal). However, the matrix A is usually complicated. The key idea of the split Bregman method is to introduce an auxiliary variable $d = \Phi u$, and solve the constrained problem

$$\min_{d,u} E(d, u), \text{ s.t. } \frac{\lambda}{2}||d - \Phi u||_2^2 = 0 \tag{22}$$

where $E(d, u) = J(d) + H(u)$ and λ is a positive constant. Theoretically, the parameter λ can be any positive number, which is used to balance $||d - \Phi u||_2^2$ with other terms in the following iterations. In practice, a suitable choice of this parameter would result in fast convergence. By this splitting technique, the l_1 norm and the matrix A applies onto d and u respectively and thus alternative updating scheme can be applied.

We then apply the original Bregman method to the problem (22). Since we have two variables (u and d), we perform the Bregman iteration for each of the variables. Namely, 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 define $b^k = p_d^k/\lambda$ and find that $p_d^k = \lambda b^k$ and $p_u^k = -\lambda\Phi^T b^k$. The iterations then become:

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

The iterates d^{k+1} and u^{k+1} can be updated alternately. 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:

$$d^{k+1} = \arg\min_d \frac{1}{\lambda}J(d) - \langle b^k, d - d^k\rangle + \frac{1}{2}||d - \Phi u^k||_2^2 \tag{23}$$

$$\begin{aligned}u^{k+1} =& \arg\min_u \frac{1}{\lambda}H(u) + \langle b^k, \Phi(u - u^k)\rangle \\ & + \frac{1}{2}||d^{k+1} - \Phi u||_2^2\end{aligned} \tag{24}$$

$$b^{k+1} = b^k - (d^{k+1} - \Phi u^{k+1}) \tag{25}$$

If J is the l_1 norm, the subproblem (23) has explicit solutions. The subproblem (24) 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 [30], and the result is:

Theorem III.2. *Assume that there exists at least one solution u^* of (22). Then we have the following properties for the split Bregman iterations (23), (24), and (25):*

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

Furthermore,

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

if u^* is the unique solution.

B. Implementation of CSE for moderate reverberations

In this subsection, we implement our proposed CSE 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:

$$d^{k+1} = \arg\min_d \frac{\mu}{\lambda}||d||_1 - \langle b^k, d - d^k\rangle + \frac{1}{2}||d - u^k||_2^2 \tag{26}$$

$$\begin{aligned}u^{k+1} =& \arg\min_u \frac{1}{2\lambda}||Au - f||_2^2 + \langle b^k, u - u^k\rangle \\ & + \frac{1}{2}||d^{k+1} - u||_2^2\end{aligned} \tag{27}$$

$$b^{k+1} = b^k - (d^{k+1} - u^{k+1}) \tag{28}$$

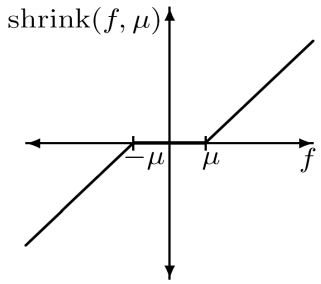


Fig. 1. *Demonstration of shrink operator in subsection III-B*

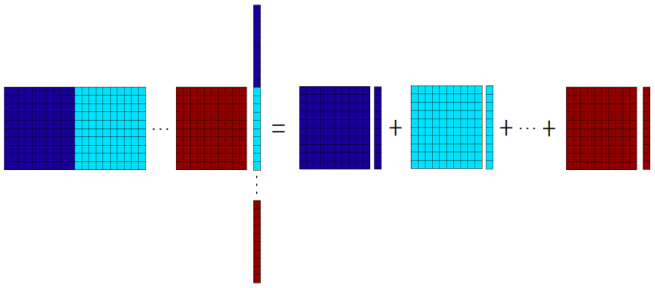


Fig. 2. *Matrix and vector decomposition:* $Au = \sum_{i=1}^{r} A_i u_i$. *Therefore the size of* $A_i^T A_i$ *is* $\frac{1}{r}$ *of* $A^T A$.

Explicitly solving (26) and (27) 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 $\tau_t(x) = \text{sign}(x)\max\{|x| - t, 0\}$ for $v = (v_1, v_2, \cdots, v_n) \in \mathbb{R}^n$ and $t > 0$, (see Fig. 1 for 1 dimensional case). 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 efficient 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, so the above algorithm runs fast.

C. 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 $A = [A_1, A_2, \cdots, A_r]$ (Fig. 2). 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).

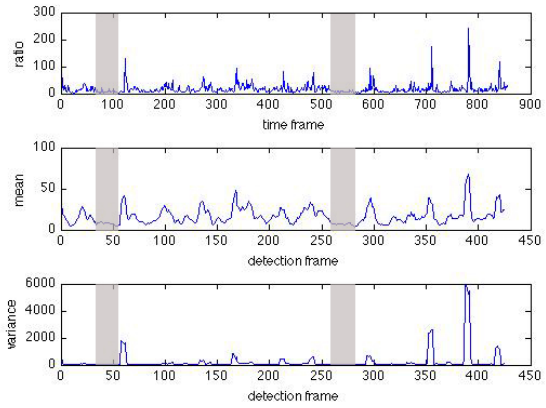


Fig. 3. *Source activity detection (mixture of speech and music). Top: ratio* $R(\tau)$*; middle: mean of* $R(\tau)$*; bottom: variance of* $R(\tau)$*. Detection frame size is 10 with shift as 2. The range of detection frame is half of time frame. Segments marked by the shadows are selected regions for D where the target speech signal is weak.*

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. Thus the total time for matrix inversion is $\frac{1}{r^2}$ as that for split Bregman method. 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 CSE 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 [31], a fast method of blind speech separation without resolving RIRs. It is known that the eigenvalue distribution of the spatial correlation matrix calculated from a microphone array input reflects the information of the number and power of sources. When the difference of the relative power of sources is small, the number of dominant eigenvalues roughly corresponds to the number of active sound sources. Previously, classification on eigenvalues by Support Vector Machines (SVM) and Support Vector Regression (SVR) [32] was proposed for detecting the overlapping speech segments. However, these methods share the common limitation that the number of microphones is required to be larger than the number of sources. In practice, the number of speakers in a meeting environment is not known ahead of time. In addition, the eigenvalue distribution has to be trained in order to carry out the classification. Reverberant and noisy conditions often degrade these methods by affecting the eigenvalue distribution of both the training and test data. Since these methods rely on the statistics of the data, they require large data usage. The detection frame size would be large (0.5 sec. in [32]), which makes it hard to find the desired segments with smaller lengths. 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). In addition, instead of simply detecting the number of active sources, source activity detection by DUET can distinguish sources from the mixtures by phase difference. 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 linear 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 ω_s are first converted into frequency-domain time-series signals $X_j(\omega, \tau)$ 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 $a(\omega, \tau) = |r(\omega, \tau)|$ and $\delta(\omega, \tau) = \frac{-1}{\omega} \angle r(\omega, \tau)$, where $r(\omega, \tau) = \frac{X_2(\omega, \tau)}{X_1(\omega, \tau)}$, $|\cdot|$ denotes the magnitude and $\angle \cdot$ denotes the phase angle of a complex number. Sufficient values of $a(\omega, \tau)$ and $\delta(\omega, \tau)$ 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. The values of $a(\omega, \tau)$ and $\delta(\omega, \tau)$ at that peak are the feature parameters (or components of the feature point) 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 location lies closest to that point in the feature space of a and δ. The individual separated signal spectrogram $Y_n(\omega, \tau)$ is estimated based on the clustering result. The TF binary mask for the n-th source signal is:

$$\mathcal{M}_n(\omega, \tau) = \begin{cases} 1 & (\omega, \tau) \in \text{cluster } C_n \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

Then $Y_n(\omega, \tau) = \mathcal{M}_n(\omega, \tau) X_J(\omega, \tau)$, where $n = 1, ..., N$ and J is a selected sensor index. Finally, inverse STFT (iSTFT) is applied to $Y_n(\omega, \tau)$ with overlap-add method [33] 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 silent 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. The fuzzy points elimination was used in our work [34] for post-processing of musical noise reduction. Specifically, at each TF point (ω, τ), the confidence coefficient of $(\omega, \tau) \in C_n$ is defined by

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

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

$$\mathcal{M}_n(\omega, \tau) = \begin{cases} 1 & (\omega, \tau) \in C_n \ \& \ CC(\omega, \tau) \leq \rho \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

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 CSE algorithm procedure is listed in Algorithm 1.

Algorithm 1: CSE 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 Enhancement: Calculate $\hat{s}_n = \sum_{j=1}^{M} u_{jn} * x_j$.

Time-invariant acoustic environment and fixed speakers render this algorithm as a mini-batch algorithm (or called intermediate method) with the online processing in place once we complete the activity detection and obtain the sparse filters.

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 CSE are chosen as $\mu = \epsilon = 10^{-3}$, $\eta = 1$, $\lambda = 2\mu$ and $\rho = 1/2$ throughout the evaluation.

The principle of the performance measures is to decompose a given estimate $\hat{s}(t)$ of a source $s(t)$ as a sum

$$\hat{s}(t) = s_{target}(t) + e_{interf}(t) + e_{noise}(t) + e_{artif}(t) \quad (31)$$

where $s_{target}(t)$ is an allowed deformation of the target source $s(t)$, $e_{interf}(t)$ is an allowed deformation of the sources which accounts for the interferences of the unwanted sources, $e_{noise}(t)$ is an allowed deformation of the perturbating noise (but not the sources), and $e_{artif}(t)$ is an "artifact" term that may correspond to artifacts of the separation algorithms such as musical noise, etc. or simply to deformations induced by the separation algorithm that are not allowed. Given such a decomposition, one can compute performance criteria as follows [44].

The Signal to Distortion Ratio

$$SDR \triangleq 10\log_{10}\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2} \quad (32)$$

The Signal to Interferences Ratio

$$SIR \triangleq 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2} \quad (33)$$

Besides these objective measures, the average PESQ [39] score was computed as a measure of performance. PESQ stands for "Perceptual Evaluation of Speech Quality". This algorithm was designed to provide a way to estimate the subjective quality of speech for telephony systems. The output from the algorithm is an estimate of the Mean Opinion Score (MOS), which is a number between 1 and 5. The meanings assigned to the scores in relation to the speech quality are: 1-Bad, 2-Poor, 3-Fair, 4-Good and 5-Excellent.

In the following subsections, (A) we compare our CSE model (6) solved by the split Bregman method with the model (5) solved by LS method; (B) verify the accuracy of the solution u^* according to the DOA estimation of the background interference based on u^* since u_1^* and u_2^* resolved the spatial difference between the two channels from the background interference to microphones; (C) study the relationship between the length of selected silent speech duration D and the enhancement quality; (D) compare the split Bregman algorithm with the subdivided split Bregman algorithm in terms of enhancement quality and computational speed; (E) investigate the proposed source activity detection method under various reverberant conditions; (F) evaluate the proposed CSE method and compare it with other BSS methods. The evaluations were done in both synthetic environment and real environment. For the synthetic acoustic environment, impulse responses are created using the Roomsim simulator [41] to synthesize the mixtures of speech signals. The simulated room size is $6 \times 4 \times 3$ m, with the microphones placed in the position $(2.5, 2, 1)$ m and the spacing 2 cm. Speakers are of the same height as microphones and 1 m away from microphones. Clean speech utterances from IEEE database [42] together with the simulated impulse responses under various acoustic conditions (Fig. 4), synthesized 50 groups of mixtures of male and female audio samples for the evaluation, each with 3 sec. duration and 16k Hz sampling frequency. For real recordings tested in subsection V-H, the room environment and the setup are illustrated in Fig. 18.

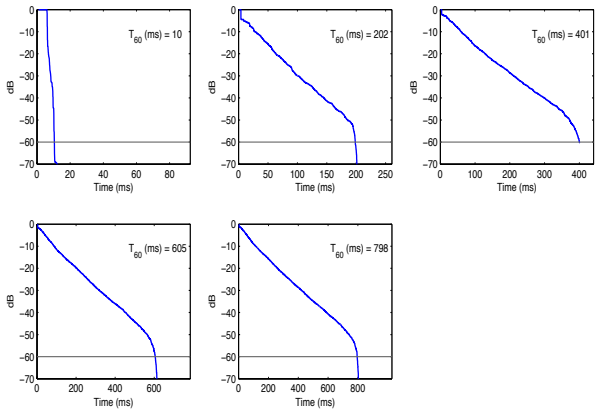


Fig. 4. *Demonstration of Schroeder curves for simulated impulse responses in different reverberant conditions.*

A. Model Comparison

In this subsection, the convex cross-channel cancellation model (5) is compared with our proposed CSE model in terms of signal to interference ratio (SIR) [44]. The 50 group of synthetic mixtures of audio samples are used for evaluation under five reverberant conditions (T_{60} = 0 ms, 200 ms, 400 ms, 600 ms, and 800 ms, see Fig. 4) with L = 64, 256, 512, 1024 and 1024 respectively. Each group contains two mixtures of two sources with the azimuth angles 30° and 70°. The detection step is skipped by knowing roughly about 0.5 sec. silent duration D (e.g. 0 sec. - 0.5 sec. (2.3 sec. - 2.8 sec.) for the speech source in the up-left (up-right) panel of Fig. 5) of each speech source ahead of time, since the purpose is to examine the optimization framework. Fig. 6 demonstrates the spectrograms of the enhanced speech sources based on the two different optimization frameworks under the reverberant condition ($T_{60} = 200$ ms). The estimated filters by CSE shows their sparsity compared with those based on (5), as seen in Figs. 7 and 8. The performance of speech enhancement is illustrated in Fig. 9 in terms of average SIR improvement, according to which the proposed CSE method outperforms the method based on the least-squares framework (5).

B. Investigation of Solution Accuracy

The question about the accuracy of the sparse solution u (u_j, $j = 1, 2$ in two sources case) arose in section II. Since u_j ($j = 1, 2$) are supposed to represent the spatial difference between the two channels (h_{11} and h_{21}) from the background interfering source to two microphones, empirically

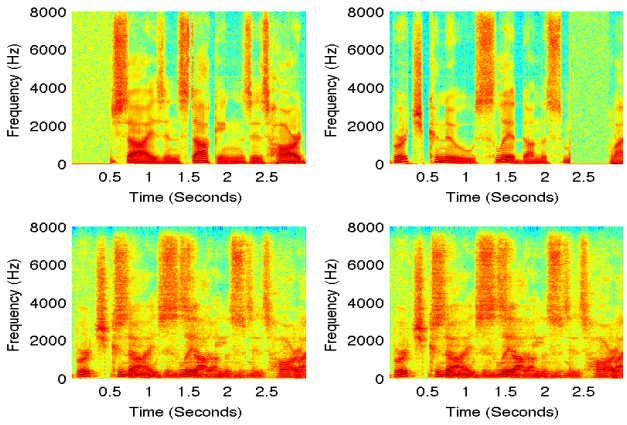


Fig. 5. *Spectrograms of clean speech sources (upper panels), and two corresponding synthetical mixtures with $T_{60} = 200$ ms (lower panels).*

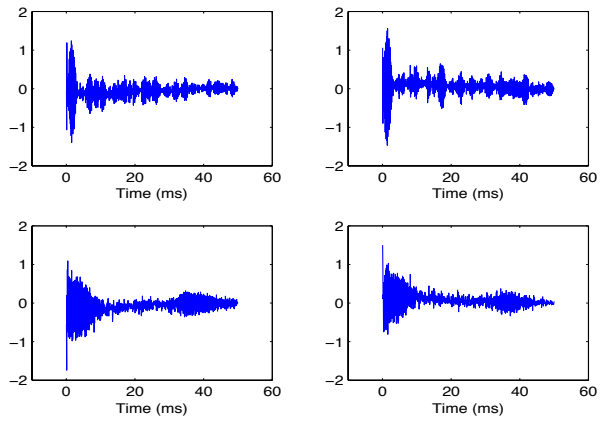


Fig. 7. *Filters u's (by LS) with 50 ms long, u_{11} and u_{21} (u_{12} and u_{22}) are used to estimate source 1 (source 2) in Fig. 6.*

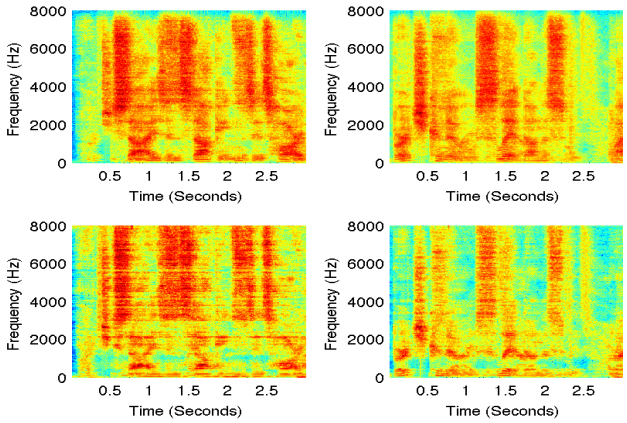


Fig. 6. *Enhancement of the two speech sources based on the two mixtures in Fig. 5 by CSE (upper panels) and LS (lower panels).*

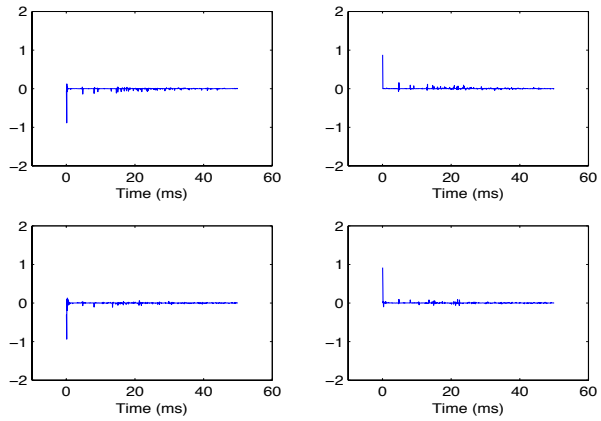


Fig. 8. *Sparse filters u's (by CSE) with 50 ms long, u_{11} and u_{21} (u_{12} and u_{22}) are used to estimate source 1 (source 2) in Fig. 6.*

the DOA estimation of the interfering source indicates the accuracy of the spatial difference represented by the solved sparse filters u_j ($j = 1, 2$). In other words, "good" solutions u_j, $j = 1, 2$, which suppress the interfering source by cross-channel cancellation, can accurately indicate the direction of the interfering source. In this experiment, based on the 50 pairs of synthetic mixtures, the azimuth angles of the interfering source are $105°$, $120°$, $135°$, $150°$ and $165°$ respectively, while the target source is fixed with the azimuth angle $70°$. The evaluation is conducted under noisy conditions (diffused noise with signal to noise ratio (SNR) [44]): 5 dB, 10 dB, 15 dB and 20 dB (while reverberation time T_{60} is fixed at 0), and reverberant conditions (T_{60}): 0 ms (anechoic), 200 ms, 400 ms, 600 ms and 800 ms respectively (while the SNR is fixed at 20 dB). The filter length L is set as 64, 256, 512, 1024 and 1024 for the five noisy and reverberant conditions respectively. The DOA is calculated by the time delay of arrival and the geometric configuration of the source and microphones. The time delay estimate (TDE) is determined (after interpolation) as the difference between two direct paths, that is TDE $= \arg\max_l |u^*_{1,l}| - \arg\max_l |u^*_{2,l}|$, [36].

It can be seen from Fig. 10 and Fig. 11 that the proposed sparse filter based single speaker DOA estimation method outperforms the GCC-SCOT [35] under various noisy and reverberant conditions except for the azimuth angle $105°$. In this case, the peak resolution is low and peak delay is small because the azimuth angle is close to $90°$ and the distance from source to microphone is short (1 m in this setup). With increasing amount of reverberation and noise, the peak delay between the two sparse filters u_j ($j = 1, 2$) becomes harder to distinguish. Nevertheless, the performance of the estimated filters u_j ($j = 1, 2$) is better than GCC-SCOT on average (last panel in each of Fig. 10 and Fig. 11) and helped to achieve the goal of cross-channel cancellation.

C. Length of Duration D v.s. Enhancement Performance

The 50 pairs of synthetic mixtures of two sources with the azimuth angles $30°$ and $70°$ in the simulated room continue to be used in this experiment. The setting of filter length L according to the reverberant conditions is same as previous experiments. With different lengths of selected silent speech durations D, CSE achieves different enhancement qualities accordingly, seen in Fig. 12. Basically, the enhancement effect

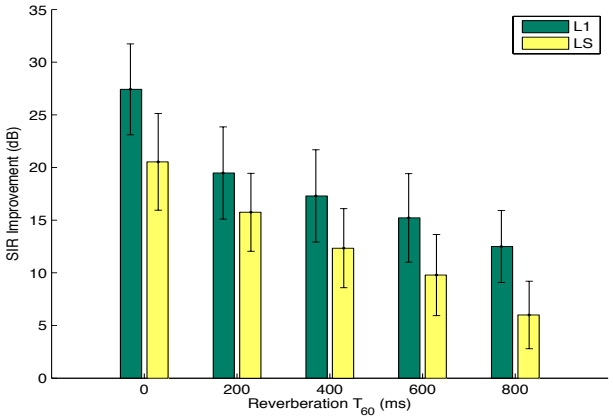


Fig. 9. *Comparison between (5) and our proposed CSE (6) in terms of SIR improvement. L1 stands for the proposed CSE model while LS stands for the least-squares framework.*

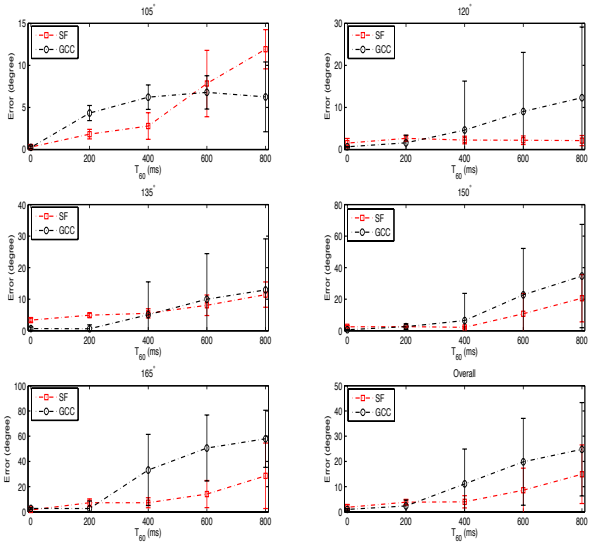


Fig. 11. *Average error and standard deviation of DOA estimation by two methods under various reverberations. SF stands for the DOA calculated by u^* and GCC stands for the GCC-SCOT method.*

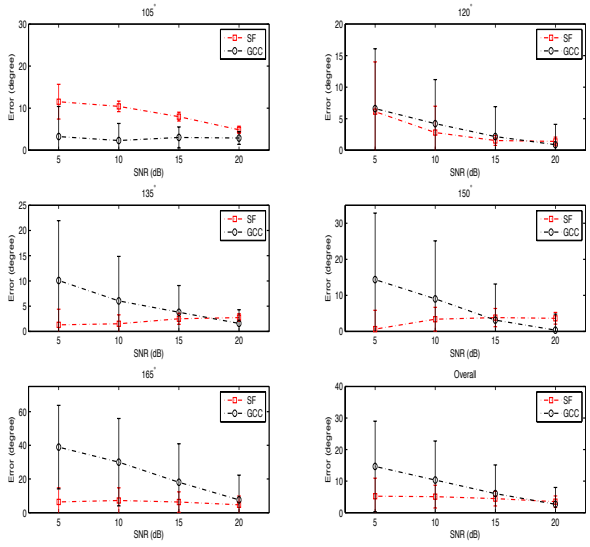


Fig. 10. *Average error and standard deviation of DOA estimation by two methods under various noisy conditions. SF stands for the DOA calculated by u^* and GCC stands for the GCC-SCOT method.*

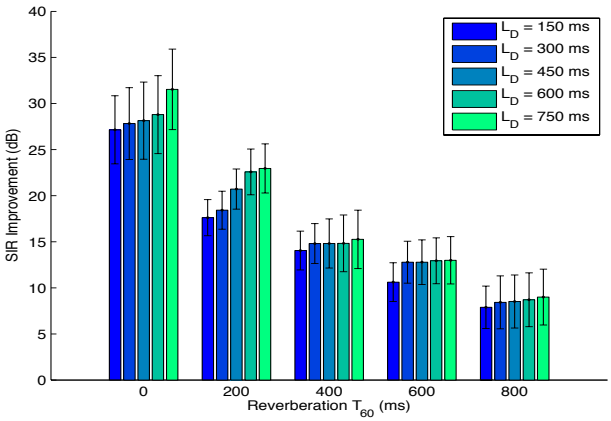


Fig. 12. *The relationship between average separation effect — SIR improvement (SIRI) and the length of selected silent speech duration with different reverberant conditions. The input SIR is -5 dB.*

is consistently improved with the increase of the silence region D (0.15 s, 0.30 s, 0.45 s, 0.60 s and 0.75 s). The enhancement reaches a plateau at 300 - 450 ms, which balances the computational speed and separation quality, while 150 msec silence region still achieves the satisfactory results. Since Au in (7) stands for the convolution of mixtures x_j and filters u_j ($j = 1, 2$) inside the duration D, small length of the duration D would be insufficient for computing convolution, which makes the solution u less optimal. As a result, the extremely small size of D degrades the enhancement performance.

D. Split Bregman v.s. Subdivided Split Bregman Method

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 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 computational costs. From Table I, $L = 800$ already achieves a good separation. Since the proposed method performs well if the estimate of spatial difference is accurate, the real impulse response filters are not necessarily fully resolved (unlike the dereverberation). In addition, the sparsity regularization encourages the solution to resolve the spatial difference by focusing on the direct path and early reverberation parts. Therefore, the filter length L is not

TABLE I
Comparison of the (subdivided) split Bregman algorithms

	Split Bregman			Subdivided Split Bregman (r=2)			Subdivided Split Bregman (r=4)		
L	Iteration	Time [s]	SIRI [dB]	Iteration	Time [s]	SIRI [dB]	Iteration	Time [s]	SIRI [dB]
100	**50**	**0.058**	6.214	50	0.531	6.221	50	0.695	**6.666**
200	**42**	**0.209**	6.766	43	0.796	6.776	44	1.134	**7.217**
400	44	**0.780**	8.069	43	1.565	8.111	**42**	2.218	**8.543**
800	62	4.386	9.107	**50**	**4.064**	9.195	50	5.370	**9.667**
1200	63	10.994	10.364	**41**	**7.019**	10.401	42	7.394	**10.888**
1600	71	21.684	11.379	66	14.820	11.265	**50**	**12.127**	**11.698**
2000	103	38.161	12.306	77	23.132	12.159	**76**	**22.475**	**12.598**
2800	-	-	-	**104**	48.245	12.984	111	**46.377**	**13.458**
3600	-	-	-	**123**	83.295	13.466	127	**72.020**	**13.937**

necessarily approaching the length of room impulse responses.

E. Comparison of L1 Optimization Algorithms

We compared several l_1 norm regularized optimization algorithms in terms of computation time and SIR improvement. Proximal forward backward splitting (PFBS) can be effectively applied to the proposed optimization problem (7). This splitting procedure, which goes back to [45], appears in many applications. Some examples include classical methods such as gradient projection and more recent ones such as the iterative thresholding algorithm FPC [46] and the framelet inpainting algorithm [47]. From Fig. 13, we see that for short sparse filters, PFBS and split Bregman algorithms are a little bit faster than the subdivided split Bregman algorithm ($r = 4$) since the computation time saved in the matrix inversion step by subdivided split Bregman algorithm is not significantly much. However, when the filter length L is larger than 1000, the subdivided split Bregman algorithm is the most efficient one. The SIR improvement increases as the filter length L goes up as shown in Fig. 14. The PFBS achieves a little higher SIRI than split Bregman and subdivided split Bregman algorithms, while it is dominated by split Bregman and subdivided split Bregman algorithms for L larger than 1000 or so.

F. Performance of The Speech Enhancement Framework

The 50 pairs of synthetic mixtures of two sources with the azimuth angles 30° (target source) and 70° (interfering source) in the simulated room continue to be used in this experiment. The setting of filter length L according to the reverberant conditions is same as previous experiments. In order to illustrate the enhancement quality of our proposed framework (6), we simplify the detection step by knowing roughly about 0.5 sec. silent duration D (e.g. 0 sec. - 0.5 sec. (2.3 sec. - 2.8 sec.) for the speech source in the up-left (up-right) panel of Fig. 5) of target speech source ahead of time. The other source is either speech or background music. The average output SIRs achieved by CSE under the various reverberant conditions and input SIRs are shown in Fig. 15.

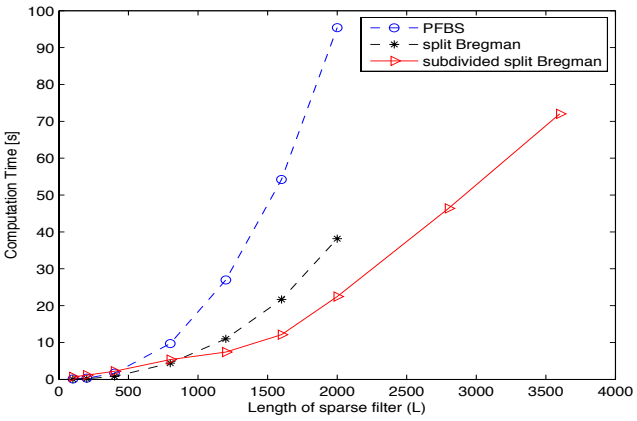


Fig. 13. *Comparison of computation time among L1 optimization algorithms. Split Bregman and PFBS run out of memory once L is above 2000 taps.*

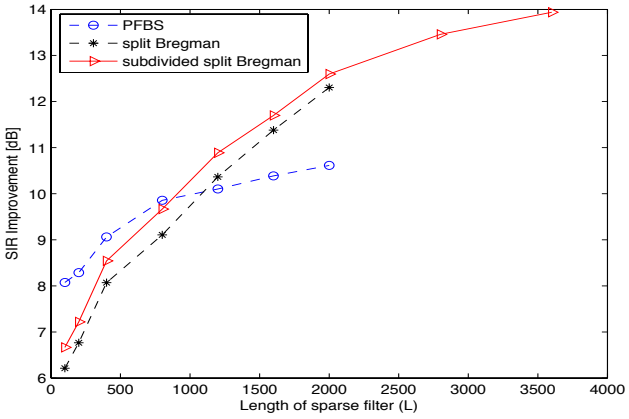


Fig. 14. *Comparison of SIR improvement among L1 optimization algorithms. Split Bregman and PFBS run out of memory once L is above 2000 taps.*

G. Activity Detection Performance

The performance of the proposed silence detection method (or detection of silent duration D) in section IV is evaluated under various reverberant conditions by computing the Recall

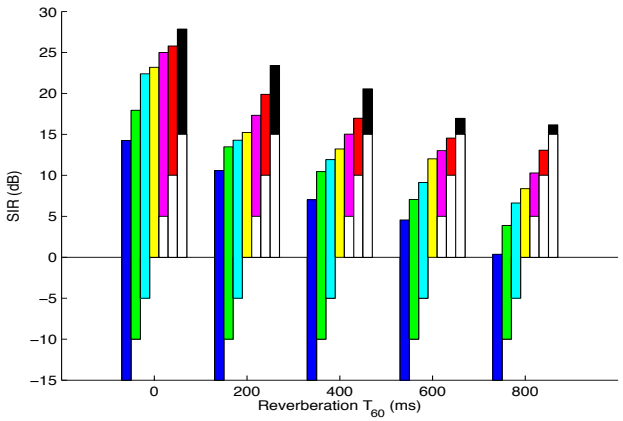


Fig. 15. *Output SIR vs. input SIR for the proposed CSE method under different reverberant conditions. The bottom of each colored bar is the input SIR while the top of each colored bar indicates the output SIR. Color bars in each reverberant condition indicate different input SIRs (-15 dB for Blue, -10 dB for Green, -5 dB for Cyan, 0 dB for Yellow, 5 dB for Purple, 10 dB for Red and 15 dB for Black).*

Rate R_r, the Precision Rate R_p, the False Alarm Rate R_f, and the F-measure as defined below. As the F-measure is the harmonic mean of the precision rate and recall rate, this score can be considered as the overall measure of the test. In addition, since the CSE method expects high correct detection rate while target missing is not a serious problem, the Precision Rate R_p provides a good reference to the whole speech enhancement algorithm.

$$R_r \triangleq \frac{Number\ of\ Correctly\ Detected\ SS\ Frames}{Total\ Number\ of\ SS\ Frames}$$

$$R_p \triangleq \frac{Number\ of\ Correctly\ Detected\ SS\ Frames}{Number\ of\ Detected\ SS\ Frames}$$

$$R_f \triangleq \frac{Number\ of\ Incorrectly\ Detected\ SS\ Frames}{Total\ Number\ of\ NS\ Frames}$$

$$F \triangleq \frac{2R_rR_p}{R_r + R_p},$$

where SS Frames stands for the *target source silent* frames, the rest frames are denoted as NS frames.

TABLE II
Single source silent frame detection rate (%) for the synthetic mixtures of two sources under various reverberant conditions. (SIR = 0 dB)

Reverberation	R_r	R_p	R_f	F
Anechoic	92.57	**90.88**	5.20	**91.71**
200 ms	88.48	**88.81**	6.24	**88.64**
400 ms	87.36	**86.08**	9.97	**86.72**
600 ms	85.87	**83.39**	9.56	**84.62**
800 ms	75.84	**80.63**	10.19	**78.16**

As shown in Table II and III (50 groups of synthetic mixtures of 2 source (3 sources) in 5 reverberant conditions are used for the evaluations in Table II (III)), the overall

TABLE III
Single source silent frame detection rate (%) for the synthetic mixtures of three sources under various reverberant conditions. (SIR = -5 dB)

Reverberation	R_r	R_p	R_f	F
Anechoic	93.04	**90.18**	7.37	**91.59**
200 ms	91.14	**85.21**	11.52	**88.07**
400 ms	89.29	**84.36**	11.54	**86.75**
600 ms	88.64	**80.77**	14.48	**84.52**
800 ms	82.81	**73.98**	17.85	**78.15**

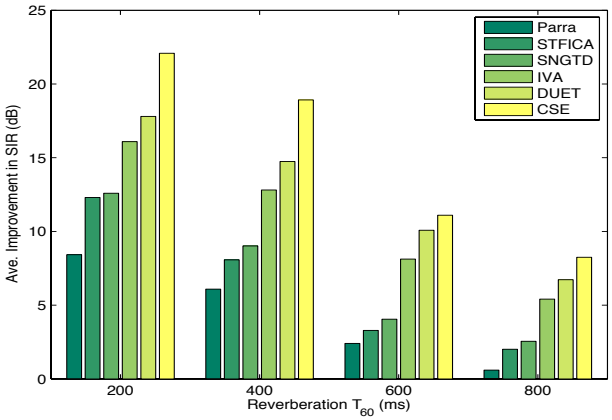


Fig. 16. *Comparison of average improvement in SIR among BSS methods*

detection accuracy (F) and the precision rate (R_p) achieve 90% under low reverberant conditions, 85% or more under medium reverberant conditions and around 80% as well under high reverberant conditions. The whole CSE algorithm relies on the good performance of the source activity detection.

H. Methods Comparisons

The comparison of a list of existing BSS methods is shown in Fig. 16 and 17 in terms of average increase in SIR, and SDR (signal to distortion ratio) [44]. These methods include Parra's decorrelation based method (Parra, [37]), spatial-temporal fast ICA (STFICA, [13]), scaled natural gradient method (SNGTD, [12]), independent vector analysis method (IVA, [38]) and DUET method. Test is based on the 50 pairs of synthetic mixtures of two speech signals used in previous subsections, with the azimuth angles 30° and 70° in the simulated room. CSE with automatic source activity detection enhances each speech source of the two as other BSS methods do. The source activity detection is first applied to detect the single silent frame for each of the two sources. Then the CSE method enhances the two sources sequentially. The filter length L is 256, 512, 1024 and 1024 for the four reverberant conditions respectively (from low to high). Fig. 16 and 17 indicate that the proposed CSE achieves the best separation quality in objective measures.

Room recorded mixtures are used to evaluate and compare the above BSS methods by the Perceptual Evaluation of Speech Quality (PESQ) [39], shown in Table IV. The speech source activity detection is completed within 2 to 3 seconds, and does not affect the efficiency of the CSE method.

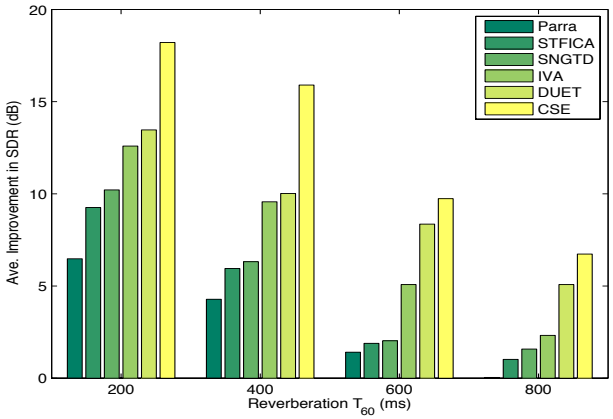


Fig. 17. *Comparison of average improvement in SDR among BSS methods*

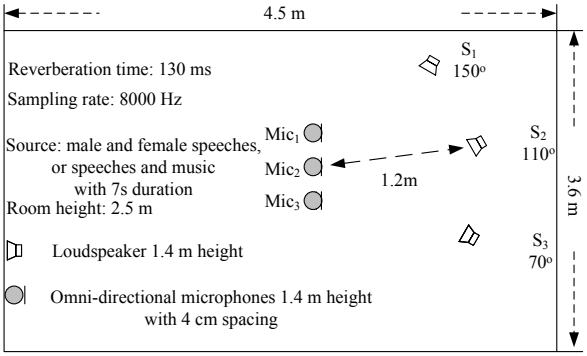


Fig. 18. *Configuration and parameters of the room recording. For the two sensors and two sources case, sources come from speaker S_1 and S_2, and Mic_2 and Mic_3 are turned on; while for the case of three sensors and three sources, all the speakers and microphones in the room are included.*

From the above objective evaluations, IVA, DUET and CSE lead other approaches. For further study, we investigate these three approaches by subjective test. Enhanced speech sources by the three methods are evaluated by 10 human subjects with normal hearing. Each subject is assigned with 30 enhanced utterances by three methods (total 90 utterances) from the above room recording dataset. We exploited the paired comparison (PC) test, which was used in [43] requiring each listener to rank the three methods according to the performance of separation quality and sound clarity. The preference percentages of our method to the other two methods are shown in Table V, and they are calculated as

$$PC_{>} = \frac{\#\ of\ pairs\ where\ CSE\ is\ better}{\#\ of\ all\ pairs\ in\ the\ test}$$
$$PC_{<} = \frac{\#\ of\ pairs\ where\ CSE\ is\ worse}{\#\ of\ all\ pairs\ in\ the\ test}$$
$$PC_{\approx} = \frac{\#\ of\ pairs\ where\ difference\ is\ not\ significant}{\#\ of\ all\ pairs\ in\ the\ test}$$

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

TABLE IV

Average PESQ of BSS methods on real recording mixtures. PRE PESQ is the average PESQ of the mixture data. Time for CSE is shown as detection time + speech enhancement time.

	2 sources (time[s])	3 sources (time[s])
PRE PESQ	1.37	1.00
Parra	1.57 (7.9)	1.44 (16.0)
STFICA	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)
CSE	2.58 (1.9+2.4)	2.15 (2.3+3.8)

TABLE V

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 clarity respectively, while $\approx$ means "hard to distinguish".

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

VI. Discussion and Conclusion

We proposed and evaluated a fast and efficient blind speech enhancement method as long as speeches make pauses. Based on the spatial difference model for cross-channel cancellation, a convex optimization problem is formulated and solved by the split Bregman method and the proposed subdivided 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 computational speed and separation quality. The proposed CSE is carried out on time domain. As discussed in the paper, with the sparsity regularization, the length of cancellation filter is not necessary to be as long as that of real impulse response for the purpose of resolving the spatial difference. In order to further speed up the algorithm, subband processing is to be evaluated. By decomposing the speech signals into different subbands, and downsampling accordingly, the length of cancellation filters in each subband is $\frac{L}{\gamma}$, where γ is the downsample factor. Signals in each subband are enhanced by CSE simultaneously, after which the enhanced time-domain signal is reassembled from the enhanced subband signals. Since each enhanced track differentiates with other tracks by different silent durations D's used in optimization, there does not exist permutation problem for the proposed CSE with subband processing. However, the CSE method with subband processing shares a common limitation, scaling ambiguity, as other subband BSS methods. This remains to be studied in our future work.

ACKNOWLEDGMENT

The authors would like to thank Yang Wang for helpful discussions. We appreciate the reviewers' constructive comments and helpful suggestions for this paper, and are also grateful for the editor's coordination of the thoughtful reviews.

REFERENCES

[1] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in Springer handbook on Speech Processing and Speech Communication, 2007.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomputing, vol. 22, pp. 21-34, July, 1998.

[3] H.Sawada,S.Araki,and S.Makino, "Frequency-domain blind source separation," in Blind Speech Separation, pp. 47-78, Springer, London, UK, 2007

[4] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," IEEE Transactions on Speech and Audio Processing, vol. 11, no. 2, pp. 109-116, 2003.

[5] L. Wang, H. Ding and F. Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," EURASIP Journal on Audio, Speech and Music Processing, 2010.

[6] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in Proceedings of the IEEE Interna- tional Symposium on Circuits and Systems (ISCAS ?7), pp. 3247-3250, May 2007.

[7] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," IEEE Transactions on Speech and Audio Processing, vol. 12, no. 5, pp. 530-538, 2004.

[8] F. Nesta, M. Omologo, and P. Svaizer, "Separating short signals in highly reverberant environment by a recursive frequency-domain BSS," in Proceedings of the 2008 Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA 2008), 2008.

[9] S. C. Douglas and X. Sun, "Convolutive blind separation of speech mixtures using the natural gradient," Speech Communication, vol. 39, no. 1-2, pp. 65-78, 2003.

[10] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments", Signal Processing, vol. 86, no. 6, pp. 1260-1277, 2006.

[11] H. Buchner, R. Aichner, W. Kellermann, W., "TRINICON-based blind system identification with application to multiple-source localization and separation", in Blind Speech Separation, Makino, S., Lee, T.-W., and Sawada, H., Eds., 2007, pp. 101-147.

[12] S.C. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," in ICASSP, Apr. 2007, vol. II, pp. 637-640.

[13] S. C. Douglas, M. Gupta, H. Sawada and S. Makino, "Spatio-temporal fastICA algorithms for the blind separation of convolutive mixtures," IEEE Trans. Audio, Speech, and Language Processing, vol. 15, no. 5, pp 1511-1520, 2007.

[14] S. Makino, "Blind source separation of convolutive mixtures," In Proceedings of The International Society for Optical Engineering, Kissimmee, FL, USA, 2006.

[15] G. Xu, H. Liu, L. Tong and T. Kailath, "A least-squares approach to blind channel identification", IEEE Trans. Audio, Speech and Language Processing, vol. 43, no. 12, pp 2982-2993, 1995.

[16] Y. Lin, J. Chen, Y. Kim and D. Lee, "Blind channel identification for speech dereverberation using l_1 norm sparse learning," Advances in Neural Information Processing Systems, 2007.

[17] D. van Compernolle, "Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings", ICASSP, pp 833-836, 1990.

[18] T. Goldstein and S. Osher, "The split Bregman algorithm for L1 regularized problems", *SIAM J. Imaging Sci.*, 2(2), 323-343, 2009.

[19] 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.

[20] Y. Lin, "L1 norm sparse Bayesian learning: theory and applications", Ph.D. Thesis, University of of Pennsylvania, 2008.

[21] T. Yoshioka, T. Nakatani, M. Miyoshi, M., "An integrated method for blind separation and dereverberation of convolutive audio mixtures", in Proceedings of the 16th European Signal Processing Conference (EUSIPCO 2008), 2008.

[22] Y. Huang, J. Benesty, J. Chen, J., "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 882-895, 2005.

[23] Y. Huang and J. Benesty, "Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization," Adaptive Signal Processing, Springer, 2003.

[24] J. Allen, D. Berkley. "Image method for efficiently simulating small-room acoustics," J. Acoustical Society America, 65:943-950, 1979.

[25] D. Duttweiler. "Proportionate normalized least-mean-squares adaptation in echo cancelers," IEEE Trans. Speech Audio Processing, 8:508-518, 2000.

[26] Y. Wang, Z. Zhou, "Background suppression in audio through learning," in preparation, 2010.

[27] 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.

[28] 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.

[29] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for l_1-minimization with application to compressed sensing," *SIAM J. Imaging Sci.*, 1(1):143-168, 2008.

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

[31] 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.

[32] K. Yamamoto, F. Asano, T. Yamada and N. Kitawaki, "Detection of overlapping speech in meetings using support vector machines and support vector regression", IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 89(8): 2158-2165, 2006.

[33] 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.

[34] W. Ma, M. Yu, J. Xin and S. Osher, "Reducing musical noise in blind source separation by time-domain sparse filters and split Bregman method," Interspeech, pp. 402-405, 2010.

[35] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay", IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 24, no. 4, pp. 320-327, 1976.

[36] J. Chen , J. Benesty and Y. Huang, "Time delay estimation in room acoustic environments: An overview", EURASIP J. Appl. Signal Process., vol. 2006, p. 7, 2006.

[37] 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.

[38] T. Kim, H. Attias, S-Y Lee, and T-W Lee, "Blind source separation exploiting higher-order frequency dependencies", IEEE Trans. Audio, Speech and Language Processing, vol. 15, no. 1, pp 70-79, 2007.

[39] 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.

[40] 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.

[41] D. R. Campbell, K. J. Palomaki, and G. Brown, "A MATLAB simulation of "shoebox" room acoustics for use in research and teaching", Computing and Information Systems J. 9, 2005.

[42] IEEE Subcommittee, "IEEE Recommended Practice for Speech Quality Measurements," IEEE Trans. Audio and Electroacoustics, AU-17(3), 225-246, 1969.

[43] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Blind sparse source separation with spatially smoothed time-frequency masking," in IWAENC, Paris, France, 2006.

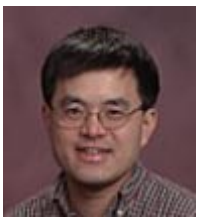
[44] C. Fvotte, R. Gribonval, and E. Vincent, "BSS_EVAL toolbox user guide–Revision 2.0," Tech. Rep. 1706, IRISA, Rennes, France, April 2005.

[45] P. L. Lions, and B. Mercier, "Algorithms for the sum of two nonlinear operators", SIAM Journal on Numerical Analysis, Vol. 16, No. 6, pp. 964-979, 1979.

[46] E. Hale, W. Yin, and Y. Zhang, "A fixed-point continuation method for l_1-regularized minimization with applications to compressed sensing", CAAM Technical Report TR07-07, 2007.
[47] J. F. Cai, R. H. Chan, and Z. Shen, "A framelet-based image inpainting algorithm", Applied and Computational Harmonic Analysis, Vol. 24, Issue 2, pp. 131-149, 2008.

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 processing.



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 Ph.D. 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.