SOURCE EXTRACTION IN AUDIO VIA BACKGROUND LEARNING

YANG WANG AND ZHENGFANG ZHOU

Abstract. Source extraction in audio is an important problem in the study of blind source separation (BSS) with many practical applications. It is a challenging problem when the foreground sources to be extracted are weak compared to the background sources. Traditional techniques often do not work in this setting. In this paper we propose a novel technique for extracting foreground sources. This is achieved by an interval of silence for the foreground sources. Using this silence interval one can learn the background information, allowing the removal or suppression of background sources. Very effective optimization schemes are proposed for the case of two sources and two mixtures.

1. Introduction

Blind source separation (BSS) is a major area of research in signal processing with a vast literature, particularly in audio signal processing. It aims to separate source signals from their mixtures without assuming detailed knowledge about the sources and the mixing process. One particular application of BSS in audio is to extract a desired audio source from mixtures involving noise, background or unwanted sources. In many practical applications such as mobile and car phones the background may in fact be stronger, even much stronger sometimes, than the desired signal. This can pose a daunting challenge.

The basic setup for BSS in audio has n audio sources $S_1(t), \ldots, S_n(t)$ and m mixtures $X_1(t), \ldots, X_m(t)$, with the model

$$X_k(t) = \sum_{i=1}^{L} \sum_{j=1}^{N_{k,i}} a_{k,i,j} S_i(t - d_{k,i,j}), \qquad k = 1, \ldots, M \tag{1.1}$$

where $a_{k,i,j}$ are the mixing coefficients and $d_{k,i,j}$ are the delays from the source i to the recording or sensing device k. The multiple delays are a result of reverberations in the ambient environment. In audio applications these mixtures are recordings made by placing m microphones in various locations. In BSS these coefficients are unknown. With the

Key words and phrases. Background audio source removal, blind source separation, convolution, reverberation, audio source cancellation, learning, quadratic programming.

The first author is supported in part by the National Science Foundation grant DMS-0813750.

presence of reverberations we often refer to the mixtures X_k in (1.1) as *convolutive mixtures*. A more concise way to formulate the model (1.1) is

$$X_k = \sum_{i=1}^{L} A_{k,i} * S_i, \qquad k = 1, \ldots, M \tag{1.2}$$

where $*$ denotes convolution and

$$A_{k,i}(t) = \sum_{j=1}^{N_{k,i}} a_{k,i,j}\delta(t - d_{k,i,j}) \tag{1.3}$$

are the convolutive kernels that represent the mixing of both the direct signals and their reverberations. BSS techniques aim to compute these convolutive kernels $A_{k,i}$.

The bulk of the studies in BSS employ the *independent component analysis (ICA)* model, where the source signals S_k, modeled as random variables, are assumed to be independent, see the reviews [3, 5] and the references therein for more details. Under this model, many techniques such as Joint Approximate Diagonalization Eigenmatrices (JADE) [2] and Information Maximization (Infomax) [1, 4], as well as their refinements, have been developed. An alternative technique is the *Degenerative Unmixing and Estimation Technique (DUET)*, which uses time-frequency separations to achieve BSS and has the advantage that it can work in the degenerative case $M < L$, see e.g. [6, 9]. All these techniques have their respective strengths and weaknesses, which are well documented in the literature so we will not go into details here. While these techniques can yield good results in BSS, they all have certain limitations that pose challenges for many source extraction and background removal applications. One of the major challenges is that they do not work well for source extraction if the desired source signals are very weak in comparison to the unwanted sources in the mixtures. Computational cost can be another issue for many practical applications such as mobile and car phone where removing background sources must be done in real time. The batch processing techniques used by many of the ICA techniques do not adapt well to dynamic real-time processing.

In source extraction problems we are given mixtures that contain several sources. We shall call those sources that we wish to extract *foreground sources* and those we wish to remove *background sources*. In this paper we propose a novel method for removing or suppressing background sources in audio mixtures, thus allowing us to bring out or enhance the foreground sources. The method requires that the foreground sources have an interval of

silence, i.e. during an interval of time the foreground sources are inactive. This interval does not need to be long. For audio signals often less than 1 second is sufficient but optimal result is achieved around 2 seconds. In most relevant applications such as mobile or car phone this requirement is not an issue. The main idea is that the convolution kernels $A_{k,i}$ in the model (1.2), although they depend on numerous factors such as building material and shape, are determined entirely by the ambient environment as long as the background sources remain stationary relative to the ambient environment. The interval of silence allows us to learn the ambient environment and estimate the convolutive kernels, from which background sources can be removed by cancelation.

2. Background Cancellation via Learning

In practice the time is discretized through sampling. It is convenient to write the convolutive mixtures in the discrete form:

$$X_k(t) = \sum_{i=1}^{L} \sum_{j=0}^{N} a_{k,i,j} S_i(t-j), \qquad k = 1, \dots, M, \tag{2.1}$$

where $t \in \mathbb{Z}$ and $T_0 \le t < T_1$. The integer N denotes the maximal delay in the mixture, which depends on the ambient environment and the sampling rate. We shall view a signal as an element in $l^\infty(\mathbb{Z})$. We shall adopt some standard notations here: For each $\mathbf{u} =\in l^\infty(\mathbb{Z})$ we use $\mathbf{u}(j)$ to denote its j-th entry and $\operatorname{supp}(\mathbf{u})$ its support, i.e. the set of indices of its nonzero entries. For and $\mathbf{u}, X \in l^\infty(\mathbb{Z})$ where $\mathbf{u}$ is finitely supported the convolution $\mathbf{u} * X \in l^\infty(\mathbb{Z})$ is defined by

$$\mathbf{u} * X(t) = \sum_{j \in \mathbb{Z}} \mathbf{u}(j) X(t-j).$$

Again we may rewrite the model (2.1) simply as

$$X_k = \sum_{i=1}^{n-1} \mathbf{u}_{k,i} * S_i, \qquad k = 1, \dots, m \tag{2.2}$$

where $\mathbf{u}_{k,i} \in l^\infty(\mathbb{Z})$ is supported on $[0, N]$ with $a_{k,i,j}$ as its j-th element.

Suppose that the sources $S_1, \dots, S_J$ are the background sources we wish to remove and $S_{J+1}, \dots, S_L$ are the foreground sources we wish to extract from the mixtures X_k, $1 \le k \le$

M. For any finitely supported $\mathbf{b}_k \in l^\infty(\mathbb{Z})$ we have

$$\mathbf{b}_k * X_k = \sum_{i=1}^{L} \mathbf{b}_k * \mathbf{u}_{k,i} * S_i.$$

It follows from (2.2) that

$$\sum_{k=1}^{M} \mathbf{b}_k * X_k = \sum_{i=1}^{L} \Big(\sum_{k=1}^{M} \mathbf{b}_k * \mathbf{u}_{k,i} \Big) * S_i. \tag{2.3}$$

If we can find $\mathbf{b}_k$ such that $\sum_{k=1}^{M} \mathbf{b}_k * \mathbf{u}_{ki} = 0$ for $1 \le i \le J$, then

$$\sum_{k=1}^{M} \mathbf{b}_k * X_k = \sum_{i=J+1}^{L} \Big(\sum_{k=1}^{M} \mathbf{b}_k * \mathbf{u}_{k,i} \Big) * S_i. \tag{2.4}$$

The unwanted background sources are now removed, leading to the extraction of foreground sources. We shall call those $\mathbf{b}_k$ *cancellation kernels*. Note that although using this technique the foreground sources are now subject to further convolutions, in general it does not degrade the foreground signals as long as the maximum delay N is not too large and the cancellation kernels not too dense. A human auditory system does not appear to be very sensitive to moderate convolutions. Furthermore, moderate convolutions do not introduce any artifacts to the extract source signals. This is a definite advantage over ICA based techniques or DUET.

One question is whether the cancellation kernels exist in general. We shall show that they do in all non-degenerative setting $M > J$. For each $X \in l^\infty(\mathbb{Z})$ we associate it with a symbolic Fourier series

$$\widehat{X}(\xi) = \sum_{n \in \mathbb{Z}} X(n) e_{-n}(\xi),$$

where $e_b(\xi) := e^{2\pi i b \xi}$. We shall refer to this series simply as the Fourier transform of X. Note that if $\mathbf{u} \in l^\infty(\mathbb{Z})$ has finite support than $\widehat{\mathbf{u}}(\xi)$ is a trigonometric polynomial. It is well known that $\widehat{\mathbf{u} * X} = \widehat{\mathbf{u}}\, \widehat{X}$. Furthermore, $\widehat{\tau_q X} = e_q(\xi) \widehat{X}(\xi)$ where $\tau_q X$ denotes X shifted to the right by q positions.

Proposition 2.1. *Let $\mathbf{u}_{k,i} \in l^\infty(\mathbb{Z})$ for $1 \le k \le M$ and $1 \le i \le J$ have finite support. Suppose that $M > J$. Then there exist finitely supported cancellation kernels $\mathbf{b}_1, \dots, \mathbf{b}_M \in l^\infty(\mathbb{Z})$ not all zero such that*

$$\sum_{k=1}^{M} \mathbf{b}_k * \mathbf{u}_{k,i} = 0, \qquad i = 1, \dots, J.$$

Proof. Taking the Fourier transform we need to show the existence of finitely supported $\mathbf{b}_k$ such that

$$\sum_{k=1}^{M} \widehat{\mathbf{b}}_k(\xi)\widehat{\mathbf{u}}_{k,i}(\xi) = 0, \qquad i = 1, \ldots, J.$$

Let $\mathcal{F}$ be the field of all real trigonometric rational functions, i.e. functions of the form f/g where both f, g are trigonometric polynomials with real coefficients. Set $\mathbf{A} = [\widehat{\mathbf{u}}_{k,i}]$ with rows indexed by i and columns indexed by k, which is $J \times M$. Since $M > J$ there exists a $Z = [f_1, \ldots, f_M]^T$ in $\mathcal{F}^M$ such that $\mathbf{A}Z = 0$. Now let $F(\xi)$ be a trigonometric polynomial that is the common denominator of all trigonometric rational functions f_k, $1 \le k \le M$ and $G_k(\xi) = F(\xi)f_k(\xi)$. Each G_k is a trigonometric polynomial with real coefficients. Let $\mathbf{b}_k \in l^\infty(\mathbb{Z})$ such that $\widehat{\mathbf{b}}_k = G_k$. Then

$$\sum_{k=1}^{M} \mathbf{b}_k * \mathbf{u}_{k,i} = 0, \qquad i = 1, \ldots, J.$$

∎

Observe that if $\mathbf{b}_1, \ldots, \mathbf{b}_M$ are cancellation kernels then so are $\tau_q\mathbf{b}_1, \ldots, \tau_q\mathbf{b}_M$ for any q. By shifting we can thus normalize the cancellation kernels so that all supp $\mathbf{b}_k$ are nonnegative and at least one $\mathbf{b}_k(0) \neq 0$. We shall call such cancellation kernels *normalized*.

The general framework for foreground source extraction we propose in this paper is to compute the cancellation kernels via background learning. This is achieved by utilizing an interval of silence for the foreground sources we wish to extract. Once we have the cancellation kernels the extraction of foreground can be made through background cancellations. Assume that the foreground sources $S_{J+1}(t), \ldots, S_L(t)$ are silent in the time interval $a \le t \le b$, i.e. $S_{J+1}(t) = \cdots = S_L(t) = 0$. Let $\mathbf{b}_1, \ldots, \mathbf{b}_M$ be the cancellation kernels. It follows from (2.4) that

$$\sum_{k=1}^{M} \mathbf{b}_k * X_k = 0$$

for $a + N \le t \le b$. Thus we can use this interval to learn the background and estimate the normalized cancellation kernels by minimizing the cost function

$$F(\mathbf{b}_1, \ldots, \mathbf{b}_M, a, b) := \sum_{a+N \le t \le b} \left| \sum_{k=1}^{M} \mathbf{b}_k * X_k(t) \right|^2, \tag{2.5}$$

subject to the constraint $\sum_{k=1}^{M} |\mathbf{b}(0)| = 1$. Once the cancellation kernels are obtained, foreground sources can be extracted in real time by (2.4) through simple convolutions

$$\sum_{k=1}^{M} \mathbf{b}_k * X_k = \sum_{i=J+1}^{L} \Big(\sum_{k=1}^{M} \mathbf{b}_k * \mathbf{u}_{k,i}\Big) * S_i.$$

Unfortunately, our numerical experiments have shown that without further constraints the cancellation kernels obtained by minimizing E do not work well. There could be several problems associated with simply minimizing the cost function. One such problem is the lack of uniqueness even when all cancellation kernels are normalized. But since any cancellation kernels can be used to remove the background sources, this may not be a major problem. A more serious problem is *over-fitting*. One possible solution to overcome the over-fitting problem is to impose sparsity on the minimizers. In our experiments the minimizers of (2.5) are almost never sparse, particularly for real life recorded mixtures. There are many ways to achieve sparsity, such as adding an l^1-norm penalty term, see e.g. [7] and the references therein. The problem with this approach is that it adds a penalty term that isn't natural, which can affect the performance. Finding the right sparsity condition to impose is the most important aspect of this approach to source extraction, and at this point is still a work in progress. In the case of two mixtures and two sources, we propose a simple solution that works extremely well in numerical experiments.

Actually in the case of two sources and two mixtures, i.e. $L = M = 2$, there is a simple and elegant way to find the cancellation kernels $\mathbf{b}_1, \mathbf{b}_2$. They can be taken as $\mathbf{b}_1 = \mathbf{u}_{2,1}$ and $\mathbf{b}_2 = -\mathbf{u}_{1,1}$, which yield

$$\mathbf{b}_1 * \mathbf{u}_{1,1} + \mathbf{b}_2 * \mathbf{u}_{2,1} = 0.$$

It follows that

$$\mathbf{b}_1 * X_1 + \mathbf{b}_2 * X_2 = (\mathbf{b}_1 * \mathbf{u}_{1,2} + \mathbf{b}_2 * \mathbf{u}_{2,2}) * S_2 = (\mathbf{u}_{2,1} * \mathbf{u}_{1,2} - \mathbf{u}_{1,1} * \mathbf{u}_{2,2}) * S_2.$$

It is not hard to show that unless the Fourier transform $\widehat{\mathbf{u}}_{1,1}, \widehat{\mathbf{u}}_{2,1}$ have a common factor, $\mathbf{u}_{2,1}, -\mathbf{u}_{1,1}$ will also be the shortest cancellation kernels. learning the cancellation kernels in this case is equivalent to learning the convolutive mixing kernels. Assume that the foreground source $S_2(t) = 0$ in the time interval $a \le t \le b$. Let

$$Y := \mathbf{u}_{2,1} * X_1 - \mathbf{u}_{1,1} * X_2.$$

Then $Y(t) = 0$ for $a + N \le t \le b$. Let $\mathbf{u}_{2,1}(n) = x_n$ and $\mathbf{u}_{1,1}(n) = y_n$ for $0 \le n \le N$ and 0 otherwise. Denote $\mathbf{x} = [x_0, x_1, \ldots, x_N]^T$ and $\mathbf{y} = [y_0, y_1, \ldots, y_N]^T$. We estimate $\mathbf{u}_{2,1}, \mathbf{u}_{1,1}$ by minimizing

$$\begin{aligned} E(\mathbf{x}, \mathbf{y}, a, b) &:= \sum_{a+N \le t \le b} |Y(t)|^2 \\ &= \sum_{t=a+N}^{b} \Big(\sum_{j=0}^{N} (x_j X_1(t-j) - y_j X_2(t-j)) \Big)^2 \\ &= \mathbf{x}^T A \mathbf{x} + \mathbf{y}^T B \mathbf{y} - 2\mathbf{x}^T C \mathbf{y}, \end{aligned} \tag{2.6}$$

where $A = [a_{ij}], B = [b_{ij}], C = [c_{ij}]$ are $(N+1) \times (N+1)$ matrices given by

$$\begin{aligned} a_{ij} &= \sum_{t=a+N}^{b} X_1(t-i) X_1(t-j), \\ b_{ij} &= \sum_{t=a+N}^{b} X_2(t-i) X_2(t-j), \\ c_{ij} &= \sum_{t=a+N}^{b} X_1(t-i) X_2(t-j) \end{aligned}$$

with $0 \le i, j \le N$. To address the over-fitting problem we propose to imposing an additional constraint that requires both $\mathbf{x} \ge 0$ and $\mathbf{y} \ge 0$, i.e. $x_j \ge 0$ and $y_j \ge 0$ for all j. Beside making the minimizers more sparse, an extra benefit of this additional constraint is that now the nonlinear constraint $\sum_{k=1}^{M} |\mathbf{b}(0)| = 1$ is reduced to the linear constraint $x_0 + y_0 = 1$, making the computation more efficient. Now $\mathbf{x}, \mathbf{y}$ are computed by solving the following quadratic programming problem:

$$\min_{(\mathbf{x},\mathbf{y})} \ \mathbf{x}^T A \mathbf{x} + \mathbf{y}^T B \mathbf{y} - 2\mathbf{x}^T C \mathbf{y} \qquad \text{Subject to} \qquad \mathbf{x} \ge 0, \ \mathbf{y} \ge 0, \ x_0 + y_0 = 1, \tag{2.7}$$

where A, B, C are as in (2.6). Furthermore, it makes sense in this particular setting. Although the kernels $\mathbf{u}_{k,i}$ do not have to be nonnegative in general, negative coefficients usually correspond to much weaker signals, and for the purpose of removing or suppressing background sources their impacts are limited. Our experiments show that the nonnegativity constraint yields superb results in the case of two sources and two mixtures and it is very robust. Numerical results will be shown in the next section.

It should be pointed out that the nonnegativity constraint only works in the case of two sources and two two mixtures. With more sources and mixtures the nonnegativity constraint no longer makes sense, and new methods will be needed.

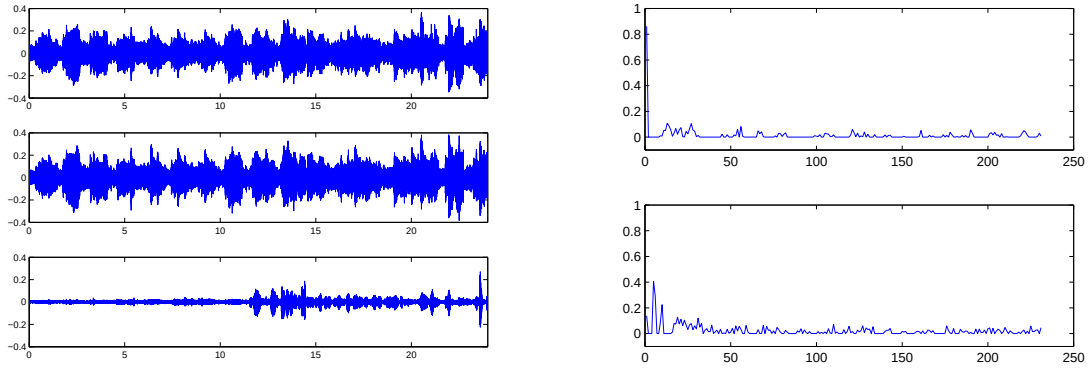
3. Numerical Examples and Future Work

We present two examples here to show the effectiveness of the algorithm for removing or suppressing background sources. Both examples are real recordings made in a regular office that had moderate reverberations at sampling rate 16kHz.

Example 1. In this example a speech is mixed with very loud background music from a boom box. Both the speaker and the boombox are about 1.5 meters from the two microphones, which are about 20cm apart. The speech, which is the desired foreground source, is silent for the first few seconds, and it is completely overwhelmed by the background music. It is rather difficult to discern the speech completely. Applying our background removal algorithm we have successfully suppressed the loud music signal so that the speech can be heard very clearly without any discernable artifact. Figure 1(a) shows the plots of the mixtures and the output signal after suppression of the background music. As one can see, it is hard to see that the foreground source even exists from the plot. After the application of our algorithm the music can still be heard, but is now suppressed to the level that it is rather faint. The strength of the foreground source, by comparison, has not been reduced. In fact, it may have been amplified.

In this example, we have restricted the size of the cancellation kernels to $N = 230$. Larger N does not seem to yield substantial improvement. In fact, when N is greater than 500 the performance seems to get a bit worse, most likely due to the additional convolutions that make the reconstructed signal more "muffled." The first 2 seconds is used to learn the cancellation kernels. In all our testings, there is no benefit for using more than 2 seconds to compute the cancellation kernels. Figure 1(b) show the cancellation kernels that have been computed. As one can see, the significant coefficient are quite sparse.

Example 2. This example shows how the background suppression algorithm can also be used as an alternative algorithm for BSS. In this example the mixtures contain speeches from a male speaker and a female speaker of approximately equal strength. Both speakers are about 2 meters from the two microphones, which are placed about 15cm apart. The male



(a) The two mixtures and the extracted source

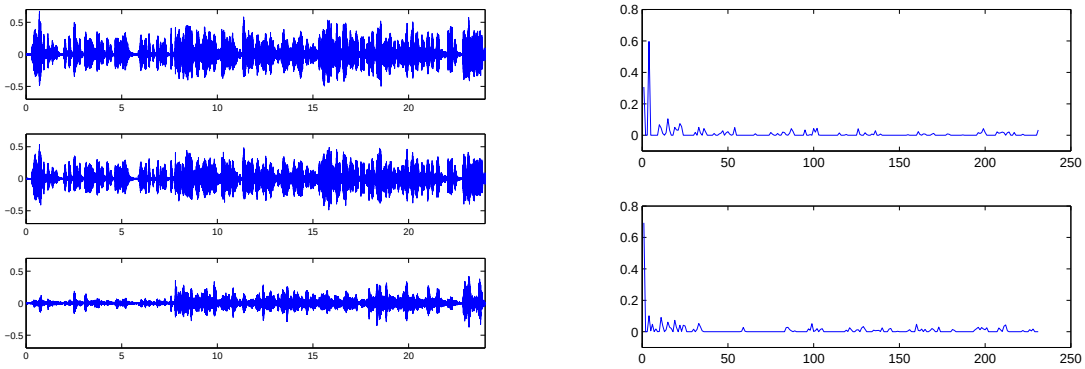
(b) The cancellation kernels

FIGURE 1

speaker, which is the foreground source, is silent for the first few seconds. The goal here is to test our algorithm as an alternative BSS method to extract the foreground source, i.e. the speech by the male speaker. Applying the interval of silence for the foreground source our background removal algorithm successfully removed the speech by the female speaker. The result is quite competitive versus other BSS algorithms. Again, the separated speech is very clear without any distortion and artifact, although some residue of the background source has remained, but it is not more than other algorithms. Figure 2(a) shows the plots of the mixtures and the separated signal, while Figure 2(b) show the cancellation kernels that have been computed, which are quite sparse as one can see. Like in the previous example, we have chosen $N = 230$ and used the first 2 seconds to compute the cancellation kernels.

While the approach of using interval of silence for background source signal removal is a novel and promising method, the quadratic programming approach with nonnegativity constraint that we use here no longer applies to cases where more than one background source need to be removed. As a plan for future work one idea is to use sparse principal component analysis (sparse PCA, see e.g. [8]) to obtain sparse cancellation kernels. Recently Yu et al [10] propose to use the split Bregman algorithm to compute sparse cancellation kernels, which is another promising solution.

The authors thank Xun Wang, Sean Wu and Na Zhu for very helpful discussions.



(a) The two mixtures and the extracted source

(b) The cancellation kernels

FIGURE 2

REFERENCES

[1] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[2] J.F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE proceedings-f*, 1993.

[3] S. Choi, A. Cichocki, H.M. Park, and S.Y. Lee. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews*, 6(1):1–57, 2005.

[4] A. Cichocki and S. Amari. *Adaptive blind signal and image processing: learning algorithms and applications*. Wiley, 2002.

[5] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[6] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING*, volume 5. IEEE; 1999, 2000.

[7] J. Liu, J. Xin, and Y. Qi. A dynamic algorithm for blind separation of convolutive sound mixtures. *Neurocomputing*, 72(1-3):521–532, 2008.

[8] Y. Wang and Q. Wu. Sparse PCA by iterative elimination algorithm. Preprint.

[9] Y. Wang, O. Yilmaz, and Z. Zhou. Phase aliasing correction for robust blind source separation using DUET. Preprint.

[10] M. Yu, W.-Y. Ma, J. Xin, and S. Osher. Convexity and fast speech extraction by split Bregman method. Preprint.

DEPARTMENT OF MATHEMATICS, MICHIGAN STATE UNIVERSITY, EAST LANISNG, MI 48824, USA.

E-mail address: ywang@math.msu.edu

DEPARTMENT OF MATHEMATICS, MICHIGAN STATE UNIVERSITY, EAST LANISNG, MI 48824, USA.

E-mail address: zfzhou@math.msu.edu