Convexity and Fast Speech Extraction by Split Bregman Method

Meng Yu[1], *Wenye Ma*[2], *Jack Xin*[1], *Stanley Osher*[2]

[1]Department of Mathematics, University of California, Irvine, USA
[2]Department of Mathematics, University of California, Los Angeles, USA
myu3@uci.edu, mawenye@math.ucla.edu, jxin@math.uci.edu, sjo@math.ucla.edu

Abstract

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

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

10.21437/Interspeech.2010-169

1. Introduction

Blind speech separation (BSS) aims to recover source signals from their mixtures without detailed knowledge of the mixing process. The time domain scaled natural gradient method [1] is often time consuming in the regime of long reverberations. Moreover, small divisors and divergence occur in silent durations of mixture signals. Though a nonlocally weighted soft constraint natural gradient method [2] resolves such issues and renders the method asymptotically consistent, it is still slow in convergence. Fundamentally, all time domain methods based on independent component analysis attempt to minimize non-convex objectives, for which no global convergence is mathematically guaranteed. Time-frequency domain method by spectral data clustering is very efficient and more speedy in separation because it does not resolve the impulse responses [7]. However, its binary sparseness hypothesis in the time-frequency domain deteriorates in reverberant conditions [7] and separation often comes with musical noise in the output.

In this paper, a new fast time domain speech extraction (FSE) method is proposed based on the assumption that the speech signal contains pauses. During silent durations of the speech signal, information of the interference (background) is collected and allows us to formulate a convex optimization problem for part of the impulse response functions which suffice to estimate the target speech. A sparse solution is then computed by l_1 regularization and the split Bregman method for which fast convergence was recently studied [6].

This paper is organized as follows. In section 2, the convex optimization problem for FSE is introduced. In section 3, computational framework by l_1 regularization and the split Bregman method is shown. In subsections 3.1 and 3.2, algorithms for moderately and highly reverberant acoustic environments are illustrated. The subdivided split Bregman method is proposed for FSE with long reverberations and large number of sources. In section 4, an onset-offset detection method of speech is outlined. In section 5, evaluations of FSE show its merits in both speed and separation quality in comparison with existing methods. Discussion and conclusions are in section 6.

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

2. Fast Speech Extraction Model

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

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

where t is time; $i = 1, 2$; and $*$ is linear convolution. Instead of finding an unmixing filter W such that $W * (x_1, x_2)$ recovers (s_T, s_B), we extract speech signal s_T by eliminating (not recovering) interference s_B. Suppose that the target speech contains pauses. Then there is a union D of disjoint time intervals where $s_T \approx 0$, while interference s_B is active. It follows from (1) that $h_{21} * x_1(t) - h_{11} * x_2(t) \approx 0$ for $t \in D$. The elimination by cross multiplication was known in blind channel identification [3] and background suppression [4]. Inside D, we seek a pair of sparse filters u_i $(i = 1, 2)$ to minimize the energy of $u_2 * x_1 - u_1 * x_2$ in the region D. Ideally, $u_1 \approx h_{11}$ and $u_2 \approx h_{21}$. Filter sparseness is achieved by l_1-norm regularization. The resulting convex optimization problem for $t \in D$ is:

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

where the second term $\frac{\eta^2}{2}(\sum_{i=1}^{2} u_i(1) - 1)^2$ is to fix scaling and prevent zero (trivial) solution. Denote the length of D by L_D and that of u_i by L. D can be as short as even 0.25 s' duration, which makes FSE method efficient on the data usage and different from other BSS methods that are based on the high order statistics of data. Since the solution u_i is l_1 regularized, the surplus length of it would be 0 while solving (2). In matrix form, convex objective (2) becomes:

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

Copyright © 2010 ISCA

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

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

When $t \notin D$, cross multiplication of (1) shows that $\hat{s}_T = u_2^* * x_1 - u_1^* * x_2 \approx h_{21} * x_1 - h_{11} * x_2 = (h_{21} * h_{12} - h_{11} * h_{22}) * s_T$. Interference s_B is eliminated and $\hat{s}_T$ sounds same as s_T to human ear. Here we assumed that the acoustic environment does not change much so that estimates of h_{11} and h_{21} during D still apply when $t \notin D$. For a convex objective with non-negativity filter constraints for sparsity, see [4].

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

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

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

3. Split Bregman Method

The split Bregman method was introduced in [6] as an efficient tool for solving optimization problems involving total variation or l_1 regularizations. It solves the unconstrained problem:

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

where J is convex but not necessarily differentiable such as the l_1 norm, H is convex and differentiable, and Φ is a linear operator. The key idea of the split Bregman method is to introduce an auxiliary variable $d = \Phi u$, and try to solve the constrained problem:

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

In [5, 6], it is proved that this kind of problem can be solved by the following iterations:

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

where $\langle \cdot, \cdot \rangle$ is the inner product. For simplicity, we introduce a new variable $b^k = p_d^k/\lambda$, and notice that $p_d^k = \lambda b^k$ and $p_u^k = -\lambda\Phi^T b^k$. Then d^{k+1} and u^{k+1} can be updated alternatively. The general split Bregman iteration is:

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

3.1. Split Bregman for Moderate Reverberations

In the case of (3), $J(u) = \mu||u||_1$, $\Phi = I$, and $H(u) = \frac{1}{2}||Au - f||_2^2$. The iterations are

$$d^{k+1} = \arg\min_d \frac{\mu}{\lambda}||d||_1 - \langle b^k, d - d^k\rangle + \frac{1}{2}||d - u^k||_2^2 \quad (4)$$

$$\begin{aligned}u^{k+1} =& \arg\min_u \frac{1}{2\lambda}||Au - f||_2^2 + \langle b^k, u - u^k\rangle \\ & + \frac{1}{2}||d^{k+1} - u||_2^2 \end{aligned} \quad (5)$$

$$b^{k+1} = b^k - (d^{k+1} - u^{k+1}) \quad (6)$$

Explicitly solving (4) and (5) gives the simple algorithm

Initialize $u^0 = d^0 = b^0 = 0$

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 $\text{shrink}(v, t) = (\tau_t(v_1), \tau_t(v_2), \cdots, \tau_t(v_{NL}))$ with $\tau_t(x) = \text{sign}(x)\max\{|x| - t, 0\}$. 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 fast as a result. For moderate reverberation, the length of room impulse response (RIR) is not too long. The size of matrix $\lambda I + A^T A$ is $NL \times NL$, N being the number of sources. The computational cost for matrix inversion is not high. The above algorithm runs fast for the purpose of FSE.

3.2. Subdivided Split Bregman for Long Reverberations

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

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

Apply the split Bregman method to update each subdivided part of u sequentially (update u_i by fixing the other $r-1$ u_j's). The 3-step algorithm in the While loop is the same except step 2 is modified as:

(2) **For** i from 1 to r

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

end For

where d_i and b_i are the subdivided parts of d and b. We precalculate inverse matrices $(\lambda I + A_i^T A_i)^{-1}$, each $\frac{NL}{r}$ dimensional. With proper choice of the number r, the computation speed can be improved significantly, as shown in section 5.

4. Source Activity Detection

The necessary preparation for FSE is silence detection of the speech sources. To maintain the overall speed of the proposed method, silence detection is based on the binary mask (BM) separation method [7], a fast method of blind speech separation without resolving RIRs. Though musical noise may occur due to binary operation in time-frequency domain and wide enough sensor spacings, BM appears reliable for identifying silence periods of a target speech from a mixture (a robust speech feature). A brief review of BM algorithm is given here. First, time domain signals $x_j(t)$, $j = 1, ..., M$, are transformed into time-frequency domain signals $X_j(f, \tau)$ by short-time Fourier transform (STFT). Next, time-frequency points are grouped into clusters such that within each cluster n, time-frequency points are dominated by the source n. The feature $\mathbf{\Theta}(f, \tau)$ was defined in [7] by direction of arrival (DOA) and distance. Then K-means clustering method finds N clusters from time-frequency points (f, τ). The separated signals $Y_n(f, \tau)$ are estimated by $Y_n(f, \tau) = M_n(f, \tau)X_J(f, \tau)$, where J is a selected sensor index, and $M_n(f, \tau)$ is the binary mask for cluster n.

The ratio $R_n(\tau) = \frac{||Y_n(\cdot,\tau)||_2^2}{||Y_B(\cdot,\tau)||_2^2}$ is used for detecting the silence part of source n, where Y_B is the sum of background sources. Though the separation quality may degrade if reverberation is long, the onset-offset feature is robust and detectable if we delete certain "fuzzy points" and reduce binary masking errors. For each time-frequency point (f, τ), the confidence coefficient of $\mathbf{\Theta}(f, \tau) \in C_n$ is defined by $CC(f, \tau) = \frac{d_n}{\min_{j \neq n} d_j}$, where d_j is the distance between $\mathbf{\Theta}(f, \tau)$ and j-th cluster centroid. The binary mask is redefined as

$$M_n(f, \tau) = \begin{cases} 1 & \mathbf{\Theta}(f, \tau) \in C_n \ \& \ CC(f, \tau) \leq \rho \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

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

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 $u_{jn}, j = 1, ..., M$
end
else
  Apply **subdivided split Bregman** method to obtain filters $u_{jn}, j = 1, ..., M$
end
Speech Extraction: Calculate $\hat{s}_n = \sum_{j=1}^{M} u_{jn} * x_j$.

5. 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. We first evaluate the proposed FSE method, and compare the split Bregman algorithm with subdivided split Bregman algorithm

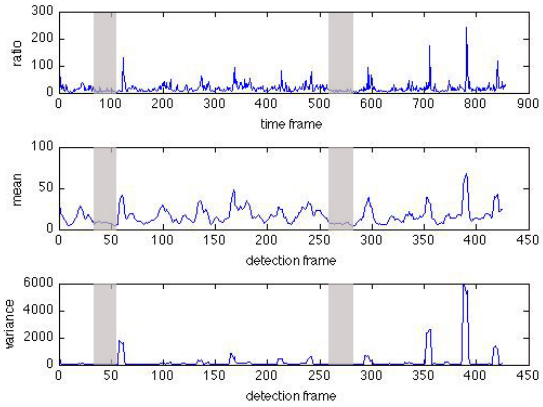

Figure 1: *Source activity detection (mixture of speech and music). Top: ratio $R(\tau)$; middle: mean of $R(\tau)$; bottom: variance of $R(\tau)$. Detection frame size is 10 with shift as 2. The range of detection frame is half of time frame. Segments marked by the shadows are selected regions for D where the speech is weak.*

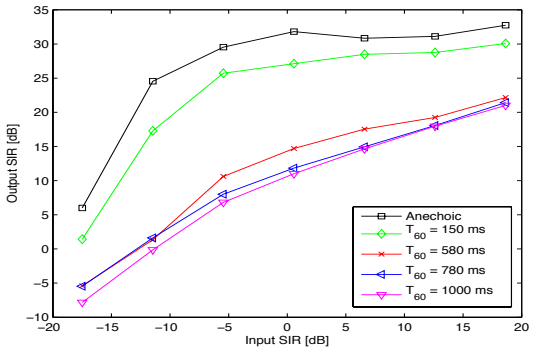

Figure 2: *Output SIR vs. input SIR for the proposed FSE method with different reverberation times.*

with synthetically mixed data (two sensors and two sources). **[Setup 1]**: The room size is $5 \times 9 \times 3.5$ m, and the impulse responses are measured by two omni-directional microphones (middle of the room and 1.5 m above the floor) with the spacing 15 cm. The sources are 1 m away from the sensors with the azimuth 30° and 90°, and the same height as sensors. The reverberation times of impulse responses are from 0 s (anechoic) to 1.0 s. In order to illustrate the separation quality and speed of our proposed method, we simplify the detection step by knowing 0.5 s' silent duration D of target speech source ahead of time. The other source is either speech or background music. The duration of the sources is 5 s and the sampling rate is 16000 Hz. Two mixtures are synthesized by measured RIRs as (1). The parameters for FSE are chosen as $\mu = \epsilon = 10^{-3}$, $\eta = 1$, and $\lambda = 2\mu$ throughout the evaluation. As the reverberation time goes up, the length of solution u increases accordingly from 40 taps to 2000 taps. Shown in Fig. 2 are the average output signal to interference ratios (SIRs) achieved by FSE for the various reverberation times and input SIRs.

Table 1 illustrates the average iterations, computation time [s] and SIR improvement (SIRI [dB]) of the split Bregman algo-

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

Table 1: *Comparison of (subdivided) split Bregman algorithms*

	Split Bregman			Subdivided Split Bregman		
L	Ite.	Time	SIRI	Ite.	Time	SIRI
100	**50**	**0.058**	6.21	50	0.531	**6.22**
200	**42**	**0.209**	6.77	43	0.796	**6.78**
400	44	**0.780**	8.07	**43**	1.565	**8.11**
800	62	4.386	9.11	**50**	**4.064**	**9.20**
1200	63	10.994	10.36	**41**	**7.019**	**10.40**
1600	71	21.684	**11.38**	**66**	**14.820**	11.27
3600	-	-	-	**123**	**83.295**	**13.47**

Table 2: *Comparison of BSS methods on synthetic mixture data*

	Time [s]	SIR [dB]	SDR [dB]	SAR [dB]
Parra[9]	7.16	5.55	1.62	5.34
IVA[10]	42.72	14.59	7.21	9.52
SNGTD[1]	122.35	11.28	4.67	7.21
FastICA[1]	1.32	9.31	4.12	7.05
FSE	1.56	26.60	15.35	16.39

The comparison of a list of existing BSS methods is shown in Table 2 in terms of computation time, SIR, signal to distortion ratio (SDR) and signal to artifact ratio (SAR). The data are synthetic mixtures of two speech sources as in **[Setup 1]** with reverberation time $T_{60} = 150\ ms$ and input SIR $\approx -5.9\ dB$. To compare the computation time of the algorithms directly and fairly, the proposed FSE method extracts two speech sources sequentially with the silent unions for the two speech sources known ahead of time. Table 2 indicates that proposed FSE achieves the best separation quality in objective measures at almost the speed of FastICA.

Room recorded mixture data are used to evaluate and compare the above BSS methods by the Perceptual Evaluation of Speech Quality (PESQ) [8]. **[Setup 2]**: The room size is $4.4\ \times 3.5\ \times 2.5$ m with reverberation time $T_{60} = 130\ ms$. The loudspeakers and omni-directional microphones are 1.3 m high from the floor. The sensors are set in the middle of the room with 4 cm spacing linearly arranged. For the two sensors and two sources case, the two loudspeakers are set 1.2 m from the sensors with the azimuth at 40° and 80° respectively. For the case of three sensors and three sources, the third loudspeaker is set 1.2 m from the sensors with the azimuth 120°. The mixture data are male and female speeches with the duration about 7 s and sampling rate 8000 Hz. Now with the source activity detection added, the separation quality of the proposed FSE exceeds those of the known methods, as seen from Table 3. The speech sources activity detection is done within 2 to 3 seconds, and does not affect the efficiency of the FSE method. DUET BSS method [11] is included in Table 3 as the microphone spacing is small enough so that there is no phase-wrap ambiguity to degrade its performance.

Table 3: *Average PESQ of BSS methods on real recording mixture data. PRE PESQ is the average PESQ of the mixture data. Time for FSE is shown as detection time + speech extraction time.*

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

6. Discussion and Conclusion

We proposed and evaluated a fast and efficient blind speech extraction method as long as target speeches contain pauses. A convex optimization problem is formulated and solved by the split Bregman method to yield sparse unmixing filters. Binary mask blind speech separation method is modified to detect the speech source onset-offset activity. Experimental results indicate that the proposed method outperforms conventional blind speech separation methods in terms of the overall computation speed and separation quality. The limitation of the proposed method is that it relies on a robust silence detection in a long reverberation multi-talker environment which will be studied further in future work.

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

7. References

[1] S. Makino *et al. (eds.)*, Blind Speech Separation, 3-45, 217-241 Springer 2007.

[2] J. Xin, M.Yu, Y. Qi, H. Yang, F-G Zeng, "A nonlocally weighted soft-constrained natural gradient algorithm for blind source separation of reverberant speech", IEEE Workshop on Application of Signal Processing to Audio and Acoustics, 81-84, Oct. 2009.

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

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

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

[6] T. Goldstein and S. Osher, "The Split Bregman Algorithm for L^1 Regularized Problems", SIAM J. Imaging Sci. 2:323-343, 2009.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.