Blind Source Separation of Speech Signals using Mixing Matrix Estimation and Subspace Method

R. Gokila
ECE Department, JCET
Trichy, India

M. Geetha
ECE Department, JCET
Trichy, India

***Abstract* - Blind source separation (BSS) is the separation of sources without having prior information about the mixtures. This is a major problem in real time world whether we have to identify a particular person in the crowd or it is an area of biomedical signal extraction like Electroencephalogram (EEG). To solve the speech separation problems in the underdetermined cases, "Two-step" method is used here, which estimates the mixing matrix first and then separates the sources. This paper provides subspace projection method with enhanced functionality, which can be used for Underdetermined Blind Source Separation (UBSS). The model takes into account both projection and size of the signal's subspace, without estimating the real source numbers at the TF points. Simulation results show that the proposed method overcomes the shortage of conventional subspace method and achieves higher separation performance.**

Index Terms – Speech separation, Speech Enhancement, Convex model, Mixing matrix Estimation.

I.INTRODUCTION

Blind source separation (BSS) aims to extract the original source signals from their mixtures observed by a set of sensors with no, or very limited, knowledge about the source signals and the mixing channel. Potential applications of BSS include speech processing, biomedical signal processing, analysis of astronomical data or satellite images, etc.

If there are fewer mixtures than sources, we have a challenging UBSS problem, where estimating the mixing channel is not sufficient for the recovery of sources. Nowadays, a large amount of algorithms for UBSS start from the assumption that the sources are sparse, i.e., the signals are mostly close to zero with the exception of several large values. Then the algorithms consist of two steps: first estimating the mixing matrix and then estimating the sources.The main contribution of this paper is to propose a new method for underdetermined blind speech separation under the "two-step" framework.

Many natural signals can achieve satisfied sparsity in transform domains, such as by wavelet packet transform or by short-time Fourier transform (STFT). Yilmaz et al.[9] exploit sparsity in the STFT domain, and assume that the sources are disjoint, i.e., there exists only one source at any TF point. Later, Aissa-El-Bey et al. [16] relax the condition and propose an efficient linear algorithm based on subspace projection using STFT, which assumes that the sources can be non-disjoint in the TF domain, i.e., the number of the active sources that coexist at any TF point is less than that of the mixtures. Both algorithms work well on audio signals.

Mixing matrix estimation is the first and crucial step in the whole process of UBSS. For this task, there have been many researches, such as Time-Frequency Ratio of Mixtures (TIFROM) [11] algorithm and its extensions. Among them, Single Source Point (SSP) detection [12]-[14] is a kind of simple and efficient method for mixing matrix estimation, which achieves good performance in both over and underdetermined cases. Specifically, it sets no conditions on the stationary, independence or non-Gaussianity of the sources, which made it be widely used. Source separation algorithms are applied after estimating the mixing matrix. However, in the underdetermined cases, the separation is more difficult even though the mixing matrix has been estimated because it is an ill-conditioned model when the number of sources is larger than that of mixtures. In order to solve the UBSS problems, additional constraints or assumptions must be set. For example, Degenerate Unmixing and Estimation Technique (DUET) [15] works under the constraints of Wdisjoint orthogonal, but the introduced time-frequency masking can only separate single source.

In general, subspace-based method for blind speech signals separation is simple and effective, but it does not estimate the number of active sources at the TF point, which limits its performance. On the other point of view, extra step means more computation complexity. In this paper, firstly, the mixing matrix is estimated using SSP detecting based on prior exploitation and a new automatic clustering method. After that, the shortage of conventional subspace method is indicated, and a new robust convex-based objective model is developed to combine both subspace projection and source number at any TF point. Compared with its predecessors, the proposed algorithm does not need to estimate the exact source number at every TF point before applying subspace method, which reduces the computation and makes it well adapted in the complex environments.

Speech enhancement is very important for applications of speech processing and communications; in our daily environs we always encounter some kind of noise or disturbance. For example, it is very difficult to

communicate with someone effectively in a train station or in a car moving at high speed. Therefore it will be imperative to study speech signals, noise and their mixtures in order to develop a technique that will effectively separate the signals or just extract the desired signal. There are two basic types of interference considered in Speech enhancement studies, one is an interference that is uncorrelated with the desired speech signal and the other is the one that is correlated with the source otherwise known as reverberation or literally called "echo". In other to achieve success in speech separation in form of mixing matrix estimation and signal separation.

II. PROBLEM FORMULATION

A. Signal Mixing Model

Blind source separation based on sparse representation of observed data matrix A. The proposed approach is also suitable for the case in which the number of sensors is less than the number of sources, as well as the case in which the source number is equal to sensor number.

In this section, the following noise free mixing model is considered,

$$X = AS \qquad (1)$$

where the mixing matrix A is unknown, the matrixS is composed of the m unknown sources, the only observable X is a data matrix with its rows being linear mixtures of sources, n≤m. The task of blind source separation is to recover the sources only using the observed data matrix X.

In this paper we consider SCA as a special model of BSS problem in the over complete case (m < n i.e. more sources than sensors), where the additional information compensating the lack of sensors is the sparseness of the sources. The task of the SCA problem is to represent the given data X as in equation (1) such that the matrix S (sources) is sparse in sense that each column of S has at least one zero element.

We present conditions on the data matrix X (SCA-conditions on the data), under which the representation in equation (1) is unique up to permutation and scaling of the sources. The task of BSS problem is to estimate the unknown sources S using the available data matrix X only. We describe under which this is possible uniquely up to permutation and scaling of the sources, which is the usual condition in the complete BSS problems using ICA.

B. ASSUMPTIONS

In this section we present a method for solving the BSS problem if the following assumptions are satisfied:

Assumption 1:

- The mixing matrix A Є R^{mxn} has the property that any square mxm submatrix of it is nonsingular

Assumption 2:

- Each column of the source matrix S has at least one zero element.

Assumption 3:

- The sources are sufficiently rich represented in the following sense: for any index set of n-m + 1 elements $I=\{i_1 \ldots\ldots 1_{n-1}\}$there exist at least m column vectors of the matrix S such that each of them has zero elements in places with indexes in I and each m-1 of them are linearly independent. Columns of X for which A2) is not satisfied are called outliers. We can detect them in some cases and eliminate from the matrix X, if the condition A3) is satisfied for a big number of columns of S.

C. Short- Time Fourier Transform

The Short Time Fourier Transform, evaluates the frequency (and possibly the phase) change of a signal over time.To achieve this, the signal is cut into blocks of finite length, and then the Fourier transform of each block is computed.It is adopted to exploit the sparsity of speech signals in TF domain. The STFT of source and mixture signals are defined as

$$S_n(t,f) = \int_{-\infty}^{\infty} s_n(\tau)\, g*(t-\tau) \exp(-j2\pi f\tau)\, d\tau (1 \le n \le N) \qquad (2)$$

$$X_m(t,f) = \int_{-\infty}^{\infty} x_m(\tau)\, g*(t-\tau) \exp(-j2\pi f\tau)\, d\tau (1 \le m \le M) \qquad (3)$$

Where g(t) is the window function.

Considering that **A** is a constant, and applying STFT to equation (1) at both sides, the mixing model in time-frequency domain becomes

$$X(t,f)=AS(t,f)+W(t,f) \qquad (4)$$

To solve the underdetermined time delay speech source separation problem in equation (1), a detailed two-step method is proposed in Fig. 1. After applying STFT to the mixtures, SSP detection method based on prior exploitation is adopted to estimate the mixing matrix. Then, for the second step, the speech signals are separated by using the new subspace projection method based on convex model. Finally, the inverse STFT is used to reconstruct the time-domain signals.

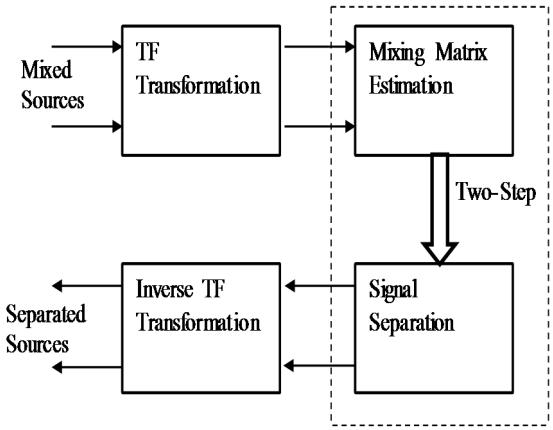

Fig. 1. Two Step method for UBSS

III. MIXING MATRIX ESTIMATION

Mixing matrix Estimation is the first step in the whole process of UBSS. For this task, Single Source Point (SSP) detection is a kind of simple and efficient method, which achieves good performance in both over and underdetermined cases.

A. IDENTIFICATION OF SSP

We are interested in performing mixing matrix estimation from the time- frequency representation of the mixtures. As described earlier, SSPs are TF points at which only one of the sources is active. As an example scenario, let us consider the case with 2 sources and 2 sensors.

Let g1 and g2 denote the columns of the mixing matrix G $\in R^{2\times2}$. At an SSP (t1, k1) in the TF domain where only the first source is active, i.e., s1(t1, k1) ≠ 0 and s2 (t1, k1) = 0, we have

$$c(t_1, k_1) = g_1 s_1(t_1, k_1) \quad (5)$$

This implies that the real and imaginary parts can be equated as

$$\mathrm{Re}\{c(t_1, k_1)\} = g_1 \mathrm{Re}\{s_1(t_1, k_1)\} \quad (6)$$
$$\mathrm{Im}\{c(t_1, k_1)\} = g_1 \mathrm{Im}\{s_1(t_1, k_1)\} \quad (7)$$

Here, it can be easily seen that the absolute directions of the real and imaginary components of c(t1, k1) are the same as that of g1. Now, consider another time-frequency point (t2, k2) where both the sources are active.

$$\mathrm{Re}\{c(t_2, k_2)\} = g_1 \mathrm{Re}\{s_1(t_2, k_2)\} + g_2 \mathrm{Re}\{s_2(t_2, k_2)\} \quad (8)$$

$$\mathrm{Im}\{c(t_2, k_2)\} = g_1 \mathrm{Im}\{s_1(t_2, k_2)\} + g_2 \mathrm{Im}\{s_2(t_2, k_2)\} \quad (9)$$

For the absolute directions of the real and imaginary components to be samein this case, the following relation needs to be satisfied,

$$\frac{\mathrm{Re}\{s1(t2,k2)\}}{\mathrm{Im}\{s1(t2,k2)\}} = \frac{\mathrm{Re}\{s2(t2,k2)\}}{\mathrm{Im}\{s2(t2,k2)\}} \quad (10)$$

Extending this to the case of multiple sources, or the absolute directions of Re{c(t, k)} and Im{c(t, k)} at any TF point (t, k) to be the same, it should be a single source point or the ratios of the real and imaginary components of the short-term Fourier transform (STFT) of all source signalsmust be the same. The points in the TF domain where theangle between the absolute directions of Re{c(t, k)} and Im{c(t, k)} is less than a predefined threshold Δθ is considered as an SSP. Typically, both the real and the imaginary parts of the SSPs, identified using above condition is stacked together in order to estimate the mixing matrix.

B. AGGLOMERATIVE HIERARCHICALCLUSTERING

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. It starts with every single object in a single cluster. Then, in each successive iteration, it agglomerates the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster.At last, the final clusters are determined after removing the scattered classes caused by noises or other disturbances. Here, two parameters are adopted, *th_d* denotes the minimum Euclidean distance to determine when the clustering is completed, and *th_N* is used to drop discrete error classes.P_l denotes different classes with $N_l$elements,

Define the center of $P_l$as

$$C_l = \frac{1}{N_l}\sum_{k=1}^{N_l} U_k \quad (11)$$

Distance between every two classes is calculated by

$$D_{l_1 l_2} = \left\| C_{l_1} - C_{l_2} \right\|_2 \quad (12)$$

However, Lihui *et al.* [14] have proved that the general SSP based algorithm for mixing matrix estimation is not limited to narrowband signals. It can also be applied for the models using wideband sources.

IV. SIGNAL SEPARATION

In the underdetermined cases, theseparation is more difficult even though the mixing matrix has been estimated because it is an ill-conditioned model when the number of sources is larger than that of mixtures. In order to solve the UBSS problems, additional constraints or assumptions must be set.

A. ORTHOGONAL PROJECTIONS

The orthogonal projection matrix Q for $\{\alpha_1 \ldots\ldots\alpha_k\}$ projecting to noise space is calculated by,

$$Q = I - A_k(A_k^H A_k^{-1}) A_k^H \quad (13)$$

Which full fills the following features ,

$$\begin{cases} Qa_i = 0, & i \in \{\alpha_1 \ldots\ldots\alpha_k\} \\ Qa_i \neq 0, & i \in \{1,2\ldots\ldots N\}, i \neq \{\alpha_1 \ldots\ldots\alpha_k\} \end{cases} \quad (14)$$

In real environment, because of the noises and calculation errors, Qa_i are not strictly equal to 0, whereas very close to 0 instead. Considering that A has been estimated, the column vectors of A_k can be detected by minimizing,

$$\{\beta_1,\ldots..,\beta_2\} = \mathrm{argmin}\{\|QX(t_0.f_0)\|A_k\} \quad (15)$$

Where $A_k$is actual mixing matrix.

The STFT values of K sources at (t,f) are estimated by pseudo-inverse calculation

$$S(t,f) = A_k^{-1} X(t,f) \quad (16)$$

At last, transform S(t,f) back to time domain using inverse STFT.

V. SIMULATION RESULTS

In the simulation, estimation error and average Signal to Interference Ratio (SIR) are adopted to evaluate the performance of mixing matrix estimation and signal separation respectively. They are defined as,

$$E_A = 10\log_g\left(\frac{1}{N}\|A - A1\|\right)_F \quad (dB) \qquad (17)$$

$$SIR = \frac{1}{N}\sum_{n=1}^{N} iolg \frac{e\{|S_n(t)|^2\}}{e\{|S_n(t) - S_{n1}(t)|^2\}} \qquad (18)$$

where **A** is the actual mixing matrix with normalized column vectors and **A1** is its estimation, $S_n(t)$ is the estimation of $S_{n1}(t)$, N is the number of sources.

For simulation, the time window length of STFT is 32, overlapping with 0.75, Fast Fourier Transformation (FFT) size is 256, and the threshold for SSP detecting is 0.001. Threshold *th_d* and *th_N* are set to 0.3 and 90 respectively. Four speech signals are taken as input, which is shown in Fig.2.

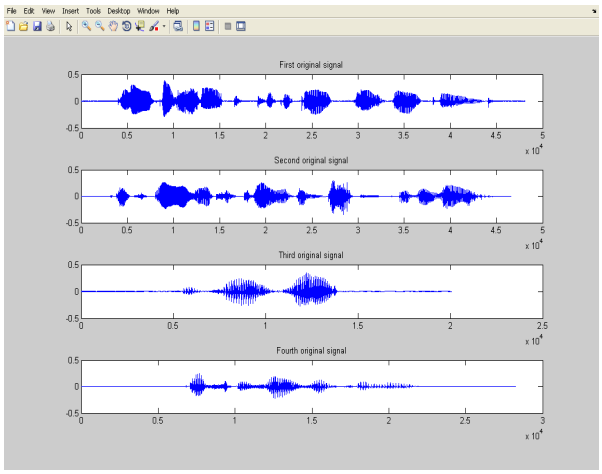

Fig.2. Source Signals

Then the signals are mixed together to obtain mixed signals, which is showned by Fig.3.

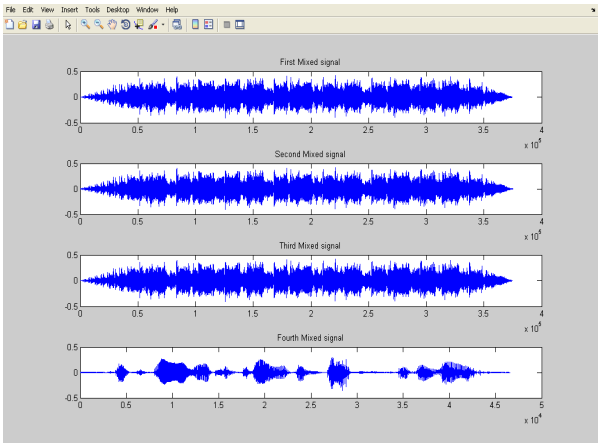

Fig.3. Mixed Signals

Fig.4. shows the scatter figures of detected SSPs. Clearly, all the SSPs locate on a unit circle with obvious clustering characteristic according to the density of distribution.The proposed AHC method can cluster the SSPs into four groups accurately, and the estimated centers almost have the same position as the actualcenters. The estimated mixing matrix is A, with $E_A$equals to -19.5298dB.

$$A = \begin{pmatrix} 0 & 2.9155 & 1.0000 & 3.0414 & 3.0414 \\ 2.9155 & 0 & 2.5495 & 3.3541 & 2.5000 \\ 1.0000 & 2.5495 & 0 & 2.0616 & 2.0616 \\ 3.0414 & 3.3541 & 2.0616 & 0 & 1.0000 \\ 3.0414 & 2.5000 & 2.0616 & 1.0000 & 0 \end{pmatrix}$$

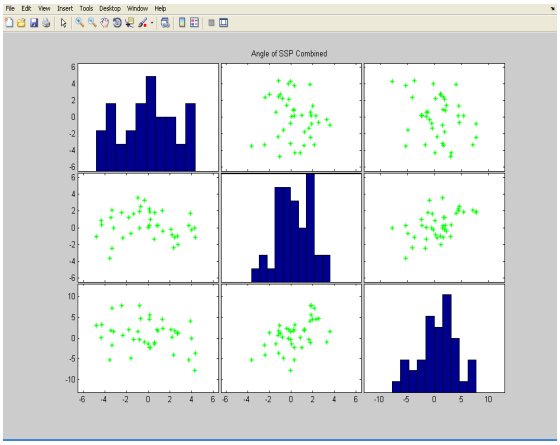

Fig.4. Performance of AHC for SSPs

Fig. 5 shows the separation results of source signals. Additive white Gaussian noises is added with SNR=30dB. It is clear that all the 4 source signals have been separated successfully byusing the proposed subspace method. The recovered wavesalmost have the same shapes as the original sources.

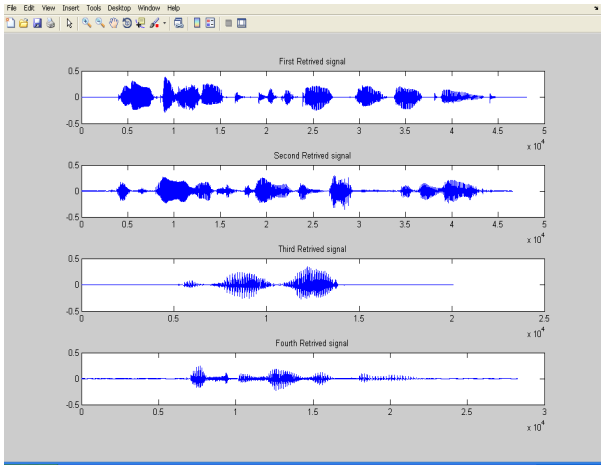

Fig.5. Separated Signals

Fig.6. demonstrates the performance comparison of SIR values for different variables.In summary, the proposed method achieves highest SIRs across all SNRs. It avoids adding steps to estimate the source numbers, and achieves robust best performance especially in the low SNR environments.In particular, conventional subspace method with almost has the same performance as proposed algorithm, that is because the assumptions ensure which avoids the case of "over-estimated" actually. But even so, in low SNRs, the proposed algorithm has better performance.

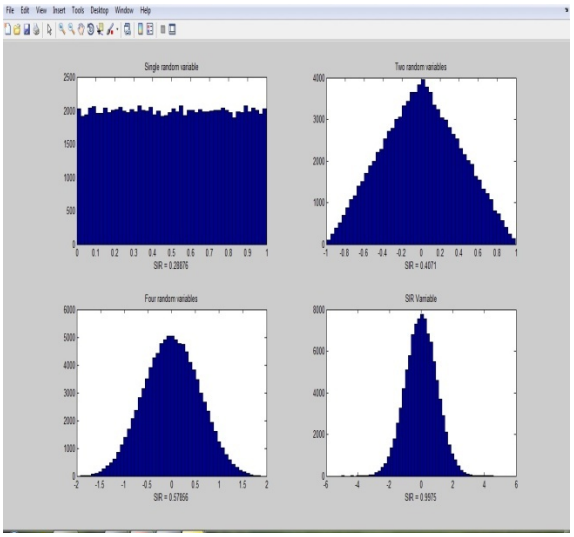

Fig.6. Performance Comparison

VI. CONCLUSION

This paper develops the underdetermined blind speechseparation method based on single source detection and a new modified subspace method. Blind source separations of four signals are done here. Instead of fixing the active source number as a constant or calculating the real source number at every TF point, the proposed algorithm introduces a convex model, taking into account both projection and size of subspace, which avoids increasing the computation cost. This paper employs two step methods for blind source separation. In first step the mixing matrix is estimated by Single source point detection and Agglomerative Hierarchical clustering method, and then the sources are separated in the second step by subspace method.Simulation results indicate that the improved method overcomes the shortage of classical methods and separates the speech signals with higher SIR than the conventional algorithms under the same conditions. So the proposed method achieves higher separation performance. Moreover, the advantage of performance is more obvious in low SNR environments.

REFERENCES

[1] Y. K. Lee and O. W. Kwon, "Application of shape analysis techniques for improved CASA based speech separation," *IEEE Trans. ConsumerElectron.,* vol. 55, no. 1, pp. 146-149, Feb. 2009.

[2] S. Han, J. Hong, S. Jeong, and M. Hahn, "Robust GSC-based speech enhancement for human machine interface," *IEEE Trans. ConsumerElectron.,* vol. 56, no. 2, pp. 965-970, May. 2010.

[3] J. S. Park, G. J. Jang, J. H. Kim , and S. H. Kim, "Acoustic interference cancellation for a voice-driven interface in smart TVs," *IEEE Trans.Consumer Electron.,* vol. 59, no. 1, pp. 244-249, Feb. 2013.

[4] S. J. Anderson, A. C. M. Fong, and J. Tang, "Robust tri-modalautomatic speech recognition for consumer applications," *IEEE Trans.Consumer Electron.,* vol. 59, no. 2, pp. 352-360, May. 2013.

[5] N. Cho and C. C. J. Kuo, "Enhanced speech separation in roomacoustic environments with selected binaural cues," *IEEE Trans.Consumer Electron.,* vol. 55, no. 4, pp. 2163-2171, Nov. 2009.

[6] Y. K. Lee, I. S. Lee, and O. W. Kwon, "Single-channel speechseparation using phase-based methods," *IEEE Trans. ConsumerElectron.,* vol. 56, no. 4, pp. 2453-2459, Nov. 2010.

[7] A.Hyvarinen, J. Karhunen, and E. Oja, *Independent ComponentAnalysis*, John Wiley&Sons: New York, 2001, pp. 407-440.

[8] P. Comon and C. Jutten, *Handbook of Blind Source Separation:Independent Component Analysis and Applications*, Academic Press ofElsevier: New York, 2010, pp. 683-814.

[9] P. Bofill and M. Zibulevsky, "Underdetermined blind source separationusing sparse representations," *Signal Process.*, vol. 81, no. 11, pp.
2353-2362, Nov. 2001.

[10] F. Abrard and Y. Deville, "A time-frequency blind signal separationmethod applicable to underdetermined mixtures of dependent sources,"*Signal Process.*, vol. 85, no. 7, pp. 1389-1403, Jul. 2005.

[11] M. Puigt and Y. Deville, "Time-frequency ratio-based blind separation methods for attenuated and time-delayed sources," *Mech. Syst. AndSignal Process.*, vol.19, no. 6, pp. 1348-1379, Nov. 2005.

[12] V. G. Reju, S. N. Koh, and I. Y. Soon, "An algorithm for mixing estimation in instantaneous blind source separation," *Signal Process.*,vol. 89, no. 9, pp. 1762-1773, Sep. 2009.

[13] V. G. Reju, S. N. Koh, and I. Y. Soon, "Underdetermined convolutiveblind source separation via time-frequency masking," *IEEE Trans.Audio, Speech, and Language Process.*, vol. 18, no. 1, pp. 101-116, Jan.2010.

[14] H. Li, Y. H Shen, J. G. Wang, and X. S. Ren, "Estimation of thecomplex-valued mixing matrix by single-source-points detection withless sensors than sources," *Trans. Emerging Tel. Tech.*, vol. 23, no. 2,pp. 137-147, Mar. 2012.

[15] S. Rickard and F. Dietrich, "DOA estimation of many W-disjointorthogonal sources from two mixtures using DUET," *in Proc. IEEEWorkshop on Statistical Signal and Array Processing*, Pocono Manor,PA, pp. 311-314, Aug. 2000.

[16] A. Aissa-El-Bey, N. Linh-Trung, K. Abed-Meraim, A. Belouchrani,and Y. Grenier, "Underdetermined blind separation of nondisjointsources in the time-frequency domain," *IEEE Trans. Signal Process.*,vol. 55, no. 3, pp. 897-907, Mar. 2007.

[17] H. Li, Y. H. Shen, M. Cao, and J. G. Wang, "Underdetermined blindseparation using modified subspace-based algorithm in the timefrequencydomain," *Prz. Elektrotechniczny*, vol. 87, no. 7, pp. 280-283,Jul. 2011.

[18] F. B. Lu, Z. T. Huang, and W. L. Jiang, "Underdetermined blindseparation of non-disjoint signals in time-frequency domain based onmatrix diagonalization," *Signal Process.*, vol. 91, no. 7, pp. 1568-1577,Jul. 2011.

[19] J. Yang, L. J. Zhang, K. W. Lu, and Q. N. Zhang, "Underdeterminedblind source separation using modified subspace method based onconvex model," *in Proc. IEEE International Conference on ConsumerElectronics*, Las Vegas, USA, pp. 135 136, Jan. 2014.