Convolutive Blind Speech Separation using Cross Spectral Density Matrix and Clustering for Resolving Permutation

C.Prabhu S.Pradeep R.Baskaran C.Chellappan
Department of Computer Science & Engineering
Anna University
Chennai, India

ABSTRACT

The problem of separation of audio sources recorded in a real world situation is well established in modern literature. The method to solve this problem is Blind Speech Separation (BSS).The recording environment is usually modeled as convolutive (i.e. number of speech sources should be equal to or less than number of microphone arrays). In this paper, we propose a new frequency domain approach to convolutive blind speech separation. Matrix Diagonalization method is applied on cross power spectral density matrices of the microphone inputs to determine the mixing system at each frequency bin up to a permutation ambiguity. Then, we propose an efficient algorithm to resolve permutation ambiguity, where we group vectors of estimated frequency responses into clusters in such a way that each cluster contains frequency responses associated with the same source. The inverse of the mixing system is then used to find the separate sources. The performance of the proposed algorithm is demonstrated by experiments conducted in real reverberant rooms.

General Terms

Blind Speech Separation, Frequency-domain Convolution, Speech Signals

Keywords

Cross-Power Spectral Density; Matrix Diagonalization; Blind Speech Separation; Permutation ambiguity; Cluster

1. INTRODUCTION

Blind Speech Separation (BSS) is the problem to separate independent sources from given mixed signals where the mixing process is unknown. The classical example is the “cocktail party problem”, where a number of people are talking simultaneously in a room, and one is trying to follow one of the discussions. It is 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 the 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”.

There are two major categories of speech separation technique using multiple microphones, they are;

- Beam Forming and
- Blind Source Separation.

Beam forming is a form of speech separation technique which enhances signal from one direction and attenuates signals from other directions, which means that a beam former enhances only the speech source of interest and suppresses others. Traditional Beam forming technique has a back drop in the the sense that it relies on the position of the speaker which is not always available for its performance. Also, errors are unavoidable when estimating the position of the speaker using microphone output analysis.

Blind source separation is a form of speech separation technique which blindly estimates individual source components from their received mixtures at sensors. The estimation is performed without prior knowledge about each source location and time activity distribution. In its application, like in speech enhancement, teleconferencing, hearing aids etc, signals are mixed in a convolutive manner causing reverberation thereby making the Blind Source Separation problem a difficult one. Our work dwells on the problem of Blind Source Separation.

In recent years few BSS method has been proposed [1]-[8]. The approaches for convolutive case can be divided into frequency domain [9]-[13] and time domain methods [14], [15].

The advantage of using frequency-domain methods is that a time-domain problem with a large number of parameters is decomposed into multiple, independent estimation problems at each frequency bin, each with fewer parameters to be estimated. As a result, in general the frequency-domain estimation algorithms have a simpler implementation and better convergence properties over their time-domain counterparts. The main difficulties with frequency-domain BSS of convolutive mixtures, however, are the arbitrary permutation and scaling ambiguities of the estimated frequency response of the un-mixing system at each frequency bin.

The method proposed in this paper is a frequency domain technique which uses the Cross-Power Spectral Density (CPSD) matrix of the microphone inputs. The proposed method is not limited to mixing systems with fixed dimensions or to ones with the same number of outputs and inputs. The only requirement is that the number of outputs for the mixing system is greater than or equal to the number of inputs.

Various methods have been proposed for resolving permutation ambiguity. Imposing constraints on the demixing filters [18] and direction of arrival (DOA) estimation [18] are some of the previous methods used for resolving permutation. Each of these methods has its own disadvantage. The former method cannot be used when the mixing filter length is too long and the latter method is not adequate for the case where the sampling rate of the observed frequency is high.

In this paper, we propose an ICA based clustering method for resolving permutation. This method is based on the method in [18], where prior information on the maximal distance between the sensors is required and sophisticated normalization method is used. In our method no prior information or the normalization method is used. The basic idea of our method is to exploit basis vectors of representing estimated frequency responses, in order to group them into clusters, each of which contains frequency responses associated with the same source.

This paper is organized as follows: in section 2, we state the BSS problem. In section 3 and 4, we deal with calculation of CPSD matrix and using matrix diagonalization on the CPSD matrix to find the mixing system. In section 5, we propose a basic algorithm for resolving permutation ambiguity. In section 6, we show the performance of our method using experimental results. In section 7 and 8, we give the conclusion and possible future work for our method.

2. BLIND SPEECH SEPARATION PROBLEM

Blind source separation is a technique for estimating individual source components from their mixtures at sensors. This is called blind because, the estimation is done without prior information on the sources, that is their spatial location and time activity distribution; and on the mixing function, i.e. information about the mixing process. The problem has become increasingly important in the area of signal and speech processing due to their prospective application in speech recognition,teleconferencing,hearing aids, telecommunication and medical signal processing, etc.In these applications, signals are mixed in a convolutive manner, at times with reverberation otherwise literary called echo. This makes blind source separation a very difficult problem.

In Blind Speech Separation, the objective is to separate multiple sources, mixed through an unknown mixing system (channel), using only the system output data (observed signals) and in particular without using any (or the least amount of) information about the sources of the system.

Consider N source and M sensor model. Then BSS problem is

$$X(t) = H(t)\,S(t) + n(t) \tag{1}$$

Where,

$X(t) = \{x_1(t), x_2(t),.....,x_M(t)\}^T$, is the vector of observed signal.

$S(t) = \{s_1(t), s_2(t),.....,s_N(t)\}^T$, is the vector of sources.

H (t) = M x N impulse response of the mixing system.

$n(t) = \{n(1),n(2),.....,n_M\}^T$, is the noise Vector.

The objective of BSS is to determine W (t), the inverse of H (t). In our paper, we use frequency-domain approach rather than time-domain approach because the frequency-domain algorithms have simpler implementation. And we also assume that the noise vector n (t) is zero mean (i.e.,) the noise is assumed independent of the sources.

So the objective of the BSS in frequency-domain approach is to determine W (ω), such that

$$W(\omega)\,H(\omega) = \Pi D(\omega) \tag{2}$$

Where,

Π = N x N permutation matrix.

D (ω) = N x N diagonal matrix with diagonal elements which are rational functions of the radian frequency ω.

In frequency-domain methods, the matrix Π is dependent on ω, thereby introducing frequency-dependent permutation errors in the output frequency response. So for improving performance, Π must be made independent of frequency.

Fig. 1 shows the flow of the proposed method. The microphone inputs are used to calculate CSPD matrices and after convolution the original sources are obtained at the output.

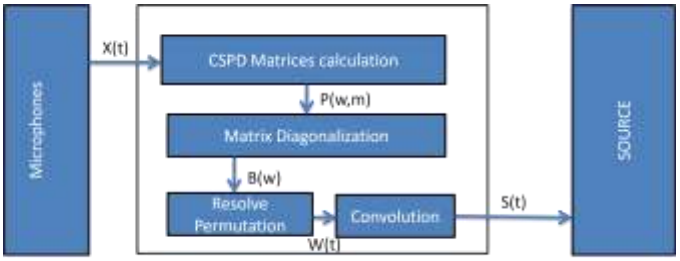

Figure 1. Overview of the proposed method

3. CPSD MATRIX CALCULATION

Cross Power Spectral Density describes us how the power of the signals is distributed over each frequency. The CPSD matrix is the Fourier transform of cross covariance between two signals.

$$P(\omega_k,m) = \sum R_{xy}(k)\, e^{-j\omega k} \tag{3}$$

Where,

$$R_{xy} = \sum X(a)\, \mathrm{conj}(Y(b)) \tag{4}$$

CPSD matrix is calculated for each epoch. “Epoch” means duration of time for which the source signals can be considered stationary within the epoch, but non-stationary between two epochs.

For convenience, calculated cross spectral density matrix is normalized.

$$C_x(\omega_k,m) = \frac{C(\omega_k,m)}{\|C(\omega_k,m)\|_F} \tag{5}$$

F stands for Frobenius norm.

4. MATRIX DIAGONALIZATION PROBLEM

After we find the CPSD matrix $P(\omega_k,m)$ m=0....M-1 , we have to apply matrix diagonalization algorithm for the matrices at each frequency ω_k , over M epochs, to estimate the mixing system, up to a permutation ambiguity at each frequency bins.

Matrix Diagonalization problem was first introduced by Flury [2] and has been adopted for BSS by [16], [17]. The problem is to find the matrix Λ_1.... Λ_r , which jointly diagonalizes the set of matrices A_1.....A_r.

$$\Lambda_k = B\ A\ B^+ \tag{6}$$

The common criterion used in matrix diagonalization to determine cost function is

$$\arg \min_{A,\Lambda(m)} \sum_{m=1}^{M} \| R_m - A\Lambda(m)A^{\dagger} \|_F^2 \tag{7}$$

The matrix diagonalization criterion used in our algorithm is

$$\sum_{\|b_i(\omega_k)\|_2=1\ k=0}^{K-1} \sum_{m=0}^{M-1} \| P_X^{\wedge}(\omega_k, m) - B(\omega_k)\Lambda(\omega_k, m)B^{+}(\omega_k) \|_F^2 \tag{8}$$

Where,

$$R_m - \ P_X^{\wedge}(\omega_k, m) \tag{9}$$

$$A - \ B(\omega_k) \tag{10}$$

$$\Lambda\,(m) - \ \Lambda\,(\omega_k, m) \tag{11}$$

Λ (ω_k, m) - Diagonal matrix representing the unknown cross-spectral density matrix of the source at each epoch m.

To initialize the algorithm, select two matrices P (ω_k,m_1), P (ω_k,m_2), m1 $\neq$ m2.Then choose the initial estimate for B (ω_k) to be the matrix consisting of the generalized eigenvectors of the matrix P (ω_k,m_2) P^{-1} (ω_k,m_1). So after finding B (ω_k) and Λ (ω_k, m), the next step is to determine W (ω_k) using those values [6].

$$W\,(\omega_k) = B\,(\omega_k)^{+} \tag{12}$$

B $(\omega_k)^{+}$ - pseudo inverse of matrix B (ω_k), (i.e.) A B A = A

For now W (ω_k) has permutation ambiguity.

5. RESOLVING PERMUTATION

It is possible that frequency components of the same source are recovered with arbitrary other. This result in permutation ambiguity and can lead us to a wrong reconstruction of the spectrum of the recovered sources. There are many methods developed by researcher to mitigate this problem, but the fact is that they have some disadvantage. The methods and their set back as also enumerated by [18] are as follows:

- Imposing constraints on demixing filters such as smoothing, gives a good result but only in a simple case, it cannot be used when mixing filter length is too long.
- Sawada et al [19], [20] used both the correlation and localization method to form what he called integrated method, but this increases the complexity and computational demand.
- In [18], Kim uses a method of grouping the vectors of estimated frequency responses into clusters in such a way that each cluster contains frequency responses associated with the same source. He did this grouping, otherwise called clustering, by applying ICA on estimated frequency responses.

Of all these methods and many other methods that have been proposed by many researches and are not mentioned here, the ICA based clustering by [18], looks simpler and less computational demanding, and in our approach to solving the permutation problem we used the method.

The normalization method in [9] uses the following near-field model for channel impulse responses.

$$h_{ij}(f)\frac{q(f)}{d_{ij}}\exp\{j2\pi f c^{-1}(d_{ij} - d_{Rj})\} \tag{13}$$

The normalization used was

$$\overline{a_{ik}}(f) \leftarrow |a_{ik}(f)| exp\left\{ j\frac{arg\left({}^{a_{ik}(f)}/_{a_{Rk}(f)} \right)}{4fc^{-1}d_{max}} \right\} \tag{14}$$

Taking the frequency permutation in account, and incorporating the normalization (13) into the near-field model (14), leads to

$$\overline{a_{ik}}(f) \approx \left\{ j\frac{\pi}{2}\frac{(d_{ik} - d_{Rk})}{d_{max}} \right\} \tag{15}$$

Eq. (15) is free of frequency-dependent factors and depends only on position of sensors and sources.

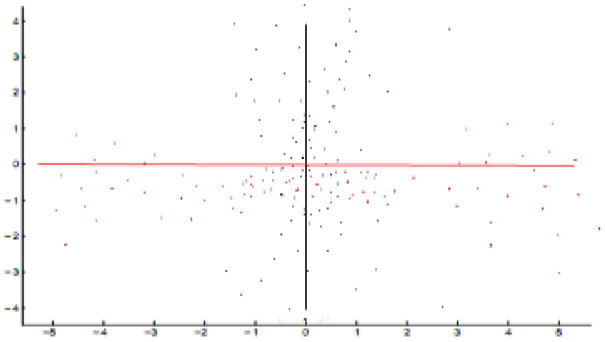

Figure 2. Two basis vectors a1 (f) and a2 (f) form two intrinsic directions that can be determined by the basis vectors computed by ICA as taking A (f) as input

Fig. 2 is the plot of the mixing vectors a_i (f), i=1, 2 with respect to frequencies and associated normalized mixing vectors.

Although the normalization in (15) eliminates the frequency dependent factors in mixing vectors, which allows us to easily group mixing vectors into clusters, each of which contains the mixing vectors associated with the same source. But the performance depends on what to choose as reference sensor.

From the Fig. 2 it is understood that a (f) associated with the same source, lies in the same direction. This direction is called intrinsic direction. Our proposed method is to find this intrinsic direction.

From the estimated W (f), we can calculate

$$A\,(f) = W^{-1}\,(f) \tag{16}$$

Where A (f) = {a_1 (f), a_2 (f), ..., a_n (f)}.

Consider a data matrix $\hat{X}$ = {A_1, A_2,..., A_M}.

$$A_k = A\left(\frac{(k-1f_s)}{M}\right) \tag{17}$$

Now using our BSS method we can find the following decomposition

$$\hat{X} = \hat{A}\,\hat{S} \qquad (18)$$

$\hat{A}$ - Mixing matrix of $\widehat{X}$.

$\hat{S}$ – Encoding Variable Matrix.

Clustering is done by considering absolute values of encoding variables that represent the contribution of mixing vectors.

Consider the case with two sensors and two sources. Let the data matrix be $\hat{X}$= [$\hat{x}_{l}$, $\hat{x}_{l+1}$]T and its associated variable matrix be $\hat{S}$ = [$\hat{s}_{1,l}$, $\hat{s}_{2,l}$]T. If $|\hat{s}_{1,l}| > |\hat{s}_{2,l}|$, $\hat{x}$l is assigned to cluster 1. Otherwise it is assigned to cluster 2. $\hat{x}$l+1 is assigned to other cluster.

Sometimes there may be a case, where $|\hat{s}_{1,l}| > |\hat{s}_{2,l}|$ and $|\hat{s}_{1,l+1}| > |\hat{s}_{2,l+1}|$ would satisfy. In times like that we take ratio of encoding variables into account and assign $\hat{x}$l and $\hat{x}$l+1 to cluster 1 and 2 respectively, if $|\hat{s}_{1,l}|$ / $|\hat{s}_{2,l}|$ > $|\hat{s}_{1,l+1}|$ / $|\hat{s}_{2,l+1}|$. Otherwise $\hat{x}$l and $\hat{x}$l+1 is assigned to cluster 2 and 1 respectively.

6. EXPERIMENTAL RESULTS

In this section, we demonstrate the performance of the proposed algorithm. First, the separate sources are mixed on the computer and then the algorithm is applied on the mixed data. Since the true sources are available, we can calculate the performance of our algorithm.

We evaluate the estimation performance by using the index defined as,

$$SNR = 10log\left(\frac{|s\,(t)|^2}{|s\,(t)-\,\hat{s}\,(t)|^2}\right) \qquad (19)$$

s (t) – Initial Individual Source

$\hat{s}$ (t) – Output Source

The estimation performance gets better as SNR (Signal to Noise Ratio) gets larger.

Recordings are conducted on a closed room of 5.2 m x 3.4 m x 2.9 m size in a less noisy environment. Experiments are conducted with various voice samples. For all experiment, sampling frequency is taken as 10 KHz, and epoch as 5000 data samples (i.e.) half second.

First we gave different types of voices to the system as input. All inputs were of 8 seconds.

For example: One Male and One Female

6.1 Individual Sources

Fig. 3 shows the speech of a male saying 'One Two Three …' for eight seconds.

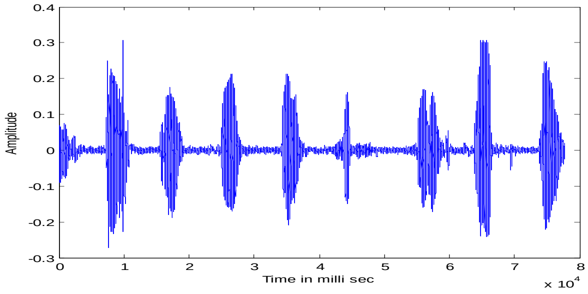

Figure 3. Source 1

Fig. 4 shows the speech of a female saying 'One Two Three …' for eight seconds.

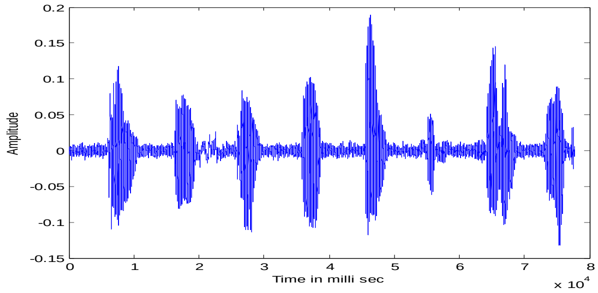

Figure 4. Source 2

6.2 Mixed Sources

Fig. 5 and Fig. 6 shows the mixed input.

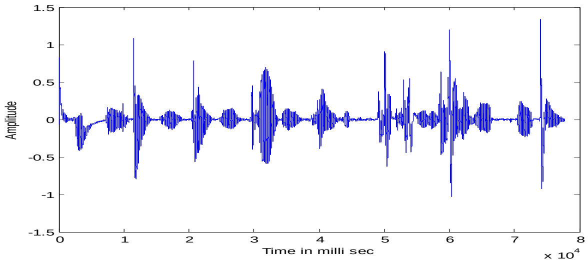

Figure 5. Mixed Source 1

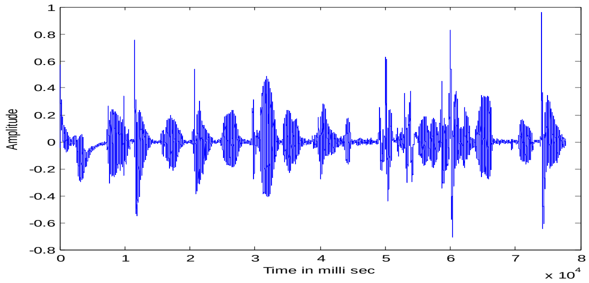

Figure 6. Mixed Source 2

6.3 Separated Sources

Fig. 7 shows the output speech of a male saying 'One Two Three …' for eight seconds.

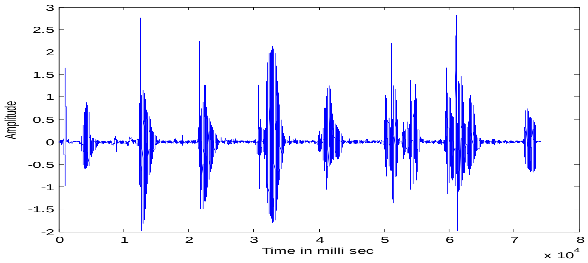

Figure 7. Separated Source 1

Fig. 8 shows the output speech of a female saying 'One Two Three …' for eight seconds.

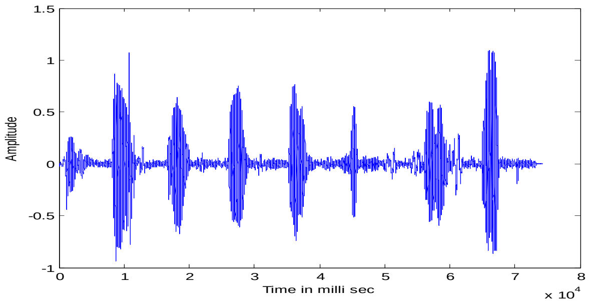

Figure 8. Separated Source 2

The table 1 shows the SNR value of two estimated sources for different types of input.

Table 1. SNR for estimated source signals

Voices	SNR (dB)	
	Source 1	*Source 2*
One Male and One Female	15.10	15.06
Two Male	15.19	15.14
Two Female	15.09	15.06

Then we gave same two voices at each time as input to the system, but with different time length at different times. The table 2 shows the SNR value of two estimated sources, where the time length at each experiment varies.

Table 2. SNR for estimated source signals for varying time length.

Time (Sec)	SNR (dB)	
	Source 1	*Source 2*
7	14.69	14.62
10	15.19	15.14
15	15.35	15.34
20	15.57	15.53
25	15.67	15.64

Fig. 9 shows the variation of performance as time increases.

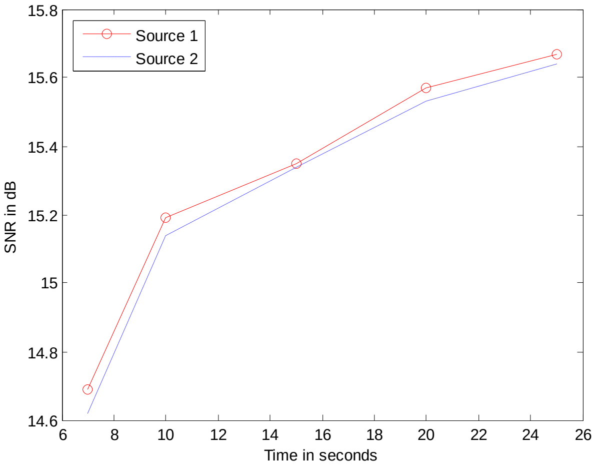

Figure 9. Performance variation with time

7. CONCLUSION

In this paper, we have studied and implemented a new approach for convolutive BSS, which can be used in applications like teleconference, hearing-aids etc. This is a very attractive method in solving convolutive mixture signals, which is the only form of mixtures applicable in real time applications. We used the cross-spectral density matrix of the microphone input, to separate the voices in a room. We used clustering technique for solving permutation ambiguity in the mixing system.

The performance of the algorithm is demonstrated by several experiments with different types of input.

The significance of this method compared to other method is that it does not make any assumption on the mixing system and it does not give any constraint on the input.

8. FUTURE WORK

This paper addresses the BSS without the consideration of noises at the input. In this method, as the number of sources and sensors increases, the computational complexity also increases. So separating voices with noise at the input and reducing the computational complexity will be investigated in the future papers.

9. ACKNOWLEDGMENT

This work is supported by the NTRO, Government of India. NTRO provides the fund for collaborative project “Smart and Secure Environment” and this paper is modeled for this project. Authors would like to thank the project coordinators and the NTRO members.

10. REFERENCES

[1] Kamran Rahbar and James P.Reilly, “A Frequency Domain method for Blind Source Separation of convolutive audio sources”, IEEE Transaction on speech and audio processing, vol.13, no.5, September 2005.

[2] Bernhard Flury and Walter Gautchi, “An Algorithm for simultaneous orthogonal transformation of several positive definite matrices to nearly diagonal form” IEEE 1997.

[3] Kolossa and R. Orglmeister, “Nonlinear post-processing for blind speech separation”, in: Proc. 5th Intl. Symposium. on ICA and BSS (ICA 2004), pp. 832–839, 2004.

[4] Aapo Hypernan and Erki Ojha, Independent Component Analysis and its Application. Neural Networks 2000, pp.411-430.

[5] Wei Liu, Danilo, P.Mandic andAndrzej Cichocki., “A Class of novel Blind Source Extraction Algorithms based on a linear predictor” ,IEEE 1997.

[6] R.H. Lambert, “Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures”, Univ. Southern California, La Jolla, CA 1996.

[7] Bulek, S and Erdol, N, “Blind speech separation using fractional order moments”, Statistical Signal Processing, IEEE/SP 15th Workshop 2009, pp. 509 – 512 .

[8] Hua Cai, Junxi Sun and Shifeng Ou, “Blind Speech Separation Employing Laplacian Normal Mixture Distribution”, Model.Mechatronics and utomation.International Conference 2007, pp. 3185 – 3189.

[9] L.Parra and C.Spence, “Convolutive blind separation of non stationary Sources”, IEEE Trans. Speech Audio Process., vol.8, no.3, pp.320–327, May2000.

[10] K.Rahbar and J.Reilly, "Blind source separation algorithm for MIMO Convolutive mixtures." in Int. Workshop on Independent Component Analysis and Signal Separation, San Diego, CA 2001, pp.242–247.

[11] Reju, V.G, Soo Ngee Koh and Ing Yann Soo, "A robust Correlation Method for Solving Permutation problem in Frequency Domain Blind Source Separation of Speech Signal", Circuits and Systems 2006, pp. 1891 – 1894.

[12] Solvang, H.K, Nagahara, Y, Araki, S, Sawada, H, Makino, S, " Frequency-Domain Pearson Distribution Approach for Independent Component Analysis (FD-Pearson-ICA) in Blind Source Separation", Audio, Speech, and Language Processing, IEEE Transactions on vol. 17, pp. 639 – 649, 2009

[13] Yu-Lin Liu, Shun Xu, Ming-Qi Li, " A Second-Order Feature Window Method for Blind Separation of Speech Signals Corrupted by Color Noise.Machine Learning and Cybernetics", International Conference on vol. 6 , pp. 3454 – 34, 2007.

[14] C.T.Ma,Z.Ding and S.F.Yau, "A two-stage algorithm for MIMO Blind deconvolution of non stationary colored signals," IEEE Trans. Signal Process., vol.48,no.4,pp.1187–1192 , 2000.

[15] H.Sahlin and H.Broman, "MIMO signal separation for FIR channels: A criterion and performance analysis," IEEE Trans. Signal Process, vol.48, no.3, pp.642–649, 2000.

[16] Fadaili, E.M. Moreau N.T. and Moreau E, "Non orthogonal Joint Diagonalization/Zero Diagonalization for Source Separation Based on Time-Frequency Distributions" Signal Processing, IEEE Transactions on vol. 55, pp. 1673 – 1687, 2007.

[17] Wenwu Wang, Sanei, S and Chambers, J.A, " Penalty function-based joint diagonalization approach for convolutive blind separation of non stationary sources", Signal Processing, IEEE Transactions vol. 53, pp. 1654 – 1669, 2005.

[18] Minje Kim and Seungjin Choi, "ICA-Based Clustering for Resolving Permutation Ambiguity in Frequency-Domain Convolutive Source Separation" IEEE 18th International conference on Pattern Recognition, 2006.

[19] J.Benesty, S.Makino and J.Chen, Speech Enhancement. Springer, 2005.

[20] Hiroshi Sawada, Ryo Mukai, Shoko Araki and Shoji 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, September 2004.