

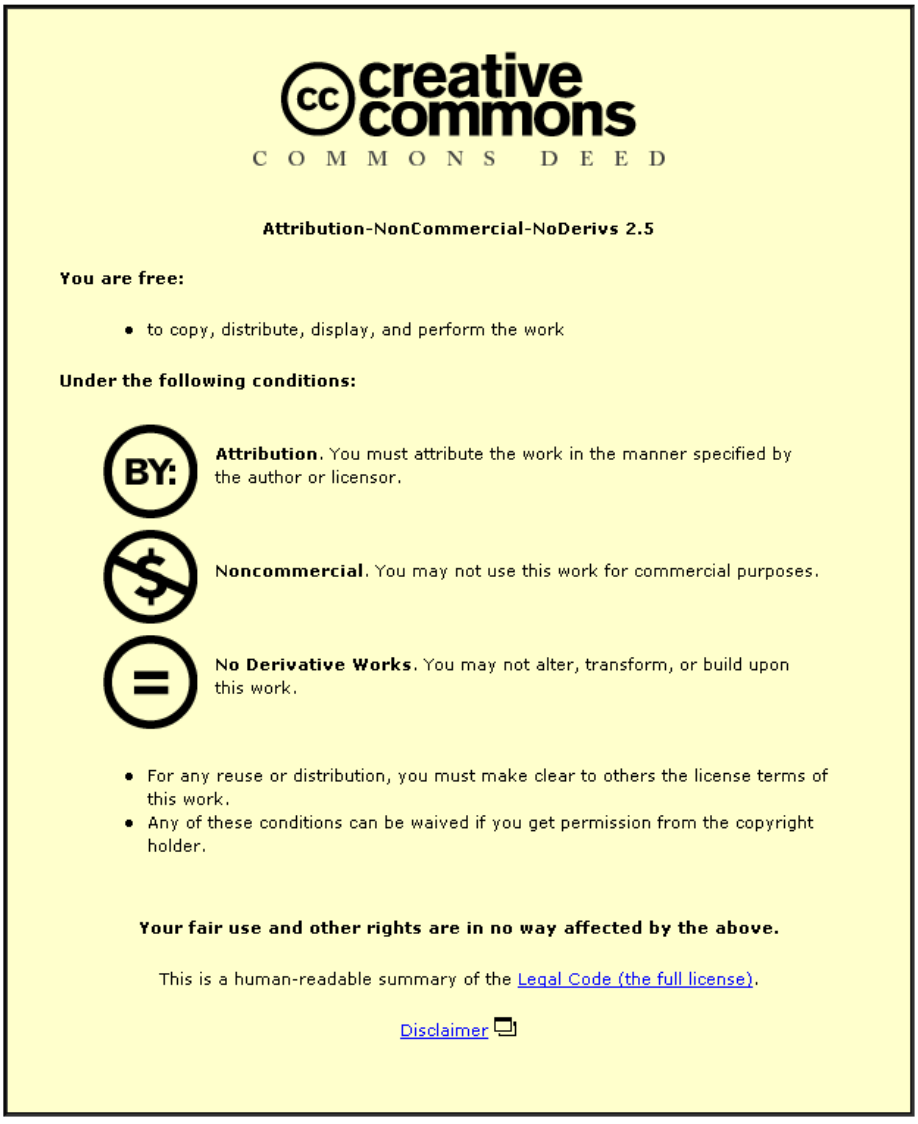

creative commons

COMMONS DEED

**Attribution-NonCommercial-NoDerivs 2.5**

**You are free:**

- to copy, distribute, display, and perform the work

**Under the following conditions:**

**Attribution**. You must attribute the work in the manner specified by the author or licensor.

**Noncommercial**. You may not use this work for commercial purposes.

**No Derivative Works**. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

**Your fair use and other rights are in no way affected by the above.**

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

# EVALUATION OF EMERGING FREQUENCY DOMAIN CONVOLUTIVE BLIND SOURCE SEPARATION ALGORITHMS BASED ON REAL ROOM RECORDINGS

*S. M. Naqvi†, Y. Zhang†, T. Tsalaile†, S. Sanei‡ and J. A. Chambers†*

†Advanced Signal Processing Group, Department of Electronic and Electrical Engineering
Loughborough University, Loughborough LE11 3TU, UK.
‡Centre of Digital Signal Processing, Cardiff University Cardiff, CF24 3AA, UK.
Email: {s.m.r.naqvi, y.zhang5, t.k.tsalaile, j.a.chambers}@lboro.ac.uk, saneis@cf.ac.uk

## ABSTRACT

This paper presents a comparative study of three of the emerging frequency domain convolutive blind source separation (FDCBSS) techniques i.e. convolutive blind separation of non-stationary sources due to Parra and Spence, penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources due to Wang et al. and a geometrically constrained multimodal approach for convolutive blind source separation due to Sanei et al. Objective evaluation is performed on the basis of signal to interference ratio (SIR), performance index (PI) and solution to the permutation problem. The results confirm that a multimodal approach is necessary to properly mitigate the permutation in BSS and ultimately to solve the cocktail party problem. In other words, it is to make BSS semiblind by exploiting prior geometrical information, and thereby providing the framework to find robust solutions for more challenging source separation with moving speakers.

***Index Terms***— Frequency domain (BSS), geometrical constraints, orthogonal/nonorthogonal constraints, penalty functions, cocktail party problem and multimodal signal separation.

## 1. INTRODUCTION

During the past decade there has been considerable research performed in the field of convolutive blind source separation due to its potential wide applications [1]. BSS is used to recover unknown sources from the observed mixtures with only limited assumptions such as the sources are independent. Many methods have been proposed to solve the BSS problem [2][3][4][5] [6] and still much work is required to solve the cocktail party problem [7]. In frequency domain convolutive blind source separation (FDCBSS) the time-domain convolutive mixing is converted into a number of independent complex instantaneous mixing operations. When the sources are reconstructed using the outputs for all the frequency bins we however face the permutation and scaling problems. The scaling problem can be easily solved by matrix normalization [8]. In most BSS algorithms the geometrical information which is available from video information is not utilized. The permutation problem degrades their performance and has to be solved. The advantage of multimodal BSS algorithms, for example, [8] is that the permutation problem can be mitigated, as will be shown by later simulations. The convolutive mixing system can be described as follows: assume $n$ statistically independent sources as $\mathbf{s}(t) = [s_1(t), \ldots, s_n(t)]^T$ where $[.]^T$ denotes the transpose operation and $t$ the discrete time index. The sources are convolved with a linear model of the physical medium (mixing matrix) which can be represented in the form of a multichannel FIR filter $\mathbf{H}$ with memory length $P$ to produce $m$ sensor signals $\mathbf{x}(t) = [x_1(t), \ldots, x_m(t)]^T$ as

$$\mathbf{x}(t) = \sum_{\tau=0}^{P} \mathbf{H}(\tau)\mathbf{s}(t-\tau) + \mathbf{v}(t) \quad (1)$$

$$\mathbf{y}(t) = \sum_{\tau=0}^{Q} \mathbf{W}(\tau)\mathbf{x}(t-\tau) \quad (2)$$

where $\mathbf{y}(t) = [y_1(t), \ldots, y_n(t)]^T$ contains the estimated sources, the sources can be estimated using a set of unmixing filter matrices $\mathbf{W}(\tau), \tau = 0, .., Q$, and $Q$ is the memory of the unmixing filters. In FDBSS the problem is transferred into the frequency domain using the short time Fourier transform STFT. Equations (1) and (2) then change respectively to:

$$\mathbf{X}(\omega, t) \approx \mathbf{H}(\omega)\mathbf{S}(\omega, t) + \mathbf{v}(\omega, t) \quad (3)$$

$$\mathbf{Y}(\omega, t) \approx \mathbf{W}(\omega)\mathbf{X}(\omega, t) \quad (4)$$

where $\omega$ denotes discrete normalized frequency. An inverse STFT is then used to find the estimated sources $\hat{\mathbf{s}}(t) = \mathbf{y}(t)$. In the following section the three approaches will be briefly studied. In Sec.3 performance measures are defined. In Sec.4 the simulation results for real world data confirm that the incorporation of prior geometric information (similarly highlighted by S. Haykin in [9]) mitigates the permutation problem and paves the path to the solution of the cocktail party problem even in a dynamic environment with moving sources, and finally conclusions are drawn.

## 2. OVERVIEW OF THE THREE FDCBSS ALGORITHMS

Parra and Spence [10] utilized second order statistics by exploiting the non-stationarity of speech to perform BSS. The non-stationarity of speech is used to provide multiple covariance matrices to be diagonalized. The solution to the permutation problem is achieved by imposing a smoothness constraint on the unmixing filters. The unmixing matrix $\mathbf{W}(\omega)$ is found across all frequency bins from

$$\mathbf{R}_Y(\omega, t_k) = \mathbf{W}(\omega)[\mathbf{R}_x(\omega, t_k) - \mathbf{\Lambda}_v(\omega, t_k)]\mathbf{W}^H(\omega) \quad (5)$$

where $(.)^H$ is the Hermitian transpose operator, $\mathbf{R}_x(\omega, t_k)$ and $\mathbf{\Lambda}_v(\omega, t_k)$ are covariance matrices of $\mathbf{X}(\omega, t_k)$ and $\mathbf{V}(\omega, t_k)$. The

Work supported by the Engineering and Physical Sciences Research Council (EPSRC) of the UK.



978-1-4244-2241-8/08/$25.00 ©2008 IEEE

covariance matrices are estimated using an averaged cross-power spectrum.The cost function based on the off-diagonal elements of $\mathbf{R}_Y(\omega, t_k)$ estimated at $t_k = kTN$, $k = 1, 2, \ldots, K$, $K$ being the number of matrices to diagonalize, is

$$\mathbf{J}(\mathbf{W}(\omega)) = \sum_{\omega=1}^{T} \sum_{k=1}^{K} \|off[\mathbf{R}_Y(\omega, t_k)]\|_F^2 \tag{6}$$

where $\|.\|_F^2$ is the squared Frobenius norm. An optimal $\mathbf{W}(\omega)$ that minimizes this cost function subject to certain constraints can be formulated as a least squares (LS) estimation problem as

$$\mathbf{W}(\omega) = \min_{W(\tau)=0, \tau>Q \ll T, W_{ii}=1} \sum_{\omega=1}^{T} \sum_{k=1}^{K} \|off[\mathbf{R}_Y(\omega, t_k)]\|_F^2 \tag{7}$$

and can be solved numerically by using a constrained gradient decent algorithm.

Wang et al. [11] proposed a solution to the BSS problem by incorporating a penalty function into the cost function (6). A penalty function is a non-negative function that is zero in the region where all constraints are satisfied (feasible region of solution space) and positive when any of the constraints are not satisfied (infeasible region of solution space). As the constraints in [10] are equality constraints, an exterior penalty function is used to convert it to an unconstrained optimization problem. We can write the modified cost function after incorporating a penalty term as

$$\mathbf{J}(\mathbf{W}(\omega)) = \sum_{\omega=1}^{T} \sum_{k=1}^{K} [\mathbf{J}_M(\mathbf{W})(\omega, k) + \lambda \mathbf{J}_C(\mathbf{W})(\omega, k)] \tag{8}$$

where $\lambda$ is a penalty weighting factor, $\mathbf{J}_M(\mathbf{W})(\omega, k)$ is the original cost function in (6) and $\mathbf{J}_C(\mathbf{W})(\omega, k)$ is the penalty function.

The key issue for successful implementation of this approach is the choice of a suitable penalty function, which depends on the nature of the constraint. We choose the penalty function as $\mathbf{J}_c(\mathbf{W})(\omega, k) = \|diag[\mathbf{W}(\omega) - \mathbf{I}]\|$. where $diag[.]$ is an operator that equates all off-diagonal elements to zero.

We can now formulate the cost to find the optimal $\mathbf{W}(\omega)$ as

$$\mathbf{J}(\mathbf{W}(\omega)) = \arg\min_{W} \sum_{\omega=1}^{T} \sum_{k=1}^{K} [\mathbf{J}_M(\mathbf{W})(\omega, k) + \lambda \mathbf{J}_C(\mathbf{W})(\omega, k)] \tag{9}$$

which has the form of a least squares problem and can be solved by an unconstrained gradient descent method with the update equation

$$\mathbf{W}^{+}(\omega) = \mathbf{W}(\omega) - (\mu_{J_M} \frac{\partial \mathbf{J}_M}{\partial \mathbf{W}^*(\omega)} + \mu_{J_C} \lambda \frac{\partial \mathbf{J}_C}{\partial \mathbf{W}^*(\omega)}) \tag{10}$$

where $(.)^*$ denotes the complex conjugate operator, $\mu_{J_M}$ and $\mu_{J_C}$ are the normalized step sizes. The permutation indeterminacy is potentially mitigated by limiting the unmixing filter lengths.

Multimodal signal processing can offer non-trivial improvements in performance over standard uni-modal signal processing. Sanei et. al. in [8] incorporated video information regarding the geometrical position of both the speakers and the microphones as shown in Figure 1, and the video information describing the geometrical position of the $j$th speaker and the $i$th microphone, the distance and the time delay between them can be calculated as $d_{ij}$ and $\tau_{ij}$ respectively. The positions and directions information used in simulations was obtained from video cameras.

$$\alpha_{ij} = \frac{\kappa}{d_{ij}^2} \cos(\theta_{ij}/r) \qquad \tau_{ij} = \frac{f_s}{C} d_{ij} \tag{11}$$

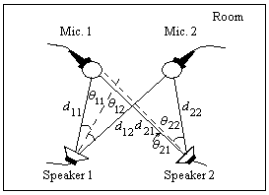


**Fig. 1**. A two-speaker two-microphone setup for recording within a reverberant (room) environment; only distances and angles between sources and microphones are shown.

where $\kappa$ is the attenuation per unit length in the medium and $r > 2$ which depends on the type of loudspeakers, $f_s$ is the sampling frequency and $C$ is velocity of sound. Considering only the direct-paths, the mixing filter $\hat{H}(\omega)$ in the frequency domain can be formulated as

$$\hat{\mathbf{H}}(\omega) = \begin{bmatrix} \alpha_{11} e^{-j\omega\tau_{11}} & \alpha_{12} e^{-j\omega\tau_{12}} \\ \alpha_{21} e^{-j\omega\tau_{21}} & \alpha_{22} e^{-j\omega\tau_{22}} \end{bmatrix} \tag{12}$$

We can now formulate a penalty function based on the distance between the unmixing filter $\mathbf{W}$ and biased estimate of the mixing filter $\hat{\mathbf{H}}$ as

$$J_C = \|\mathbf{W} - \mathbf{P}\hat{\mathbf{H}}^{-1}\|_F^2 = \|vec(\mathbf{W} - \mathbf{P}\hat{\mathbf{H}}^{-1})\|_2^2 \tag{13}$$

where $\mathbf{P}$ is the permutation matrix and $vec(.)$ converts a matrix column-wise into a vector. The optimum $\mathbf{W}$ and $\mathbf{P}$ with the overall cost function are

$$\begin{aligned} \mathbf{W}_{opt}(\omega) &= \arg\min_{\mathbf{W}} \{J_M(\mathbf{W}(\omega)) + \lambda J_C(\mathbf{W}(\omega))\} \\ \mathbf{P}_{opt}(\omega) &= \arg\min_{\mathbf{P}} \{J_C(\mathbf{W}(\omega))\} \end{aligned} \tag{14}$$

which can be found with a gradient descent algorithm and the adaptation of the permutation matrix $\mathbf{P}$ is also used in solving the permutation problem.

## 3. PERFORMANCE MEASUREMENT

In this section, we briefly describe objective performance measures that will be used for evaluation of separation. In BSS, the objective evaluation is possible only if true system parameters are known. In this paper initially the algorithms are objectively evaluated based on the real recorded room impulse responses, i.e. the mixing filters are real recorded room impulse responses, and the observed mixture signals are obtained by convolving the source signals with these real room impulse responses. Finally, the performance on the real room recordings of the same room geometry were confirmed subjectively by listening tests. In this paper the performance of the algorithms is evaluated on the basis of two criteria on real room recordings.

The SIR is calculate as in [8]

$$SIR = \frac{\Sigma_i \Sigma_\omega |H_{ii}(\omega)|^2 \langle |s_i(\omega)|^2 \rangle}{\Sigma_i \Sigma_{i \neq j} \Sigma_\omega |H_{ij}(\omega)|^2 \langle |s_j(\omega)|^2 \rangle} \tag{15}$$

where $H_{ii}$ and $H_{ij}$ represents respectively, the diagonal and off-diagonal elements of the frequency domain mixing filter, and $s_i$ is the frequency domain representation of the source of interest.

The PI [3] as a function of the overall system matrix $\mathbf{G} = \mathbf{WH}$ is given as

$$PI(G) = \Big[\frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{k=1}^{m}\frac{abs(G_{ik})}{max_k abs(G_{ik})} - 1\Big)\Big] + \Big[\frac{1}{m}\sum_{k=1}^{m}\Big(\sum_{i=1}^{n}\frac{abs(G_{ik})}{max_i abs(G_{ik})} - 1\Big)\Big] \quad (16)$$

where $G_{ik}$ is the $ik$th element of $\mathbf{G}$. The motivation for selecting this criterion is the evaluation of performance at bin level.

## 4. EXPERIMENTAL RESULTS

The simulations were performed on real recorded speech signals generated for a room geometry as illustrated in Figure 2.

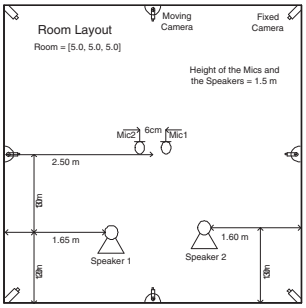


**Fig. 2**. A two-speaker two-microphone layout for recording within a reverberating (room) environment. Room impulse response length is 130 ms.

The important variables were selected as: FFT length $T = 1024$ and filter length $Q = 512$ half of $T$, $r = 4$, the sampling frequency was 8KHz and the room impulse duration was 130ms, $\lambda$ was empirically chosen (here $\lambda = 0.15$) and the learning rates $\mu$ and $\eta$ were gradually decreased with respect to the iteration index $j$

$$\mu_j = \eta_j = \gamma\frac{0.02}{1-(0.98)^j} \quad (17)$$

where $\gamma$ is a constant with $\gamma = 0.01$. In the first experiment the positions of the sensors and speakers are Mic1 $= [2.47, 2.50, 1.5]$, Mic2 $= [2.53, 2.50, 1.5]$, Speaker1 $= [1.0, 2.0, 1.5]$ and Speaker2 $= [3.5, 2.0, 1.5]$. The resulting performance indices are shown in Figure 3.

Since we know that performance index calculated by (16) is insensitive to permutation. We therefore introduce another criterion for the two sources case which is sensitive to permutation and shown for the real case for convenience, i.e. in the case of no permutation, $H = W = I$ or $H = W = [0, 1; 1, 0]$ then $G = I$ and in the case of permutation if $H = [0, 1; 1, 0]$ then $W = I$ and vice versa therefore, $G = [0, 1; 1, 0]$. Hence for a permutation free FDCBSS $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$. The results of calculating this criterion for the first experiment are shown in Figure 4.

In the second experiment we only changed the positions of the speakers. Mic1 $= [2.47, 2.50, 1.5]$, Mic2 $= [2.53, 2.50, 1.5]$, Speaker1 $= [2.00, 1.20, 1.5]$ and Speaker2 $= [3.25, 1.20, 1.5]$. The resulting performance indices are shown in Figure 5, and for this experiment we also evaluated the permutation on the basis of the criterion mentioned above and the results are shown in Figure 6.

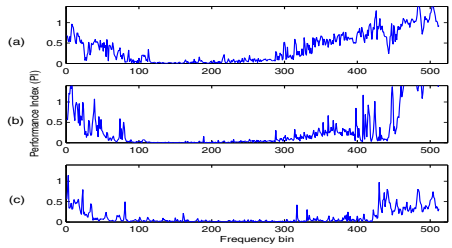


**Fig. 3**. Performance index at each frequency bin for the (a) Parra and Spence algorithm [10], (b) Wang et al. algorithm [11], and (c) multimodal FDCBSS algorithm [8]. A lower PI refers to a superior method.

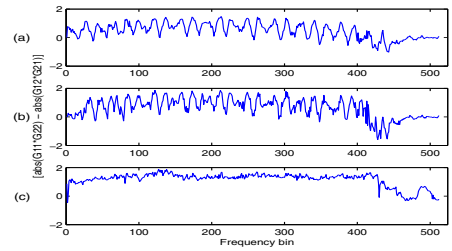


**Fig. 4**. Evaluation of permutation in each frequency bin for the (a) Parra and Spence algorithm [10], (b) Wang et al. algorithm [11], and (c) multimodal FDCBSS algorithm [8]. $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

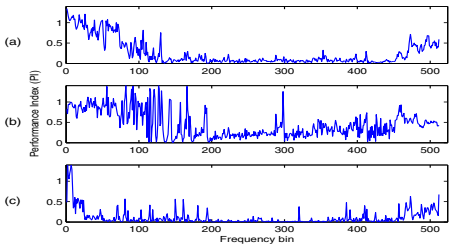


**Fig. 5**. Performance index at each frequency bin for the (a) Parra and Spence algorithm [10], (b) Wang et al. algorithm [11], and (c) multimodal FDCBSS algorithm [8].

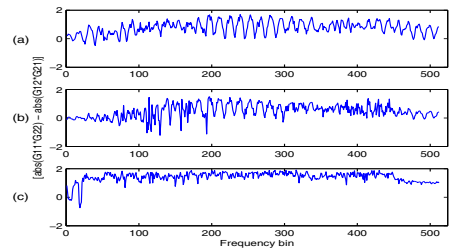


**Fig. 6**. Evaluation of permutation in each frequency bin for the (a) Parra and Spence algorithm [10], (b) Wang et al. algorithm [11], and (c) multimodal FDCBSS algorithm [8].

In the third and last experiment we reduced the distance between the microphones. Mic1 $= [2.48, 2.50, 1.5]$, Mic2 $= [2.52, 2.50, 1.5]$, Speaker1 $= [2.00, 1.20, 1.5]$ and Speaker2 $= [3.25, 1.20, 1.5]$. The resulting performance indices are shown in Figure 7, and the evaluation for permutation on the basis of the criterion mentioned above is shown in Figure 8. Figures 3(c), 5(c) & 7(c) show good performance i.e. close to zero across the majority of the frequency bins since this is due to the multimodal approach. Figures 4(c), 6(c) & 8(c) show that the multimodal FDCBSS method mitigate the permutation. Actually, in unimodal BSS no priori assumptions are typically made on the source statistics or the mixing system. On the other hand, in a multimodal approach a video system can capture the approximate positions of the speakers and the directions as done in [8]. Such video information can thereby help to estimate the unmixing matrices more accurately and ultimately increase the separation performance. We highlight that the convergence time of [11] and [8] is slightly higher than [10].

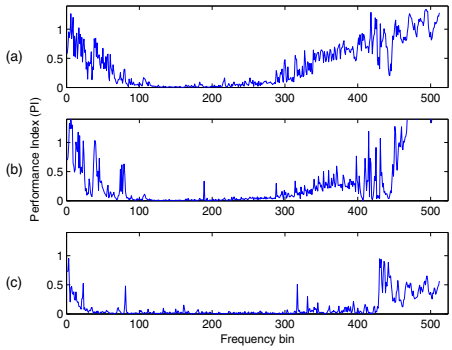


**Fig. 7**. Performance index at each frequency bin for the (a) Parra and Spence algorithm [10], (b) Wang et al. algorithm [11], and (c) multimodal FDCBSS algorithm [8]. A lower PI refers to a superior method.

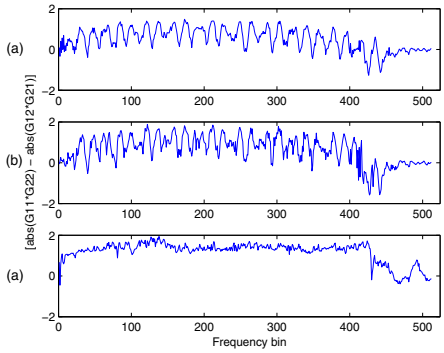


**Fig. 8**. Evaluation of permutation in each frequency bin for the (a) Parra and Spence algorithm [10], (b) Wang et al. algorithm [11], and (c) multimodal FDCBSS algorithm [8]. $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

Finally, the SIR (15) is calculated and results are shown in Table-1, and the results have been confirmed subjectively by listening tests.

**Table 1**. Comparison of SIR-Improvement between algorithms for different sets of mixtures.

| Algorithms | SIR-Improvement/dB |
|---|---|
| Parra's Method | 6.8 |
| Wang et al. Method | 9.2 |
| Multimodal FDCBSS Method | 9.8 |

## 5. CONCLUSION

In this work the separation of emerging FDCBSS algorithms was evaluated objectively by the performance indices with solution for permutation at frequency bin level and overall SIR-Improvement at different conditions of sources and subjectively by listening tests, which confirms the advantage of the multimodal FDCBSS algorithm, and the need for a multimodal solution in frequency domain BSS. This will particularly be the case with moving speakers in the cocktail party scenario (for moving speakers, the direction of the source can also be obtained from the cameras, and can be used to improve the performance of BSS), and as such the multimodal approach provides the foundation for such work.

## 6. REFERENCES


[1] S. Haykin and Ed., *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*, John Wiley, New York, 2000.

[2] A. S. Bregman, *Auditory scence analysis*, MIT Press, Cambridge, MA, 1990.

[3] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, John Wiley, 2002.

[4] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[5] S. Makino, H. Sawada, R. Mukai, and S.Araki, "Blind separation of convolved mixtures of speech in frequency domain," *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, pp. 1640–1655, Jul 2005.

[6] T. Tsalaile, S. M. Naqvi, K. Nazarpour, S. Sanei, and J. A. Chambers, "Blind source extraction of heart sound signals from lung sound recordings exploiting periodicity of the heart sound," *Proc. IEEE ICASSP, Las Vegas, USA*, 2008.

[7] C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal Of The Acoustical Society Of America*, vol. Vol 25, no. 5, pp. 975–979, September 1953.

[8] S. Sanei, S. M. Naqvi, J. A. Chambers, and Y. Hicks, "A geometrically constrained multimodal approach for convolutive blind source separation," *Proc. IEEE ICASSP, Hawaii, USA*, 2007.

[9] S. Haykin and Ed., *New Directions in Statiatical Signal Processing: From Systems to Brain*, The MIT Press, Cambridge, Massachusetts London, 2007.

[10] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. On Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[11] W. Wang, S. Sanei, and J.A. Chambers, "Penalty function based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Trans. Signal Processing*, vol. 53, no. 5, pp. 1654–1669, 2005.