Partial separation method for solving permutation problem in frequency domain blind source separation of speech signals

Reju, Vaninirappuputhenpurayil Gopalan; Koh, Soo Ngee; Soon, Ing Yann

2008

Reju, V. G., Koh, S. N., & Soon, I. Y. (2008). Partial Separation Method for Solving Permutation Problem in Frequency Domain Blind Source Separation of Speech Signals. *Neurocomputing*, 71, 2098-2112.

<https://hdl.handle.net/10356/94136>

<https://doi.org/10.1016/j.neucom.2007.08.030>

© 2008 Elsevier. This is the author created version of a work that has been peer reviewed and accepted for publication by *Neurocomputing*, Elsevier. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: [DOI: <http://dx.doi.org/10.1016/j.neucom.2007.08.030>].

Downloaded on 26 Apr 2023 16:40:33 SGT

Partial Separation Method for Solving Permutation Problem in Frequency Domain Blind Source Separation of Speech Signals

V.G.Reju, Soo Ngee Koh, Ing Yann Soon

*School of Electrical and Electronic Engineering, Nanyang Technological University,
Singapore 639798*

Abstract

This paper addresses the well known permutation problem in frequency domain blind source separation. The proposed method uses correlation between two signals in each DFT bin to solve the permutation problem. One of the signals is partially separated by a time domain blind source separation method and the other is obtained by the frequency domain blind source separation method. Two different ways of configuring the time and frequency domain blocks, i.e., in parallel or cascade, have been studied. The cascaded configuration not only achieves a better separation performance but also reduces the computational cost as compared to the parallel configuration.

Key words: Blind source separation, convolutive mixture, direction of arrival, independent component analysis, permutation problem.

Email addresses: PE2630688@ntu.edu.sg (V.G.Reju), esnkoh@ntu.edu.sg (Soo Ngee Koh), eiyysoon@ntu.edu.sg (Ing Yann Soon).

URLs: <http://www.ntu.edu.sg/eee/home/esnkoh> (Soo Ngee Koh),
<http://www.ntu.edu.sg/home/eiyysoon> (Ing Yann Soon).

20 Blind source separation (BSS) is the technique of separating the sources from
21 their mixtures without any prior knowledge about the sources or the mix-
22 ing process. Independent component analysis (ICA) is one of the techniques
23 commonly used for blind source separation. There are many algorithms for
24 separation of linear instantaneous mixtures where the mixing matrix consists
25 of simple scalars and so is the unmixing matrix. However in real acoustic
26 environments, the mixing is very complex convolutive mixing, so that finite
27 impulse response (FIR) filter matrix is required in place of simple matrix of
28 scalars. This makes the problem of blind source separation a challenging task.
29 For example, the speech signal mixed in a real acoustic environment normally
30 requires thousands of filter taps for the unmixing filter.

31 There are two main approaches for the BSS of convolutive mixtures. The first
32 one is the time domain approach and the second is the frequency domain
33 approach. Also there is a combination of the above two, where the compu-
34 tational complexity of the time domain method is reduced by implementing
35 the time domain convolution operation in the frequency domain [13,8]. The
36 main disadvantage of the time domain method is that it is computationally
37 intensive and convergence speed is low for long filters. Whereas, the frequency
38 domain method is computationally efficient as the convolution in time do-
39 main becomes simple element-wise multiplication in the frequency domain
40 and hence the complex valued ICA algorithm can be applied to each discrete
41 Fourier transform (DFT) bin. However, the method has the disadvantage of
42 inconsistent permutation in the DFT bins after separation by ICA. This is the
43 popularly known permutation problem of frequency domain BSS. The algo-

44 rithms used for solving the permutation problem will align the permutation
45 in the DFT bins so that the separated signal in the time domain will contain
46 the frequency components from the same source signals.

47 Various methods have been proposed to solve the permutation problem in
48 the frequency domain BSS. L. Parra et al [18] suggested a method which
49 constrains the length of the filter but this is not suitable for real acoustic envi-
50 ronments where the length of the separation filter is of the order of thousands.
51 Smoothing of the separation matrix is another method [24,5]. The property
52 that for speech signals, the adjacent bands are highly correlated is utilized
53 in [15] to solve the permutation problem. In [19] the reliability of the cor-
54 relation method is improved with the dyadic permutation sorting algorithm.
55 Direction of arrival (DOA) is another popular method, where the directivity
56 pattern formed by the separation matrix in the frequency domain is used to
57 estimate the direction of the sources and hence the permutation problem is
58 solved [14,22,11,12]. The DOA method cannot solve the permutation problem
59 in all the DFT bins especially for lower frequency bins. H. Sawada et al [23]
60 combined the DOA method with the correlation methods to form a robust
61 method for solving the permutation problem which is suitable for all the cases
62 except when the sources are very close to each other or collinear.

63 In Section 2, the problem of BSS is introduced and Section 3 summarizes
64 the drawbacks of the existing methods and a new method is proposed in
65 Section 4, which is applicable to collinear sources also, provided a time domain
66 BSS method which can separate the signals at least partially so that the
67 spectra of the separated signals will be closer to those of the sources. Section 5
68 summarizes the experimental results and conclusions are drawn in Section 6.

70 The problem of BSS can be explained as follows. Suppose there are P sources
 71 mixed convolutively to obtain Q mixtures, then the Q mixtures at the sensor
 72 output can be written as

$$\mathbf{x}(t) = \mathbf{H} * \mathbf{s}(t) \quad (1)$$

73 where $\mathbf{x}(t) = [x_1(t), x_2(t), \dots, x_Q(t)]^T$ are the sensor outputs and $\mathbf{s}(t) =$
 74 $[s_1(t), s_2(t), \dots, s_P(t)]^T$ are the mutually independent sources. The super-
 75 script T represents the matrix transpose and the operator $*$ denotes the con-
 76 volution. The matrix $\mathbf{H}$ is the mixing matrix of order $Q \times P$ whose qp^{th}
 77 element, $h_{qp}(l)$ is the impulse response from source p to sensor q so that
 78 $x_q(t) = \sum_{p=1}^P \sum_{l=0}^{\infty} h_{qp}(l) s_p(t-l)$, for $q = 1, \dots, Q$. Even though in real acous-
 79 tic environments the length of the mixing filter is infinity, the filter coefficients
 80 will decay to negligibly small values after a certain time and hence the length
 81 of the unmixing filter can be limited to a finite value.

82 Unlike the blind deconvolution, the objective of BSS is to separate the sources
 83 $\mathbf{s}(t)$ from the mixtures $\mathbf{x}(t)$ in such a way that they are mutually independent
 84 [26]. Hence the separated signals $y_r(t)$, $r = 1, \dots, P$ will be the filtered version
 85 of the original sources $s_{\Pi(r)}(t)$ [23]

Fig. 1. BSS for convolutive mixtures in the case of two sources and two sensors

$$y_r(t) = \sum_l \alpha_r(l) s_{\Pi(r)}(t-l) \quad (2)$$

where $\alpha_r(l)$ is a filter, $\Pi : \{1, \dots, P\} \rightarrow \{\Pi(1), \dots, \Pi(P)\}$ is the permutation and $s_{\Pi(r)}(t)$ represents r^{th} signal with permutation. Since the mixing filters can be considered as finite length filters, the sources can be separated using the unmixing filters of finite length L and hence the output will be

$$\mathbf{y}(t) = \mathbf{W} * \mathbf{x}(t) \quad (3)$$

where $\mathbf{y}(t) = [y_1, y_2, \dots, y_P]^T$ are the separated signals and $\mathbf{W}$ is a $P \times Q$ matrix of FIR filters with elements $w_{rq}(l)$, $r = 1, \dots, P$, $q = 1, \dots, Q$ and $l = 0, 1, \dots, L$. The mixing and unmixing processes are depicted in Fig.1, for the case of two sources and two sensors. Using the convolution multiplication property of DFT, (1) and (3) can be written as shown in (4) and (5).

$$\mathbf{X}(f, t) = \mathbf{H}(f) \mathbf{S}(f, t) \quad (4)$$

$$\mathbf{Y}(f, t) = \mathbf{W}(f) \mathbf{X}(f, t) \quad (5)$$

where $\mathbf{X}(f, t)$, $\mathbf{S}(f, t)$ and $\mathbf{Y}(f, t)$ are the discrete time-frequency transformation vectors of $\mathbf{x}(t)$, $\mathbf{s}(t)$ and $\mathbf{y}(t)$ respectively. $\mathbf{H}(f)$ and $\mathbf{W}(f)$ are respectively the mixing and unmixing filters in the frequency domain. By observing (5), we can see that, conceptually any linear ICA algorithm for complex numbers can now be applied to each DFT bin. In this paper, the information maximization approach [6] combined with natural gradient (NG) [3] and FastICA algorithms [7] are used in each DFT bin. The information maximization

combined with NG algorithm is used to initialize the unmixing matrix for the FastICA algorithm, which can improve the separation performance. The NG algorithm is initialized with unit matrix.

It is a well known fact that, in BSS the sources can at best be obtained up to a scaling and permutation [27], i.e.,

$$\mathbf{W}(f) \mathbf{H}(f) = \Gamma(f) \mathbf{D}(f) \quad (6)$$

where $\Gamma(f)$ is the permutation matrix and $\mathbf{D}(f)$ is the scaling diagonal matrix. Many methods have been proposed to solve the scaling problem. In this paper, the method suggested in [15] is used. The permutation problem in the time domain method, called global permutation, is not a serious issue, because it is not affecting the quality of the separated signals. In contrast, the permutation problem in the frequency domain, called local permutation, is complicated as it is in each of the frequency bins, which will affect the separation performance. Hence the permutations in each of the DFT bins is to be aligned in such a way that the separated signals in the time domain will contain the frequency components of the same source signals. The following section summarizes the drawbacks of the two popular methods, out of the many existing methods.

3 Drawbacks of the existing methods

The most commonly used approaches for solving the permutation problem are the DOA approach and the correlation approach. For the convenience of readers, the basic principles and pitfalls of these methods are summarized

below.

3.1 Direction Of Arrival approach

The Fourier transform of the impulse response from source $s_p(t)$ at direction θ_p to the q^{th} sensor, $h_{qp}(t)$, if we neglect the room reverberation, can be approximated as [14,22,23]

$$H_{qp}(f) = e^{j2\pi f \tau_{qp}}. \quad \tau_{pq} = c^{-1} d_q \sin \theta_p \quad (7)$$

where τ_{qp} is the time lag for the source with respect to the q^{th} sensor placed at position d_q (assuming that the sensors are placed in a linear array and the direction orthogonal to the array is 0°) and c is the velocity of the signal. Using (4) and (5), the frequency response from p^{th} source to r^{th} separated signal can

Fig. 2. Directivity pattern of the two sources shown in Fig.5 at two different frequencies.

131 be written as [14,22,23]

$$\begin{aligned}
 U_{rp}(f) &= \sum_{q=1}^Q W_{rq}(f) H_{qp}(f) \\
 &= \sum_{q=1}^Q W_{rq}(f) e^{j2\pi f c^{-1} d_q \sin \theta_p}
 \end{aligned} \tag{8}$$

132 If we consider the direction θ_p as a variable, say θ , the modified equation will
 133 be [14,22,23]

$$U_r(f, \theta) = \sum_{q=1}^Q W_{rq}(f) e^{j2\pi f c^{-1} d_q \sin \theta} \tag{9}$$

134 The gain $|U_r(f, \theta)|$ will vary according to the direction of the angle θ and
 135 hence it is called the directivity pattern. The value of the gain $|U_r(f, \theta)|$ will
 136 be sufficiently high when θ is equal to the direction of the source corresponding
 137 to the separated signal y_r and low in the directions of the other sources. Hence
 138 the directions of all the sources can be estimated by the rows of the unmixing
 139 filter, $W_{rq}(f)$, $r = 1, \dots, P$, for all the frequency bins. The permutation
 140 problem in each bin can be fixed by proper clustering of the above estimated
 141 directions. However, as shown in Fig.2, for lower frequencies we will not be
 142 able to estimate the directions of the sources and hence we are unable to
 143 fix the permutation for these bins. Also, under heavy reverberant conditions
 144 we may not be able to solve the permutation problem for all the frequency
 145 bins, even for the higher frequencies and hence, the permutations in these bins
 146 have to be fixed by some other methods. Another major disadvantage of using
 147 DOA method is that, when the sources are collinear or at very small angle,

148 the method will fail.

149 3.2 Correlation approach

150 For speech signals there is a strong correlation between the adjacent frequency
 151 bands. This inter-frequency correlation is utilized to solve the permutation
 152 problem in correlation methods (see [15] for more details). In [23], the algo-
 153 rithm proposed in [15] is simplified so that the envelop $v_r^f(t) = |S_{\Pi(r)}(f, t)|$ is
 154 used to measure the correlation between the neighboring bands. The correla-
 155 tion between the envelopes in the neighboring frequency bins will be high, if the
 156 separated signal belongs to the same source signal. If Π_f is the permutation
 157 corresponding to $\Gamma^{-1}(f)$, the permutation at frequency f , can be estimated
 158 by maximizing the sum of the correlation between the neighboring frequencies
 159 within the frequency distance δ [23].

$$\Pi_f = \operatorname{argmax}_{\Pi} \sum_{|g-f| \leq \delta} \sum_{r=1}^P \operatorname{cor}(v_{\Pi(r)}^f, v_{\Pi_g(r)}^g) \quad (10)$$

160 where Π_g is the permutation at frequency g . The correlation between two
 161 signals $x(t)$ and $y(t)$ is defined as [23]

$$\operatorname{cor}(x, y) = \frac{(\mu_{xy} - \mu_x \mu_y)}{\sigma_x \sigma_y} \quad (11)$$

162 where μ_x is the mean and σ_x is the standard deviation of x . The main dis-
 163 advantage of this method is that a mistake in one frequency bin will lead to
 164 complete misalignment beyond that frequency bin.

166 In the work of H. Sawada et al [23], the DOA and correlation approaches were
 167 combined so that the permutation is fixed for certain frequency bins where the
 168 confidence of DOA method was sufficiently high. For the remaining bins the
 169 permutation was decided based on the correlation between the neighboring
 170 or harmonic frequency bins without changing the permutation fixed by the
 171 DOA method. Harmonic correlation method utilizes the fact that, for speech
 172 signals, the envelop $v_r^f(t)$ at frequency f has strong correlation to that at its
 173 harmonic frequencies $2f, 3f$ and so forth.

174 Hence, if the permutation is not fixed for frequency f , but fixed for its har-
 175 monic components, then the permutation at f, Π_f , can be fixed by maximizing
 176 the correlation between $v_r^f(t)$ at frequency f and its harmonic bins [23], i.e.,

$$\Pi_f = \operatorname{argmax}_{\Pi} \sum_{g \in (\text{Harmonics of } f)} \sum_{r=1}^P \operatorname{cor}(v_{\Pi(r)}^f, v_{\Pi_g(r)}^g) \quad (12)$$

177 The reason for using the harmonic correlation method in addition to the neigh-
 178 boring bands correlation method is that, the DOA approach does not provide
 179 sufficient confidence to fix the permutation, continuously for a certain range
 180 of frequencies, which is at the lower frequency bins. In such case, the use of
 181 neighboring correlation method alone may result in consecutive misalignment.

183 The proposed system basically consists of two blocks, namely, the time do-
 184 main and frequency domain blocks. The time domain block partially separates
 185 the signal mixture using the time domain algorithm for BSS, whereas the fre-
 186 quency domain block separates the signals using the frequency domain BSS
 187 algorithm. The permutation problem in the DFT bins of the frequency domain
 188 BSS block is solved by utilizing the correlation between the envelop, $|\hat{S}_q(f, t)|$,
 189 in the DFT bins of the partially separated signal (partially separated signal
 190 is converted to the DFT domain time series signals, $\hat{S}_q(f, t)$) and that of the
 191 frequency domain method.

192 The input to the frequency domain stage can either be the output from the
 193 microphones or the partially separated signals from the time domain stage.
 194 Correspondingly there are two configurations for the proposed system, namely
 195 the parallel configuration [20] and the cascaded configuration [21]. The detailed
 196 explanations for each of these configurations are given in the following sections.
 197 Before proceeding to the next section, the time domain BSS algorithm used
 198 in this paper is summarized below for the convenience of the readers.

199 The time domain algorithm used in this paper is the computationally efficient
 200 implementation [1] of the algorithm proposed in [8,9]. The algorithm is based
 201 on the second order statistics which utilizes the non-stationarity and non-
 202 whiteness property of the speech signals. The cost function for the algorithm
 203 is [8,1,9].

$$J(m) = \sum_{i=0}^m \beta(i, m) \left\{ \log \left(\det \left(\text{bdiag} \left(\mathbf{Y}^T(i) \mathbf{Y}(i) \right) \right) \right) - \log \left(\det \left(\mathbf{Y}^T(i) \mathbf{Y}(i) \right) \right) \right\} \quad (13)$$

204 where β is the weighing function, m is the block index (the speech signal is
 205 divided into different blocks) and bdiag represents the block diagonal oper-
 206 ation. $\mathbf{Y}(m) = [\mathbf{Y}_1(m), \dots, \mathbf{Y}_P(m)]$ is the block output signal matrix, the
 207 columns of $\mathbf{Y}_r(m)$ contains blocks of the r^{th} output signal, $y_r(t)$, of length N
 208 samples and each column is delayed by one sample as

$$\mathbf{Y}_r(m) = \begin{bmatrix} y_r(mL_t) & \cdots & y_r(mL_t - L_t + 1) \\ y_r(mL_t + 1) & \cdots & y_r(mL_t - L_t + 2) \\ \vdots & \ddots & \vdots \\ y_r(mL_t + N - 1) & \cdots & y_r(mL_t - L_t + N) \end{bmatrix} \quad (14)$$

209 where N is the length of the output block and L_t is the length of the time
 210 domain unmixing filter. Hence the size of the matrix $\mathbf{Y}(m)$ will be of the order
 211 of $N \times L_t P$. The natural gradient of (13) with respect to the unmixing filter
 212 in time domain, $\mathbf{W}_t$, gives [1,9]:

$$\nabla_{\mathbf{W}_t}^{NG} J(m) = 2 \sum_{i=0}^m \beta(i, m) \mathbf{W}_t(i) \{ \mathbf{R}_{\mathbf{Y}\mathbf{Y}}(i) - \text{bdiag}(\mathbf{R}_{\mathbf{Y}\mathbf{Y}}(i)) \} \text{bdiag}^{-1}(\mathbf{R}_{\mathbf{Y}\mathbf{Y}}(i)) \quad (15)$$

213 where the $L_t P \times L_t P$ correlation matrix $\mathbf{R}_{\mathbf{Y}\mathbf{Y}}$ consists of the correlation ma-

214 trices $\mathbf{R}_{\mathbf{y}_p \mathbf{y}_q}(m) = \mathbf{Y}_p^T(m) \mathbf{Y}_q(m)$ of size $L_t \times L_t$. When the output signals
 215 are mutually independent $\mathbf{R}_{\mathbf{y}_p \mathbf{y}_q}(m) = 0$, for $p \neq q$. Since the temporal cor-
 216 relation of the speech signal is also taken into consideration, the algorithm is
 217 free from whitening problem. The direct computation of the (15) is complex
 218 but an efficient and approximated version of the algorithm, which is given in
 219 [1], is used in this paper.

220 4.1 Parallel configuration

221 The block diagram of the proposed method for the parallel configuration with
 222 two sources is shown in Fig.3. For simplicity, we assume that the number of
 223 sources is equal to the number of mixed signals, i.e., $P = Q$. The time domain
 224 signals at the microphone outputs are first converted to the frequency domain
 225 time series signals, $X_q(f, t)$, $q = 1, \dots, Q$, using K point DFT. Complex
 226 valued ICA algorithm for linear mixtures is then applied to all the K frequency

Fig. 3. Block diagram for the proposed method (parallel configuration).

bins so as to obtain the separated signals in each bin. The separated signals in different frequency bins may have different permutations that have to be solved before they are combined by the inverse discrete Fourier transform (IDFT) to obtain the separated signals, $y_r(t)$, $r = 1, \dots, P$. To solve this permutation problem, we use partially separated signals, $\hat{s}_p(t)$, $p = 1, \dots, P$, separated in the time domain as shown in Fig.3. The listening quality of the partially separated signal need not be very good but the spectra of the separated signals must be closer to the corresponding sources.

Each partially separated signal in the time domain, $\hat{s}_p(t)$, is then transformed to the frequency domain using DFT of the same length as that for the microphone output, i.e., K . Now k^{th} bin of p^{th} partially separated signal, $\hat{S}_p(f_k, t)$, will be more correlated to the corresponding bin of the fully separated signal $S_{\Pi(p)}(f_k, t)$. So the frequency bin among, $S_{\Pi(q)}(f_k, t)$, $q = 1, \dots, P$, having the highest correlation with $\hat{S}_p(f_k, t)$ is identified, which is $S_{\Pi(q)}(f_k, t)$, for $q = p$, and is assigned to $Y_p(f_k, t)$. Since the adjacent bins of a speech signal are highly correlated, instead of taking a single bin from the partially separated signal, the average of k^{th} and its adjacent bins is used. Hence the permutation, $\Pi_{f_k}(p)$, for the k^{th} bin is

$$\Pi_{f_k}(p) = \underset{\Pi(q), q=1, \dots, P}{\operatorname{argmax}} \operatorname{cor} \left(\frac{1}{2\Delta k + 1} \sum_{b=k-\Delta k}^{k+\Delta k} |\hat{S}_p(f_b, t)|, |S_{\Pi(q)}(f_k, t)| \right) \quad (16)$$

The same procedure is repeated for all $\hat{S}_p(f_k, t)$, $p = 1, \dots, P$, and $k = 1, \dots, K$. Subsequently, these assignments $Y_p(f_k, t)$ have to be converted back to the time domain signal $y_p(t)$ using IDFT. One problem with this method

248 is that we may not be able to solve the permutations for all the bins, with full
 249 confidence, which can be explained as follows:
 250 Let $c_{pq,k}$, $p, q = 1, \dots, P$, be the correlation between k^{th} bin of the p^{th} partially
 251 separated signal and the q^{th} fully separated signal. For P sources there will be
 252 $P \times P$ such correlations for each bin. Considering $c_{pq,k}$ is an element in the
 253 p^{th} row and q^{th} column of a $P \times P$ matrix and if the highest P values are
 254 distributed in that matrix in such a way that there is no column or row, com-
 255 mon to each other, we can say that each partially separated signal has one and
 256 only one correlation with the fully separated signal. In that case $S_{\Pi(q)}(f_k, t)$
 257 is assigned to $Y_p(f_k, t)$ with confidence. If not, leave that bin as it is and use
 258 the correlation between neighboring or harmonic bins to solve the permuta-
 259 tion problem. For these bins also, permutations are fixed after checking the
 260 confidence. To check the confidence we use the same procedure that has been
 261 explained above. For the remaining bins, the neighboring correlation approach
 262 [23], which is shown in (10), is used to solve the permutation problem.

263 For clarification the confidence checking procedure can be illustrated with an
 264 example as follows: Let the number of sources be $P = 2$ and the correlations
 265 at bin k are $c_{11,k} = 0.6$, $c_{22,k} = 0.7$, $c_{12,k} = 0.3$ and $c_{21,k} = 0.2$. Here, the
 266 highest two correlations are $c_{11,k}$ and $c_{22,k}$, which are the diagonal elements, if
 267 we consider the correlations as the elements of a matrix and hence there is no
 268 common row and column for the highest elements. In this case we can solve
 269 the permutation problem with confidence. Instead, if $c_{11,k} = 0.8$, $c_{22,k} = 0.1$,
 270 $c_{12,k} = 0.3$ and $c_{21,k} = 0.4$, we leave that bin unaltered during the first round of
 271 solving the permutation (the round which uses confidence check), because the
 272 highest two correlations, $c_{11,k}$ and $c_{21,k}$, lie in the same column of the matrix.
 273 During the second round, where we use the correlation without confidence

check, we only check either $c_{11,k} + c_{22,k} \geq c_{12,k} + c_{21,k}$ or $c_{11,k} + c_{22,k} < c_{12,k} + c_{21,k}$, and solve the permutation problem accordingly, which is the method adopted in [23]. Hence the proposed method consists of the following three steps:

- Fix the permutation using the correlation between the partially separated signal and the fully separated signal for the bins where it can be fixed with confidence.
- Fix the permutation either using the adjacent bins correlation or harmonic correlation method with confidence, without changing the permutation fixed in the previous step.
- For the remaining bins, fix the permutation using (10).

4.2 Cascaded configuration

Certain time domain algorithms will distort the spectrum of the separated signals. For example, the algorithm which does not consider the temporal correlation of the speech signal (e.g. [2]) may whiten the spectrum of the sep-

Fig. 4. Block diagram for the proposed method (cascaded configuration).

288 arated signal. For such time domain algorithms the parallel configuration will
 289 be better [20], otherwise the output of the frequency domain stage also will
 290 remain distorted because of its distorted input. On the other hand, if the al-
 291 gorithm is not distorting the spectrum of the separated signals, the cascaded
 292 configuration will be the best option as it not only reduces the computa-
 293 tional cost but also improves the overall separation performance. The block
 294 diagram for the cascaded configuration is shown in Fig.4. By comparing the
 295 two block diagrams, parallel and cascaded, it can be seen that the number
 296 of fast Fourier transform (FFT) blocks for the cascaded configuration is only
 297 2, whereas for the parallel configuration is 4. Hence the computational cost
 298 is reduced. Since the input signal to the frequency domain stage in the case
 299 of cascaded configuration is the partially separated signals, which is already
 300 separated to a certain level, the overall performance is higher than that of the
 301 parallel configuration. In [16,17] it is shown that the cascaded configuration of
 302 the time domain and the frequency domain stages will improve the separation
 303 performance. Unlike in [16,17], where the DOA method is used to solve the
 304 permutation problem in the frequency domain stage, the proposed algorithm
 305 uses the signals from the time domain stage to solve the permutation problem
 306 and hence the computational cost is optimally utilized.

307 **5 Experimental results**

308 For performance analysis of the proposed algorithms, both simulated and mea-
 309 sured room impulse responses are used. The simulated room impulse responses
 310 are used in Sections 5.1 and 5.2. The wall reflections up to 10^{th} order is taken
 311 and humidity, temperature, absorption of sound due to air etc., are consid-

Table 1
Experimental Conditions

Source signals	Speech of 15sec, except in Section 5.4
Direction of sources	As shown in the respective figures
Distance between two microphones	20cm for PS and 4cm for DOA
Sampling rate f_s	16kHz
Frame size of DFT	$K = 4096$
Frequency resolution	$\Delta f = f_s/K = 3.90625\text{Hz}$
Distance Δk	$18\Delta f$
Distance δ	$9\Delta f$
Number of filter taps for time domain algorithm	512 (unless otherwise specified)

Fig. 5. The source-microphone configuration for the room impulse responses simulation.

312 ered for calculating the impulse responses [28]. Whereas, in Sections 5.3, 5.4
313 and 5.5, the measured real room impulse response is used. For all the experi-
314 ments, the average performance of 10 speech utterances are used to evaluate
315 the performance. The algorithms are also tested with real recorded speech
316 mixtures.

Fig. 6. Separation performance of the proposed method (reverberation time $TR_{60} = 86\text{ms}$): Partial Separation approach PS, Correlation approach C1 and the combined approaches PS+C1, PS+C2+C1 and PS+C2+Ha +C1

Fig. 7. NRR at different frequencies for the 4th set of speech utterances in Fig.6.

5.1 Performance evaluation for collinear and non-collinear sources

The major disadvantages of the time domain BSS method for convolutive mixtures are the statistical interdependence among the filter coefficients, which

hinders the convergence, and the heavy computational cost involved for long filter taps [25]. Furthermore, good separation of convolutive mixture using time domain method when the sources are collinear is a formidable task [10]. To show that the proposed algorithm can solve the permutation problem even when the quality of the partially separated signal is very poor, we simulated the impulse response between the sources and microphones for the configurations shown in Fig.5. The sources are placed in two different configurations: *i*) collinear (Sources 2 and 2'), for which the time domain separation is very poor *ii*) non-collinear (Sources 3 and 3'), for which the time domain separation is good. The mixed signals are generated by convolving the impulse responses obtained for the above configurations with the speech signals so that the performance can be evaluated by calculating the noise reduction rate (NRR) [16]

$$NRR = \frac{1}{P} \sum_{q=1}^P (SNR_q^{(O)} - SNR_q^{(I)}) \quad (17)$$

$$SNR_q^{(O)} = 10 \log_{10} \frac{\sum_f |A_{qq}(f)S_q(f)|^2}{\sum_f |A_{qp}(f)S_p(f)|^2} \quad (18)$$

$$SNR_q^{(I)} = 10 \log_{10} \frac{\sum_f |H_{qq}(f)S_q(f)|^2}{\sum_f |H_{qp}(f)S_p(f)|^2} \quad (19)$$

where $SNR_q^{(O)}$ and $SNR_q^{(I)}$ are the output SNR and the input SNR respectively, $q \neq p$. $A_{qp}(f)$ is the q^{th} row and p^{th} column of the matrix $\mathbf{A}(f)$, as given in (20)

$$\mathbf{A}(f) = \mathbf{W}(f)\mathbf{H}(f) \quad (20)$$

For each configuration of the source positions shown, both parallel and cascaded configurations of the proposed methods are used to solve the permuta-

tion problem. In each case, the permutation is solved in five different ways:

- Partial Separation method (PS). Bins whose permutation could not be solved with confidence are left unaltered.
- Correlation between adjacent bins (C1) using (10)
- Combination of PS and C1 (PS+C1), i.e., PS followed by C1.
- PS followed by C2, that solves permutation by using correlation between adjacent bins with confidence, and finally followed by C1 (i.e., PS+C2+C1)
- Combination of PS and C2, followed by Ha which solves permutation using correlation between the harmonic components with confidence, and finally followed by C1 (i.e., PS+C2+Ha+C1).

Fig.6 shows the performance of the proposed method for both collinear and non-collinear sources. For clarity, only the average performance of the 10 sets of speech utterances are shown, except for the collinear sources in the parallel configuration where the performances of all the 10 sets of speech utterances are shown. For the parallel configuration, since the time domain NRR is not directly observable at the output, it is shown with ‘dash-dot’ line. As reported in [23] and from Fig.6, it can be seen that C1 can solve the permutation problem in most of the cases but it is not stable and sometimes it results in very poor performance. The performance of PS alone is not good but stable as PS solves the permutation for the frequency bins where the confidence is sufficiently high. However, for PS+C1, the performance is improved in terms of NRR compared to PS and stability compared to C1 alone. After solving permutation by PS followed by C2 and before C1, the performance is again improved. Also it can be seen that PS+C2+Ha+C1 offers almost the same performance as that without Ha. This is because Ha is used to solve permutation problem at lower frequencies where the problem is already solved by

363 PS method. The NRR measured at each frequency bin for a pair of speech
 364 utterances (set 4) using different methods are shown in Fig.7. It shows the
 365 effectiveness of the proposed method in solving the permutation misalignment
 366 problem. There are large regions of permutation misalignment when C1 alone
 367 is used. Since PS can provide correct permutation for certain frequencies in
 368 these regions, the problem is solved when PS is combined with C1. Unlike
 369 the DOA approach, where it is very hard to estimate the direction of arrival
 370 at lower frequencies and hence difficult to solve the permutation with con-
 371 fidence, PS solves the permutation problem almost uniformly irrespective of
 372 the frequencies as shown in Fig.7. Note that the case shown in Fig.7 is for the
 373 worst case, where the NRR for the time domain method is only 3.9812dB and
 374 also it is for the parallel configuration. From Fig.6, it can be seen that even
 375 for collinear sources with poor separation in time domain, if we use cascaded
 376 configuration the performance will be significantly better. In Fig.6 for the cas-
 377 caded configurations, the performance shown for the correlation method alone
 378 (C1) is better than the all other methods, because C1 was successful for all
 379 the 10 sets of speech utterances. Whereas for the parallel configuration, the
 380 method failed for some of the sets. This is not because of the configuration but
 381 because of the reason that the correlation method alone is highly unreliable.

382 5.2 *Performance evaluation under different reverberation times*

383 For further analysis, the performance of the algorithm is compared with the
 384 DOA method. Since DOA utilizes the direction of arrival of the signal for
 385 solving permutation problem, the distance between microphones must be less
 386 than $c/2f$, where c is the velocity and f is the frequency of the signal, to avoid

387 spatial aliasing. However, when the microphones are very close, the problem
388 reduces to the single channel mixing case. For the time domain method used
389 in this paper more spacing between the microphones is required to achieve
390 separation. Table 2 shows the performance of the time domain and DOA
391 method for different microphone spacings, while keeping all other parameters
392 such as the position of the sources, central point of the microphones, room
393 surface absorption, size of the room etc., same as that in other experiments.
394 Since the performance of the PS method depends on the quality of the partially
395 separated signal, the NRR of the time domain method is shown instead of
396 the NRR of the PS method. Also while calculating the NRR for the DOA
397 method, the bins where the DOA method could not solve the permutation
398 with confidence are left unaltered. This is because, the robustness of the DOA
399 method followed by correlation methods depends on the success of the DOA
400 method alone in solving the permutation problem. From the table we can see
401 that 4cm is the optimum distance for the DOA method. For the PS method
402 20cm is taken. Note that this comparison is done only to find out the optimum
403 microphone distance for the DOA method. Since the distances between the

Table 2

NRR for the time domain method and DOA method for different microphone spacings (Room surface absorption = 0.5)

Spacing (cm)	Time domain (dB)	DOA (dB)
2	1.9785	6.3786
4	2.7985	6.8414
8	4.2190	6.2397
12	6.2419	5.2858
16	7.8860	4.3331
20	8.2101	3.6543

Fig. 8. Room impulse response for different values of surface absorption: 0.3 ($TR_{60} = 235\text{ms}$), 0.5 ($TR_{60} = 130\text{ms}$), 0.7 ($TR_{60} = 86\text{ms}$) and 0.9 ($TR_{60} = 63\text{ms}$). Only the impulse responses from Source 1 to Microphone 1 are shown.

Fig. 9. Performance comparison of PS method alone with DOA method alone as a function of room surface absorption.

404 microphones are not the same, the following comparisons are not exact but
 405 they will give an approximate result.

406 For the experiments in this section, room impulse response is simulated [28]
 407 for different values of room surface absorption. The impulse responses thus
 408 obtained are shown in Fig.8. The other experimental conditions are shown in
 409 Table. 1. The positions of the sources are shown in Fig.5 (Sources 1 and 1').

Fig. 10. Performance comparison of PS method alone without confidence check with PS method after confidence check followed by the methods which utilizes the correlation between adjacent and harmonic bins, for parallel and cascaded configurations. The DOA method after confidence check followed by correlation methods are also shown.

410 The performance of the two methods, PS and DOA, when used alone for
 411 different values of reverberation time are shown in Fig.9. For both the cases of
 412 ‘DOA only’ and ‘PS only with confidence check’ the bins whose permutation
 413 could not be fixed with confidence are left unaltered.

414 Unlike the correlation approach discussed in Section 3.2, which utilizes the
 415 inter-frequency correlation of the signal for permutation alignment [15], the PS
 416 method utilizes the correlation with the partially separated signal for solving
 417 permutation. So, even if the permutation fixed in one bin is wrong, the per-
 418 mutation alignment in the remaining bins will not be affected by the wrongly
 419 aligned one. Hence the reliability of the method is high and we can fix the
 420 permutation in all the bins without confidence check. This is shown as ‘PS
 421 only without confidence check’ in Fig.9, where the permutation of all the bins
 422 are fixed without checking the confidence as

Fig. 11. The source-microphone configuration for the measurement of real room impulse response

$$\Pi_{f_k} = \underset{\Pi}{\operatorname{argmax}} \sum_{p=1}^P \operatorname{cor} \left(\frac{1}{2\Delta k + 1} \sum_{b=k-\Delta k}^{k+\Delta k} \left| \hat{S}_p(f_b, t) \right|, \left| S_{\Pi(p)}(f_k, t) \right| \right) \quad (21)$$

for all the bins, where Π_{f_k} is the permutation at k^{th} frequency bin and $2\Delta k + 1$ is the total number of adjacent bins of the partially separated signal taken to obtain the average. From Fig.10, it can be seen that the difference in performance of PS alone without confidence check ('PS only') and PS with confidence check followed by the methods utilizing correlation between adjacent and harmonic bins ('PS followed by others') are very small. Therefore, the PS method alone can be used to solve the permutation problem which reduces the computational cost at the expense of very small performance reduction. In Fig.10, the performance of the DOA method followed by other correlation methods ('DOA followed by others') are also shown for comparison. The reasons for the high deviation in performance of the time domain method, for the changes in room reverberation time, as shown in Fig.9 and 10 are explained in Section 5.5.

Fig. 12. NRR for various algorithms using real room impulse response. PS - Partial Separation method with confidence check, C1 - Correlation between the adjacent bins without confidence check, C2 - Correlation between adjacent bins with confidence check, Ha - Correlation between the harmonic components with confidence check, PS1 - Partial separation method alone without confidence check.

5.3 Performance evaluation using the measured real room impulse response

The performance of the PS method is also evaluated using the measured impulse response of a real furnished room. The reverberation time of the room (TR_{60}) is 187ms and is measured with the help of an acoustic impulse response measuring software “Sample Champion” [29]. The microphone and loud speaker transfer function are neglected in the measurement. The positions of the sources and sensors are shown in Fig.11. The other experimental conditions are same as those used in the previous sections. The calculated performances are shown in Fig.12, from where it can be seen that the performance of PS alone without confidence check is very close to that of PS+C2+Ha+C1. Therefore the PS method alone, which is reliable and independent of the positions of the sources and better than the DOA approach, can be used to solve the permutation problem in frequency domain BSS of speech signals. Note that the impulse responses for the PS method and DOA methods are not ex-

Fig. 13. NRR for different lengths of speech utterances when the NRR of the partially separated signals used for solving permutation problem are of different levels.

450 actly the same because of the difference in microphone spacing and hence, for
 451 C1, the NRRs are different.

452 5.4 Robustness test for short speech utterances

453 The accuracy of the correlation method for solving the permutation problem
 454 depends on the length of the speech utterances [22,14] because as the speech
 455 length becomes shorter they tends to resemble each other. Since the proposed
 456 method is based on the correlation between the partially separated and fully
 457 separated signals in the frequency bins, the robustness of the algorithm for
 458 shorter speech segments is tested in this experiment. The algorithm is tested
 459 for 1, 2, 3, 4, 5 and 10 seconds speech utterances using different levels of partial
 460 separation (2, 4, 6 and 8 dB). Fig.13 shows the test results and it can be seen
 461 that the algorithm works even for smaller speech lengths and 3 sec speech is
 462 sufficient to get a good separation. For these experiments the length of the
 463 data in the frequency bins is fixed at 500 for all speech lengths, by adjusting

Fig. 14. Performance variation for various filter lengths of the time domain stage.

Fig. 15. Performance of the frequency domain stage followed by time domain stage configuration for different lengths of filter taps as well as for different lengths of the data for learning.

the overlapping of the sliding window, while converting into frequency domain.
The unmixing filter length of the time domain stage was 512 in all the cases
and the number of iterations are adjusted to get different levels of partial
separation.

Fig. 16. Effect of permutation in the frequency bins, for time domain separation. NRR for the mixture due to the permutation of clean signals is indicated by “clean permuted” and that of the mixture is shown by “mixture permuted”. For example if multiple = 8, the permuted bins are 8, 16, 24,...,4096; similarly for other multiples.

5.5 Effect of combination order in cascaded configuration

In [16,17] the time domain and frequency domain stages were cascaded for improving the NRR. Even though the objectives of [16,17] were not to solve the permutation problem, to improve the NRR, we study the performance difference between the orders of the time domain and frequency domain stages when they are swapped. To find out the optimum filter length for the time domain stage in the proposed cascaded configuration, we have conducted experiments for different values of filter lengths. The result are shown in Fig.14. To model the complete reflections in the room, the length of the unmixing filter must be atleast equal to the length of the mixing filter [9]. From Fig.14 it can be seen that, smaller the filter length, poorer the performance. This is because of the reason that the length of the unmixing filter is less than the required. However as the unmixing filter length increases due to the interdependency of the filter coefficients [25], the convergence will become poorer, which is ev-

482 ident from the Fig.14. As the filter length increases from 64 to 512, the NRR
 483 is increasing and then decreasing. Here, out of the different filter lengths used,
 484 the best performance is achieved at 512. Hence for the remaining parts of the
 485 experiments 512 taps are used for the time domain stage, unless otherwise
 486 it is specified. The NRR for the cascaded configuration of frequency domain
 487 stage followed by the time domain stage with different values of filter length
 488 for the time domain stage is shown in Fig.15, the 100% data learning case (the
 489 remaining cases will be explained later). For the frequency domain stage the
 490 permutation problem is solved using the proposed parallel configuration with
 491 only a difference that, instead of using the partially separated signals, the clean
 492 signals are used to solve the permutation problem in the best possible way,
 493 which is the ideal case. The results showed that for the cascaded configuration
 494 with frequency domain stage followed by the time domain stage, the NRR is
 495 poorer than that of the frequency domain stage alone. This is due to: *i*) For
 496 further separation, after the frequency domain separation, fine tuning of the
 497 unmixing filter coefficients with longer filter taps is needed. But as discussed
 498 above, when the filter taps increases, because of the interdependency of the
 499 filter coefficients, the convergence is poor. *ii*) In the frequency domain stage
 500 even if we perfectly separate the signals in all the bins, fails to completely
 501 solve the permutation problem, the resultant mixing will be too complex for
 502 the time domain algorithm to separate and is even worse, if the signals in the
 503 frequency bins remain mixed in addition to the permutation. This is clear from
 504 the simulation results shown in Fig.16, where in one case the clean signals are
 505 first converted into frequency domain and every B^{th} bin is permuted, where
 506 $B = 2^b$, $b = 2, 3, \dots, 11$, so that the signals in all the bins are fully separated
 507 and then after converting into time domain they are mixed signals, because of
 508 the permutation. In the second case, instead of using clean signals, the mixed

509 (convolutive mixing using real room impulse response) signals are used so that
 510 after converting into the time domain, the resultant signals are more complex
 511 mixtures than that of the simple convolutive mixtures. When applied these
 512 two sets of signals to the time domain algorithm, it is found that even for
 513 the case where the mixing was only because of the permutation, the separation
 514 is very difficult and when the mixing is because of the convolutive as
 515 well as that of the permutation, the resultant is worse. In Fig.16, for smaller
 516 values of B , the NRR for the “mixture permuted” (permuted using mixed
 517 signals) signals is higher than that of the “clean permuted” (permuted using
 518 clean signals) signals. This is because, at the k^{th} frequency bin, $X_1(f_k, t)$ and
 519 $X_2(f_k, t)$ are more dependent than $S_1(f_k, t)$ and $S_2(f_k, t)$ [4], where $X_n(f_k, t)$
 520 and $S_n(f_k, t)$ are the data in the k^{th} frequency bin for n^{th} mixed and clean
 521 signals respectively. Hence the mixing effect caused due to the permutation of
 522 $S_1(f_k, t)$ and $S_2(f_k, t)$ is more than that of permuting $X_1(f_k, t)$ and $X_2(f_k, t)$.
 523 In the cascaded configuration of frequency domain stage followed by the time
 524 domain stage, even if the permutation is fully solved and the signals are not
 525 completely separated in the frequency bins, the time domain stage requires
 526 fine tuning with long filter taps, which is a difficult task because of the reasons
 527 mentioned in Section 5.1. Moreover, if the separation in each bin is not perfect,
 528 the chance for getting wrong permutations is high and this will further
 529 worsen the result.

530 The length of the data in the frequency bin will affect the separation of the frequency
 531 domain stage, because, as the data length decreases the independence
 532 assumption will collapse [4] and hence the separation will be poor. Fig.15 also
 533 shows the variation in NRR for different data lengths. The maximum length
 534 of the data in each bin is 117 (15 sec speech, 16 kHz sampling frequency, 4098

point FFT and 50 % overlap). Since the mixing filter is constant over the full
 length of the speech signals, only a fraction of the data (20%, 40%, 60% and
 80%) in the frequency bins is used for learning, ie., to estimate the unmixing
 filter. Then the full length data is separated using the estimated filter. As ex-
 pected [4], smaller the data length for learning, poorer the separation, which
 is shown in Fig.15. It can also be seen that the time domain stage could not
 improve the separation further, for any of these cases. Hence we conclude that
 it is better to place the time domain stage in front of the frequency domain
 stage because *i)* the time domain block used to generate the partially sepa-
 rated signals for solving permutation problem can be fully utilized, otherwise
 we need a separate time domain block, only for solving permutation *ii)* The
 overall performance is better than that of the individual blocks *iii)* Efficient
 and approximate algorithms like [1] can be used with smaller taps, as we need
 only partial separation.

In [16,17], the performance of the time domain stage followed by the frequency
 domain stage configuration was poor compared to its swapped configuration
 because to initialize and to solve the permutation problem, direction of arrival
 and beam forming techniques were used. Hence when the time domain block
 was in the first stage the directional information was not available for the
 frequency domain stage.

6 Conclusions

We have proposed a method for solving the permutation problem in frequency
 domain BSS of speech signals. The method uses correlation between the signals
 in each frequency bin, one of which is partially separated by a time domain

559 BSS method and the other is obtained by a frequency domain BSS method.
560 The algorithm does not require the knowledge of the directions of arrival of
561 the sources. Hence, if the time domain BSS algorithm can partially separate
562 the signals so as to make the spectrum of the resultant signals closer to their
563 corresponding sources, the proposed method can be used to separate collinear
564 sources also. Unlike the correlation method which utilizes the inter-frequency
565 correlation of the signals, the reliability of the proposed method is high as it
566 utilizes the correlation with partially separated signals. The additional com-
567 putational cost of the proposed method in the cascaded configuration is due
568 to the time domain stage. Since it is cascaded with the frequency domain
569 stage, the performance improvement is due to the correctly fixed permutation
570 as well as due to the achievable cascaded configuration. Hence, the additional
571 computational cost is optimally utilized.

572 **References**

- 573 [1] R. Aichner, H. Buchner and W. Kellermann, Real-time convolutive blind
574 source separation based on a broadband approach, Fifth Int. Symposium on
575 Independent Component Analysis and Blind Signal Separation (2004) 840-847.
- 576 [2] R. Aichner, S. Araki, S. Makino, T. Nishikawa and H. Saruwatari, Time
577 domain blind source separation of non-stationary convolved signals by utilizing
578 geometric beamforming, Workshop on Neural Networks for Signal Processing
579 (2002) 445- 454.
- 580 [3] S. Amari, Natural gradient works efficiently in learning, Neural Comput. 10(2)
581 (1998) 251-276.
- 582 [4] S. Araki, R. Mukai, S. Makino, T. Nishikawa and H. Saruwatari, The

- 583 fundamental limitation of frequency domain blind source separation for
 584 convolutive mixtures of speech, *IEEE Trans. Speech Audio Process.*, 11(2)
 585 (2003) 109-116.
- 586 [5] F. Asano, S. Ikeda, M. Ogawa, H. Asoh and N. Kitawaki, A combined approach
 587 of array processing and independent component analysis for blind separation of
 588 acoustic signals, In *Proc. ICASSP (2001)* 2729-2732.
- 589 [6] A. J. Bell and T. J. Sejnowski, An information-maximization approach to blind
 590 separation and blind deconvolution, *Neural Comput.* 7(6) (1995) 1129-1159.
- 591 [7] E. Bingham and A. Hyvärinen, A fast fixed-point algorithm for independent
 592 component analysis of complex-valued signals, *Int. J. of Neural Systems* 10(1)
 593 (2000) 1-8.
- 594 [8] H. Buchner, R. Aichner and W. Kellermann, A generalization of a class of blind
 595 source separation algorithms for convolutive mixtures, In *Proc. ICA (2003)* 945-
 596 950.
- 597 [9] H. Buchner, R. Aichner and W. Kellermann, A generalization of blind source
 598 separation algorithms for convolutive mixtures based on second-order statistics,
 599 *IEEE Transactions on speech and audio processing* 13(1) (2005) 120-134.
- 600 [10] J. R. Hopgood, P. J. W. Rayner and P. W. T. Yuen, The effect of sensor
 601 placement in blind source separation, *IEEE Workshop on Applications of Signal*
 602 *Processing to Audio and Acoustics (2001)* 95-98.
- 603 [11] M. Z. Ikram and D. R. Morgan, A beamforming approach to permutation
 604 alignment for multichannel frequency domain blind speech separation, In *Proc.*
 605 *ICASSP (2002)* 881-884.
- 606 [12] M. Z. Ikram and D. R. Morgan, Permutation inconsistency in blind speech
 607 separation: Investigation and solutions, *IEEE Transactions on speech and audio*
 608 *processing* 13(1) (2005) 1-13.

- 609 [13] M. Joho and P. Schniter, Frequency domain realization of a multichannel blind
610 deconvolution algorithm based on the natural gradient, In Proc. ICA (2003)
611 543-548.
- 612 [14] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda and F. Itakura, Evaluation
613 of blind signal separation method using directivity pattern under reverberant
614 conditions, In Proc. ICASSP (2000) 3140-3143.
- 615 [15] N. Murata, S. Ikeda and A. Ziehe, An approach to blind source separation based
616 on temporal structure of speech signals, *Neurocomputing*, 41(1- 4), (2001) 1-24.
- 617 [16] T. Nishikawa, Blind source separation based on multistage independent
618 component analysis, PhD thesis, Nara Institute of Science and Technology, 2005.
- 619 [17] T. Nishikawa, H. Saruwatari and K. Shikano, Blind source separation of acoustic
620 signal based on multistage ica combining frequency-domain ica and time-domain
621 ica, *IEICE Trans. Fundamentals*, E86-A(4) (2003) 846-858.
- 622 [18] L. Parra and C. Spence, Convolutional blind separation of non-stationary sources,
623 *IEEE Transactions on speech and audio processing* 8(3) (2000) 320-327.
- 624 [19] K. Rahbar and J. P. Reilly, A frequency domain method for blind source
625 separation of convolutional audio mixtures, *IEEE Transactions on Speech and*
626 *Audio Processing*, 13(5) (2005) 832-844.
- 627 [20] V. G. Reju, S. N. Koh, I. Y. Soon and X. Zhang, Solving permutation problem
628 in blind source separation of speech signals: A method applicable for collinear
629 sources, In Proc. ICICS (2005) 1461-1465.
- 630 [21] V. G. Reju, S. N. Koh and I. Y. Soon, A Robust Correlation Method for Solving
631 Permutation Problem in Frequency Domain Blind Source Separation of Speech
632 Signals, In Proc. APCCAS (2006) 1893-1896.
- 633 [22] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano,
634 Blind Source Separation Combining Independent Component Analysis and

- 635 Beamforming, EURASIP Journal on Applied Signal Processing, No.11, (2003)
636 1135-1146.
- 637 [23] H. Sawada, R. Mukai, S. Araki and S. Makino, A robust and precise method for
638 solving the permutation problem of frequency domain blind source separation,
639 IEEE Transactions on Speech and Audio Processing 12(5) (2004) 530-538.
- 640 [24] P. Smaragdis, Blind separation of convolved mixtures in the frequency domain,
641 Neurocomputing 22 (1998) 21-34.
- 642 [25] P. Smaragdis, Efficient blind separation of convolved sound mixtures, IEEE
643 Workshop on Applications of Signal Processing to Audio and Acoustics 1997.
- 644 [26] X. Sun and S. C. Douglas, A natural gradient convolutive blind source
645 separation algorithm for speech mixtures, In Proc. ICA (2001) 59-64.
- 646 [27] L. Tong, R. W. Liu, V. C. Soon and Y. F. Huang, Indeterminacy and
647 identifiability of blind identification, IEEE Trans. Circuits Syst., 38(5) (1991)
648 499-509.
- 649 [28] <http://bass-db.gforge.inria.fr/BASS-dB/?show=browse&id=filters>
- 650 [29] <http://www.purebits.com/>