A Novel Hybrid Approach to the Permutation Problem of Frequency Domain Blind Source Separation

Wenwu Wang[1], Jonathon A. Chambers[1], and Saeid Sanei[2]

[1] Communications and Information Technologies Research Group
Cardiff School of Engineering, Cardiff University, Cardiff, CF24 0YF, UK
wenwu.wang@ieee.org, chambersj@cf.ac.uk
[2] Centre for Digital Signal Processing Research
King's College London, Strand London, WC2R 2LS, UK
saeid.sanei@kcl.ac.uk

Abstract. We explore the permutation problem of frequency domain blind source separation (BSS). Based on performance analysis of three approaches: exploiting spectral continuity, exploiting time envelope structure and beamforming alignment; we present a new hybrid method which incorporates a psychoacoustic filtering process for the misaligned permutations unable to be delt with by these approaches. We use a subspace based method (MUSIC) rather than conventional beamforming for the accurate estimation of the direction of arrivals (DOAs) of the source components, and a frequency dependent distance for the correlation of time envelopes. The proposed methods are compared with other approaches by signal to interference ratio (SIR) evaluation, and the new hybrid approach is shown to have the best performance.

1 Introduction

Convolutive BSS has recently received extensive interest within the signal processing community due to its potential applications in communications, speech processing, and medical imaging. An effective method of addressing this problem is to transform it into the frequency domain so that a series of complex-valued instantaneous BSS problems is solved separately using a conventional instantaneous mixing independent component analysis (ICA) framework. A crucial limitation associated with such a transformation is the permutation indeterminacy which is induced inherently by the general ICA approach. That is, the reconstructed source signals in the time domain will remain distorted if the permutations of the recovered frequency domain source components are not consistent with each other.

To address this problem, several approaches have been developed, which can be approximately classified as: (1) exploiting the continuity of the spectra of the recovered signals or the separation matrix [1] [2]; (2) Exploiting the time structure of the source components [3]; (3) applying beamforming techniques to

C.G. Puntonet and A. Prieto (Eds.): ICA 2004, LNCS 3195, pp. 532–539, 2004.

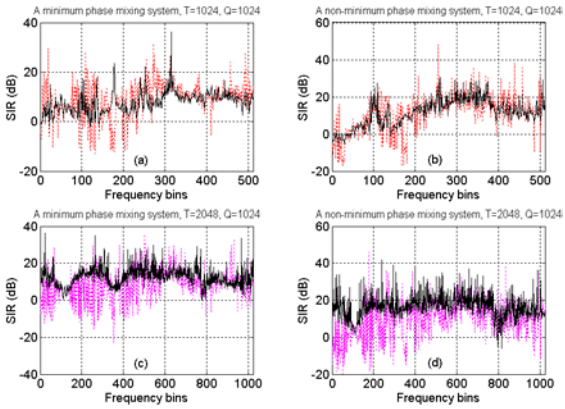

Fig. 1. SIR improvement across frequency axis before (dotted line) and after permutation alignment (solid line) using two methods: the separation matrices coupling over neighboring frequency bins (a) (b) and filter length constraint (c) (d).

the permutation alignment [4] [5]. These approaches may work well for carefully defined situations, but not necessarily for others. A recent work in [6] suggests that it is possible to combine the different properties of these approaches for developing a more robust and precise solution. In this paper, building upon this idea, we aim at developing a new hybrid approach, which is expected to benefit from some established results but have better performance. Additionally, we introduce some results of psychoacoustic research for reducing the permutation effect.

The remainder of the paper is organized as follows. Frequency domain BSS (FDBSS) together with its associated permutation problem is briefly described in Section 2. The various solutions are investigated in Section 3, which includes the introduction of the psychoacoutic filtering technique for the permutation problem. Section 4 summarizes the new hybrid approach and evaluates its performance. Finally, Section 5 concludes the paper.

2 Frequency Domain BSS and Permutation Problem

Assume that N source signals are recorded by M microphones (here we are particularly interested in acoustic applications), where $M \geq N$. The output of the j-th microphone is modeled as a weighted sum of convolutions of the source signals corrupted by additive noise, that is, $x_j(n) = \sum_{i=1}^{N} \sum_{p=0}^{P-1} h_{jip} s_i(n-p) + v_j(n)$, where h_{jip} is the p-th element of the P-point impulse response from source i to microphone j ($j = 1, \cdots, M$), s_i is the signal from source i, x_j is the signal received by microphone j, v_j is the additive noise, and n is the discrete time index. All signals are assumed zero mean. Using a discrete Fourier transformation (DFT), a frequency domain implementation of the mixing system is denoted as $\mathbf{X}(\omega, t) = \mathbf{H}(\omega)\mathbf{S}(\omega, t) + \mathbf{V}(\omega, t)$, where $\mathbf{S}(\omega, t)$ and $\mathbf{X}(\omega, t)$ are the time-frequency representations of the source vector and the mixture vector

Table 1. Overall SIR improvement before and after (B/A) applying the methods of filter constraint (FC) and separation matrices coupling (MC) respectively.

Systems/Methods	1/MC	1/FC	2/MC	2/FC	3/MC	3/FC
SIR in dB (B/A)	3.99/5.30	1.85/9.50	1.74/0.87	0.82/8.76	-1.31/-0.57	-0.10/10.50

respectively. Using the conventional ICA framework, $\mathbf{X}(\omega, t)$ can be separated at each frequency bin as $\mathbf{Y}(\omega, k) = \mathbf{W}(\omega)\mathbf{X}(\omega, k)$, where $\mathbf{Y}(\omega, k)$ is the time-frequency representation of the estimated source vector (assumed to be mutually independent), and k is the discrete time block index. Due to the inherent permutation ambiguity at each frequency bin, the recovered source components may have different permutations along the frequency axis so that the reconstructed source signals are still distorted in the time domain if the permutations are not correctly aligned. In the following discussion, we will use the penalty function based FDBSS algorithm developed in [9] for the separation of mixtures $\mathbf{X}(\omega, t)$, which exploits second order statistics (SOS) of nonstationary signals. We choose the penalty function to be in the form of a non-unitary constraint. The cost function is minimized by the gradient adaptation. Due to the limited space in this paper, we omit the implementation details which can be seen in [9].

3 Solutions to Permutation Problem

In this section, we will investigate some approaches briefly described in Section 1 and show some new results. We will use the SIR [2] as the performance index for the following evaluation, i.e.

$$SIR = 10\log\{(\sum\nolimits_{\omega}\sum\nolimits_{i}|H_{ii}(\omega)|^2\left\langle|s_i(\omega)|^2\right\rangle)/(\sum\nolimits_{\omega}\sum\nolimits_{i\neq j}|H_{ij}(\omega)|^2\left\langle|s_j(\omega)|^2\right\rangle)\}.$$

3.1 Exploiting Spectral Continuity

For this approach, either the recovered source components or separation matrices are assumed to have spectral similarities between neighboring frequency bins [1] [2]. In [1], an adaptive scheme was presented to apply frequency coupling for the unmixing matrices between neighboring frequency bins, that is $\Delta W_f \leftarrow \Delta W_f + k\Delta W_{f-1}$, where $0 < k < 1$. This intuitive scheme implicitly assumes that the permutations have been slightly changed during mixing, however it has limited performance for many cases, such as in Fig. 1 (a) and (b), where we can only identify a small SIR improvement along the frequency axis. In [2], a smoothness constraint was imposed on the unmixing filters in the time domain, that is, $Q < T$, and hence forced the solutions to be continuous in the frequency domain. As shown in Fig. 1 (c) and (d), compared with [1], this approach has a superior average performance along the frequency axis which is nevertheless, not consistent at every frequency, especially for some low frequencies. From Table 1, we find that the filter constraint approach is more robust with respect to the

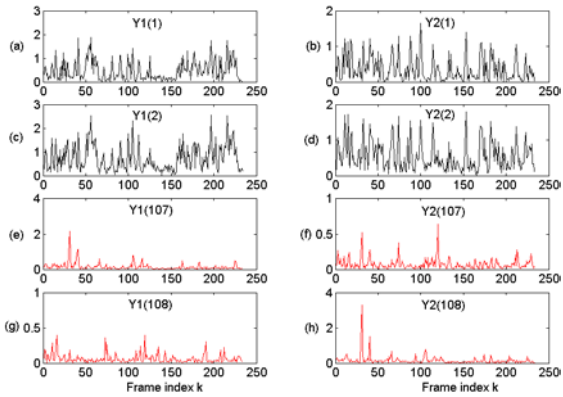

Fig. 2. The time envelopes of two separated source components at four different frequency bins; the upper four plots (a, b, c, d) represent two adjacent lower frequency bins, the lower four plots (e, f, g, h) represent two adjacent higher frequency bins.

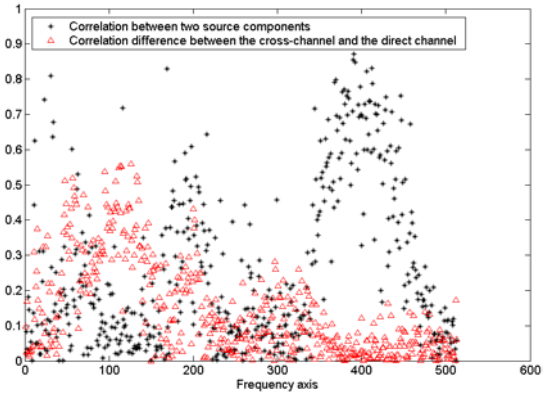

Fig. 3. Correlation value distribution along frequency axis.

mixing systems as compared with [1]. However, it is observed in [5] that the filter constraint may not be appropriate for a reverberant environment where a long filter may otherwise have a better performance. A merit of exploiting spectral continuity is that uniformity of the spectrum of the source signals has been preserved, which may not be shared by other approaches e.g. [3], where the frequencies have been processed separately. The identified drawbacks can be compensated by the approaches discussed in the following sections.

3.2 Exploiting Time Envelope Structure

This method was motivated by the time structure of speech signals [3] [6]. It is known that the source components at different frequency bins belonging to the same source signal should have similar shape in amplitude if they are modulated in a similar way. As a result, by measuring the correlation between the recovered

source components at each frequency bin, we can determine the right order of the components in order to group them to the corresponding source. Mathematically, we define the time envelope of each extracted source component as $\mathcal{Y}_i(\omega, k) = |Y_i(\omega, k)|$, $i = 1, ..., N$. Fig. 2 shows an example of the time envelopes of the source components separated by the algorithms described in Section 2. From Fig. 2, we see that: 1) the envelopes from the same source signal at adjacent frequency bins are more similar to each other, such as (a) and (c), (e) and (h); 2) there exists the permutation problem since (e) corresponds to (h) but not (g). Therefore, by testing the correlations between the envelopes, we can determine the permutation for each frequency bin. A crucial problem in implementing this approach is, however, the selection of the frequency distance $\triangle d_\omega$ for the envelope correlation. In [3], the sum of the aligned frequencies is taken as the reference for the decision of the unpermuted frequencies, which unfortunately suffers from the fact that the envelopes with longer frequency distance do not necessarily have similar shapes (see Fig. 2 (a) and (g)). As a result, the permutation of the higher frequencies would not be accurately aligned since the correlation difference is small in this case (see Fig. 3). An alternative method for reducing this effect is to consider the correlation between the envelopes at neighboring frequency bins [6], however, it is sensitive to any misaligned frequency bins. To overcome this shortcoming, we propose to use the sum of the correlations as an approximate reference and conduct the correlations between neighboring frequency bins. Fig. 3 indicates that a fixed frequency distance is not appropriate for the envelope correlation. Therefore, we start the process from the frequency with the smallest correlation between the source components and adjust the distances to the correlation value at the current frequency between the source components.

3.3 Beamforming Alignment

Beamforming techniques have shown to be another promising approach for solving the permutation problem [4] [5], which is essentially motivated by the similarities between convolutive BSS and array signal processing. Comparatively, the model of convolutive BSS can be denoted by a phase and amplitude response, i.e., $\mathbf{y}(k) = e^{j\omega k}\mathbf{r}(\omega, \theta)$, where $\mathbf{r}(\omega, \boldsymbol{\theta}) = \mathbf{W}^H(\omega)\mathbf{D}(\omega, \boldsymbol{\theta})$, $\mathbf{D}(\omega, \boldsymbol{\theta}) = [\mathbf{d}(\omega, \theta_1), \cdots, \mathbf{d}(\omega, \theta_M)]$, $\mathbf{d}(\omega, \theta_j) = [e^{j\omega\tau_i(\theta_j)}]^H$ are steering vectors, and τ_i, $i = 1, \cdots, N$ denote propagation delays. The separation matrices for each frequency bin ω are analogously regarded as beamformers. Therefore, the DOAs of source components can be observed from every row of $\mathbf{W}(\omega)$ by plotting the directivity pattern, i.e. $F_i(\omega, \theta) = \sum_{k=1}^{M} W_{ik}(\omega)e^{j\omega(k-1)\tau_{ki}}$, where $\tau_{ki} = d_k \sin\theta_i / c$ is the time delay with respect to the ith source signal from the direction of θ_i, observed at the kth microphone with distance d_k, and c is the velocity of the sound. By estimating the DOAs at each frequency bin, the permutations can be determined in a straightforward way, sweeping or keeping the rows in $\mathbf{W}(\omega)$.

It has been suggested in [5] to use a low frequency range [1 $c/2d$) for the estimation of the DOAs of the sources (null directions) since their accurate estimates can not be guaranteed due to the existence of grating lobes at higher frequencies. However, it is also shown in [6] that for very low frequencies, null directions

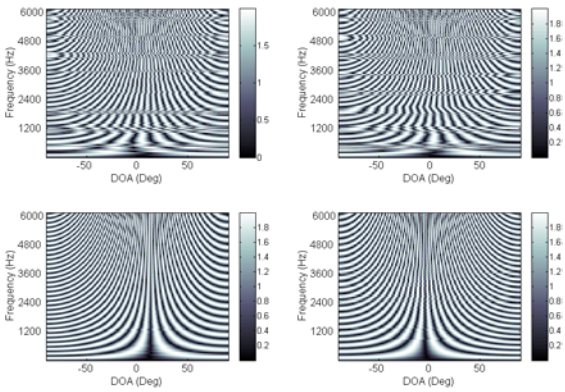

Fig. 4. The directivity pattern as a function of frequency before (upper two) and after (lower two) alignment by the MUSIC approach.

cannot be accurately estimated, due to the *flatness* of the directivity patterns. Another downside is that, unlike BSS which does not suffer from the prior information about the source location, it requires the two sources to be located up to a desired power resolution [7]. To give a more accurate estimate of the DOAs, we suggest to resort to subspace-based methods such as MUSIC [7]. To this end, we define the following MUSIC operator, $\bar{F}_i(\omega,\theta) = 1/\left\|\tilde{\mathbf{P}}(\omega,\theta)\hat{\mathbf{a}}(\omega,\theta_i)\right\|_2$, where $\tilde{\mathbf{P}}(\omega,\theta)$ is the noise subspace formed by the estimate $\hat{\mathbf{A}}(\omega) = \mathbf{W}^{-1}(\omega)$, and $\hat{\mathbf{a}}(\omega,\theta_i)$ is the ith column of $\hat{\mathbf{A}}(\omega)$. Fig. 4 shows an example of the beam pattern of $\mathbf{W}(\omega)$ using $\bar{F}_i(\omega,\theta)$.

3.4 Psychoacoustic Post-filtering

To compensate for misaligned bins, a potential method is to exploit human perception for acoustic signals. Psychoacoutic studies reveal that, although human hearing ranges from about 20Hz to 20KHz, most of the energy of speech lies in the lower frequency band (with bandwidth normally less than 5KHz) [8]. The just-audible thresholds and critical bandwidths are not constant but non-uniform, non-linear across all frequencies and dependent on different sounds. This means that the average human does not have the same perception at all frequencies. This fact suggests that some frequencies can be cut due to the limitation of the human auditory system and the masking effect, however without loss of necessary information contained in speech. Based on this point, we propose to use a psychoacoustic model as a post-filter after the permutations initially aligned by the aforementioned approaches. This model exploits two properties of the human auditory system: absolute threshold of hearing (ATH) (also known as threshold of quiet) and auditory masking (AM). The tone masker and noise masker are calculated respectively and the maskers that are weaker than another masker within one critical bandwidth are attenuated, and the ATH is used as a reference for determining the global threshold. An experiment result by apply-

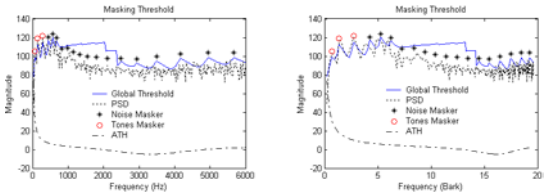

Fig. 5. Psychoacoutic post-filtering of one reconstructed speech signal from FDBSS output using threshold masking.

Table 2. SIR improvement of the various approaches

Methods	No alignment	[2]	[5]	[3]	[6]	**proposed DOA**	**proposed hybrid**
SIR_{av} (dB)	−0.33	9.59	10.04	6.23	11.35	10.89	14.12

ing this model to separate speech components is shown in Fig. 5, which clearly shows that there exists enough redundant information (including noise, see the masker above the global threshold) in the recovered source components that can be removed.

4 Approach Summary and Numerical Experiment

Based on the discussions of the above sections, our proposed hybrid approach for solving the permutations of $\mathbf{W}(\omega)$ is summarized as: 1) *performing filter constraint*; 2) *performing DOA alignment and detecting confidence*; 3) *retaining the frequency bins with high confidence, performing envelope correlation for the remaining frequencies, detecting confidence again;* 4) *performing psychoacoustic filtering for all the remaining frequency bins.* It should be noted that the procedure of confidence detection is to ensure a sufficiently high confidence for the permuted frequencies, which can be conducted in the same way as in [6].

We perform an experiment to evaluate the overall averaged performance of the proposed approach for three mixing systems which are identical to those used in Table 1. The result is compared with the method in [2] (using spectral continuity), [5] (using conventional beamforming), [3] (using time envelope), and [6] (using a combined approach). We artificially mix two speech signals (sampled at 12kHz with length of 9 seconds). $Q = 1024$, $T = 1024$ (for [2], $T = 2048$). For [6], $\triangle d_\omega = 3\Delta\omega$, where $\Delta\omega$ is the frequency resolution. The penalty function parameter is $\kappa = 0.1$ and the number of intervals used to estimate each cross-power-matrix is 7 (see [9]). The distance between two sensors is $1m$, the directions of the sources are respectively $19.68°$ and $-5.35°$. For the proposed method (step 3), $\triangle d_\omega$ decreases with a linear regulation from $10\Delta\omega$ to $\Delta\omega$ as frequency increases. From Table 2, we know that: 1) MUSIC has a superior performance over conventional beamforming (such as [5]) for the permutation alignment; 2) calculating the correlation over the whole frequency does not give an accurate alignment (see [5]) as compared with neighboring frequency coupling

in [6]; 3) The proposed hybrid approach has a significantly improved performance due to the introduction of the psychoacoustic perception together with a more accurate DOA estimation and a dynamic frequency distance for envelope correlation.

5 Conclusion

A hybrid approach for solving the permutation problem of FDBSS has been presented. A psychoacoustic filtering technique has been effectively introduced to incorporate the human perception of sound in order to reduce the permutation effect at some frequency bins which are not accurately aligned. The subspace based MUSIC method has also been introduced to provide more accurate beam patterns along frequency bins. By varying the frequency intervals for envelope correlation, the nonstationarity of speech signals is nicely exploited. More extensive evaluations for the proposed approach including subjective tests using the mean opinion score (MOS) are currently under consideration.

References

1. P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol.22, pp. 21–34, 1998.
2. L. Parra and C. Spence, "Convolutive blind source separation of nonstationary sources," *IEEE Trans. on Speech Audio Proces.*, pp. 320–327, May 2000.
3. N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1-24, Oct. 2001.
4. S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," *Proc. ICASSP*, pp.3140-3143, 2000.
5. M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," *Proc. ICASSP*, pp. 881-884, May 2002.
6. H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *Proc. ICA*, Nara, Japan, Apr. 1-4, 2003.
7. H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE SP Mag.*, pp. 67-94, Jul. 1996.
8. E. Zwicker and H. Fastl, "Psychoacoustics: facts and models," Springer, 2nd Ed.,1999.
9. W. Wang, J. A. Chambers, and S. Sanei, "Penalty function approach for constrained convolutive blind source separation," *Proc. ICA*, Granada, Spain, Sept. 22-24, 2004 (accepted).