

SOLVING THE PERMUTATION PROBLEM USING PHASE LINEARITY AND FREQUENCY CORRELATION

Keisuke Toyama^{1,2}, Andrew Nesbit¹, Maria G. Jafari¹, and Mark D. Plumbley¹

¹ Centre for Digital Music, Queen Mary University of London
Mile End Road, London, E1 4NS, U.K.

² Technology Development Group, Sony Corporation
1-7-1 Konan, Minato-ku, Tokyo, 108-0075, Japan
keisuke.toyama@jp.sony.com

ABSTRACT

This paper describes a method for solving the permutation problem in blind source separation (BSS) by frequency-domain independent component analysis (FD-ICA). FD-ICA is a well-known method for BSS of convolutive mixtures. However, FD-ICA has a source permutation problem, where estimated source components can become swapped at different frequencies. Many researchers have suggested methods to solve the source permutation problem including using correlation between adjacent frequencies. In this paper, we discuss a new method for solving the permutation problem, based on the linearity of the phase response of the FD-ICA de-mixing matrix, and a combination method of the proposed phase linearity method and the inter-frequency correlation method. Initial results indicate that our methods can provide an almost perfect solution to the permutation problem in an anechoic environment, and better performance than the inter-frequency correlation method alone in an echoic environment.

Keywords: Blind source separation (BSS), independent component analysis (ICA), permutation problem, spatial aliasing, linearity, phase response.

1 INTRODUCTION

Blind Source Separation (BSS) is defined as the problem of recovering each of a set of source signals from a given set of mixture signals. One of the main methods for BSS is independent component analysis (ICA) [1] which can separate the sources without any prior information if they are independent of each other. For separation of convolutive mixtures, such as anechoic or echoic audio mixtures, the frequency-domain ICA (FD-ICA) method [3] is often used. However, FD-ICA has a source permutation problem [3]–[6] which is an ambiguity in the ordering of the separated sources in each frequency bin. The two main approaches to solve the permutation problem are to use

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

inter-frequency correlation [3, 5, 6], or the direction of arrival (DOA) [4, 5]. The former approach can solve the permutation problem if sources have high correlation between energies in adjacent frequencies. However, there are no guarantees that such a condition is always satisfied. On the other hand, the DOA can suffer from a spatial aliasing problem above a certain frequency limit. In this paper, we discuss these issues and propose a new method based on phase linearity which might solve these problems.

2 BSS FOR CONVOLUTIVE MIXTURES

In the time-frequency domain, the observed signals at microphones $X_l(f, t)$ are expressed as

$$X_l(f, t) = \sum_{k=1}^K H_{lk}(f) S_k(f, t), \quad l = 1, \dots, L \quad (1)$$

where f represents frequency, t is the frame index, $H_{lk}(f)$ is the frequency response from source k to microphone l , and $S_k(f, t)$ is a time-frequency-domain representation of a source signal. Equation (1) can also be expressed as $\mathbf{X}(f, t) = \mathbf{H}(f)\mathbf{S}(f, t)$ where $\mathbf{X}(f, t) = [X_1(f, t), \dots, X_L(f, t)]^T$ is the observed signal vector, $\mathbf{S}(f, t) = [S_1(f, t), \dots, S_K(f, t)]^T$ is the source signal vector, and

$$\mathbf{H}(f) = \begin{bmatrix} H_{11}(f) & \cdots & H_{1K}(f) \\ \vdots & \ddots & \vdots \\ H_{L1}(f) & \cdots & H_{LK}(f) \end{bmatrix} \quad (2)$$

is the complex-valued mixing matrix.

In frequency-domain ICA, we perform signal separation from $\mathbf{X}(f, t)$ using the complex-valued de-mixing matrix

$$\mathbf{W}(f) = \begin{bmatrix} W_{11}(f) & \cdots & W_{1L}(f) \\ \vdots & \ddots & \vdots \\ W_{K1}(f) & \cdots & W_{KL}(f) \end{bmatrix} \quad (3)$$

so that the reconstructed output signals $\mathbf{Y}(f, t) = [Y_1(f, t), \dots, Y_K(f, t)]^T = \mathbf{W}(f)\mathbf{X}(f, t)$ become mutually independent.

In this paper, we perform this ICA step using the information maximization approach combined with the natural gradient [2], whereby the de-mixing matrix \mathbf{W} is updated by the learning rule,

$$\mathbf{W}^{(n+1)} = \mu[\mathbf{I} - \langle \phi(\mathbf{Y})\mathbf{Y}^H \rangle_t] \mathbf{W}^{(n)} + \mathbf{W}^{(n)} \quad (4)$$

where μ is a step-size parameter, $\langle \cdot \rangle_t$ denotes the averaging operator over time, and $\phi(\cdot)$ is a nonlinear function for

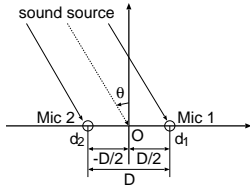


Figure 1: Direction of arrival.

a complex signal. We use $\phi(Y_k) = \tanh(|Y_k|) \exp(j\angle Y_k)$ as the nonlinear function. Hereafter, we suppose we have two sources ($K = 2$) and two microphones ($L = 2$) for simplicity.

3 PERMUTATION PROBLEM

FD-ICA has an ambiguity in the order of the rows of $\mathbf{W}(f)$, such that permuted matrix is also the solution for FD-ICA. This problem is called as the *permutation problem* [3]–[6].

One possible approach to solve the permutation problem is to use correlation between adjacent frequencies [3, 5, 6]. In this approach, we use the magnitude of the envelope of output signals $v_k^f(t) = |Y_k(f, t)|$ of the separated signal $Y_k(f, t)$. Here, we define the correlation of two magnitudes $\alpha(t)$ and $\beta(t)$ as

$$0 \leq \text{cor}(\alpha, \beta) = \frac{\text{cov}(\alpha, \beta)}{\sigma_\alpha \cdot \sigma_\beta} \leq 1 \quad (5)$$

where $\text{cov}(\cdot)$ is the covariance and σ is the standard deviation. If α and β are uncorrelated, $\text{cor}(\alpha, \beta) = 0$. We would expect magnitudes of adjacent frequency bins to be highly correlated within a given signal and less correlated with different signals. To use this idea, we calculate

$$D_{f\text{cor}}(f) = \sum_{g \in \mathcal{F}} (\text{cor}S_{f,g}(t) - \text{cor}C_{f,g}(t)) \quad (6)$$

where

$$\text{cor}S_{f,g}(t) = \text{cor}(v_1^f(t), v_1^g(t)) + \text{cor}(v_2^f(t), v_2^g(t))$$

$$\text{cor}C_{f,g}(t) = \text{cor}(v_1^f(t), v_2^g(t)) + \text{cor}(v_2^f(t), v_1^g(t)). \quad (7)$$

If $D_{f\text{cor}}(f) < 0$, we assume that the permutation has occurred at frequency f , whereas if $D_{f\text{cor}}(f) > 0$, the permutation has not occurred at frequency f . The simplest set of the frequencies g is $\mathcal{F} = \{g : f - \delta_{f_i} \leq g \leq f + 1\}$, where δ_{f_i} is a distance of the frequency for calculating the correlation. However, this strategy has a problem in which an error at one frequency will be propagated to the others. To avoid this problem, Murata [6] proposed that \mathcal{F} is a set of frequencies for which the permutation problem has been solved. However, Murata's method has a drawback that the signals at frequency f and g have few correlations in the case of a long distance between these two frequencies. To tackle this drawback, Sawada [5] proposed that \mathcal{F} is a set of frequencies which are harmonics of the frequency f . This idea is not suggested to use by itself but with the direction of arrival (DOA) method which we explain at the next paragraph.

Another approach for the permutation problem is to use the DOA [4, 5]. Here, we suppose a signal with frequency f comes from a source in the direction of θ as shown in the Figure 1. When the signal $\exp(j2\pi ft)$ is observed at point O , the observed signals at the microphones are $X_l(f, t) = \exp(j2\pi f[t - d_l \sin(\theta_k(f))/c])$, where d_l is the position of the microphone ($d_1 = -d_2 = D/2$)

and c is the speed of sound. The frequency response of the de-mixing process between the observed signals and the separated signals is expressed by the ratio of them, $Y_k(f, t)/\exp(j2\pi ft)$. Thus, we can obtain the gain of the frequency response with respect to the direction as

$$\begin{aligned} G_k(\theta_k(f)) &= |Y_k(f, t)/\exp(j2\pi ft)| \\ &= |W_{k1}(f) \exp(-j2\pi f(d_1 \sin(\theta_k(f))))/c \\ &\quad + W_{k2}(f) \exp(-j2\pi f(d_2 \sin(\theta_k(f))))/c|. \end{aligned} \quad (8)$$

If $f < c/2D$, the gain $G_k(\theta_k(f))$ has one peak and one null point at a maximum in a half period of $\theta_k(f)$ where $|\theta_k(f)| \leq \pi/2$ [4, 5]. The direction where the gain has the minimum value (null point) could be regarded as the direction of source signal. Therefore, we can solve the permutation to compare the direction of two sources, $\theta_1(f)$ and $\theta_2(f)$. For more details of this process see [4, 5].

However, if $f > c/2D$, the gain $G_k(\theta_k(f))$ has two or more local minimum points so that we cannot decide the magnitude relationship between $\theta_1(f)$ and $\theta_2(f)$ uniquely. This problem is called the *spatial aliasing problem*. For example, if the distance between two of microphones is 4 cm and the speed of sound is 343 m/sec, the spatial aliasing problem occurs for $f > 4287.5$ Hz.

4 PROPOSED METHOD

If the recording environment is anechoic, the coefficient of the mixing matrix $H_{lk}(f)$ is a delayed impulse. Thus, the phase response $\angle H_{lk}(f)$ is linear. Therefore, the phase response of the de-mixing matrix $\angle W_{kl}(f)$ should ideally be linear. Here, we consider the difference of the phase responses of the de-mixing matrix $\tilde{W}_k(f) = \angle W_{k1}(f) - \angle W_{k2}(f)$. The difference should be also linear phase, so we can represent the difference by the following equation:

$$\tilde{W}_k(f) = a_k f + b_k. \quad (9)$$

To solve the permutation problem, we utilise this linear phase property by following six steps.

[Step 1] Smooth $\angle W_{kl}(f)$ by a moving-average filter to reduce fluctuation as follows:

$$\angle \tilde{W}_{kl}(f) = \frac{1}{2M} \sum_{m=-M}^{M-1} \angle W_{kl}(f + m) \quad (10)$$

where M is the length of the moving-average filter, and could be decided by users. Hence, we obtain the difference of the phase responses as

$$\tilde{W}_k(f) = \angle \tilde{W}_{k1}(f) - \angle \tilde{W}_{k2}(f). \quad (11)$$

[Step 2] Estimate a_k and b_k by using the method of least squares, as

$$a_k = \frac{1}{C_f} \left[(f_h - f_l + 1) \sum_{f=f_l}^{f_h} f \tilde{W}_k(f) - \sum_{f=f_l}^{f_h} f \sum_{f=f_l}^{f_h} \tilde{W}_k(f) \right] \quad (12)$$

$$b_k = \frac{1}{C_f} \left[\sum_{f=f_l}^{f_h} f^2 \sum_{f=f_l}^{f_h} \tilde{W}_k(f) - \sum_{f=f_l}^{f_h} f \tilde{W}_k(f) \sum_{f=f_l}^{f_h} f \right] \quad (13)$$

where

$$C_f = (f_h - f_l + 1) \sum_{f=f_l}^{f_h} f^2 - \left(\sum_{f=f_l}^{f_h} f \right)^2 \quad (14)$$

and f_l and f_h are chosen from a low frequency range where the two curves of the equation (11) do not cross. The frequencies f_l and f_h are the low and high limits of the frequency range used to estimate a_k and b_k . For ex-

Table 1: Comparison of average SIR, SAR, and SDR [7] obtained with the inter-frequency correlation method and the proposed methods. All values are expressed in decibels (dB).

Anechoic	Inter-frequency correlation method	Proposed method A	Proposed method B
SIR	29.11	29.13	29.11
SAR	12.61	12.61	12.61
SDR	12.45	12.45	12.45
Echoic	Inter-frequency correlation method	Proposed method A	Proposed method B
SIR	17.90	23.61	24.16
SAR	9.39	9.92	10.04
SDR	8.72	9.67	9.81

ample, f_l is chosen to avoid the effect of low frequencies such as bins 5–20, and f_h could be calculated as

$$f_h = \min_f \left(\left| \tilde{W}d_1(f) - \tilde{W}d_2(f) \right| < \frac{\pi}{2} \right). \quad (15)$$

In this range, $f_l \leq f \leq f_h$, the two curves of the equation (13) are not expected to cross.

[Step 3] Calculate the estimated linear curve $\hat{W}d_k(f)$.

[Step 4] Wrap the values of $Wd_k(f)$ and $\hat{W}d_k(f)$ into $-\pi$ to π to avoid the effect of circular jump:

$$Wd_k(f) \leftarrow \text{mod}(Wd_k(f), 2\pi) - \pi \quad (16)$$

$$\hat{W}d_k(f) \leftarrow \text{mod}(\hat{W}d_k(f), 2\pi) - \pi. \quad (17)$$

[Step 5] Calculate the distance between $Wd_k(f)$ and $\hat{W}d_k(f)$ of all combinations,

$$D_{prop}(f) = [|Wd_1(f) - \hat{W}d_1(f)| + |Wd_2(f) - \hat{W}d_2(f)|] \\ - [|Wd_1(f) - \hat{W}d_2(f)| + |Wd_2(f) - \hat{W}d_1(f)|]. \quad (18)$$

[Step 6] If $D_{prop}(f) < 0$, consider that a permutation has occurred at the frequency f , whereas if $D_{prop}(f) > 0$, a permutation has not occurred at the frequency f .

This method is similar to the DOA method from the aspect of using the difference of the phase response of the de-mixing matrix. The difference between the DOA method and this proposed method is that this method does not use the position of the microphones. Thus, this method does not suffer from the spatial aliasing problem to the same extent.

However, this method has a drawback around the points where two linear curves $\hat{W}d_k$ cross. Around those points, this method cannot distinguish two sources, because the two sources have the same phases at these points so that D_{prop} might be almost 0. To overcome this problem, we consider combining this method with another approach which is not based on the phase information of the de-mixing matrix. Here, we combine the proposed method and the inter-frequency correlation method using the following additional steps:

[Step 7] Search the points where the two linear curves cross. These cross points f_c could be calculated by the following equation:

$$f_c = \{f : (\hat{W}d_1(f) - \hat{W}d_2(f)) \\ \times (\hat{W}d_1(f-1) - \hat{W}d_2(f-1)) < 0\}. \quad (19)$$

[Step 8] Replace the permutation obtained by Step 6 in the range where $f_c - \delta_{f_c} \leq f \leq f_c + \delta_{f_c}$ with the result of the inter-frequency correlation method. δ_{f_c} is a constant value which user can define.

5 EXPERIMENTS

To confirm these approaches, we performed two experiments to separate two speech signals (5 sec of speech at 44.1 kHz) in a simulated anechoic environment ($T_{60} = 0$ msec) and echoic environment ($T_{60} = 420$ msec) using the RIR tool box¹. To obtain the simulated mixture observations, we set 30 as the number of reflections and -3 dB as the reflection coefficient in the echoic environment. The simulated dimension of the room is 500x800x300 cm, the location of sources are 100x400x100 and 300x400x100 cm, and the location of the microphones are 248x200x100 and 252x200x100 cm. Thus, the distance between the two microphones is 4 cm. For the FD-ICA part, we adopt 2048 as the length of FFT window, 0.01 as the step size μ , and 300 as the number of iterations. In these experiments, we compared the performance of our method to the inter-frequency correlation method. For the inter-frequency correlation method, we adopt the simplest set of frequencies where we set 6 as the parameter δ_{f_i} . For the proposed method A (Step 1–6), we use 5 as the length of the moving-average filter M , and 15 as the lowest frequency bin number f_l to estimate a_k and b_k . For the proposed method B (Step 1–8), we use 20 as the parameter δ_{f_c} . These user defined parameters, δ_{f_i} , M , f_l , and δ_{f_c} are decided by an exploratory experiment.

Here, we define “correct” permutation data to evaluate the performance of the inter-frequency correlation method and our proposed methods. The “correct” data are obtained by the correlation between the input signal observed at microphone ($U_{lk}(f, t) = H_{lk}(f)S_k(f, t)$) and the separated signal which is projected to the microphone by the inverse matrix of the de-mixing matrix ($Z_{lk}(f, t) = W_{lk}^{-1}(f)Y_k(f, t)$) at each frequency as following equations:

$$D_{crt}(f) = corS_c(f) - corC_c(f) \quad (20)$$

where

$$corS_c(f) = \sum_{l=1}^2 \sum_{k=1}^2 \{cor(|U_{lk}(f, t)|, |Z_{lk}(f, t)|) \\ + cor(\angle(U_{lk}(f, t)), \angle(Z_{lk}(f, t)))\} \quad (21)$$

$$corC_c(f) = \sum_{l=1}^2 \{cor(|U_{l1}(f, t)|, |Z_{l2}(f, t)|) \\ + cor(|U_{l2}(f, t)|, |Z_{l1}(f, t)|) \\ + cor(\angle(U_{l1}(f, t)), \angle(Z_{l2}(f, t))) \\ + cor(\angle(U_{l2}(f, t)), \angle(Z_{l1}(f, t)))\}. \quad (22)$$

If $D_{crt}(f) < 0$, we assume that the permutation has occurred at the frequency f , whereas if $D_{crt}(f) > 0$, the permutation has not occurred at the frequency f .

The results are shown in Figures 2–5 and Table 1. The proposed method A can solve the permutation problem almost perfectly in the anechoic environment. However, the proposed method A can have errors at frequencies near where two estimated linear phase lines cross (f_c : e.g. frequency bin 820). The result of the proposed method B is worse than that of the proposed method A in the anechoic environment, the method has mistakes at frequency bin number 411 and 820, because the inter-frequency correlation method also happens to have mistakes at those points.

In the echoic environment, the result of the proposed

¹<http://www.2pi.us/rir.html>

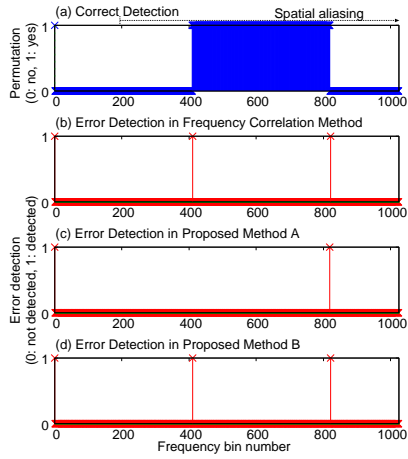


Figure 2: Detection of permutation in the anechoic environment; (a) correct detection, (b) detection errors in inter-frequency correlation method (error rate: 0.3% ($\approx 3/1025$)), (c) detection errors in proposed method A (0.2%), (d) detection errors in proposed method B (0.3%).

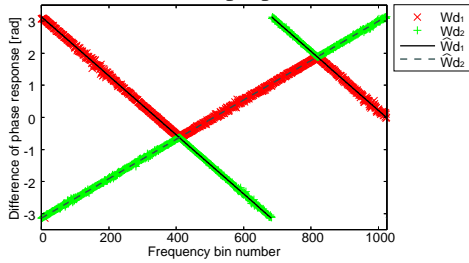


Figure 3: The de-mixing matrix phase difference in the anechoic environment, showing (i) observed points Wd_k and (ii) estimated lines $\hat{W}d_k$ from equation (9).

method A is not perfect but better than the inter-frequency correlation method. However, the proposed method A has permutation errors around the cross points (f_c : frequency bins 422 and 849) in the echoic environment as in the anechoic environment. The result of the proposed method B appears to be better than the proposed method A, in that some of the permutation errors are reduced, but, with only a marginal improvement in SDR, so there is still room for further improvement.

6 CONCLUSION

We have proposed a method which uses the linearity of the phase response of the de-mixing matrix to tackle the permutation problem in blind audio source separation. We have also proposed a method which combines the proposed method which used the linearity and the inter-frequency correlation method. These proposed methods can solve the problem well in an anechoic environment, and in our example, for a reverberant environment, the proposed methods give slightly better performance than that of the inter-frequency correlation method.

However, the proposed methods have a difficulty around the points where the two linear curves of the difference of the phase response cross, and the methods cannot solve the permutation at those points perfectly. We have used the inter-frequency correlation method to reduce this problem, nevertheless this method does not guarantee to solve the permutation problem around the cross points. In future work, we are considering combining with alterna-

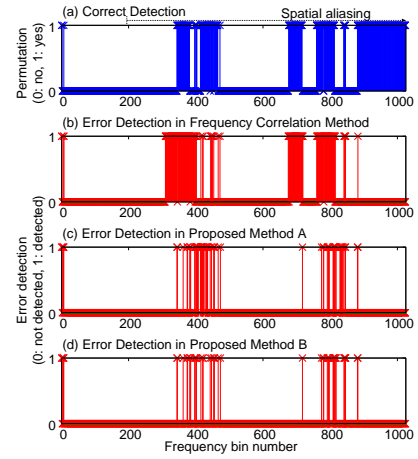


Figure 4: Detection of permutation in the echoic environment; (a) correct detection, (b) detection errors in inter-frequency correlation method (error rate: 20.2%), (c) detection errors in proposed method A (7.3%), (d) detection errors in proposed method B (4.9%).

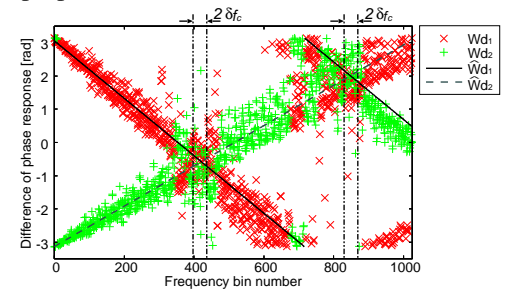


Figure 5: The de-mixing matrix phase difference in the echoic environment, showing (i) observed points Wd_k and (ii) estimated lines $\hat{W}d_k$ from equation (9).

tive method based on amplitude information of sources to solve the permutation problem around those points.

References

- [1] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, pp. 287–314, 1994.
- [2] T. Lee, "Independent Component Analysis," *Norwell, MA: Kluwer*, 1998.
- [3] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp.21–34, 1998.
- [4] S. Kurita, *et al.*, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," *Proc. ICASSP 2000*, vol. 5, pp. 3140–3143, 2000.
- [5] H. Sawada, *et al.*, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech and Audio Proc.*, vol. 12, pp. 530–538, 2004.
- [6] N. Murata, *et al.*, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp.1–24, 2001.
- [7] E. Vincent, *et al.*, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 14, pp.1462–1469, 2006.