USING PHASE LINEARITY TO TACKLE THE PERMUTATION PROBLEM IN AUDIO SOURCE SEPARATION

Keisuke Toyama[1,2]*, Andrew Nesbit*[1]*, Maria G. Jafari*[1]*,* and *Mark D. Plumbley*[1]

[1] Centre for Digital Music, Queen Mary University of London, Mile End Road, London, E1 4NS, U.K.
[2] 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. Initial results indicate that our method can provide an almost perfect solution to the permutation problem in an anechoic environment, and better performance than the method based on correlation between adjacent frequencies in an echoic environment.

Index Terms— 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. ICA for convolutive mixtures has two main approaches, time-domain ICA (TD-ICA) [2] and frequency-domain ICA (FD-ICA) [3]. The first approach, TD-ICA, separates the given set of mixture signals by convolution operations with FIR filters. However, in many cases, the de-mixing FIR filters for convolutive mixtures require a large number of coefficients [4]. Thus, it takes high computing costs to obtain the coefficients until the algorithm converges. The second approach, FD-ICA, is a frequency-domain BSS method, where complex-valued ICA for instantaneous mixtures is applied in each frequency bin. In the frequency domain, the convolution operation in the time domain is expressed as a multiplication, so that FD-ICA can separate mixture signals which are recorded in a real, convolutive environment. However, FD-ICA has a source permutation problem [3]–[7] 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 inter-frequency correlation [3, 4, 6, 7], or the direction of arrival (DOA) [5, 6]. The former approach can solve the permutation problem if sources have high correlation between energies in adjacent frequencies. However, there are no guarantees such condition is always satisfied. On the other hand, the latter approach works well in frequencies up to a certain limit. Above this limit the DOA method suffers from a spatial aliasing problem. In this paper, we discuss these problems and propose a new method which might solve these problems.

2. BSS FOR CONVOLUTIVE MIXTURES

Let $s_1, ..., s_K$ be source signals and $x_1, ..., x_L$ be sensor observations. These convolutive mixture observations are formulated as

$$x_l(t) = \sum_{k=1}^{K} \sum_{\tau} h_{lk}(\tau) s_k(t-\tau), \quad l = 1, ..., L \qquad (1)$$

where t represents time and $h_{lk}(\tau)$ represents the impulse response from source k to microphone l. De-mixing operations to obtain separated signals $y_k(t)$ are formulated as

$$y_k(t) = \sum_{l=1}^{L} \sum_{\tau} w_{kl}(\tau) x_l(t-\tau), \quad k = 1, ..., K \qquad (2)$$

where $w_{kl}(\tau)$ represents the de-mixing coefficients of FIR filters for deconvolution. In a real situation, the length of the impulse responses $h_{lk}(\tau)$ and $w_{kl}(\tau)$ may be thousands of taps, so it can be very difficult to solve the convolutive BSS problem. A reasonable approach to this issue is frequency-domain BSS, where a short-time Fourier transform (STFT) is applied to the microphone observations $x_l(t)$. In the frequency domain, the convolutive model in the equation (1) can be approximated as an instantaneous mixture model at each frequency

$$X_l(f,t) = \sum_{k=1}^{K} H_{lk}(f) S_k(f,t), \quad l = 1, ..., L \qquad (3)$$

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 $s_k(t)$. The equation (3) can also be expressed as $\mathbf{X}(f,t) = \mathbf{H}(f)\mathbf{S}(f,t)$ where $\mathbf{X}(f,t) = [X_1(f,t), ..., X_L(f,t)]^T$ is the observed signal vector, $\mathbf{S}(f,t) = [S_1(f,t), ..., S_K(f,t)]^T$ is the source signal vector, and

$$\mathbf{H}(f) = \begin{bmatrix} H_{11}(f) & \cdots & H_{1K}(f) \\ \vdots & \ddots & \vdots \\ H_{L1}(f) & \cdots & H_{LK}(f) \end{bmatrix} \quad (4)$$

is the mixing matrix which is complex-valued. Next, we perform signal separation using the complex-valued de-mixing matrix

$$\mathbf{W}(f) = \begin{bmatrix} W_{11}(f) & \cdots & W_{1L}(f) \\ \vdots & \ddots & \vdots \\ W_{K1}(f) & \cdots & W_{KL}(f) \end{bmatrix} \quad (5)$$

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

In this paper, we adopt the information maximization approach combined with the natural gradient as the ICA algorithm for instantaneous mixtures [2]. The de-mixing matrix $\mathbf{W}$ is updated by the learning rule,

$$\mathbf{W}^{(n+1)} = \mu[\mathbf{I} - \langle\phi(\mathbf{Y})\mathbf{Y}^H\rangle_t]\mathbf{W}^{(n)} + \mathbf{W}^{(n)} \quad (6)$$

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 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]–[7].

One possible approach to solve the permutation problem is to use correlation between adjacent frequencies [3, 4, 6, 7]. 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

$$0 \le cor(\alpha,\beta) = \frac{cov(\alpha,\beta)}{\sigma_\alpha \cdot \sigma_\beta} \le 1 \quad (7)$$

where $cov(\cdot)$ is the covariance and σ is the standard deviation. If α and β are uncorrelated, $cor(\alpha,\beta) = 0$. 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

$$D_{fcor}(f) = \sum_{g\in\mathcal{F}} (corS_{f,g}(t) - corC_{f,g}(t)) \quad (8)$$

where

$$corS_{f,g}(t) = cor(v_1^f(t), v_1^g(t)) + cor(v_2^f(t), v_2^g(t))$$

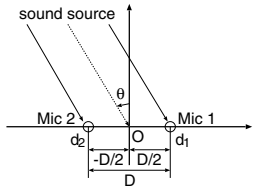

Fig. 1. Direction of arrival.

$$corC_{f,g}(t) = cor(v_1^f(t), v_2^g(t)) + cor(v_2^f(t), v_1^g(t)). \quad (9)$$

If $D_{fcor}(f) < 0$, we assume that the permutation has occurred at frequency f, whereas if $D_{fcor}(f) > 0$, the permutation has not occurred at frequency f. The simplest set of the frequencies g is $\mathcal{F} = \{g : f - \delta_f \le g \le f - 1\}$, where δ_f 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 [7] 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 [6] 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 [5, 6]. 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

$$\begin{aligned} G_k(\theta_k(f)) &= |Y_k(f,t)/\exp(j2\pi ft)| \\ &= |W_{k1}(f)\exp(-j2\pi f(d_1\sin(\theta_k(f)))/c) \\ &\quad + W_{k2}(f)\exp(-j2\pi f(d_2\sin(\theta_k(f)))/c)|. \end{aligned} \quad (10)$$

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 $|\theta_k(f)| \le \pi/2$ [5, 6]. 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 [5, 6].

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 $Wd_k(f) = \angle W_{k1}(f) - \angle W_{k2}(f)$. The difference should be also linear phase, so we can represent the difference by the following equation:

$$\hat{Wd}_k(f) = a_k f + b_k. \tag{11}$$

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:

$$\angle \tilde{W}_{kl}(f) = \frac{1}{2M} \sum_{m=-M}^{M-1} \angle W_{kl}(f+m) \tag{12}$$

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

$$\tilde{Wd}_k(f) = \angle \tilde{W}_{k1}(f) - \angle \tilde{W}_{k2}(f). \tag{13}$$

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

$$a_k = \frac{1}{C_f}\left[(f_h - f_l + 1)\sum_{f=f_l}^{f_h} f\tilde{Wd}_k(f) - \sum_{f=f_l}^{f_h} f \sum_{f=f_l}^{f_h} \tilde{Wd}_k(f)\right] \tag{14}$$

$$b_k = \frac{1}{C_f}\left[\sum_{f=f_l}^{f_h} f^2 \sum_{f=f_l}^{f_h} \tilde{Wd}_k(f) - \sum_{f=f_l}^{f_h} f\tilde{Wd}_k(f) \sum_{f=f_l}^{f_h} f\right] \tag{15}$$

where

$$C_f = (f_h - f_l + 1)\sum_{f=f_l}^{f_h} f^2 - \left(\sum_{f=f_l}^{f_h} f\right)^2 \tag{16}$$

and f_l and f_h are chosen from a low frequency range where the two curves of the equation (13) 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 example, f_l is chosen to avoid the effect of low frequencies such as bins 5–20, and f_h could be calculated as

$$f_h = \min_f \left(\left| \tilde{Wd}_1(f) - \tilde{Wd}_2(f) \right| < \frac{\pi}{2} \right). \tag{17}$$

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 $\hat{Wd}_k(f)$.

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

$$Wd_k(f) \leftarrow \mathrm{mod}\,(Wd_k(f), 2\pi) - \pi \tag{18}$$

$$\hat{Wd}_k(f) \leftarrow \mathrm{mod}\,(\hat{Wd}_k(f), 2\pi) - \pi. \tag{19}$$

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

Table 1. Comparison of average SIR, SAR, and SDR [8] obtained with the inter-frequency correlation method and the proposed method. All values are expressed in decibels (dB).

	Anechoic		Echoic	
	Inter-freq. Correlation	Proposed Method	Inter-freq. Correlation	Proposed Method
SIR	28.99	29.06	26.02	26.18
SAR	13.62	13.69	11.74	11.87
SDR	13.43	**13.50**	11.56	**11.68**

$$D_{prop}(f) = [|Wd_1(f) - \hat{Wd}_1(f)| + |Wd_2(f) - \hat{Wd}_2(f)|] - [|Wd_1(f) - \hat{Wd}_2(f)| + |Wd_2(f) - \hat{Wd}_1(f)|]. \tag{20}$$

[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.

5. EXPERIMENTS

To confirm this approach, 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 box[1]. 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 two microphones is 4 cm. For 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 2 as the parameter δ_f. For the proposed method, 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.

The results are shown in Figures 2–5 and Table 1. The proposed method can solve the permutation problem almost perfectly in the anechoic environment. However, the method can have an error at frequencies near where two estimated linear phase lines cross (e.g. frequency bin number 820). At such points, this method cannot distinguish two sources, because the two sources have the same phases at these points. In an echoic environment, the result of the proposed method

[1] http://www.2pi.us/rir.html

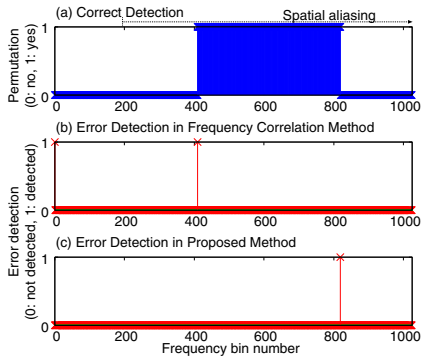

Fig. 2. Detection of permutation in the anechoic environment; (a) correct detection, (b) detection errors in inter-frequency correlation method (error rate: 0.2% (=2/1025)), (c) detection errors in proposed method (0.1%).

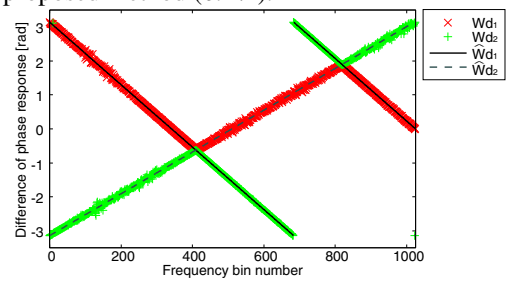

Fig. 3. The de-mixing matrix phase difference in the anechoic environment, showing (i) observed points Wd_k and (ii) estimated lines $\hat{W}d_k$ from equation (11).

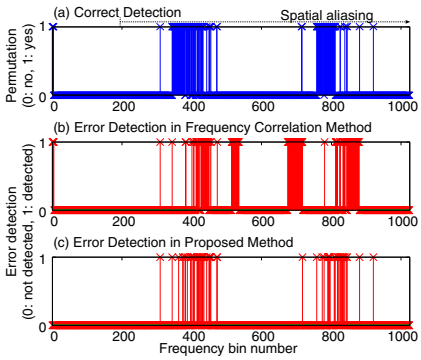

Fig. 4. Detection of permutation in the echoic environment; (a) correct detection, (b) detection errors in inter-frequency correlation method (error rate: 16.0%), (c) detection errors in proposed method (6.2%).

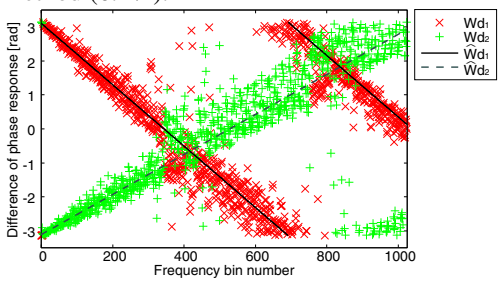

Fig. 5. The de-mixing matrix phase difference in the echoic environment, showing (i) observed points Wd_k and (ii) estimated lines $\hat{W}d_k$ from equation (11).

is not perfect but better than the inter-frequency correlation method. However, our method has the same problem around the cross points (here, frequency bin numbers 418 and 839) in the echoic environment as in the anechoic environment.

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. The proposed method can solve the problem well in an anechoic environment, and in our example, for a reverberant environment, the proposed method gives slightly better performance than that of the inter-frequency correlation method.

However, the proposed method has a difficulty around the points where the two linear curves of the difference of the phase response cross, and the method cannot solve the permutation at those points perfectly. In future work, we are considering combining with another method such as the inter-frequency correlation method to solve the permutation around those points.

7. 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] N. Murata, *et al.*, "An on-line algorithm for blind source separation on speech signals," *Proc. 1998 Int. Symp. Nonlinear Theory and Its Application*, vol. 3, pp.923–926, 1998.

[5] 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.

[6] 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 Processing*, vol. 12, pp. 530–538, 2004.

[7] N. Murata, *et al.*, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp.1–24, 2001.

[8] E. Vincent, *et al.*, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, pp.1462–1469, 2006.