A ROBUST APPROACH TO THE PERMUTATION PROBLEM OF FREQUENCY-DOMAIN BLIND SOURCE SEPARATION

Hiroshi Sawada *Ryo Mukai* *Shoko Araki* *Shoji Makino*

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{sawada,ryo,shoko,maki}@cslab.kecl.ntt.co.jp

ABSTRACT

This paper presents a robust and precise method for solving the permutation problem of frequency-domain blind source separation. It is based on two previous approaches: the direction of arrival estimation approach and the inter-frequency correlation approach. We discuss the advantages and disadvantages of the two approaches, and integrate them to exploit the both advantages. We also present a closed form formula to calculate a null direction, which is used in estimating the directions of source signals. Experimental results show that our method solved permutation problems almost perfectly for a situation that two sources were mixed in a room whose reverberation time was 300 ms.

1. INTRODUCTION

Blind source separation (BSS) is a technique for estimating original source signals using only sensor observations, which consist of mixtures of the original signals. If the mixture is instantaneous, we can directly apply independent component analysis (ICA) [1, 2] to separate mixed signals. In a real room environment, however, signals are mixed in a convolutive manner with reverberations. This makes the BSS problem difficult since we need a set of filters, not just scalars, to separate signals. One of the major methods to obtain such separating filters is frequency-domain BSS [3–10], where a convolutive mixture in the time domain is converted into multiple instantaneous mixtures. Thus, we can apply ICA to instantaneous mixtures in every frequency bin.

The problem with frequency-domain BSS is the indeterminacy of permutation that is inherent to ICA. We need to map a separated signal at each frequency to a target source signal so that we properly reconstruct a separated signal in the time domain. Various approaches have been proposed to the permutation problem. Making separating matrices smooth in the frequency domain is one solution. This has been realized by averaging separating matrices with adjacent frequencies [3], limiting the filter length in the time domain [4], or considering the coherency of separating matrices at adjacent frequencies [5]. Another approach is based on direction of arrival (DOA) estimation in the beamforming theory [6, 7]. If source signals are speech, we can employ the inter-frequency correlations of signal envelopes to align permutations [8, 9]. Each of these approaches has different characteristics. They may perform well under certain specific conditions but not others. Therefore, we believe that integrating some of these approaches is one way of obtaining better performance.

In this paper, we propose a new method for solving the permutation problem, which incorporates two of the previous approaches. The first is the DOA approach, which is described in Sec. 3. The second is based on inter-frequency correlations, which is discussed in Sec. 4. Our new method is proposed in Sec. 5. The experimental results reported in Sec. 6 are promising.

The second contribution of this paper is a closed form formula for calculating a null direction, which is used in estimating the directions of source signals (Sec. 3). It dramatically reduces the calculation cost of null directions compared with the conventional method by searching for the minimum of a directivity pattern.

2. FREQUENCY-DOMAIN BSS

Suppose that P source signals $s_p(t)$ are mixed and observed at Q sensors $x_q(t) = \sum_{p=1}^{P} \sum_k h_{qp}(k) s_p(t-k)$, where $h_{qp}(k)$ represents the impulse response from source p to sensor q. The goal of BSS is to obtain separated signals $y_1(t), \ldots, y_P(t)$ that are estimates of the source signals $s_1(t)$, $\ldots, s_P(t)$. The separating system typically consists of a set of FIR filters $w_{rq}(k)$ that produces separated signals $y_r(t) = \sum_{q=1}^{Q} \sum_k w_{rq}(k) x_q(t-k)$.

This paper employs a frequency-domain approach where frequency responses $W_{rq}(f)$ of the separating filter $w_{rq}(k)$ are first calculated. By L-point short time DFT, time-domain signals $x_q(t)$ are converted into frequency-domain time-series signals $X_q(f,m)$, where $f = 0,\ f_s/L,\ \ldots,\ f_s(L-1)/L$ (f_s: sampling frequency), and m is the frame index. Assume that $\mathbf{X}(f,m)$ is a Q-dimensional vector $\mathbf{X}(f,m) = [X_1(f,m),\ \ldots,\ X_Q(f,m)]^T$. To obtain frequency responses $W_{rq}(f)$, we solve ICA problem $\mathbf{Y}(f,m) = \mathbf{W}(f)\mathbf{X}(f,m)$, where $\mathbf{Y}(f,m) = [Y_1(f,m), \ldots, Y_P(f,m)]^T$ and $\mathbf{W}(f)$ is a $P \times Q$ matrix whose elements are $W_{rq}(f)$. $Y_r(f,m)$ is a frequency-domain representation of $y_r(t)$.

The ICA algorithm we use is the information maximization approach [1] combined with the natural gradient [2]. A

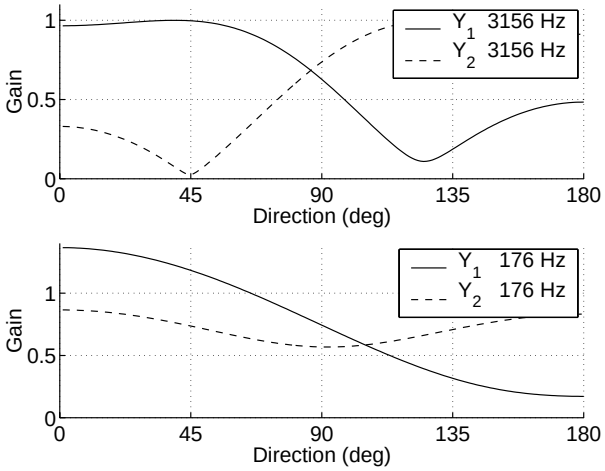

Fig. 1. Directivity patterns

separating matrix $\mathbf{W}$ is gradually improved by the learning rule $\Delta\mathbf{W} = \mu\left[\mathbf{I} - \langle\Phi(\mathbf{Y})\mathbf{Y}^H\rangle\right]\mathbf{W}$, where μ is a step-size parameter, $\langle\cdot\rangle$ denotes the averaging operator, and $\Phi(\cdot)$ is a nonlinear function for a complex signal $Y_r = |Y_r|\, e^{j\cdot phase(Y_r)}$. We use $\Phi(Y_r) = -\frac{\partial}{\partial|Y_r|}\log p(|Y_r|)\ e^{j\cdot phase(Y_r)}$ assuming that the density $p(Y_r)$ is independent of the phase [10].

An ICA solution has an ambiguity on permutation: if we permute the rows of $\mathbf{W}(f)$, it is still a solution. Thus, we have to align the rows of $\mathbf{W}(f)$ so that $Y_r(f,m)$ at all frequencies correspond to the same source $s_p(t)$. This is the permutation problem. After solving the problem, we obtain separating filters $w_{rq}(k)$ by applying inverse DFT to $W_{rq}(f)$.

3. THE DIRECTION OF ARRIVAL APPROACH

In this section, we first review the method [6, 7] for solving the permutation problem by estimating the directions of source signals. If the sensor spacing is appropriately narrow (e.g., conditions in Table 1), each row of $\mathbf{W}(f)$ usually forms spatial nulls in the directions of jammer signals and extracts a target signal in another direction [11]. By analyzing the null directions, we can estimate the directions $\Theta(f) = [\theta_1(f), \ldots, \theta_P(f)]^T$ of target signals that every row of $\mathbf{W}(f)$ extracts. Then, we can align permutations according to $\Theta(f)$.

The null directions can be analyzed by plotting the directivity pattern of each output $Y_r(f,m)$. Let d_q be the position of sensor q (we assume linearly arranged array sensors), and θ_p be the direction of source s_p (the direction orthogonal to the array is $90°$). In the beamforming theory [12], the frequency response of an impulse response $h_{qp}(t)$ is approximated as $H_{qp}(f) = e^{j2\pi fc^{-1}d_q\cos\theta_p}$, where c is the velocity of propagation. In this approximation, we assume a plane wavefront and no reverberation. The frequency response $B_{rp}(f)$ from a source s_p to a separated signal y_r can

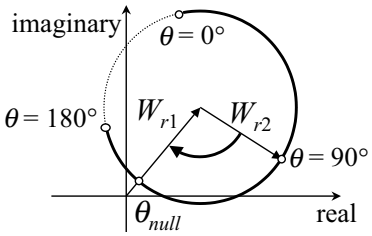

Fig. 2. Directivity pattern on a complex plane

be expressed as $B_{rp}(f) = \sum_{q=1}^{Q} W_{rq}(f)\cdot e^{j2\pi fc^{-1}d_q\cos\theta_p}$. If we regard θ_p as a variable θ, the formula is expressed as $B_r(f,\theta) = \sum_{q=1}^{Q} W_{rq}(f)\cdot e^{j2\pi fc^{-1}d_q\cos\theta}$. It changes according to the direction θ, and thus is called a directivity pattern.

Figure 1 shows directivity patterns for two sources. The upper part (3156 Hz) shows that output Y_1 extracts a source signal originating from around $45\,°$ and suppresses the other signal coming from around $125°$. With a similar consideration on Y_2, we estimate the directions $\Theta(3156) = [45, 125]^T$ of the target signals. A simple way to solve the permutation problem is to permute $\mathbf{W}(f)$ at each frequency so that $\Theta(f)$ are sorted. However, not every frequency bin gives us such an ideal directivity pattern. The lower part of Fig. 1 is the pattern at a low frequency (176 Hz). We see that a null is not well formed for Y_1 and the null of Y_2 is in an obscure direction. In fact, we cannot estimate $\Theta(176)$ or decide a permutation for this frequency with confidence.

Now we state two problems with this method: 1) directions of arrival cannot be well estimated at some frequencies, especially at low frequencies where the phase difference caused by a sensor spacing is very small, 2) the calculation of null directions by plotting directivity patterns is time consuming. The first problem will be solved in Sec. 5.

For the second problem, here we provide a closed form formula for calculating a null direction (only for two sensors). The directivity pattern $B_r(f,\theta)$ can be simplified as $B_r(f,\theta) = W_{r1}(f) + W_{r2}(f)\cdot e^{j2\pi fc^{-1}d\cos\theta}$ if we assume $d_1 = 0$ and $d_2 = d$. As $\cos\theta$ changes from 1 to -1 in accordance with the change of direction θ, the frequency response $B_r(f,\theta)$ rotates around a circle whose center and radius are W_{r1} and W_{r2}, respectively (Fig. 2). We see that spatial aliasing does not occur if $2\pi fc^{-1}d < \pi \Leftrightarrow f < c/(2d)$. When the distance from the origin is minimized, the plot corresponds to the null direction θ_{null}. The situation is one where W_{r2} rotated by $e^{j2\pi fc^{-1}d\cos\theta_{null}}$ points to the opposite direction of W_{r1}: $angle(W_{r2}) + 2\pi fc^{-1}d\cos\theta_{null} = angle(W_{r1}) \pm \pi$. Therefore,

$$\theta_{null} = \cos^{-1}\frac{angle(W_{r1}) - angle(W_{r2}) \pm \pi}{2\pi fc^{-1}d}$$

The sign of $\pm\pi$ is selected so that the numerator is in the $-\pi$ to π range. There are some cases where the absolute value of the argument of $\cos^{-1}$ is larger than 1. This corresponds to a situation where there is no null in $0\,° \le \theta \le 180°$.

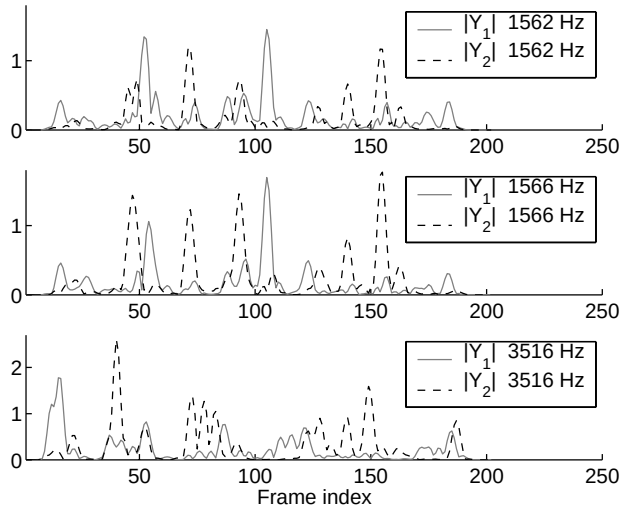

Fig. 3. Envelopes at different frequencies

4. THE CORRELATION APPROACH

This section discusses an approach to permutation alignment based on inter-frequency correlation [5, 8, 9]. We take the envelope $v_r^f(m) = |Y_r(f,m)|$ of a separated signal $Y_r(f,m)$ to measure correlations. Let us define the correlation of two signals $x(m)$ and $y(m)$ as $\mathrm{cor}(x,y) = [\langle x \cdot y \rangle - \langle x \rangle \cdot \langle y \rangle]/(\sigma_x \cdot \sigma_y)$, where $\langle \cdot \rangle$ is an averaging operator and σ is a standard deviation. Based on this definition, $\mathrm{cor}(x,x) = 1$, and $\mathrm{cor}(x,y) = 0$ if x and y are uncorrelated. Envelopes have high correlations at neighboring frequencies if separated signals correspond to the same source signal [8, 9]. Figure 3 shows an example. Two envelopes v_1^{1562} and v_1^{1566}, as well as v_2^{1562} and v_2^{1566}, are highly correlated. Thus, calculating such correlations helps us to align permutations.

Henceforth let π denote a permutation: $\{1,\ldots,P\} \rightarrow \{1,\ldots,P\}$. A simple criterion to decide a permutation π_f of frequency f is to maximize the sum of correlations between neighboring frequencies within distance D:

$$\pi_f = \mathrm{argmax}_\pi \sum_{|k-f|\le D} \sum_p \mathrm{cor}(v^f_{\pi(p)}, v^k_{\pi_k(p)}), \qquad (1)$$

where π_k is the permutation at frequency k. This criterion is based on local infomation and has a drawback that mistakes in a narrow range of frequencies may lead to the complete misalignment of the frequencies beyond that point. To avoid this problem, the method proposed in [9] does not limit the frequency range in which correlations are calculated. It decides permutations one by one based on the criterion:

$$\pi_f = \mathrm{argmax}_\pi \sum_p \mathrm{cor}(\, v^f_{\pi(p)}, \sum_{k\in F} v^k_{\pi_k(p)}\,), \qquad (2)$$

where F is a set of frequencies in which the permutation is decided. This method assumes high correlations of envelopes even between frequencies that are not close neighbors, although this is not always the case. As shown in Fig. 3, v_r^{1566} and v_r^{3516} do not have a high correlation. Therefore, this method still has a drawback in that permutations may be misaligned at many frequencies.

```
for (∀f) Θ(f) = DOA(f, W(f))
m_Θ = ⟨Θ(f)⟩_f    /* averaged directions */
F = ∅    /* the set of fixed frequencies */
/* Fix permutations by the DOA approach */
for (∀f) {
  if ( confident(Θ(f), m_Θ, W(f)) ) {
    π_f = getPermutation(Θ(f))
    F = F ∪ {f}
  }
}
/* Fix permutations by the correlation approach */
while (∃f ∉ F) {
  for (∀f ∉ F) {
    c_f = max_π Σ_{|k−f|≤D, k∈F} Σ_p cor(v^f_{π(p)}, v^k_{π_k(p)})
    π_f = argmax_π Σ_{|k−f|≤D, k∈F} Σ_p cor(v^f_{π(p)}, v^k_{π_k(p)})
  }
  i = argmax_f c_f
  F = F ∪ {i}
  c_i = 0
}
```

Fig. 4. Pseudo-code for the integrated method

5. A ROBUST INTEGRATED APPROACH

In this section, we propose a robust and precise approach which integrates the two approaches discussed above. We consider their characteristics below.

robustness The direction of arrival (DOA) approach is robust since a misalignment at a frequency does not affect other frequencies. The correlation approach is not robust as discussed in Sec. 4.

preciseness The DOA approach is not precise since the evaluation is based on the approximation of a mixing system as explained in Sec. 3. The correlation approach is precise as long as signals are well separated by ICA since the measurement is based on separated signals.

Our approach benefits from both advantages: the robustness of the DOA approach and the preciseness of the correlation approach. Figure 4 shows the pseudo-code.

We first fix permutations at some frequencies where the confidence of the DOA approach is sufficiently high. The procedure *confident* decides whether the confidence is high enough. Our criteria for the decision are: 1) the number of estimated directions is the same as the number of sources, 2) the directions $\Theta(f)$ do not largely differ from the averaged directions $\mathbf{m}_\Theta$, 3) the SNR calculated by the frequency responses $B_r(f,\theta_p)$ for each direction is sufficiently large.

Then, we decide the permutations for the remaining frequencies by the correlation method without changing the permutations fixed by the DOA approach. The permutations are decided in order of the sum of correlations with fixed frequencies $k \in F$ within distance $|k-f| \le D$.

Table 1. Experimental conditions

Length of source signal	6 sec
Direction of sources	50° and 120° (2 sources)
Distance between 2 sensors	d = 4 cm
Reverberation time	T_R = 300 ms
Sampling rate	f_s = 8 kHz
Frequency resolution	$L = 2048$
Nonlinear function	$\Phi(Y_r) = e^{j \cdot phase(Y_r)}$
Distance to take correlation	$D = 3 \cdot f_s / L$

This approach does not result in a large misalignment as long as the permutations fixed by the DOA approach are correct. Moreover, the correlation part compensates for the lack of preciseness of the DOA approach.

6. EXPERIMENTAL RESULTS

We performed experiments to separate speech signals in a reverberant environment whose conditions are summarized in Table 1. We generated mixed signals by convolving a speech signal $s_p(t)$ and an impulse response $h_{qp}(t)$ so that we can calculate SNRs (signal-to-noise ratios) by

$$10\log[\textstyle\sum_{r=p}\sum_t y_{rp}(t)^2] - 10\log[\sum_{r\neq p}\sum_t y_{rp}(t)^2],$$

where $y_{rp}(t) = \sum_{q=1}^{Q}\sum_k w_{rp}(k)x_{qp}(t-k)$ and $x_{qp}(t) = \sum_k h_{qp}(k)s_p(t-k)$. We separated 12 combinations of speech signals with 4 different methods for comparison: the DOA approach, the correlation approach based on (1), the correlation approach based on (2), and our proposed method. Figure 5 shows the results, where "ave" shows the average result of 12 combinations for each method.

The performance with "DOA" is stable, but not sufficient. The results with "Cor (1)" and "Cor (2)" are not stable and sometimes very poor, although most of the time they are very good. The "Integrated" method offers stable and very good results. The percentages of the permutations fixed by the DOA approach with confidence were around 45%. We additionally obtained optimal permutations by maximizing the SNR at each frequency. Although this is not a realistic solution, we can estimate the upper bounds of performance. The average of "Optimal" was 16.8 dB. Since the average performance with "Integrated" was 16.3 dB, we consider that the integrated method performed very well.

7. CONCLUSIONS

We proposed a robust and precise method for solving permutation problem. It integrates two previous approaches: the DOA approach and the correlation approach. The criterion of the DOA approach is directions which is absolute. This makes the approach robust. By contrast, the criterion of the correlation approach is calculated from the separated signals themselves. This makes the approach precise. Our proposed method benefits from both advantages. In our experiments, the proposed method solved permutation problems almost perfectly under conditions whereby 2 sources were mixed in a room where T_R = 300 ms.

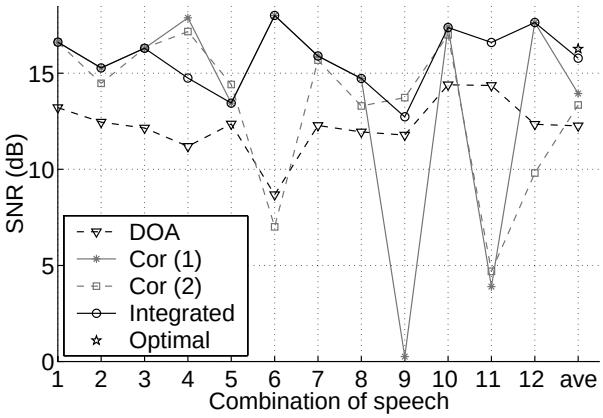

Fig. 5. Experimental results

8. REFERENCES

[1] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[2] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

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

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

[5] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "A combined approach of array processing and independent component analysis for blind separation of acoustic signals," in *Proc. ICASSP 2001*, May 2001, pp. 2729–2732.

[6] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. ICASSP 2000*, June 2000, pp. 3140–3143.

[7] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. ICASSP 2002*, May 2002, pp. 881–884.

[8] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proc. ICA 2000*, June 2000, pp. 215–220.

[9] 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.

[10] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," in *Proc. ICASSP 2002*, May 2002, pp. 1001–1004.

[11] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind source separation with different sensor spacing and filter length for each frequency range," in *Proc. Workshop on Neural Networks for Signal Processing*, Sept. 2002, pp. 465–474.

[12] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, pp. 2–24, Apr. 1988.