A New Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation

Xuebin HU†, Nonmember and Hidefumi KOBATAKE†, Member

SUMMARY Frequency domain blind source separation has the great advantage that the complicated convolution in time domain becomes multiple efficient multiplications in frequency domain. However, the inherent ambiguity of permutation of ICA becomes an important problem that the separated signals at different frequencies may be permuted in order. Mapping the separated signal at each frequency to a target source remains to be a difficult problem. In this paper, we first discuss the inter-frequency correlation based method [1], and propose a new method using the continuity in power between adjacent frequency components of same source. The proposed method also implicitly utilizes the information of inter-frequency correlation, as such has better performance than the previous method.

1. Introduction

Blind source separation (BSS) has received extensive attention in signal and speech processing, machine intelligence, and neuroscience communities. The goal of BSS is to recover the unobserved original sources without any prior information given only the sensor observations that are the unknown linear mixtures of the independent source signals. If the mixture is instantaneous, we can directly employ independent component analysis (ICA) to achieve the task. In real environment, due to multi-path propagation and reverberation, the signals impinged on an array of microphones are the convolutive mixtures of sources. BSS may be implemented in time domain by learning the time-domain coefficients of the unmixing filter. However, the filter may need to be thousands of taps long to properly invert the mixing. Computationally, it is lighter to move to the frequency domain as convolution with long filter in the time domain becomes efficient multiplications in the frequency domain under certain conditions [1], [2]. This has the great advantage that ICA still could be directly used to achieve the separation. Frequency domain BSS brings out the problems that the standard ICA indeterminacies of scaling and permutation appear at each output frequency bin. The scaling problem could be easily solved by putting the separated frequency components back to the microphones with the inverse matrices. However, permutation remains to be a difficult problem. We need to map a separated component at each frequency to a target source signal so as to properly reconstruct the separated signals in the time domain. Various proposals have been reported using different continuity criteria to overcome the permutation problem [2]. Nevertheless, it is still open to satisfying and rigorous solution.

There are some inherent limitations on various proposals. For example, one method makes use of the coherency of separating matrices at neighbor frequencies. However, the coherency only exists in very simple environment, it does not hold in most case. Another approach is based on direction of arrival (DOA) estimation in array signal processing [3]. By analyzing the directivity patterns formed by the separating matrix, source directions can be estimated and therefore permutation can be solved. The inherent limitation of this approach is that the sources must be spaced apart away. Otherwise it could not work well due to the variation of DOA at different frequencies [4] and the unavoidable error in DOA estimation. Ikeda et al. proposed an approach employing the inter-frequency correlation of signal envelopes to align permutation because source signals are speech [1]. It seems a sound solution as inter-frequency correlation does exist at adjacent frequencies for speech signals. This approach is related with our method and is to be discussed in Sect. 3. A recently proposed method [5] combines the DOA estimation with inter-frequency correlation, and also incorporates the harmonic structure of speech signals. It seems to achieve the best performance up to now, but has the disadvantage that the location of microphones should be known in advance. In other words, it is not completely blind.

This paper proposes a new method based on a similar but different assumption from Ikeda’s method. We assume that there exists continuity in power (amplitude) between the waveforms of adjacent frequency components of same source. Based on the assumption, we propose to use the distance between the signals at adjacent frequencies to align the separated signals. As the information from the distance implicitly includes the information of inter-frequency correlation, the proposed method does not conflict with the Ikeda’s method but includes more helpful information. Consequently, it has a better performance than the previous method.

Section 2 briefly describes the frequency domain BSS system. In Sect. 3, we review the inter-frequency correlation based method and present the proposed method. Section 4 gives the comparison test results, and followed by the con-
where \( X \) becomes, \( \text{verse matrices separately. The unmixing filter matrix then}
\)
\[
x(t) = A(t) \ast s(t) = \sum_q a_{qk}(t) \ast s_q(t),
\]
(1)
where \( A \) is an unknown polynomial matrix, \( a_{qk} \) is the impulse source \( q \) to microphone \( k \), and the asterisk symbol \( \ast \) refers to convolution operation. The obtained mixed sources are decomposed into frequency domain by performing short-term discrete Fourier transform. Then the convolutive mixing problem becomes multiple instantaneous mixing problems.

\[
X(\omega, t) = A(\omega)S(\omega, t),
\]
(2)
where, \( X(\omega, t) = [X_1(\omega, t), \cdots, X_N(\omega, t)]^T \), \( A(\omega) \) is a \( K \times Q \) matrix, and \( S(\omega, t) = [S_1(\omega, t), \cdots, S_Q(\omega, t)]^T \). The unmixing filter matrix \( W(\omega) \) is derived using the Infomax algorithm [6]. The learning rule is defined as follows,

\[
W_{i+1}(\omega) = W_i(\omega) + \eta[I - \varphi(Y(\omega, t))]Y^H(\omega, t)W_i(\omega)
\]
\[
\varphi(Y(\omega, t)) = 2 \tanh(\text{Re}(Y(\omega, t))) + 2j \tanh(\text{Im}(Y(\omega, t)))
\]
(3)
where, \( \eta \) is a factor that determines the convergence speed, \( \varphi(\cdot) \) is a nonlinear score function, and \( Y(\omega, t) = W_i(\omega)X(\omega, t) \). The ambiguity of scaling is solved by filtering the individual output of the unmixing filter using the inverse matrices separately. The unmixing filter matrix then becomes,

\[
W'_i(\omega) = W^{-1}(\omega)\delta(i, i)W(\omega),
\]
(4)
where \( W(\omega) \) denotes the derived unmixing filter, \( W'_i(\omega) \) denotes the unmixing that outputs the \( \cdot \)th source signal, and \( \delta(i, i) \) denotes a “delta matrix” of which only the \( i \)

element equals to one and all the remaining elements are zeros.

The permutation problem is the main topic of this paper and is to be described in Sect. 3. After solving the ambiguity of scaling and permutation, the derived unmixing filters are then transformed back to time domain through inverse Fourier transform. The time domain unmixing filter is derived as follows,

\[
W_{i,t} = F^{-1}[W'_i(\omega_1)H(\omega_1), \cdots, W'_i(\omega_N)H(\omega_N)]
\]
\[
\cdot \text{ham}(t),
\]
(5)
where \( F^{-1} \) denotes inverse Fourier transformation, \( \text{ham}(t) \) denotes the Hamming window, and \( H(\omega_k) \) is a circular time shift operator. When the window length is \( N \), a time shift of \( N/2 \) is experimentally good, i.e., \( H(\omega_0) = e^{j\pi n} \) [7]. At the last, the mixed signals are separated by the derived time-domain unmixing filter.

### 3. Continuity Based Method

In this section, we first review the inter-frequency correlation based method proposed by Ikeda et al. [1], and then we introduce a looks similar but different method based on a new assumption of inter-frequency continuity in power between adjacent frequency components.

#### 3.1 Correlation Based Method

In the correlation based method [1], it is assumed that if the split band-passed signals originate from the same source signal, they are under the influence of a similar modulation in amplitude. In other words, there is correlation exist between the envelopes of Fourier components of same source.

The operator to take the envelope, \( \varepsilon \) is defined as

\[
\varepsilon \tilde{s}_i(t; \iota; i) = \frac{1}{2M} \sum_{t'=t-M}^{t+M} \sum_{k=1}^{K} [\tilde{s}_{k,i}(t'; i)],
\]
(6)
where \( \tilde{s}_i(t'; i) \) denotes the frequency component of the \( i \)
th source, and \( \tilde{s}_{k,i}(t'; i) \) denotes the input of the \( i \)
th source component into the \( k \)-th \( (k = 1, \cdots, K) \) sensor. \( M \) is the number of time steps for taking the moving average, and \( \iota \) refers to the sequence number of windows.

The permutation is solved using the correlation between the envelopes of separated signals. First, the sequence of frequency to solve the permutation is determined by sorting the similarity between the separated components in an increasing order. The similarity is defined as follows,

\[
sim(\varepsilon Y_i(\omega, t), \varepsilon Y_j(\omega, t)) = \frac{\varepsilon Y_i(\omega, t) \cdot \varepsilon Y_j(\omega, t)}{||\varepsilon Y_i(\omega, t)|| ||\varepsilon Y_j(\omega, t)||}
\]
(7)
where, \( Y_i(\omega, t) \) and \( Y_j(\omega, t) \) are the separated frequency components of source \( i \) and \( j \), “ . ” denotes inner product, and \( || \) \( || \) denotes the norm. For \( \omega_1 \), assign the order as it is. For \( \omega_N \), find the alignment that maximizes the correlation between the envelope with the aggregated envelopes from \( \omega_1 \) through
of the aligned sources.

In Ikeda’s method, because permutation is solved in the increasing order of similarity, it is implemented in a random frequency sequence. This implies that the aligned frequencies may be apart away from the frequency to be decided. Figure 2 gives an example showing that the envelopes of same source have high correlation at adjacent frequencies (see the top and middle rows of source 1 and source 2 in Fig. 2). However, correlation does not hold when frequencies are apart away (see the bottom rows of sources 1 and 2 in Fig. 2). Consequently, using the sum of envelopes of decided frequencies can not ensure a good job. The inter-frequency correlation should only be used within an adjacent frequency band.

3.2 Distance-Based Method

The inter-frequency correlation based method only uses the correlation between signal envelopes. From Fig. 2, we see that the envelopes of same source have similarity both in waveform and amplitude. It is also reasonable to assume that there exists continuity in power between adjacent frequency components. In other words, the power will not change dramatically between neighbor frequencies. Here, the power is defined as \( \ln |s_i(\omega, t)|^2 \), the natural logarithm of the squared amplitude of the frequency component at a certain bin. Based on the assumption, we proposed to use the distance between the signal vectors at adjacent frequencies to align the separated signals. The new assumption does not conflict with the correlation assumption but includes more helpful information. The distance criterion implicitly utilizes the information of inter-frequency correlation. As such, it should have a better performance than the previous method. For example, it is possible that the Fourier components of different sources may be relatively correlated. In such case, the previous method might be difficult to deal with. However, if the powers of the highly correlated envelopes of different sources are quite different, the continuity based method will be competent.

We use the continuity of power within a neighboring frequency band. This has the advantage that allows a separation failure at adjacent frequencies, which is sometimes unavoidable due to reasons like the low-independence between original source components [8]. Additionally, solving the permutation within a short band instead of only the immediate neighboring frequency eliminates the risk to transfer a misalignment to all the subsequent frequencies.

The distance between two signal vectors \( s_f(\omega, t), s_j(\omega, t) \), is defined as,

\[
d_{i,j}(\omega_k, \omega_r) = \left( \sum_t |v_i(\omega_k, t) - v_j(\omega_r, t)|^p \right)^{1/p}
\]

where, \( \omega_k \) denotes the frequency at which permutation is to be decided, and \( \omega_r \) denotes the frequency to be used as reference. Note we use an alternative form of power, \( \ln |s_i(\omega, t)|^2 \), instead of the conventional form of \( |s_i(\omega, t)|^2 \), to avoid the metric of distance become numerically too big. \( p \) is a constant. If \( p \) equals one, \( d_{i,j} \) is the sum of the absolute difference, and when \( p \) equals two, it is the Euclidean distance. Before calculating the distance, natural logarithm is taken for reducing the effect from the variation of amplitude.

The procedures of the proposed approach are as follows:

- Using Eq. (7) to find \( \omega_1 \) at which the separated signals are the most uncorrelated. The order of \( \omega_1 \) is set as it is. Here, the subscript of \( \omega \) refers to the solution sequence. \( \omega_1 \) is used as the start point. The permutations are to be solved in the monotonously increasing and decreasing directions from \( \omega_1 \).

- For \( \omega_k \), we first find the most reliable reference frequency \( \hat{\omega}_r \) in the band of \( [\omega_{k-L}, \ldots, \omega_{k-1}] \) when spread to increasing direction, or in \( [\omega_{k+1}, \ldots, \omega_{k+L}] \) when spread to decreasing direction. \( L \) is the band width.

\[
\hat{\omega}_r = \arg \max_{\omega_1} F
\]

\[
F = |d_{i,j}(\omega_k, \omega_1) - d_{i,j}(\omega_k, \omega_r)|
\]

where, \( F \) denotes a relative distance between one source to another. It is a measure of reliability of a permutation decision. A higher \( F \) means the signals at \( \omega_k \) and \( \omega_r \) of same source have stronger continuity in power, and the signals of different sources are more apart away. In other words, the decision made with \( \omega_r \) as the reference is more reliable.
4. Experimental Results

4.1 Reliability Comparison

We use two sound signals, decomposing them into their frequency domain representations, and compare the proposed method with Ikeda’s method with regard to their reliabilities of decisions. For the proposed method, we set \( p \) equals 2, band width \( L \) equals 5, and evaluate \( d_{ij} \) and \( d_{ij} \) for each \( \omega \) using Eqs. (8), (9). The reliability is reflected by the difference between them. For the correlation-based method, we evaluate \( c_{ij} \) and \( c_{ij} \), which are the correlation coefficients between the envelopes of frequency \( \omega \) and the aggregated envelopes of aligned frequencies. Figure 3 shows the analysis result. The distance index \( d_{ij} \) appears much more stable than the correlation coefficient \( c_{ij} \), which implies the proposed criterion is a better measure than the correlation in making the decision whether two components belong to the same source or not. In the upper drawing of Fig. 3, \( c_{ij} \) and \( c_{ij} \) are very close to each other at some frequencies, while this is not the case between \( d_{ij} \) and \( d_{ij} \). With the proposed approach, a more reliable and better performance could be expected.

- Assigning \( s_j(\omega_k, t) \) to \( j \)-th source if \( d_{ij}(\omega_k, \hat{\omega}_r) > d_{ij}(\omega_k, \hat{\omega}_r) \), else, assigning it as it is.
- The above two procedures are iteratively executed until the permutation at all frequency bins have been decided.

4.2 Separation Tests

We evaluate the proposed method in a two-source two-microphone system. The configuration was as follows (see Fig. 4): The space between the two microphones was 5.66 cm. The two sound sources were about 2 m away from the center of the microphones, and were in the direction of -40 and 40 degrees, respectively. Twenty pairs of different sound signals were used to simulate the observed mixtures using the impulse response from RWCP Sound Scene Database in Real Acoustic Environment. The sound signals were about two seconds in length. The reverberation time was 300 msec. The performance was evaluated under the reverberation time changes from 20 to 100 msec (Using the beginning 20 to 100 msec of the real response. Note: more than 99 percent of energy of the impulse response is within the first 100 msec). The truncated impulse responses were multiplied by the Hamming window, of which the center was corresponding to the peak of the impulse response and covered the whole time decays. The sampling rate was 8 kHz, DFT length was 64 msec, Hamming window was used, and window shift was 2 msec. The proposed method was tested at \( p \) set to 1 and 2, respectively. For investigating the up-limit of permutation solution, we use the original source signal as the reference for solving the permutation. The separating performance is evaluated using the following defined signal-to-noise ratio,

\[
SNR_{ij} = 10 \log_{10} \frac{\sum_{t} signal_i(t; j)^2}{\sum_{t} error_i(t; j)^2}
\]

\[
signal_i(t; j) = a_{ij} s_j(t)
\]

\[
error_i(t; j) = y(t; j) - signal_i(t; j)
\]

Figure 5 shows the separation performance at different reverberation times. The result is the average of twenty trials. At the reverberation time of 20 ms, compare with the previous method, the proposed method achieved about 6.8 dB improvement when \( p \) equals 1, and 7.4 dB improvement when \( p \) equals 2, and closed to the up-limit. When the reverberation time gets longer, because the separation performance itself decreased (SNR of perm4 the permutation solution using real signal decreased with reverberation

![Figure 3](image1.png)  
**Fig. 3** Comparison of the proposed method with Ikeda’s method on the difference between \( (c_{ij}, c_{ij}) \), \( (d_{ij}, d_{ij}) \). The difference reflects the reliability of alignment. In the proposed method, the band width \( L \) is set to 5.

![Figure 4](image2.png)  
**Fig. 4** Illustration of the configuration.
time), the overall performance decreased gradually. Nevertheless, the proposed method still achieved better performance. At the reverberation time of 100 ms, about 2.0 and 1.6 dB higher performances were achieved when $p$ equals 1 or 2, respectively. When the reverberation time equals 300 msec, the improvement is minor. This is because when the reverberation becomes longer, the frequency-domain separation itself becomes worse. If the separation cannot reach a certain level, the proposed method will be confused (it is same for other methods) and can hardly contribute to the performance.

In this paper, we say a permutation error occurred when the order made by the method is inconsistent with the perfect solution, which uses the real signal as the reference. The error rate is equal to the number of errors divided by the number of whole frequency bins. Figure 6 shows the error rates of previous method and the proposed method at different reverberation times. Figure 7 shows the error distributions of Ikeda’s method and the new method when reverberation time is equal to 20 or 100 msec, respectively. The vertical coordinate is the sum of the errors in the twenty trails. These figures demonstrated the improvement achieved by the proposed method, especially at the low frequencies.

5. Conclusion

This paper proposed a new permutation solution for frequency domain blind source separation. It is proposed based on a sound assumption that there exists continuity in amplitude between the waveforms of adjacent frequency components of same source.

The proposed method has the advantage that it keeps the useful information from correlation and introduces new favorable information from the continuity of amplitude. This makes the proposed method to do a better job than the previous method in aligning the separated components in frequency domain blind source separation.

References


Xuebin Hu received the B.S. degree in Engineering Physics from Tsinghua University, China in 1989, and the M.E. and Ph.D. degrees in Bio-applications and systems engineering from Tokyo University of Agriculture and Technology in 2001 and 2003, respectively. He worked at the Dongfang Electric Co., China from 1989 to 1998. Currently, he is a research associate at the Graduate School of Bio-Applications and Systems Engineering, Tokyo University of Agriculture and Technology, Japan. His research interests include array signal processing and medical image processing.

Hidefumi Kobatake received the B.E., M.E., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1967, 1969 and 1972, respectively. From 1972 to 1975 he was an Research Associate at the Institute of Space and Astronautical Science, University of Tokyo, Tokyo, Japan. Since 1975 he has joined the Tokyo University of Agriculture and Technology, Tokyo, Japan, where he is now a Professor at the Graduate School of Bio-Applications and Systems Engineering. During 1983–1984 he was a visiting Associate Professor at the School of Electrical Engineering, Georgia Institute of Technology, Atlanta. His research activities include areas of speech processing, image processing, and the applications of digital signal processing. He is the author of books on speech recognition, digital signal processing, mathematical morphology, etc. He is the member of IEEE, SICE, etc.