EFFECT OF FREQUENCY OVERSAMPLING AND CASCADE INITIALIZATION ON PERMUTATION CONTROL IN FREQUENCY DOMAIN BSS

Christopher Osterwise, Steven L. Grant[†]

Missouri University of Science and Technology
Rolla, MO USA

ABSTRACT

This paper presents a new algorithm for addressing the permutation ambiguity in convolutive blind source separation. The proposed algorithm seeks to prevent permutations by frequency oversampling, and then exploiting the induced correlation between bins. Any remaining permutation is then corrected by beam pattern estimation. Cascade initialization is shown to improve system performance while decreasing processing time, while frequency oversampling is shown to increase performance with slight increases in computational time. The algorithm is proven robust against isotropic noise, producing a SIR improvement of 12.9 dB when the reverberation time is 260 ms, and the signals at the input are only 5 dB above the noise floor.[*]

Index Terms— Blind source separation (BSS), convolutive mixture, frequency-domain ICA, permutation problem

1. INTRODUCTION

Blind Source Separation (BSS) has been a topic of considerable interest for many years. The goal of the technique is to reconstruct individual sources from observed mixtures of different sources, without a priori knowledge of the individual sources or the mixing system. The general mixing model for the Blind Source Separation problem is given as:

$$\mathbf{x}(n) = \mathbf{A} \bullet \mathbf{s}(n) + \boldsymbol{\xi}(n) \tag{1}$$

The solution, or separation equation, to this, is a multiplication with matrix **W**:

$$\mathbf{y}(n) = \mathbf{W} \bullet \mathbf{x}(n) \tag{2}$$

where $\mathbf{s}(n) = [s_1(n) \quad \dots \quad s_N(n)]^T$ is the source vector of N sources at time n (unknown to the algorithm), $\mathbf{x}(n) = [x_1(n) \quad \dots \quad x_M(n)]^T$ is the observed vector of M simultaneous samples, $\boldsymbol{\xi}(n) = [\xi_1(n) \quad \dots \quad \xi_N(n)]^T$ is the isotropic noise at the microphones, and $\mathbf{y}(n) = [y_1(n) \quad \dots \quad y_N(n)]^T$ is the separated output signals. For well-defined systems, it must be that $M \geq N$. For instantaneous mixtures, **A** and **W** are matrices of size MxN and NxM, respectively, and the operator $\bullet$ indicates matrix multiplication. When dealing with convolutive mixtures, the same notation is broadened to describe the new conditions. Each element of the **A** and **W** matrices is an FIR filter, and the symbol $\bullet$ instead indicates convolution of each element with each signal, followed by summation of the rows.

A more complete and detailed introduction to the BSS problem can be found in Mazur and Mertins [1] and in Parra & Spence [2].

Inherent to the BSS problem is the permutation ambiguity. Since the input is a mixture of N sources with an unknown mixing matrix, the order of inputs cannot be known with absolute certainty. That is, if input s_1 and s_N were interchanged, and row 1 and row N of **A** likewise swapped, the observed signal would be identical. For convolutive mixtures handled in the frequency domain, this becomes a great concern: to ensure separation, if the component of one signal at frequency bin ω_a arrives at output k, it must be ensured that the corresponding component of the *same* signal at bin ω_b arrives at output k. If this is not met, the signals (or portions thereof) remain mixed.

The use of ICA introduces the scaling ambiguity. It cannot be known for sure that the output is of the same magnitude as its corresponding source; it could be scaled by a constant. Using a frequency-domain BSS algorithm, the effect of this uncertainty is more noticeable than in the time-domain. Assuming perfect separation, and that the permutation ambiguity is resolved, each output signal may be scaled by an arbitrary constant in each frequency bin. The resulting effect is that an additional filter is applied to the output signal. The usual consequence of this is increased reverberation, which in the case of speech signals reduces intelligibility. Fortunately, a number of different algorithms have been shown to adequately correct the scaling ambiguity. These include filter shortening [3], shaping [4], or the popular Minimal Distortion Principle [5]. This last algorithm removes any additional filtering on the signal produced by the demixing algorithm, leaving only what is caused by the mixing environment. This is accomplished by applying postfilters on the separated signals while still in the time-frequency domain. The postfilters take the form of the diagonal matrix $\boldsymbol{\Gamma}^{-1}$, calculated in [5] as:

$$\boldsymbol{\Gamma}_{\omega}^{-1} = \mathrm{diag}\{\mathbf{W}_{\omega}^{-1}\} \tag{3}$$

where $\mathbf{W}_{\omega}$ is the separation matrix at frequency bin ω.

To correct the permutation ambiguity is a task far more vital for the success of the separation technique. It has been the primary focus of a great number of BSS papers.

Permutation alignment algorithms are usually divided into two categories. The first are approaches based on the geometry of the microphones relative to the sources. These utilize methods such as time difference of arrival (TDOA) or direction of arrival (DOA) [6], [7], [8]. The permutation of each frequency is calculated on its own merit, independently. These algorithms usually present a strict requirement on the geometry of the system, such as requiring an array of microphones with constraints on spacing.

[*] This work was partially funded under the Wilkens Missouri Endowment.

[†] Formerly Steven L. Gay

Algorithms of the other classification rely on inter-frequency correlation to align permutations. This approach originated with Parra and Spence [2], who demonstrated that demixing filters with permutations would be longer in time domain, and thus less smooth in frequency. They used a projection operator calculated from the time domain to maximize the smoothness of the frequency response, which in turn controlled permutation. These techniques are usually more accurate when the signals are long enough to correlate frequency bins. However, these suffer from instability in the manner that one error in ordering can propagate through several neighboring frequencies. The result is a partially separated set of signals, where one frequency range will have one permutation, while another will be ordered differently.

Recent techniques have attempted to bridge the gap between these approaches. Mazur and Mertins [1] classify frequencies into groups known to have the same permutation, and then align the groups with each other. In [9] each frequency bin is aligned utilizing a target amplitude envelope (based on surrounding frequencies) and TDOA estimation.

Like recent algorithms, the proposed method attempts to utilize both approaches to align frequency bins. All of these algorithms, however, focus on correcting permutation errors after separation. This paper will examine a method to *prevent* permutation during the ICA stage via inter-frequency correlation, then correct the remaining errors using directivity estimation.

1.1. Interpolation in Frequency Bins

When a signal is converted to the frequency domain using an N-point FFT, where N is at least twice as large as the length of the signal, it produces a representation that is smoothed in the frequency domain. This smoothing occurs due to points inserted in the frequency response that are interpolations of the surrounding data.

When a signal is taken from the time domain to the time-frequency domain via a series of short-term Fourier transforms (STFT), and this frequency interpolation is exercised, it creates a series of subbands with overlapping frequencies. Any cross-section of this data taken at a specific point in time will appear oversampled in the frequency domain. This paper defines the value K as the frequency oversampling factor, obtained via the ratio of the FFT size to the length of the data window L. Since each of these new frequency bins are interpolations of the surrounding bins, each bin is now correlated to its neighbors.

Figure 1 shows the correlation between these subbands. An eight-second long speech signal, s(t), was converted into time-frequency domain representation, $S_K(\omega, t)$ using a series of STFTs with a Hamm window. For demonstration, the data window length was kept to $L = 8$, and the value of K varied from 1 to 8. Frequency bins after $K(1+L/2)$ were discarded. The cross-correlation matrix R_{SS} was computed from the remaining bins. The images were created by mapping the magnitude of the auto-correlation matrix to pixel brightness: a lighter shade indicates a higher correlation. For the first two images, the frequency bin for each column is labeled.

In the first image ($K = 1$), there is no oversampling in the frequency domain. Since the source file is a speech signal, there is already some inter-frequency dependence. Frequency bins 0 and 1 are correlated, as is bin 3 with both 2 and 4. However, there is effectively no correlation between bins 1 and 2.

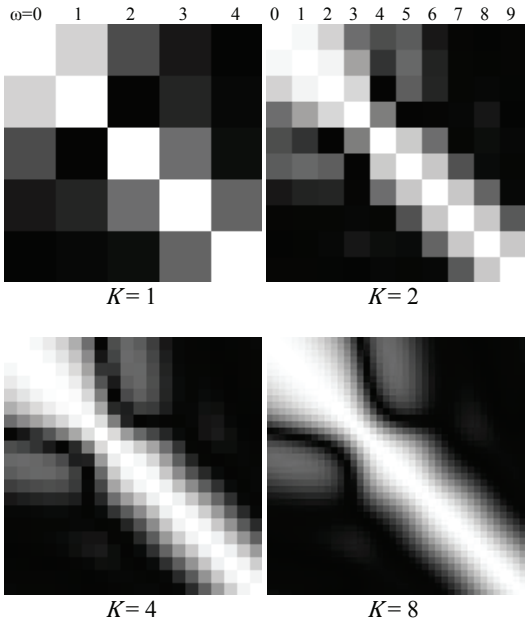

Figure 1: Magnitude of R_{SS} with increasing overlap (K). White squares are of high magnitude.

As the frequency oversampling K increases, new bins are created which interpolate between the surrounding frequencies. At $K = 2$, the data in bins 2 and 4 contain the data that was initially in bins 1 and 2 for $K = 1$.

As K increases, each frequency bin becomes increasingly correlated with its neighbors. At $K = 1$, bins 1 and 2 are statistically uncorrelated. However, at $K = 2$, bin 2 is correlated with 3, which in turn is correlated with 4. It is this statistical correlation between adjacent frequency bins that will be exploited to reduce the occurrence of permutations during separation.

2. PROPOSED ALGORITHM

The proposed BSS algorithm, Cascaded ICA with Intervention Alignment (CICAIA), makes use of the overlap in the frequency domain by first preventing, and then correcting, as many permutation errors as possible in the initial calculation of the separation matrices. Any unresolved permutations can then be corrected by a block depermutation step. The algorithm may be summarized by the following steps:

1) Convert the incoming signals into their Time-Frequency domain representations, using a short-time Fourier transform with frequency overlap.
2) In one sweep across the frequencies,
   a) For each bin after the first, initialize the ICA algorithm with the result from the previous frequency bin.
   b) Perform ICA to separate the signal.
   c) Check for possible permutations.
   d) If permutation is confirmed, attempt to correct it.
3) If needed, correct permutations in block over the entire array of frequencies.
4) Drop unnecessary frequencies.
5) Rescale remaining frequencies.
6) Reconstruct signal.

The full description of the algorithm follows:

2.1. Conversion

To begin, all input signals are converted into the time-frequency domain by means of the STFT. Data segments of length L are Hamm windowed, then converted into their frequency domain representation using a KL-point FFT. This leads to an oversampling in the frequency domain by a factor of K. As the source signals are real, the Fourier transform is a conjugate-symmetric function; thus it is sufficient to keep only the first $\frac{1}{2}KL + 1$ of these bands for perfect reconstruction. During processing, only frequencies from 128 to 3872 Hz were considered; frequencies outside this range were set to zero.

2.2. ICA with Cascaded Initialization

In this phase, a two-stage permutation control unit designed by Peng Xie [10] operates sequentially on the individual bins. In each frequency, the separation matrix is obtained by ICA, and is then checked for permutation against that of the previous subband. If a permutation is suspected, the order of outputs is determined relative to the previous subband, and corrected if necessary. The permutation-corrected separation matrix is then used as the initialization for the ICA algorithm of the next subband.

The proposed method utilizes the fixed-point ICA algorithm proposed by Hyvärinen, with a modification made to determine the complete demixing matrix ($\mathbf{W}_\omega$) at each update (rather than determining $\mathbf{W}_\omega$ one column at a time). The matrix is declared to be converged when it stops changing "significantly" after each iteration—that is, when the magnitude of the update vector drops below a threshold. When this occurs, the demixing matrix can be compared to the previous frequency bin to determine permutation.

To determine if the demixing matrix has been permuted, a simple, quick-to-compute metric is necessary. As Section 1.1 explains, there is some cross-correlation between most neighboring frequency bins of a real-world signal. Because of this correlation, the coefficients of separation matrices in adjacent frequency bins will not change significantly. The magnitude of distance between these coefficients thus provides a simple, quick, and efficient criterion to determine whether a permutation exists. This is defined as:

$$D(\omega) = \sum_{i,j} \left| W_{i,j;\omega} - U_{i,j;\omega-1} \right| \qquad (4)$$

where $D(\omega)$ is the distance, $W_{i,j;\ \omega}$ is the separation matrix at frequency ω, and $U_{i,j;\ \omega-1}$ is the permutation-corrected separation matrix of the previous frequency bin. A threshold ε is chosen: if $D(\omega)$ is above ε, there is a suspected permutation between ω-1 and ω, and an intervention depermutation step is taken. For the simulations shown below, the threshold was selected to be $\varepsilon = 0.2$. In the scenario of only two sources, the permutation can be removed by the directivity technique [8], [11]. This technique determines the directionality of the received signals based on the estimated demixing matrix, and re-orders the outputs based on this estimate. For the case of three or more sources, this technique becomes less reliable. Instead, the outputs should be sorted via correlation with harmonics or adjacent bins known to have good separation [12], or via a TDOA approach [6], [7]. This intervention step is not performed on every frequency, only those that are suspected of permutation. As a result, the algorithm selected for depermutation in this fashion can be more computationally complex, without significantly delaying the completion time of the entire system. Moreover, while a greater number of frequency bins are present in CICAIA, it is not inherently more computationally complex than other algorithms.

If the new order of outputs on a suspected permutation is different from the previous arrangement, the bin is labeled as an *actual* permutation. The distinction is important. Bins suspected of permutation require the intervention step—a process which slows down processing. Actual permutations will degrade performance if not detected. Therefore, minimizing actual permutations is essential; minimizing suspected permutations without losing actual permutations, is additionally beneficial.

Even without frequency overlap, the cascaded ICA initialization approach is effective at reducing the number of suspected and actual permutations in the BSS solution. Initializing the ICA algorithm in each frequency with the separation matrix from the previous bin significantly increases the chance the converged output will be in the same arrangement for most bins. This is then immediately tested, and corrected as necessary. Instead of applying a blanket permutation alignment algorithm to the output of all frequency bands, corrective efforts are only necessary at the frequencies where such a permutation is suspected. With the increased correlation between frequency bands created by oversampling in the frequency domain, the solution improves even further. As K increases, the correctly-ordered separation matrices between adjacent bands will become increasingly similar, making a permutation easier to detect.

3. SIMULATIONS

Multiple aspects of the proposed algorithm have been tested using simulated mixtures. All source signals are eight-second long samples of speech, sampled at 8 kHz. The impulse response from each source location to the microphone array is generated using the image model, using different values for room reverberation time. No mixing filter exceeded 3000 samples in length. Additional white Gaussian noise is applied to the pickup of each microphone when the mixed signals are generated.

To test the robustness of the algorithm against noise, its performance was evaluated on simulated data along with several other contemporary approaches. The simulated room was 6 x 5 x 3 meters, with a reverberation time of T_{60}= 260 ms. Two microphones were placed at the center of the room, 16 cm apart from each other. Two sources were placed one meter away from the microphones, at 45° and 135°. The isotropic noise at the microphones was increased with each simulation, starting at 50 dB below the mixture strength.

Figure 2 compares the improvement in SIR from the proposed (CICAIA) algorithm against a nonblind permutation alignment (as a theoretical maximum), and against the algorithms of Pham, Servière, and Boumaraf [13]; Rahbar & Reilly [14]; Parra and Spence [2]; and Trinicon [15]. All algorithms were run with a frame length of 1024 samples, with 75% overlap in time. CICAIA use a frequency oversampling (K) factor of 8.

At low noise levels (≥ 30 dB SNR), Pham performed the best. Rahbar & Reilly [14] produced a similar quality of result. Comparatively, CICAIA performed the second-worst, but only one dB below Trinicon. At higher noise levels, the performance of most algorithms began to taper off; however, the CICAIA algorithm did not drop in performance until 5 dB. Indeed, at very low SNR, CICAIA performed almost as well as the ideal case.

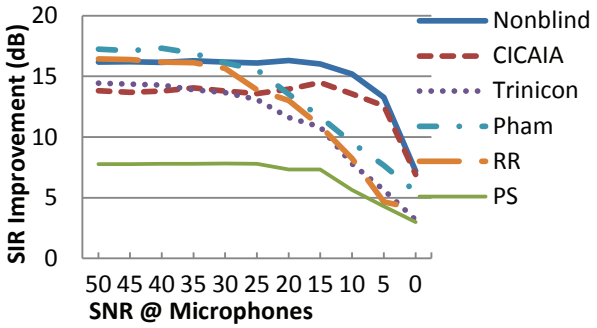

Figure 2: Performance of proposed algorithm (CICAIA) in increasing noise in comparison with other methods (Pham [13], Rahbar & Reilly [14], Parra & Spence [2], and Trinicon [15])

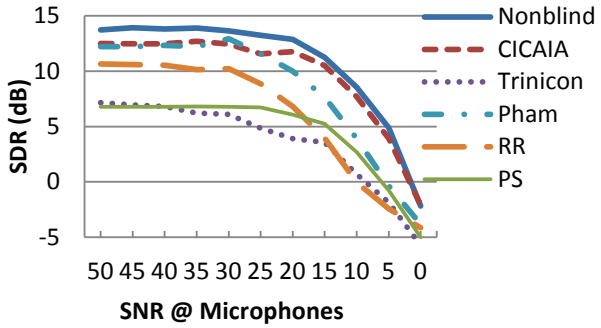

Figure 3: Separation Quality of algorithms

Figure 3 shows the SDR results from the above simulation. Under low noise conditions, CICAIA and Pham produce the best quality output (highest SDR). As the noise level increases, Pham's quality declines faster.

4. CONCLUSION

In comparison with block depermutation algorithms, progressive depermutation allows corrective efforts to focus on frequency bins where a problem is suspected. This allows the use of more computationally-intensive depermutation techniques without significantly slowing down the process as a whole. The ICA chain also means that a successful result in one frequency bin effects good separation in successive bins. Progressive depermutation is therefore a feasible replacement, or at the least supplement, for block depermutation.

In comparing the performance of different algorithms under varying SNR environments, we found that CICAIA performed the best at low SNRs (< 20 dB). In low noise environments, CICAIA produces only moderate separation. In high noise environments, CICAIA has the best separation. Under any condition, it produces output with the best Signal-to-Distortion ratio.

5. REFERENCES

[1] R. Mazur and A. Mertins, "An approach for solving the permutation problem of convolutive blind source separation based on statistical signal models," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 17, no. 1, pp 117-126.

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

[3] R. Mazur and A. Mertins, "Using the scaling ambiguity for filter shortening in convolutive blind source separation," in *Proc. IEEE Int. Conf. for Acoust., Speech, and Signal Processing*, Taipei, Taiwan, April 2009, pp. 1709-1712

[4] R. Mazur and A. Mertins, "A method for filter shaping in convolutive blind source separation," in *Independent Component Analysis and Signal Separation (ICA2009)*, ser. LNCS, vol. 5441. Springer, 2009, pp. 282-289

[5] K. Matsuoka: "Minimal distortion principle for blind source separation," in *Proceedings of the 41st SICE Annual Conference*, vol.4, pp. 2138–2143 (August 5-7, 2002)

[6] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Auido, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592-1604, July 2007

[7] F. Nesta, M. Omologo, and P. Svaizer, "A novel robust solution to the permutation problem based on a joint multiple TDOA estimation," *International Workshop for Acoustic Echo and Noise Control,* Sept. 2008.

[8] 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. IEEE Int. Conf. for Acoust., Speech, and Signal Processing*, June 2000, pp. 3140-3143.

[9] F. Nesta, T. Wada, B.H. Juang, "Coherent Spectral Estimation for a Robust Solution of the Permutation Problem," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 105-108

[10] P. Xie, S. Grant, "Frequency-domain blind source separation with permutation control," *International Workshop for Acoustic Echo and Noise Control (IWAENC 2008)*, September 2008.

[11] R. Aichner, S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Time domain blind source separation of non-stationary convolved signals by utilizing geometric beamforming," *12th IEEE Workshop on Neural Networks for Signal Processing, 2002,* pp. 445- 454.

[12] 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," *IEEE Int. Sym. on Independent Component Analysis and Blind Signal Separation*, April 2003, pp. 505-510

[13] D.T. Pham, C. Servière, and H. Boumaraf, "Blind separation of speech mixtures basxed on nonstationarity," *Proc. Sym. on Signal Processing and Its Applications*, 2003, pp 73-76.

[14] K. Rahbar and J. Reilly, "Frequency Domain Method for Blind Source Separation of Convolutive Audio Mixtures," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp 832-844, Sept. 2005.

[15] H. Buchner, R. Aichner, and W. Kellermann, "Trinicon: A versatile framework for multichannel blind signal processing," *Proc. IEEE Int. Conf. for Acoust., Speech, and Signal Processing*, May 2004.