A New Permutation Control Method for Frequency Domain BSS[1]

Christopher Osterwise and Steven L. Grant[2]
Missouri Science & Technology, Rolla, MO, USA
Email: ctoyv6@mst.edu
Email: sgrant@mst.edu

Abstract—This paper introduces a new frequency domain blind source separation algorithm: Inter-frequency Correlation with Microphone Diversity (ICMD). Here, we consider using different sets of microphones where in each set the number of microphones and sources are equal. In the frequency domain, cascaded ICA initialization (CII) is used, where the separation matrix of one bin is used to initialize the ICA iterations of the next. CII greatly reduces the number of permutation changes in successive bins. However, for a given microphone set, it is not uncommon that ICA will fail to separate some bins, thus defeating CII. This problem is addressed as follows. 1) In addition to CII the inter-frequency correlation matrix of the separated signals is used to align permutations in successive frequency bins. 2) The condition number of this matrix is monitored to determine if separation has failed for the current bin and microphone set. 3) If so, an alternate set with better separation is selected and again, inter-frequency correlation is used to align the permutations of the new set of microphones with the old. Results show a marked improvement in separation when there are three or more sources.

I. INTRODUCTION

The task of Blind Source Separation (BSS) is to separate multiple independent sources in an unknown mixing environment. The general mixing model for the system is given in matrix convolution form as:

$$\mathbf{x}(t) = \mathbf{A}(t) * \mathbf{s}(t) + \xi(t) \tag{1}$$

The model for the solution is similar in form:

$$\mathbf{y}(t) = \mathbf{W}(t) * \mathbf{x}(t) \tag{2}$$

where $\mathbf{s}(t) = \begin{bmatrix} s_1(t) & \cdots & s_M(t) \end{bmatrix}^{\mathrm{T}}$ is the source signal of M sources, $\mathbf{x}(t) = \begin{bmatrix} x_1(t) & \cdots & x_N(t) \end{bmatrix}^{\mathrm{T}}$ is the observed signal from N microphones, $\mathbf{y}(t) = \begin{bmatrix} y_1(t) & \cdots & y_M(t) \end{bmatrix}^{\mathrm{T}}$ is the output of reconstructed sources, $\xi(t) = \begin{bmatrix} \xi_1(t) & \cdots & \xi_N(t) \end{bmatrix}$ is the isotropic noise at the microphones, $\mathbf{A}(t)$ is the $N \times M$ mixing matrix, and $\mathbf{W}(t)$ is the $M \times N$ demixing matrix. Each element of $\mathbf{A}$ and $\mathbf{W}$ is an FIR filter of length L.

With a few notable exceptions like Trinicon [1], a majority of BSS algorithms operate in the frequency domain. Using the short-time Fourier transform, this converts the task of finding L-length filters in the time domain, to the more straightforward task of finding L independent demixing matrices $\mathbf{W}_\omega$ of M^2 nonzero elements each. This technique can easily be accomplished by applying ICA to each frequency bin. The inherent flaw with this approach is the ambiguity of output from ICA. 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.

Permutations are often aligned by exploiting known properties of system geometry and inter-frequency correlation. Some algorithms ([2], [3], [4], [5]) use the microphone geometry to extract direction of arrival (DOA) or time-difference of arrival (TDOA), and align outputs based on these values. These methods work well even in noise, and generally are not restricted by the properties of the inputs. They are also robust against error contamination: one frequency bin that cannot be properly ordered will not negatively impact other frequencies. They do, however, pose restrictions on the geometry of the system, such as requiring a microphone array with minimum dimensions and known spacing, or sources with sufficient spatial distribution.

Other methods ([6], [7], [8]) use inter-frequency correlation—the similarity of different frequencies from a common source—to align the order of the outputs. This principle can be extended further, such as correlating frequency bins with external patterns like lip movement [9]. These techniques generally produce good results regardless of system geometry, but are often more susceptible to noise. In addition, because of how the correlation is usually measured, inputs must be long enough to support the statistical independence of the signals' envelopes.

Several recent methods ([10], [11]) exploit both properties to align permutations. They combine the benefits of both methods, but share the system geometry restrictions of the first category as well as the signal length requirements of the second.

All algorithms that rely on inter-frequency correlation, in full or in part, must address the issue that a failed separation or improper output order of one bin can adversely affect the alignment of other bins. In most algorithms, the performance loss is caused by erosion to the basis of comparison. This leads to ambiguity in the cost function for permutation correction, resulting in more misalignment. In CICAIA [10] the effect was obvious: a permutation error in one bin could immediately cascade to the remaining frequencies. Fortunately, CICAIA's beamforming-based Intervention Alignment detected and cor-

[1] This work was partly performed under the Wilkens Missouri Endowment
[2] Formerly Steven L. Gay

rected these events. Pham, Servière, and Boumaraf [6] loop through the entire frequency range several times, with the intent that as long as the majority of bins are ordered correctly, the remaining will align with sufficient iteration. In Mazur and Mertins's approach, permutation errors lead to clusters of frequencies with a common alignment. Most of their algorithms ([12], [13]) revolve around marking the bins where separation is expected to have failed, and returning later to align the groups. In a similar fashion, the proposed algorithm detects when separation has failed; however, when this occurs, the problem is immediately rectified, by activating microphone diversity.

II. DIVERSITY

Diversity combining is a proven technique used to improve the reliability of a wireless connection. It is the concept of using more antennas than signals, and extracting the desired signal as needed. This can either be switching between antennas when the current one fails, actively polling all antennas to see which has the best signal, or some more complex form of constructing an output from the available measurements. In simplest form, multiple antennas are distributed so that if a transceiver loses sight of one, it can connect with another. This is the basic paradigm behind mobile phones. Modern cellular towers take this a step further: every facing of a tower has two antennas, one of which is only for receiving, the other both transmits and receives as needed. This setup is specifically intended to counter the effects of multipath interference, and so is copied frequently. Long before the Wireless-N protocol came out, wireless internet routers had two (or more) antennas one wavelength apart. If the signal in one antenna got too low, the base station would switch to the other to communicate with the device.

Microphone diversity is a direct extension of the technique to the audible frequencies. It is not, however, as noticeable as antenna diversity. Most conference-room speakerphones use multiple microphones, and select among them the audio with the best sound from the active speaker. Freudenberger [14] recently proposed a setup using microphone diversity to be placed in a car, to create a hands-free device for cell phones with low environmental noise pickup. In [14] the signals picked up by individual microphones are combined in the frequency domain according to estimated noise levels. The result is an output signal with better SNR than what could be achieved by single-microphone noise suppression, in a physical setup that is more flexible than beamforming.

In utilizing diversity, the proposed algorithm for BSS applies to the over-determined case—that is, extracting M sources from N microphones, where $N > M$. This paradigm is not frequently addressed in BSS literature, but is nonetheless common in application. Consider the following: a majority of BSS algorithms assume equal number of sensors as sources. In practical applications, to provide for separation of up to N sources, any hardware realization of the algorithm must include inputs from N microphones. However, rarely will the actual number of sources get that high. For most algorithms, the additional microphones will just be disabled for the remainder of the time, discarding their inputs. The proposed, on the other hand, takes advantage of their presence.

III. METHODS

The Inter-frequency Correlation with Microphone Diversity algorithm (ICMD) uses cascaded initialization to inhibit permutations, aligns permutations using energy profiles, and supplements this with direct cross-correlation to ensure the current bin is valid. If separation is suspect, it takes advantage of microphone diversity to repair the error.

As with all algorithms, we begin by converting the inputs from all microphones into the frequency domain. Assuming that the number of sources M is known, we take as many microphones to be the first set, and treat them as though handling the determined case. The algorithm then performs the following steps at each frequency bin ω:

1) Extract $\mathbf{y}_\omega(t)$ via Cascade-Initialized ICA
2) Calculate the cross-correlation matrix $\mathbf{R}_{yy}$
3) If $\mathbf{R}_{yy}$ is ill-conditioned:
   a) Sample several sets of microphones to find $\mathbf{y}_\omega(t)$
   b) Choose the set with the lowest condition value of $\mathbf{R}_{yy}$
4) Align the permutation, according to energy profiles
5) Update the history for the next frequency

A. Cross-correlation

The cross-correlation matrix is constructed as per equation (3):

$$v_\omega(k,t) = y_\omega(k,t) / \sqrt{\mathrm{E}\{v_\omega^2(k,t)\}}$$
$$\mathbf{R}_{yy} = \mathrm{E}\{|\mathbf{v}_{\omega-1}(t)||\mathbf{v}_\omega^{\mathrm{T}}(t)|\} \quad (3)$$

where $y_\omega(k,t)$ is the k[th] output at frequency ω over time, $v_\omega(k,t)$ is its unit-energy normalization, $\mathbf{v}_\omega(t)$ is the column vector of $v_\omega(k,t)$ for all k, and expectations are taken over time. The creation of $\mathbf{R}_{yy}$ depends on the length of input used in the correlation. In general, more samples give better performance, but require a longer input.

Mazur and Mertin's α-algorithm is "based on the observation, that false alignments usually happen at positions where separation performance is poor" [12]. In these circumstances, the outputs at one frequency will resemble more than one output at the previous bin—there will be no simple one-to-one correlation and thus no mapping can be made. Therefore, we monitor the condition number of the cross-correlation matrix. When this number rises above a threshold, the microphone diversity routine is triggered. For the experiments below, the threshold was chosen as 10. During this routine, several potential microphone sets are polled, and the set that would produce the lowest cross-correlation matrix is chosen. That set is then used until the next time the routine is triggered.

Ideally, we would want to test all potential collections of microphones; however, this number increases rapidly as the number of available microphones rises. Fortunately, it is usually sufficient to sample a small number of sets—say, 10—and choose the best set from that sample. It is not necessary to find the "best" set of microphones for the given frequency.

B. Profiles

Energy profiles are defined in [6] based on the concept that frequency components of sources will have amplitude over time similar to the envelope of the signal itself, but with a frequency-dependent scaling factor. Here, they are calculated as follows

$$\mathbf{e}_\omega(t) = \log\left(\left|\mathbf{y}_\omega(t)\right|\right) \tag{4}$$

$$\mathbf{e}'_\omega(t) = \mathbf{e}_\omega(t) - \mathrm{E}\left\{\mathbf{e}_\omega(t)\right\} \tag{5}$$

where $\mathbf{e}_\omega(k,t)$ is the energy profile at bin ω for output k, and $\mathbf{e}'_\omega(k,t)$ is the centered energy profile, which removes the scaling factor. The proper order of outputs is achieved by finding the permutation at each frequency bin that minimizes the summed magnitude of the difference vectors created by subtracting the profiles of the sources from the energy profiles at the outputs, as per (6):

$$\arg\min_{\Pi} \sum_k \left\|\mathbf{e}'_\omega(t) - \Pi\mathbf{e}'_\circ(t)\right\|_2 \tag{6}$$

Ideally, we would want to use the average centered profiles of our sources for our target profiles, $\mathbf{e}'_\circ(t)$, but this would violate the blind nature of the system. In [6], the target profiles are constructed by a continually-updating average over all frequencies. This average is improved as alignment is performed over the entire frequency range in successive iterations. In ICMD, since additional steps are taken to ensure separation in all bins, it is sufficient to create $\mathbf{e}'_\circ(t)$ from a moving average of the previous R frequencies. For a system running at a sampling frequency of 8000 Hz, 100 is sufficient for R.

It would be possible to obviate the energy profiles, and align permutations based solely on $\mathbf{R}_{yy}$ from (3). Doing so, however, would make each frequency bin dependent only on the previous, and would require sufficient excitation in all frequencies in all sources. If two sources had a common null frequency, where microphone diversity cannot repair the error, the system would be irrecoverable. Using profiles allows for a basis of comparison that is more robust to local perturbations.

IV. Simulations

The performance of ICMD was tested in a simulated mixture of multiple sources. Impulse responses from a simulated room 6 m by 9 m by 2.5 m in dimension (roughly the size of a classroom), with a reflectivity constant of 0.810, were constructed using the image model. The resulting RIRs had a reverberation factor of T_{60} = 440 ms, and a sparseness factor of 0.450. Eight-second-long clips of speech, sampled at 8 kHz, formed the sources. Five of these were scattered throughout the room, and nine microphones were randomly placed near the center of the room, far away from the sources. Isotropic noise was added to the microphone pickups during simulation, varying the input SNR from 40 to 5 dB. The eight-second inputs translated to 127 samples in each frequency bin, all of which were used to create $\mathbf{R}_{yy}$. Considering its performance in [10], and the ease of expanding the implementation to more than two sources, the algorithms of Pham et al. and Rahbar and Reilly [7] were tested for benchmark.

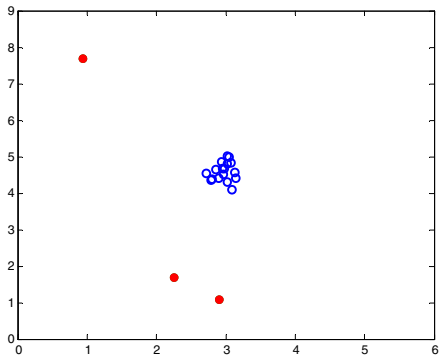

Figure 1. Position of 3 sources (.) and 9 microphones (o) in simulated room.

The algorithms were first tested in simple conditions: only two sources with low background noise. All three algorithms performed admirably: 23, 18, and 16 dB of signal-to-interference ratio (SIR) improvement for ICMD, Pham, and R&R, respectively.

A third source was added, and the behavior changed dramatically. The initial performance of Pham went from 18 dB to only 8. This did bring an improvement in its noise resilience, however, as the performance did not degrade until the input noise reached 15 dB below the signal. The performance of Rahbar and Reilly decreased dramatically to just around 2 dB of separation. ICMD produced 14 dB of SIR improvement in good conditions, but started degrading at only 20 dB of SNR. At 10 dB SNR, the performance dropped to only 10 dB SIRI, approximately the performance of Pham in good conditions. The placement of sources and microphones is shown in Figure 1. The resulting performance curve is shown in Figure 2 and Figure 3Figure 3. Quality of output for previous figure.. The proposed manages 14 dB of separation improvement in high SNR, and maintains this performance until 20 dB SNR.

The experiment was repeated with five sources, to view the flexibility of the algorithm. Its performance did not change significantly: 14 dB of SIR improvement in good conditions, diminishing gradually with background noise to roughly 5 dB SIRI at 0 dB SNR.

V. Conclusion

The proposed algorithm, Inter-frequency Correlation with Microphone Diversity, shows the capability to extract an arbitrary number of sources in moderate noise, with 14 dB of SIR improvement on a mixture of five sources in good conditions. It does not hold up as well as our previous work CICAIA in harsh noise, but has far better performance with increasing

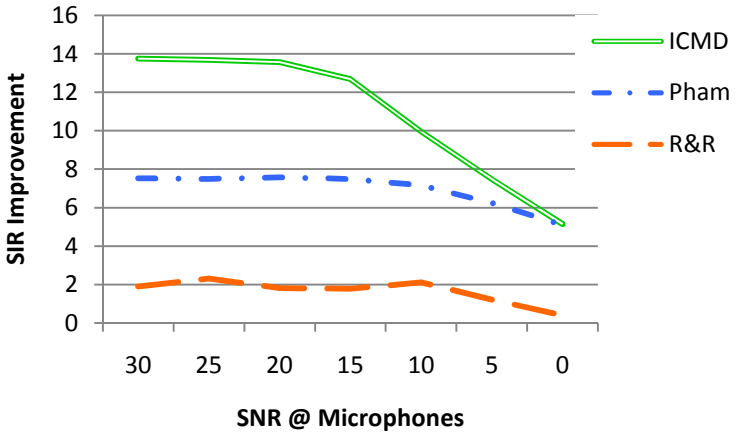

Figure 2. ICMD (proposed) vs. Pham, and Rahbar & Reilly in noise. Average performance of 5 trials is shown.

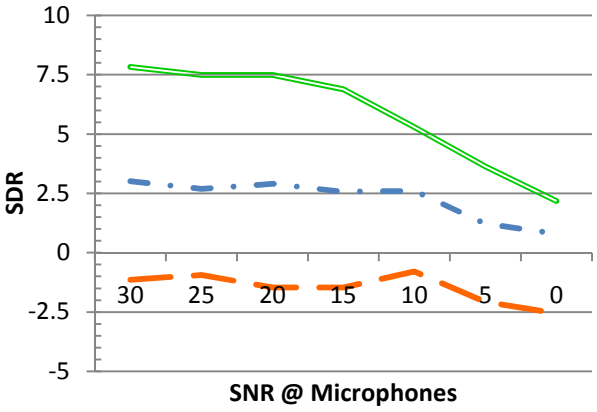

Figure 3. Quality of output for previous figure.

number of sources, and no notable restrictions on source or microphone placement.

The algorithm can automatically select microphones as appropriate, and does not need any foreknowledge of the placement of these microphones. Future work could easily expand this technique to handle moving sources in a distributed-microphone room.

REFERENCES

[1] 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

[2] 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

[3] 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.

[4] 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.

[5] T. Ono, S. Miyabe, N. Ono, S. Sagayama, "Blind source separation with distributed microphone pairs using permutation correction by intra-pair TDOA clustering," *International Workshop for Acoustic Echo and Noise Control*, Aug. 2010, Tel Aviv, Israel

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

[7] 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.

[8] 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.

[9] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Trans. on Audio, Speech, and Language Processing*, vol 15, no 1, pp. 96-108, January 2007.

[10] C. Osterwise, S. Grant, "Effect of frequency oversampling and cascade initialization on permutation control in frequency domain BSS," *Proc. IEEE Int. Conf. for Acoust., Speech, and Signal Processing*, 2012, pp 285-288

[11] 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," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 5, Sept. 2004, pp. 530-538.

[12] R. Mazur and A. Mertins, "Improving the robustness of the correlation approach for solving the permutation problem in the convolutive blind source separation," *International Workshop on Acoustic Echo and Noise Control*, Aug 2010, Tel Aviv, Israel

[13] 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, Jan. 2009, pp 117-126.

[14] J. Freudenberger, "Spectral combining for microphone diversity systems," *17th European Signal Processing Conference (EUSIPCO)*, August 2009, Glasgow, Scotland