A MULTIRESOLUTION APPROACH TO BLIND SEPARATION OF SPEECH SIGNALS IN A REVERBERANT ENVIRONMENT

Muhammad Z. Ikram

Center for Signal and Image Processing
Georgia Institute of Technology
Atlanta, GA 30332-0250
mzi@ece.gatech.edu

Dennis R. Morgan

Bell Laboratories, Lucent Technologies
700 Mountain Avenue 2D-537
Murray Hill, NJ 07974-0636
drrm@bell-labs.com

ABSTRACT

The performance of existing blind speech separation methods is limited in a realistic reverberant environment, where a need for long un-mixing filters is imperative. We first show how these methods suffer while trying to balance the competing objectives of frequency-domain permutation alignment and spectral resolution. We then propose a multistage multiresolution algorithm, which aligns the un-mixing filter permutations over the whole frequency band without sacrificing spectral resolution. We perform experiments in both real and simulated reverberant environments, and obtain improved separation results that are comparable to the ideal benchmark obtained by aligning the permutations using prior knowledge of the mixing filters.

1. INTRODUCTION

In many applications such as hand-free telephony, teleconferencing, and speech recognition, one is interested in separating independent speech signals using multiple microphones in a reverberant environment [1]. This task is accomplished using blind source separation (BSS), where the term *blind* refers to the fact that very little is known about the source signals or the way they are mixed together.

In this paper, we consider an acoustic scenario, where two microphones receive multiple filtered copies of two statistically independent speech signals $s_i(n)$, $i = 1, 2$. Mathematically, the received signals can be expressed as a convolution, i.e.,

$$x_j(n) = \sum_{i=1}^{2} \sum_{p=0}^{P-1} h_{ji}(p) s_i(n-p), \quad j = 1, 2 \tag{1}$$

where $h_{ji}(p)$ models the P-point impulse response from source i to microphone j. In a more compact matrix-vector notation, (1) can be stated as

$$\mathbf{x}(n) = \mathbf{H}(n) * \mathbf{s}(n), \tag{2}$$

where $\mathbf{x}(n) = [x_1(n) \;\; x_2(n)]^T$ is the received signal vector, $\mathbf{H}(n)$ is the (2 × 2) mixing filter matrix, $*$ is the convolution operator, and $\mathbf{s}(n) = [s_1(n) \;\; s_2(n)]^T$ is the source signal vector. The aim of the BSS method is to find a (2 × 2) un-mixing filter $\mathbf{W}(n)$ of length Q that separates the two sources up to an arbitrary filter and permutation [2], i.e.,

$$\widehat{\mathbf{s}}(n) = \mathbf{W}(n) * \mathbf{x}(n), \tag{3}$$

where $\widehat{(\cdot)}$ denotes an estimate.

In [3], we applied a frequency-domain method [4] that uses non-stationary second-order statistics to investigate the BSS problem. Throughout this paper, we will refer to this general technique as the frequency-domain, second-order statistics (FDSOS) method. Frequency-domain processing is motivated by the transformation of a computationally complex convolutive BSS problem in the time domain to multiple easier-to-solve instantaneous BSS problems in the frequency domain. We showed that a fundamental problem arises in BSS when the un-mixing filter has many taps and therefore high frequency resolution: the permutation of the recovered signals can flip back and forth across frequency. We refer to this problem as *permutation inconsistency*. The end result of such an inconsistency is that the separated speech signal quality is significantly degraded. This fundamental problem is aggravated as the length of the mixing/un-mixing filters increases. Our previous studies [3] revealed that if the un-mixing filter matrix permutations are properly aligned at all frequency bins, the performance of the source separation method is greatly improved. Several methods have been proposed in the literature to solve this problem. Among these, the simplest and yet the most effective is a length constraint on the un-mixing filter, whereby a moderate improvement in separation performance is obtained by sacrificing spectral resolution, resulting in a permutation-inconsistency/spectral-resolution tradeoff.

In this paper, we extend our study of the permutation inconsistency problem. We propose a multiresolution BSS approach, which significantly reduces the permutation misalignment over the whole frequency band while keeping the valuable spectral resolution intact. In this method, separation is done in stages, where short un-mixing filters are used to align the permutations in the initial stages. The filter length is then increased in the later stages to provide the desired spectral resolution. We show that significant performance gain is obtained using only two stages. We present implementation details of the algorithm and evaluate the performance of the proposed method in both real and simulated reverberant environments and under different microphone spacings.

2. BLIND SPEECH SEPARATION IN THE FREQUENCY DOMAIN

The time-domain convolutive mixture $\mathbf{x}(n)$ in (2) can be transformed to an instantaneous mixture in the frequency domain by computing its T-point short-time Fourier transform

$$\mathbf{x}(\omega, m) = \mathbf{H}(\omega)\mathbf{s}(\omega, m), \tag{4}$$

where m is the block index and, ideally, $T = 2P$. For a given set of received data $\mathbf{x}(n)$, $n = 0, \ldots, N-1$, we obtain

$$\mathbf{x}(\omega, m) = \sum_{\tau=0}^{T-1} w(\tau)\mathbf{x}(\beta T m + \tau)e^{-j2\pi\omega\tau/T}, \tag{5}$$

for $\omega = 1, \ldots, T$ and $m = 0, \ldots, N/(\beta T) - 1$, where $w(\tau)$ is a window function and β $(0 < \beta \leq 1)$ is the data overlap factor. The covariance matrix $\mathbf{R}_{\mathbf{x}}(\omega, k)$, assuming ergodicity of the received data, can be estimated using

$$\widehat{\mathbf{R}}_{\mathbf{x}}(\omega, k) = \frac{1}{M}\sum_{m=0}^{M-1} \mathbf{x}(\omega, Mk+m)\mathbf{x}^H(\omega, Mk+m), \tag{6}$$

for $k = 0, \ldots, K-1$, where $(\cdot)^H$ denotes conjugate transposition. Note that in (6), the frequency-domain data is averaged over $M = N/(K\beta T)$, possibly overlapping, consecutive blocks to obtain the covariance matrix at super-block index k. Under the assumption of mutually independent source signals, we seek an un-mixing filter matrix $\mathbf{W}(\omega)$ that decorrelates the estimated source signals $\widehat{s}_1(n)$ and $\widehat{s}_2(n)$, and thus diagonalizes their covariance matrix given by

$$\mathbf{\Lambda}_{\widehat{\mathbf{s}}}(\omega, k) = \mathbf{W}(\omega)\widehat{\mathbf{R}}_{\mathbf{x}}(\omega, k)\mathbf{W}^H(\omega). \tag{7}$$

As shown in [4], this second-order decorrelation criterion alone does not provide enough conditions to solve for $\mathbf{W}(\omega)$, unless the number of outputs is twice the number of inputs (four outputs for the two-input case). However, for non-stationary signals, we can write independent decorrelation equations (7) for K sufficiently separated time intervals [2]. The un-mixing filter $\mathbf{W}(\omega)$ for each frequency bin ω $(\omega = 1, \ldots, T)$ that simultaneously satisfies the K decorrelation equations can then be obtained using an over-determined least-squares solution [4]

$$\widehat{\mathbf{W}}(\omega) = \arg \min_{\mathbf{W}(\omega)} \sum_{k=1}^{K} \|\mathbf{V}(\omega, k)\|^2, \tag{8}$$

where $\|\cdot\|^2$ is the squared Frobenius norm (sum of squares of all elements) and the error

$$\begin{aligned}\mathbf{V}(\omega, k) &= \mathbf{W}(\omega)\widehat{\mathbf{R}}_{\mathbf{x}}(\omega, k)\mathbf{W}^H(\omega) \\ &\quad - \mathrm{diag}\left[\mathbf{W}(\omega)\widehat{\mathbf{R}}_{\mathbf{x}}(\omega, k)\mathbf{W}^H(\omega)\right],\end{aligned} \tag{9}$$

where $\mathrm{diag}[\cdot]$ is the diagonal matrix formed by extracting the diagonal elements of the matrix argument. The least-squares solution to (8) can be obtained using the well-known steepest descent algorithm

$$\begin{aligned}\mathbf{W}^{(l+1)}(\omega) &= \mathbf{W}^{(l)}(\omega) - \mu(\omega) \\ &\quad \cdot \frac{\partial}{\partial \mathbf{W}^{(l)\,H}(\omega)} \left\{\sum_{k=1}^{K} \|\mathbf{V}^{(l)}(\omega, k)\|^2\right\}\end{aligned} \tag{10}$$

for $\omega = 1, \ldots, T$. Following [4], we use a step size of the form

$$\mu(\omega) = \frac{\alpha}{\sum_{k=1}^{K} \|\widehat{\mathbf{R}}_{\mathbf{x}}(\omega, k)\|^2}, \tag{11}$$

where α is a normalized step size. At each iteration, we only update the off-diagonal elements of $\mathbf{W}(\omega)$, thus retaining diagonal elements at their initial values.

The average input signal-to-interference ratio is defined as the ratio of the total signal power obtained via direct channels to the total signal power received via cross channels, i.e.,

$$\mathrm{SIR_i} = \frac{\sum_{\omega=1}^{T}\sum_{i=1}^{2} |H_{ii}(\omega)|^2 |s_i(\omega)|^2}{\sum_{\omega=1}^{T}\sum_{j=1, j\neq i}^{2}\sum_{i=1}^{2} |H_{ji}(\omega)|^2 |s_i(\omega)|^2}. \tag{12}$$

Replacing $\mathbf{H}(\omega)$ by $\mathbf{W}(\omega)\mathbf{H}(\omega)$ similarly defines the average post-processing, or output, signal-to-interference ratio $\mathrm{SIR_o}$. The objective of the BSS methods is to obtain a high SIR improvement given by the ratio $\mathrm{SIR_o}/\mathrm{SIR_i}$.

3. PERMUTATION-INCONSISTENCY/SPECTRAL-RESOLUTION TRADEOFF

It is well-known that a blind estimate of $\mathbf{W}(\omega)$ at frequency ω can at best be obtained up to a scale and a permutation [5]. Therefore, at each frequency ω, the separated signal $s_1(\omega)$ may have $\widehat{s}_1(\omega) = \gamma s_1(\omega)$ or $\widehat{s}_1(\omega) = \gamma s_2(\omega)$, where γ is an arbitrary scaling factor, and the second possibility arises from a simple interchange of the rows of the un-mixing filter matrix $\mathbf{W}(\omega)$. Consequently, the recovered source signal $\widehat{s}_i$ is not necessarily a consistent estimate of s_i over all frequencies. In [4], this problem was solved by constraining the length of $\mathbf{W}(n)$ to $Q < T$, thereby forcing the solution to be smooth or continuous in the frequency domain.

Let us look at the effect of such a constraint on the performance of the BSS method in a reverberant environment. We first consider a case where the direct- and cross-channel impulse responses are measured in an actual reverberant room of dimensions 16.6 ft by 11.2 ft by 8.0 ft high, which contains two omni-directional microphones and two loudspeakers placed at locations shown in Fig. 1. Signal $s_1(n)$ is the speech from a female talker and signal $s_2(n)$ is the speech from a male talker. All the data are recorded at 8-kHz sampling rate. For our experiments, we use four different pairs of speech samples taken from a digital speech database. The duration of impulse responses is 256 ms ($P = 2048$ samples) and the two mixed signals are recorded for 51.2 s ($N = 409600$ samples). In processing the data using the FDSOS algorithm, we first compute the FFT as in (5) of size $T = 2048$ using a Hamming window and with an overlap of $T/2$ samples ($\beta = 0.5$). Based on our results in [3], we use $M = 100$ blocks to compute the covariance matrix in (6) and $K = 4$ super blocks to compute the mean-squared error in (8). We choose a normalized step size of $\alpha = 0.5$, $\mathbf{W}^{(0)} = \mathbf{I}_2$ (the 2×2 identity matrix), and allow the algorithm to run for 100 iterations.

The average output SIR obtained using the actual room impulse responses with un-mixing filters of length $Q = T = 2048$ and speech sample pair # 1 is only 1.04 dB (average input SIR = 1.37 dB). This poor performance is due to random permutations of $\mathbf{W}(\omega)$ distributed over T frequency bins, which in turn corrupts the separation performance. We use the length constraint ($Q < T$) in an attempt to align the permutations of $\mathbf{W}(\omega)$. Constraining the length of the un-mixing filters $\mathbf{W}(n)$ means insufficient spectral resolution in the frequency domain. On the other hand, for values of $Q \approx T$, the length constraint is unable to provide sufficient continuity of the un-mixing filters in the frequency domain [3]. This permutation-inconsistency/spectral-resolution tradeoff is pictured

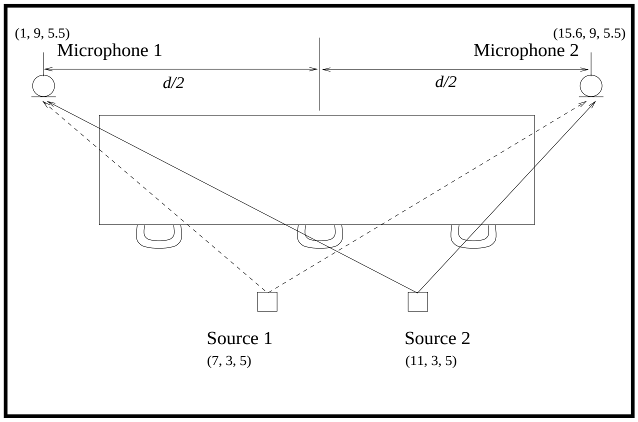

Fig. 1. Room geometry (coordinate units in ft).

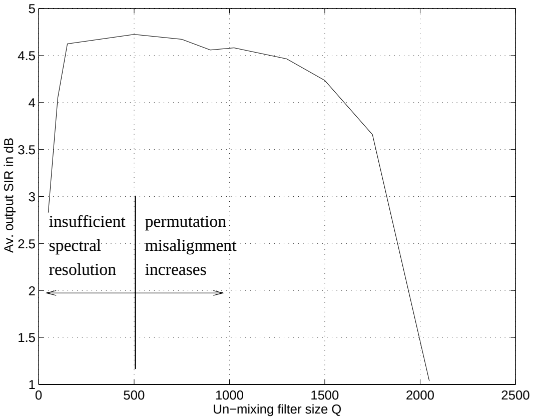

Fig. 2. Plot of the average output SIR (average input SIR = 1.37 dB) versus the un-mixing filter length Q showing the permutation-inconsistency/spectral-resolution tradeoff.

in Fig. 2, which shows the performance of the FDSOS method for different values of un-mixing filter length Q. It shows that $Q = 500$ provides the best compromise between the two competing objectives.

A natural question that now arises is whether we can do still better than the length constraint in terms of enhancing the performance of BSS via the FDSOS method. In [3], we established ideal performance benchmark by aligning the permutations of $\mathbf{W}(\omega)$ based on a prior knowledge of the mixing filters. In practice, this knowledge is not available, so this technique is really just a diagnostic tool used to determine the potential improvement that could be achieved. For the above example, permutation alignment with $Q = 2048$ increases the SIR_o to 8.31 dB [3].

4. MULTIRESOLUTION BSS

To satisfy the desired albeit conflicting requirements of permutation alignment and spectral resolution, we propose a multiresolution frequency-domain (MRFD) algorithm. In this multistage procedure, we use the FDSOS method with increasing values of filter length Q at each stage of the algorithm. Different values of Q imply varying the frequency-domain resolution of $\mathbf{W}(\omega)$ at each stage, hence the name "multiresolution." The rationale behind such an approach is to allow the permutations to align themselves using a smaller value of $Q \ll T$ in the early stages of the algorithm. Once the permutations are aligned, they tend to retain their order even if the value of Q is increased. The increase in the value of Q, however, provides the desired spectral resolution, which is lacking in the early stages.

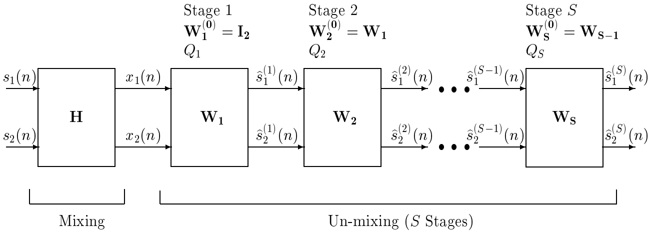

Fig. 3. A pictorial representation of the S-stage multiresolution algorithm to separate speech signals $s_1(n)$ and $s_2(n)$.

A block diagram illustrating the blind separation of speech signals using the MRFD algorithm is shown in Fig. 3. The mixing stage is followed by S un-mixing stages, each of which attempts to further separate the sources using an un-mixing filter with increased spectral resolution. Note that the separated output signals from each un-mixing stage are fed as inputs to the next stage and the final set of weights (after convergence) at each stage are carried over as the initial weights for the following stage. To initiate the separation procedure, we use $\mathbf{W}_1^{(0)} = \mathbf{I}_2$ in the first stage. Since the un-mixing of the speech signals is carried out by S different sets of weights, the SIR_o for the MRFD algorithm can be computed using the overall multichannel response

$$\mathbf{A}(\omega) = \mathbf{W}_S(\omega)\mathbf{W}_{S-1}(\omega)\cdots\mathbf{W}_1(\omega)\mathbf{H}(\omega) \qquad (13)$$

in place of $\mathbf{H}(\omega)$ in (12).

5. EXPERIMENTAL RESULTS

Even though the MRFD idea of Fig. 3 seems to be quite simple, its implementation raises many questions. First, one should be interested in finding an optimum value of the number of stages S that achieves best separation. Second, an intelligent choice should be made for the value of Q in each stage. To answer these questions and to probe the efficacy of the proposed algorithm, we carried out some experimental studies.

Using speech sample pair # 1, we perform different experiments using the MRFD algorithm by varying the number of stages and values of Q. The results are enumerated in Table 1, where it is seen that the best performance is obtained for the two-stage case with $Q_1 = 500$ and $Q_2 = 2048$. Note that the value $Q_2 = 2048$ is the maximum possible length of the un-mixing filters, which gives the best spectral resolution. More interestingly, the value $Q_1 = 500$ coincides naturally with the finite-length constraint of $Q = 500$, that resulted in maximum SIR improvement in the case of the (single-stage) FDSOS method (see Fig. 2). We can, therefore, suggest that a judicious choice of the lengths of the un-mixing filters for the two stages is $Q_1 = 500$ and $Q_2 = 2048$. Let us use these values and apply the two-stage MRFD algorithm to the four different pairs of the speech samples mixed using the actual

Stage 1		Stage 2		Stage 3		Stage 4	
Q_1	$\mathrm{SIR_o}$ (dB)	Q_2	$\mathrm{SIR_o}$ (dB)	Q_3	$\mathrm{SIR_o}$ (dB)	Q_4	$\mathrm{SIR_o}$ (dB)
300	4.63	2048	6.91				
500	4.72	2048	7.00				
700	4.71	2048	7.00				
500	4.72	1800	5.25	2048	5.80		
500	4.72	800	5.63	1000	5.74	1500	5.54

Table 1. Experimental results evaluating the average output SIR of the multistage MRFD algorithm for different values of filter length Q_s; actual room impulse responses, speech sample pair # 1.

room impulse responses. The results are given in the top section of Table 2. The second column lists the distance d between microphones, which is 14.5 ft in this case. A performance improvement in going from one stage (FDSOS method) to two stages is clearly evident. Moreover, the $\mathrm{SIR_o}$ (7.00 dB) for speech sample pair # 1 is close to the ideal benchmark of 8.31 dB obtained using permutation alignment [3].

To analyze the performance of the algorithm under varying reverberant conditions and microphone spacings d , we artificially generate the room impulse responses h_{ji} using the image method of [6]. We use a room model of the same dimensions as the actual room with the two speech sources located at the same positions (see Fig.1). The microphone spacing d, however, can now be varied. First, we use a reflection coefficient, $\rho = 0.7$, and select $d = 2$ and 10 ft to simulate the varying acoustic environment. Thus, we generate two different sets of impulse responses, one for each value of d. The average reverberation time of these impulse responses is 271 ms, which is approximately the same as that of the actual room. The two-stage ($S = 2$) MRFD algorithm is applied to the simulated mixed speech signals in both cases with $Q_1 = 500$, $Q_2 = 2048$. The results for the four pairs of speech samples are shown in the middle section of Table 2. The results for $d = 10$ ft are consistent with those for the actual room. Also, we see that reducing the microphone spacing has no significant affect on performance.

We point out here that the particular choice of Q_1 and Q_2 made above holds well only for rooms having characteristics (e.g., reflection coefficient, size) similar to that of the actual room. As we will show now, the choice varies with these parameters. Recall that our objective in the MRFD algorithm is to impart high spectral resolution to the un-mixing filters by choosing a large value of Q in the later stages. Note that in all the experiments conducted so far in this section, the reverberation time of the impulse responses is on the order of 270 ms, corresponding to 2160 samples at the 8-kHz sampling rate. This matches well with $Q_2 = 2048$, which we have used in our experiments. Similarly, as a rough guide, we may propose to choose the value of Q_2 in direct accordance with the length of the mixing filters determined by their reverberation times. For example in the next set of experiments, we use a room model with a reflection coefficient $\rho = 0.3$, which corresponds to an average reverberation time of 90 ms. We, therefore, select $Q_2 = 700$. Once Q_2 is selected, the value of Q_1 is set to be equal to a multiple of Q_2. We suggest using the multiple $1/4$ based on the ratio $Q_1/Q_2 = 500/2048$ derived from our experiments using the actual room impulse responses. The simulation results for both $d = 2$ and 10 ft are shown in the bottom section of Table 2. Note that in this case most of the SIR gain is obtained at the end of the first stage, although the second stage still further enhances the performance somewhat.

Acoustic Condition	d (ft)	$\mathrm{SIR_i}$ (dB)	Stage 1		Stage 2	
			Q	$\mathrm{SIR_o}$ (dB)	Q	$\mathrm{SIR_o}$ (dB)
Actual Room	14.5	1.37	500	4.72	2048	7.00
		1.09		4.44		5.92
		0.85		4.06		5.65
		0.99		5.15		6.99
Image Model, $\rho = 0.7$	2	0.19	500	4.18	2048	6.56
		-0.30		5.17		9.42
		-0.14		4.41		8.32
		-0.02		5.02		9.74
	10	1.37	500	5.40	2048	8.05
		0.67		5.13		7.58
		0.84		5.31		8.32
		1.24		6.13		8.64
Image Model, $\rho = 0.3$	2	0.57	150	12.67	700	13.50
		0.23		16.07		18.36
		0.31		16.49		18.00
		0.45		17.24		20.51
	10	2.65	150	9.48	700	10.58
		2.13		10.32		11.81
		2.37		12.76		18.50
		2.60		11.42		14.26

Table 2. Performance of the two-stage MRFD algorithm. The result in each line corresponds to a different speech sample pair.

ACKNOWLEDGMENT

The authors would like to thank G. W. Elko for suggesting the multiresolution approach.

6. REFERENCES

[1] S. Haykin, Ed., *Unsupervised Adaptive Filtering. Volume 1: Blind Source Separation*. New York: Wiley, 2000.

[2] E. Weinstein, M. Feder, and A. V. Oppenheim, “Multi-channel signal separation by decorrelation,” *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 405–413, Oct. 1993.

[3] M. Z. Ikram and D. R. Morgan, “Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment,” in *Proc. IEEE ICASSP*, 2000, vol. 2, pp. 1041–1044.

[4] L. Parra and C. Spence, “Convolutive blind separation of nonstationary sources,” *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 320–327, May 2000.

[5] L. Tong, R.-W. Liu, V. C. Soon, and Y.-F. Huang, “Indeterminacy and identifiability of blind identification,” *IEEE Trans. Circuits Syst.*, vol. 38, pp. 499–509, May 1991.

[6] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, Apr. 1979.