EXPERIMENTAL INVESTIGATION OF DELAYED INSTANTANEOUS DEMIXER FOR SPEECH ENHANCEMENT

Yong Xiang, Yingbo Hua, Senjian An

Dept. of Elec. and Electronic Eng.
The University of Melbourne
Victoria 3010 Australia
yxiang,yhua,senjian@ee.mu.oz.au

Alex Acero

Speech Technology Group
Microsoft Research
Redmond, Washington 98052, USA
alexac@Microsoft.com

ABSTRACT

This paper presents a delayed instantaneous demixer (DID) for speech signal separation from real recordings. Based on the fact that the original signals are colored and mutually uncorrelated, a simple algorithm is derived to estimate the parameters of the demixer. This algorithm consists of two parts: a grid searching method to estimate time delays and an alternating projection method to estimate gain coefficients. Experimental result demonstrates the performance of the model and the algorithm.

1. INTRODUCTION

Blind signal recovery from FIR (finite impulse response) and MIMO (multi input and multi output) channel outputs is an intense area of research. It has a wide range of applications such as speech enhancement, telecommunications and medical signal analyses. While many blind deconvolution methods have to use higher order statistics (HOS) for white input signals (e.g., [1]), the additional information that the input signals are colored can be exploited to design second order statistics (SOS) based algorithms [2, 3, 4]. In fact, most natural signals are temporally colored rather than white. Among the SOS-based algorithms, the BIDS (blind identification via decorrelating subchannels) algorithm [4] requires weaker identifiability condition than the matrix pencil (MP) algorithm [2] and the subspace algorithm [3].

However, in some practical applications, complete recovery of input signals is not necessary. For example, in the problem of *Cocktail Party*, speech enhancement can be done by separating the desired speech signals from interfering sources. Although the separated signals may be convolutive distorted versions of the original speech signals, this distortion is to some extent not detectable by human ears. In the blind signal separation (BSS) problem, neither the channel nor the signals are known. Many algorithms reported have been tested on computer generated signals, to separate signals from instantaneous mixtures [5, 6] and from dynamic mixtures [7, 8]. Other algorithms have been tested using real acoustically mixed speech signals (e.g., [9, 10]). But these algorithms are normally complicated because of high filter orders and some suffer from local minimum problems.

This paper focus on separating two unknown speech signals from their convolutive mixtures recorded by two microphones. A formulation of the problem is given in section 2. The delayed instantaneous demixer (DID) model for separating acoustically mixed signals is presented in section 3 with implementation details in section 4. Section 5 shows an experimental example.

2. PROBLEM FORMULATION

A noiseless 2 × 2 FIR MIMO channel can be described as

$$\mathbf{y}(n) = \mathbf{H}'(n) * \mathbf{s}(n) \triangleq \sum_{l=0}^{L_H} \mathbf{H}'(l)\mathbf{s}(n-l) \qquad (1)$$

where $\mathbf{s}(n) = [s_1(n), s_2(n)]^T$ denotes the 2×1 *unknown* input vector, $\mathbf{y}(n) = [y_1(n), y_2(n)]^T$ the 2×1 output vector and $\mathbf{H}'(n)$ the 2×2 sequence of the system impulse response of length L_H. The operators $*$ and ^{T} represent convolution and transpose, respectively. All data in the time domain are assumed, without loss of generality, to be real valued. An equivalent expression of (1) is

$$\mathbf{y}(n) = \mathbf{H}'_z(z)\mathbf{s}(n) \qquad (2)$$

where $\mathbf{H}'_z(z) = \sum_{l=0}^{L_H} \mathbf{H}'(l)z^{-l}$ denotes the channel matrix. The order of $\mathbf{H}'_z(z)$ reflects the reverberation time varying on the room size, wall absorbance, and speaker and microphone positions, etc. The blind channel identification and equalization methods can be used to estimate $\mathbf{s}(n)$ from

THIS WORK WAS SUPPORTED BY THE CENTER FOR SENSOR SIGNAL & INFORMATION PROCESSING (CSSIP), AND AUSTRALIAN RESEARCH COUNCIL.

$\mathbf{y}(n)$ under certain conditions, e.g., the BIDS algorithm assumes that the channel matrix is irreducible and the input signals are mutually uncorrelated and of sufficiently diverse power spectra. As we discussed above, blind signal separation is sufficient for most speech enhancement problems. In other word, any (diagonal) convolutive version of $\mathbf{s}(n)$ is a desired solution as long as the order of convolution is not so high. To achieve this, we need to construct a demixer $\mathbf{G}'(z)$ such that $\mathbf{G}'(z)\mathbf{H}'(z)$ is a diagonal polynomial matrix up to a permutation matrix. Obviously, selecting proper structures of $\mathbf{G}'(z)$ is important in designing simple, fast and robust algorithms.

3. DELAYED DEMIXER MODEL

A general expression of channel matrix $\mathbf{H}'_z(z)$ is

$$\mathbf{H}'_z(z) = \begin{bmatrix} h'_{11}(z) & h'_{12}(z) \\ h'_{21}(z) & h'_{22}(z) \end{bmatrix} \tag{3}$$

where $\mathbf{h}'_{ij}(z), i, j = 1, 2$ are polynomials. Under the assumption that at least one element on each row and column of $\mathbf{H}'_z(z)$ is a polynomial of minimum phase, (3) can be re-written as

$$\begin{aligned}\mathbf{H}'_z(z) &= \begin{bmatrix} 1 & \frac{h'_{12}(z)}{h'_{22}(z)} \\ \frac{h'_{21}(z)}{h'_{11}(z)} & 1 \end{bmatrix} \begin{bmatrix} h'_{11}(z) & 0 \\ 0 & h'_{22}(z) \end{bmatrix} \\ &= \mathbf{H}_z(z)\Lambda\end{aligned}$$

Note that if not all of these minimum phase polynomials are on the main diagonal, there exists permutation between two sources. Possible permutation is ignored here for convenience. Taking the advantage of indeterminacies of BSS, (2) becomes

$$\mathbf{y}(n) = \mathbf{H}_z(z)\mathbf{x}(n) \tag{4}$$

where $\mathbf{x}(n) = [h'_{11}(z)s_1(n), h'_{22}(z)s_2(n)]^T$. The *Cocktail Party* problem is a specific case of this model where each speaker is assumed to be close to a distinct microphone which leads to $h'_{11}(z) = h'_{22}(z) = 1$. Most existing algorithms choose an FIR demixer with unit main diagonal elements.

Huang *et al* carried out an experiment to measure the impulse responses of the cross-channel acoustic paths [11]. This experiment was done in a room-acoustic environment at the House Ear Institute, Los Angeles. The room size is $21ft \times 13.5ft$ with a table of size $12ft \times 4ft$ in the middle of the room. Two speakers, sitting at two sides of the table (face to face), are $4ft$ away from each other. Their experimental result shows that the lengths of the two cross-channel filters are 200 samples, corresponding to 18.7ms at a sampling rate of 10667Hz. A close looking at the cross-channel filters, one can see that only very few of the pulses have dominant magnitudes and stay with one another although the orders of the cross-channels are in general high. Based on this important observation, we can further simplify the channel matrix in (4) to be

$$\mathbf{H}_z(z) \approx \begin{bmatrix} 1 & a_1 z^{-T_1} \\ a_2 z^{-T_2} & 1 \end{bmatrix}$$

which leads to the corresponding delayed instantaneous demixer

$$\mathbf{G}(z) = \begin{bmatrix} 1 & b_1 z^{-T_1} \\ b_2 z^{-T_2} & 1 \end{bmatrix}$$

We will show next that the construction of the demixer $\mathbf{G}(z)$ is extremely simple.

4. ALGORITHM IMPLEMENTATION

The first part of the algorithm is to estimate time delays T_1 and T_2.

4.1. Estimating time delays

Define

$$\mathbf{z}(n) \triangleq \mathbf{G}(z)\mathbf{y}(n) \tag{5}$$

$$\mathbf{R}_{\mathbf{y}}(\tau) \triangleq \mathbf{E}\left[\mathbf{y}(n)\mathbf{y}^T(n-\tau)\right]$$

$$\mathbf{R}_{\mathbf{z}}(\tau) \triangleq \mathbf{E}\left[\mathbf{z}(n)\mathbf{z}^T(n-\tau)\right]$$

where $\mathbf{E}$ denotes the expectation operator. Let $r_{y_i y_j}(\tau)$ and $r_{z_i z_j}(\tau)$, $i, j = 1, 2$ denote the *ij*th elements of $\mathbf{R}_{\mathbf{y}}(\tau)$ and $\mathbf{R}_{\mathbf{z}}(\tau)$, respectively. After a simple manipulation, one has

$$\begin{aligned} r_{z_1 z_2}(\tau) &= b_2 r_{y_1 y_1}(\tau + T_2) + b_1 b_2 r_{y_2 y_1}(\tau + T_2 - T_1) \\ &\quad + r_{y_1 y_2}(\tau) + b_1 r_{y_2 y_2}(\tau - T_1) \end{aligned} \tag{6}$$

$$\begin{aligned} r_{z_2 z_1}(\tau) &= b_2 r_{y_1 y_1}(\tau - T_2) + b_1 b_2 r_{y_1 y_2}(\tau + T_1 - T_2) \\ &\quad + r_{y_2 y_1}(\tau) + b_1 r_{y_2 y_2}(\tau + T_1) \end{aligned} \tag{7}$$

The autocorrelation matrix of $\mathbf{y}(n)$ can be computed as

$$\hat{\mathbf{R}}_{\mathbf{y}}(\tau) \approx \frac{1}{N}\sum_{n=0}^{N} \mathbf{y}(n)\mathbf{y}^T(n-\tau) \tag{8}$$

Substituting (8) into (6) and (7), we obtain $\hat{r}_{z_1 z_2}(\tau)$ and $\hat{r}_{z_2 z_1}(\tau)$.

The cost function to be minimized is

$$\mathbf{J} \triangleq \sum_{\tau=0}^{K} \left(\hat{r}^2_{z_1 z_2}(\tau) + \hat{r}^2_{z_2 z_1}(\tau)\right) \tag{9}$$

For the 2×2 case, there are only 4 unknown parameters T_1, T_2, b_1, b_2 in the cost function. We first use a simple grid searching method to estimate T_1 and T_2. We set a range $[0, T]$ for T_1 and T_2 with step size 1, and a range $[-b, b]$ for b_1 and b_2 with step size Δ_b, respectively. Here, T, b and Δ_b are all positive real values. Basically, large T and b or small Δ_b corresponds to more computation time. For different combinations of these parameters, $\mathbf{J}$ is computed according to (9). Then, T_1 and T_2 can be obtained from that set of parameter combination which leads to the least $\mathbf{J}$. If one wants to obtain accurate estimates of b_1 and b_2 at the same time, then step size Δ_b must be chosen to be small enough that slows down the searching speed. We will use an alternating projection method to estimate b_1 and b_2 after T_1 and T_2 are obtained.

4.2. Estimating gain coefficients

Let $\mathbf{G}(z) = \sum_{l=0}^{L_G} \mathbf{G}(l) z^{-l}$, where $L_G = max(T_1, T_2)$. It follows from (5)

$$\mathbf{z}(n) = [\mathbf{G}(0), \mathbf{G}(1), \cdots, \mathbf{G}(L_G)] \begin{bmatrix} \mathbf{y}(n) \\ \mathbf{y}(n-1) \\ \vdots \\ \mathbf{y}(n-L_G) \end{bmatrix}$$
$$= \bar{\mathbf{G}} \bar{\mathbf{y}}$$

Denote

$$\bar{\mathbf{G}}^T \triangleq [\mathbf{g}_1, \mathbf{g}_2]$$

$$\mathbf{R}_{\bar{\mathbf{y}}}(\tau) \triangleq \mathbf{E}\left[\bar{\mathbf{y}}(n)\bar{\mathbf{y}}^T(n-\tau)\right] \qquad (10)$$

We can form a cost function as

$$\mathbf{J}' = \sum_{\tau=0}^{K} \left[\left(\mathbf{g}_1^T \hat{\mathbf{R}}_{\bar{\mathbf{y}}}(\tau) \mathbf{g}_2 \right)^2 + \left(\mathbf{g}_2^T \hat{\mathbf{R}}_{\bar{\mathbf{y}}}(\tau) \mathbf{g}_1 \right)^2 \right]$$

where $\hat{\mathbf{R}}_{\bar{\mathbf{y}}}(\tau)$ is the estimate of $\mathbf{R}_{\bar{\mathbf{y}}}(\tau)$. The calculation of $\hat{\mathbf{R}}_{\bar{\mathbf{y}}}(\tau)$ is similar to that of $\hat{\mathbf{R}}_{\mathbf{y}}(\tau)$. The cost function $\mathbf{J}'$ is a non-quadratic function of $\bar{\mathbf{G}}$. But it is quadratic with respect to each individual row of $\bar{\mathbf{G}}$.

Denote

$$\mathbf{Q}_j = \sum_{\tau=0}^{K} \left(\hat{\mathbf{R}}_{\bar{\mathbf{y}}}(\tau) \mathbf{g}_j \mathbf{g}_j^T \hat{\mathbf{R}}_{\bar{\mathbf{y}}}^T(\tau) + \hat{\mathbf{R}}_{\bar{\mathbf{y}}}^T(\tau) \mathbf{g}_j \mathbf{g}_j^T \hat{\mathbf{R}}_{\bar{\mathbf{y}}}(\tau) \right)$$

where $j = 1, 2$. By differentiating the cost function $\mathbf{J}'$, we have

$$\frac{\partial \mathbf{J}'}{\partial \mathbf{g}_i^T} = 2 \mathbf{Q}_j \mathbf{g}_i$$

where $i = 1, 2 (i \neq j)$. Let $\mathbf{g}_i'$ is $\mathbf{g}_i$ with all zero elements removed from $\mathbf{g}_i$, and $\mathbf{Q}_j'$ is $\mathbf{Q}_j$ with the columns corresponding to the zero elements of $\mathbf{g}_i$ removed from $\mathbf{Q}_j$. The alternating projection method is formulated as follows:
(1) Iteration index $l = 0$.
(2) Set $\mathbf{g}_1^{(0)} = \mathbf{g}_2^{(0)} = [1, 0, \cdots, 0]^T$. Then use b_1 and b_2 obtained from grid search method to replace the $(2T_1 + 2)$-th element of $\mathbf{g}_1^{(0)}$ and the $(2T_2 + 1)$-th element of $\mathbf{g}_2^{(0)}$, respective.
(3) $\mathbf{g}_i'^{(l+1)}$ = unit-norm least-eigenvector of $\mathbf{Q}_j'^{(l+1)}$, where $i, j = 1, 2$ but $i \neq j$, and

$$\begin{aligned} \mathbf{Q}_j^{(l+1)} &= \sum_{\tau=0}^{K} \Big(\hat{\mathbf{R}}_{\bar{\mathbf{y}}}(\tau) \mathbf{g}_j^{(k)} \mathbf{g}_j^{(k)T} \hat{\mathbf{R}}_{\bar{\mathbf{y}}}^T(\tau) \\ &\quad + \hat{\mathbf{R}}_{\bar{\mathbf{y}}}^T(\tau) \mathbf{g}_j^{(k)} \mathbf{g}_j^{(k)T} \hat{\mathbf{R}}_{\bar{\mathbf{y}}}(\tau) \Big) \end{aligned}$$

where $k = l$ if $j = 2$, or $k = l + 1$ if $j = 1$. $\mathbf{g}_i^{(l+1)}$ should also be updated at this step.
(4) If $\sum_{i=1}^{2} \mid \mathbf{g}_i'^{(k+1)} - \mathbf{g}_i'^{(k)} \mid \leq \epsilon$ (ϵ denotes the selected threshold), stop; otherwise $l = l + 1$, goto step (3).

Remark: For real speech data, it is not necessary to use a very large set of data samples to construct a demixer. We only need to pick a segment of data that contain “uncorrelated words”, i.e., the words mixed in that segment should have highly uncorrelated waveforms.

5. EXPERIMENTAL RESULT

This section demonstrates the proposed method using real recording made by T.-W. Lee (available from $http://www.cnl.salk.edu/\sim tewon/$). The recording was done in a normal office room at sampling rate 16kHz. Two Speakers have been recorded speaking simultaneously. Speaker 1 says the digits from one to ten in English and speaker 2 counts at the same time the digits in Spanish. The distance between the speakers and the microphones is about 60cm in a square ordering. The Independent Component Analysis (ICA) method was used by Lee to separate the mixed speech signals and the reconstructed waveforms are shown in Figure 1. More recently, the probabilistic Independent Component Analysis (PICA) method was proposed by Acero *et al* to enhance speech (see $http://research.microsoft.com/\sim alexac/bss/$). The PICA method is not blind because it uses a more accurate probabilistic model of speech. These methods are computationally costly as they use a large set of data samples to construct the high order channel equalizers.

Figure 2 shows the waveforms of the separated signals by using the DID method. In the experiment, we chose $T = 100$, $b = 1.2$ and $\Delta_b = 0.1$. 20000 samples (about one word length) between sample 10500 and sample 30499 were used to estimate the delayed instantaneous demixer $\mathbf{G}(z)$. We have observed that a data segment of one word length is enough to construct the demixer. By playing these separated signals, we found that the ICA method offered better separation than the the DID method but the latter yielded less noisy separated signals. While the PICA method performed the best among the three methods, the DID method delivered the least computational complexity. The DID method could be very useful for data preprocessing.

6. CONCLUSION

In this paper, we have proposed a simple delayed instantaneous demixer for signal separation. The derivation of this demixer was motivated by the following observation: in some acoustic environment, the cross-channel impulse responses were dominated by only few strong pulses that exist in a very small neighborhood. An algorithm based on grid searching and alternating projection was used to construct the demixer. Real recorded data were used to investigate the performance of the DID method with comparison to the ICA method and the PICA method. The simplicity of the DID method makes it a potential candidate for data preprocessing.

7. REFERENCES

[1] J. Tugnait and B. Huang, “On a whitening approach to partial channel estimation and blind equalization of FIR/IIR multiple-input multiple-output channels,” *IEEE-T-SP*, Vol.48, No.3, pp. 832-845, Mar. 2000.

[2] C. T. Ma, Z. Ding, and S. F. Yau, “A two stage algorithm for MIMO blind deconvolution of colored input signals”, *IEEE-T-SP*, Vol.48,No.4, Apr. 2000.

[3] A. Gorokhov and P. Loubaton, “Subspace-based techniques for blind separation of convolutive mixtures with temporally correlated sources”, *IEEE-T-CS - I*, vol.44, No.9, pp. 813–820, Sept. 1997.

[4] Y. Hua, Y. Xiang, and K. Abed-Meraim, “Blind Identification of Colored Signals Distorted by FIR Channels,” *Proc. of IEEE ICASSP’2000*, Istanbul, Turkey, June 2000.

[5] L. Tong, R. Liu, V. C. Soon, and Y. H. Huang, “Indeterminacy and identifiability of blind identification,” *IEEE-T-CS*, vol.38, No.5, pp. 499–509, May 1991.

[6] A. Belouchrani, K. Abed-Meraim, J.F. Cardoso, and E. Moulines, “Blind source separation using second order statistics,” *IEEE-T-SP*, Vol.45, No.2, pp. 434–444, Feb. 1997.

[7] U. Lindgren and H. Broman, “Source separation: Using a criterion based on second order statics,” *IEEE-T-SP*, Vol.46, No.7, July 1998.

[8] H.L. Nguyen Thi and C. Jutten, “Blind source separation for convolutive mixtures,” *Signal Processing*, 45:209-229, 1995.

[9] Konstantinos I. Diamantaras, Athina P. Petropulu, and Binning Chen, “Blind two-input-two-output FIR channel identification based on frequency domain second-order statistics,” *IEEE-T-SP*, Vol.48, No.2, Feb. 2000.

[10] T.-W. Lee and A. Ziehe, “Combining time-delayed decorrelation and ICA: towards solving the Cocktail Party problem,” *Proc. of IEEE ICASSP’98*, Vol.2, pp. 1249-1252, Seattle, May 1998.

[11] J. Huang, K.-C. Yen, and Y. Zhao, “Subband-based adaptive decorrelation filtering for co-channel speech separation,” *IEEE-T-SAP*, Vol.8, No.4, July 2000.

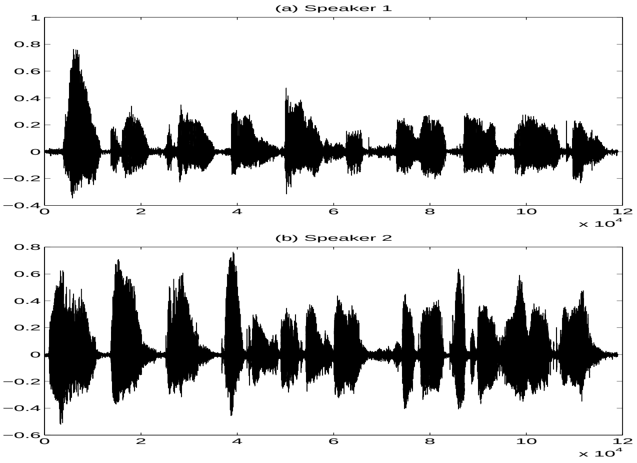

Figure 1. The reconstructed waveforms using the ICA method.

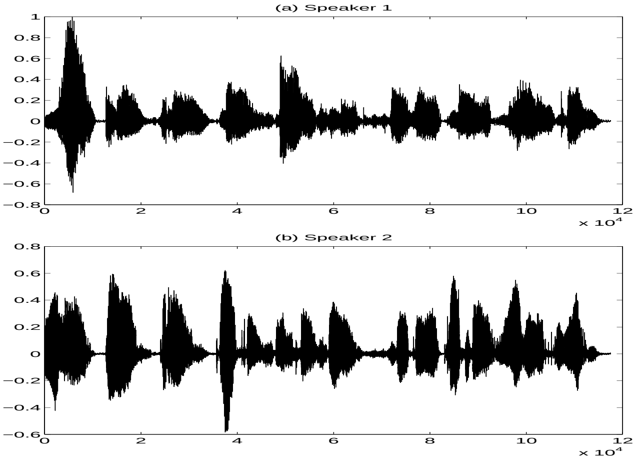

Figure 2. The reconstructed waveforms using the DID method.