BLIND SEPARATION OF L SOURCES FROM M MIXTURES OF SPEECH SIGNALS

Phillip De Leon and Yunsheng Ma

New Mexico State University
Klipsch School of Electrical and Computer Engineering
Box 30001 / Dept. 3-0
Las Cruces, New Mexico 88003-8001
{pdeleon, yuma}@nmsu.edu

ABSTRACT

In many real-world applications of blind source separation, the number of mixture signals, M available for analysis often differs from the number of sources, L which may be present. In this paper, we extend a successful and efficient kurtosis maximization algorithm used in speech separation of two sources from two linear mixtures for use in problems with arbitrary numbers of sources and mixtures. We examine three cases: underdetermined ($M < L$), critically-determined ($M = L$), and overdetermined ($M > L$). In each of these cases, we present simulation results (using the TIMIT speech corpus) and discuss observed algorithm limitations.

1. INTRODUCTION

The separation of individual speech signals (sources) from mixtures of other speech signals and noise has been actively investigated over the last few years [1],[2],[3]. Applications of this work include audio-interfaces, hearing aids, multimedia, and speech recognition systems. Given the complicated nature of speech this is a difficult problem compounded by environmental effects such as noise and reverberation. Furthermore, there is a strong desire for natural sounding separated outputs and a simple algorithm suitable for real-time operation.

In this paper, we generalize a previously published method for blind separation of two speech signals from two mixtures to separation of L sources from M mixtures [1]. We evaluate the separation performance of the generalized algorithm for three cases: undetermined ($M < L$), critically-determined ($M = L$), and overdetermined ($M > L$). In each of these cases, we present simulation results (using the TIMIT speech corpus) and discuss observed algorithm limitations.

This research is supported by the U.S. Air Force Research Laboratories Grant #F41624-99-0001.

2. TWO-SOURCE, TWO-MIXTURE KURTOSIS MAXIMIZATION ALGORITHM

2.1. Problem Formulation

In the two-source, two-mixture blind speech separation problem illustrated in Fig. 1, we assume two unknown speech signals, s_1 and s_2 are mixed in a linear fashion to produce two mixture signals x_1 and x_2. (The more realistic problem setting would assume convolutional mixtures, however, this is a much more difficult problem currently with no known solution which produces large Signal-to-Interference Ratios.) We wish to produce y_1 and y_2 which approximate s_1 and s_2.

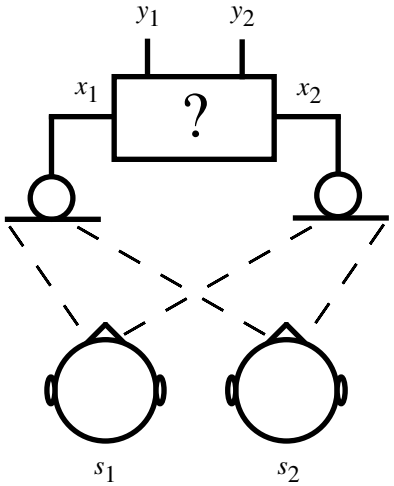

Fig. 1. Speech signal separation problem

Mathematically we may express the mixing model as

$$\mathbf{x}(n) = \mathbf{A}(n)\mathbf{s}(n) \quad (1)$$

where

$$\begin{aligned} \mathbf{s}(n) &= \left[\begin{array}{cc} s_1(n) & s_2(n) \end{array}\right]^T, \\ \mathbf{x}(n) &= \left[\begin{array}{cc} x_1(n) & x_2(n) \end{array}\right]^T \end{aligned} \quad (2)$$

are the vectors of source, mixture signals respectively and $\mathbf{A}(n)$ is the unknown, possibly time-varying, 2×2 mixing matrix composed of scalar elements. The objective is to compute a separation matrix, $\mathbf{W}(n)$ such that

$$\mathbf{y}(n) \quad = \quad \mathbf{W}(n)\mathbf{x}(n) \tag{3}$$

where

$$\mathbf{y}(n) \quad = \quad \left[\begin{array}{cc} y_1(n) & y_2(n) \end{array} \right]^T \tag{4}$$

is the vector of output signals approximating the separated sources. Clearly, choosing $\mathbf{W}$ such that $\mathbf{WA} = \mathbf{I}$ (identity matrix) or $\mathbf{J}$ (counter identity matrix) would invert the mixing process and separate the signals (assuming $\mathbf{A}$ is invertible) but $\mathbf{A}$ is not known.

In simulations, the quality of separation can be measured by examining how close the product matrix $\mathbf{WA}$ is to being diagonal or anti-diagonal. This measure simply examines the ratio of the largest element to smallest element of each row and is equivalent to measuring the power of the desired source to that of the undesired source or the signal-to-interference ratio (SIR). Informal listening evaluations indicate a separation ratio of 20dB or higher produces a fairly distinct source output. Duplicate (same) source outputs manifest themselves in product matrices which have the larger elements in the same column and thus negative SIRs. Finally, SIRs near 0dB indicate no real source separation has occurred.

2.2. Algorithm Development

The Kurtosis Maximization Algorithm (KMA) is based on the fundamental assumption that linear mixtures of speech signals have a kurtosis, defined as

$$\kappa_x \quad \equiv \quad \frac{E\left[x^4\right]}{\left\{E\left[x^2\right]\right\}^2}, \tag{5}$$

less than that for either source [1]. Under this assumption, a simple and computationally inexpensive gradient ascent algorithm, is employed to maximize kurtosis thereby separating the source speech signals from the mixture. The idea is expressed as

$$\begin{aligned} \mathbf{W}(n+1) &= \mathbf{W}(n) + \mu \bigtriangledown \kappa_{\mathbf{y}} \\ &= \mathbf{W}(n) + \mu \left[\begin{array}{cc} \frac{\partial \kappa_{y_1}}{\partial W_{11}} & \frac{\partial \kappa_{y_1}}{\partial W_{12}} \\ \\ \frac{\partial \kappa_{y_2}}{\partial W_{21}} & \frac{\partial \kappa_{y_2}}{\partial W_{22}} \end{array} \right] \\ &= \mathbf{W}(n) + \mu \mathbf{C}(n) \end{aligned} \tag{6}$$

where μ is the step size, $\bigtriangledown \kappa_{\mathbf{y}}$ is the gradient of the kurtosis of the output signals with respect to the elements of the separation matrix, and $\mathbf{C}(n)$ is the correction matrix used in the update rule. Statistical expectations in the correction matrix are approximated by instantaneous or auto-regressive (AR) estimators.

A normalized version of the algorithm has been proposed and shown to yield better performance [4]. In the normalized KMA, we scale the correction matrix, $\mathbf{C}(n)$ by its ℓ_2 norm

$$\mathbf{W}(n+1) \quad = \quad \mathbf{W}(n) + \frac{\tilde{\mu}}{||\mathbf{C}(n)||_2^2}\mathbf{C}(n) \tag{7}$$

where $\tilde{\mu}$ is the normalized step size and

$$||\mathbf{C}(n)||_2^2 \quad = \quad \max\left\{eigenvalue\left[\mathbf{C}(n)\mathbf{C}^T(n)\right]\right\}. \tag{8}$$

In the two-source, two-mixture speech separation problem, the normalized KMA has been shown to provide mean SIRs on the order of 25-40dB [4].

3. GENERALIZED, NORMALIZED KMA

In the derivation of the generalized KMA speech separation algorithm, we assume L unknown speech signals, $s_1, \ldots, s_L$ are mixed in a linear fashion to produce M mixture signals, $x_1, \ldots, x_M$. We wish to produce $y_1, \ldots, y_L$ which approximate $s_1, \ldots, s_L$ (or some other permutation of the signal set).

Mathematically we may express the generalized mixing model as

$$\mathbf{x}(n) \quad = \quad \mathbf{A}(n)\mathbf{s}(n) \tag{9}$$

where

$$\begin{aligned} \mathbf{s}(n) &= \left[\begin{array}{c} s_1(n), \ldots, s_L(n) \end{array} \right]^T, \\ \mathbf{x}(n) &= \left[\begin{array}{c} x_1(n), \ldots, x_M(n) \end{array} \right]^T \end{aligned} \tag{10}$$

are the vectors of source, mixture signals respectively and $\mathbf{A}(n)$ is the $M \times L$ mixing matrix composed of scalar elements. The objective is to compute a $L \times M$ separation matrix, $\mathbf{W}(n)$ such that

$$\mathbf{y}(n) \quad = \quad \mathbf{W}(n)\mathbf{x}(n) \tag{11}$$

where

$$\mathbf{y}(n) \quad = \quad \left[\begin{array}{c} y_1(n), \ldots, y_L(n) \end{array} \right]^T \tag{12}$$

is the vector of output signals approximating the separated sources.

As in the two source, two mixture case, we formulate a multidimensional objective function composed of the kurtoses of the output signals

$$\mathbf{J} \quad = \quad \left[\kappa_{y_1}, \ldots, \kappa_{y_L}\right]^T. \tag{13}$$

The gradient of $\mathbf{J}$ with respect to the elements of the separation matrix is given by

$$\frac{\partial J_l}{\partial W_{lm}} = 4 \times \frac{E\left[y_l^3 x_m\right] E\left[y_l^2\right] - E\left[y_l^4\right] E\left[y_l x_m\right]}{\left\{E\left[y_l^2\right]\right\}^3}. \quad (14)$$

Second order statistics in (14) are approximated at time n as

$$\begin{aligned} E\left[y_l^2\right] &\approx \sum_{i=1}^{M}\sum_{j=1}^{M} W_{li}\hat{r}_{ij}(n)W_{lj} \\ E\left[y_l x_m\right] &\approx \sum_{i=1}^{M} W_{li}\hat{r}_{im}(n) \end{aligned} \quad (15)$$

where $\hat{r}_{ij}(n)$ is the auto-regressive estimate of the cross correlation, $E\left[x_i x_j\right]$

$$\hat{r}_{ij}(n) \approx \lambda\hat{r}_{ij}(n-1) + (1-\lambda)x_i(n)x_j(n). \quad (16)$$

Fourth order statistics in (14) are approximated at time n with instantaneous estimators

$$\begin{aligned} E\left[y_l^3 x_m\right] &\approx y_l^3(n)x_m(n) \\ E\left[y_l^4\right] &\approx y_l^4(n). \end{aligned} \quad (17)$$

Substitution of (15) and (17) into (14) yields the lmth element of the $L \times M$ correction matrix, $\mathbf{C}(n)$ which is then used in the $L \times M$ version of (7).

4. RESULTS

In order to measure the performance of the generalized KMA algorithm, simulations were conducted. Source speech signals were chosen from the TIMIT speech corpus and mixtures were digitally synthesized according to the mixing matrix. Algorithm parameters were selected as $\tilde{\mu} = 0.0001$ and $\lambda = 0.99995$. The results for four cases are described below.

4.1. Critically-Determined: Three Sources, Three Mixtures

In the critically-determined case, we choose the mixing matrix at random,

$$\mathbf{A} = \begin{bmatrix} 0.9501 & 0.4860 & 0.4565 \\ 0.2311 & 0.8913 & 0.0185 \\ 0.6068 & 0.7621 & 0.8214 \end{bmatrix} \quad (18)$$

and initialize the separation matrix as

$$\mathbf{W}(0) = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}. \quad (19)$$

Fig. 2 illustrates the SIRs of the three source, three mixture simulation. We note in this simulation after a short adaptation period, excellent separation on the order of 30dB or higher.

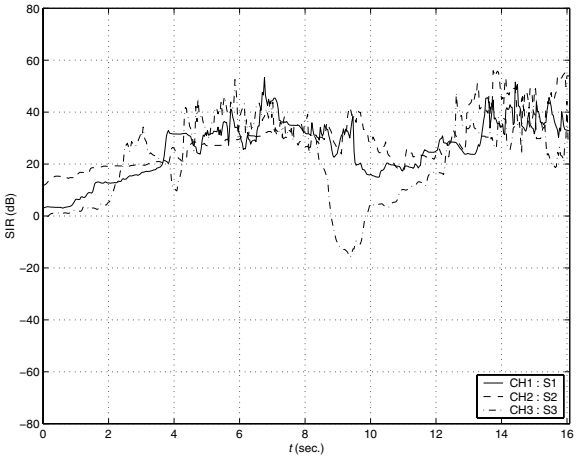

Fig. 2. Signal-to-Interference Ratios of three source, three mixture simulation.

4.2. Over-Determined: Two Sources, Three Mixtures

In the over-determined case, we choose the mixing matrix at random,

$$\mathbf{A} = \begin{bmatrix} 0.9501 & 0.4860 \\ 0.2311 & 0.8913 \\ 0.6068 & 0.7621 \end{bmatrix} \quad (20)$$

and initialize the separation matrix as

$$\mathbf{W}(0) = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \end{bmatrix} \quad (21)$$

Fig. 3 illustrates the SIRs of the two source, three mixture simulation. We note in this simulation after a short adaptation period, excellent separation on the order of 30dB or higher. This is an expected result given the results in Section 4.1.

4.3. Under-Determined: Three Sources, Two Mixtures

In the under-determined case, we choose the mixing matrix at random,

$$\mathbf{A} = \begin{bmatrix} 0.9501 & 0.6068 & 0.8913 \\ 0.2311 & 0.4860 & 0.7621 \end{bmatrix} \quad (22)$$

and initialize the separation matrix as

$$\mathbf{W}(0) = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \\ 0.5 & 0.5 \end{bmatrix} \quad (23)$$

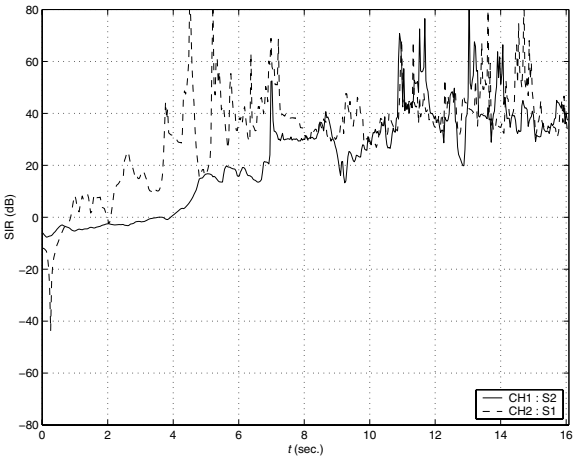

Fig. 3. Signal-to-Interference Ratios of two source, three mixture simulation.

Fig. 4 illustrates the SIRs of the three source, two mixture simulation. We note in this simulation, separation of sources is poor with two sources having a minor 5dB SIR improvement while the other source is not separated at all.

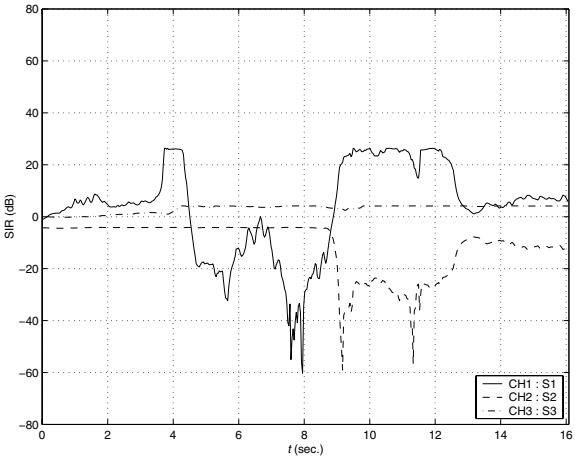

Fig. 4. Signal-to-Interference Ratios of three source, two mixture simulation.

5. CONCLUSIONS

In this paper, we have generalized a previously published algorithm for separation of L speech signals from M linear mixtures of these signals. The algorithm performs very well (SIR $\approx$ 30dB) for $L \leq M$ but does not produce satisfactory results for $L > M$.

6. REFERENCES

[1] J. LeBlanc and P. De Leon, "Speech separation by kurtosis maximization," *Proc. ICASSP*, vol. 2, pp. 1029–1032, 1998.

[2] K. Yen and Y. Zhao, "Adaptive co-channel speech separation and recognition," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 138–151, Mar. 1999.

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

[4] P. De Leon and Y. Ma, "Normalized, hos-based, blind speech separation algorithms," *Asilomar Conf. Sigs. , Sys., and Comps.*, 2000.