BLIND SEPARATION OF MIXTURES OF SPEECH SIGNALS WITH UNKNOWN PROPAGATION DELAYS

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

The ultimate goal of blind speech separation is to separate individual speech (source) signals from a set of convolutional mixtures of these signals. This is a difficult problem given the overlap in both time- and frequency-domains of the sources not to mention compounding acoustical effects. Current approaches to this problem have yielded only modest results. In this paper, we extend a blind separation algorithm (used for instantaneous mixtures) to compensate for non-time aligned sources in the mixtures. While this is a special case of the convolutional mixing model, it takes into account propagation delays and can thus be used in environments where echo is minimal.

1. INTRODUCTION

Separation of convolutional mixtures of speech signals has received considerable attention in the research community over the last two years due to broad applications in audio-interfaces, hearing aids, and speech recognition systems [1], [2], [3]. In this problem, illustrated in Fig. 1, we assume two unknown speech signals, s_1 and s_2 are filtered by $\mathbf{a}_{ji}$ and mixed to produce two mixture signals x_1 and x_2. The filter, $\mathbf{a}_{ji}$ represents the length-N, unknown room impulse response from source i to microphone j. In the blind separation problem, we wish to produce y_1 and y_2 which approximate s_1 and s_2 given only x_1 and x_2.

The blind speech separation problem is a very difficult one given the complicated nature of speech signals (non-stationary, overlap in time- and frequency-domains, etc ...). Convolutional mixing further complicates the problem due to causality and stability restrictions on the inverse filters not to mention length requirements in the FIR approximation. Research results of the general problem are modest at best with typical Signal-to-Interference Ratio Improvements (SIRIS) on the order of 10-15dB [3], [4].

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

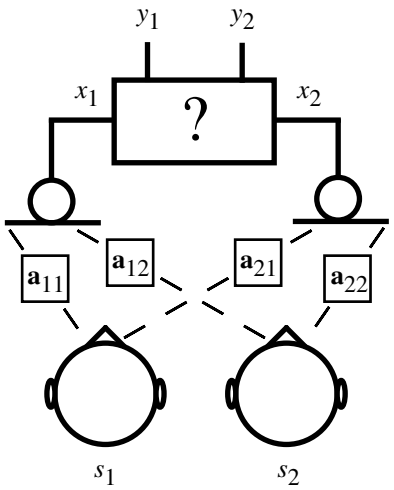

Fig. 1. Speech signal separation problem

We mathematically describe the general blind speech separation problem as follows. The mixing equation is given by

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

where

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

is the vector of input, source signals respectively, $\mathbf{A}(n)$ is 2×2 matrix of (possibly time-varying) impulse responses

$$\mathbf{A}(n) = \left[\begin{array}{cc} \mathbf{a}_{11}(n) & \mathbf{a}_{12}(n) \\ \mathbf{a}_{21}(n) & \mathbf{a}_{22}(n) \end{array} \right], \quad (3)$$

and $*$ is the convolution operator. The jth mixture signal

from (1) is given by

$$x_j(n) = \mathbf{a}_{j1}^T(n)\mathbf{s}_1(n) + \mathbf{a}_{j2}^T(n)\mathbf{s}_2(n) \quad (4)$$

where

$$\mathbf{s}_i(n) = \left[\begin{array}{cccc} s_i(n) & \ldots & s_i(n-N+1) \end{array} \right]^T. \quad (5)$$

The objective is to determine a separation matrix,

$$\mathbf{W}(n) = \left[\begin{array}{cc} \mathbf{w}_{11}(n) & \mathbf{w}_{12}(n) \\ \mathbf{w}_{21}(n) & \mathbf{w}_{22}(n) \end{array} \right] \quad (6)$$

such that

$$\mathbf{y}(n) = \mathbf{W}(n) * \mathbf{x}(n) \quad (7)$$

where

$$\mathbf{y}(n) = \left[\begin{array}{cc} y_1(n) & y_2(n) \end{array} \right]^T \quad (8)$$

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 separate the signals (assuming $\mathbf{A}$ is invertible) but $\mathbf{A}$ is unknown.

Prior to publication of research on the general separation problem of convolutional mixtures, the special case of linear mixtures was first investigated [1]. In the linear mixing model, we assume mixtures are composed of scaled and time-aligned source signals (no filtering). Thus the mixing and separation matrices are composed of scalar elements

$$\mathbf{A}(n) = \left[\begin{array}{cc} a_{11}(n) & a_{12}(n) \\ a_{21}(n) & a_{22}(n) \end{array} \right] \quad (9)$$

and

$$\mathbf{W}(n) = \left[\begin{array}{cc} w_{11}(n) & w_{12}(n) \\ w_{21}(n) & w_{22}(n) \end{array} \right]. \quad (10)$$

In the study of speech separation under linear mixing models, a computationally efficient Kurtosis Maximization Algorithm (KMA) was proposed and shown to lead excellent separation results.

In this paper, we extend KMA for separation of linear mixtures of speech signals to include arbitrary delays in the mixing model. In this way, we at least account for propagation delays from speakers to microphones and can operate in conditions where echo and reverberation are minimal.

2. KURTOSIS MAXIMIZATION ALGORITHM FOR LINEAR MIXTURES OF SPEECH

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

$$\kappa_x \equiv \frac{E\left[x^4\right]}{\{E\left[x^2\right]\}^2}, \quad (11)$$

$$\begin{array}{l} \mathbf{y}(n) = \mathbf{W}(n)\mathbf{x}(n) \\ \hat{\sigma}_i^2(n) = \lambda_2\hat{\sigma}_i^2(n-1) + (1-\lambda_2)x_i^2(n) \\ \hat{r}_{12}(n) = \lambda_2\hat{r}_{12}(n-1) + (1-\lambda_2)x_1(n)x_2(n) \\ \alpha_i = 4y_i^3(n) \\ \beta_i = -w_{i1}(n)\hat{r}_{12}(n)x_1(n) - w_{i2}(n)\hat{\sigma}_2^2(n)x_1(n) + \\ \quad w_{i1}(n)\hat{\sigma}_1^2(n)x_2(n) + w_{i2}(n)\hat{r}_{12}(n)x_2(n) \\ \gamma_i = \left[w_{i1}^2(n)\hat{\sigma}_1^2(n) + 2w_{i1}(n)w_{i2}(n)\hat{r}_{12}(n) + \right. \\ \quad \left. w_{i2}^2(n)\hat{\sigma}_2^2(n)\right]^{-3} \\ \mathbf{C}(n) = \left[\begin{array}{cc} -\alpha_1\beta_1\gamma_1 w_{12}(n) & \alpha_1\beta_1\gamma_1 w_{11}(n) \\ -\alpha_2\beta_2\gamma_2 w_{22}(n) & \alpha_2\beta_2\gamma_2 w_{21}(n) \end{array} \right] \\ \mathbf{W}(n+1) = \mathbf{W}(n) + \frac{\tilde{\mu}}{||\mathbf{C}(n)||_2^2}\mathbf{C}(n) \end{array}$$

Fig. 2. Normalized Kurtosis Maximization Algorithm (KMA) for speech separation.

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

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

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.

Better performance has been reported when using a normalized version of KMA [5]. In this case, the correction matrix, $\mathbf{C}(n)$ is scaled by its ℓ_2 norm

$$\mathbf{W}(n+1) = \mathbf{W}(n) + \frac{\tilde{\mu}}{||\mathbf{C}(n)||_2^2}\mathbf{C}(n) \quad (13)$$

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

$$||\mathbf{C}(n)||_2^2 = \max\left\{eigenvalue\left[\mathbf{C}(n)\mathbf{C}^T(n)\right]\right\} \quad (14)$$

The complete normalized KMA is given in Fig. 2.

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. SIRs near 0dB indicate no real source separation has occurred. Simulations using the normalized KMA for speech separation (linear mixtures) have shown very good performance with mean SIRs on the order of 25-40dB [5].

3. EXTENDING KMA TO INCLUDE PROPAGATION DELAYS

We now wish to extend the KMA to include arbitrary propagation delays from sources to microphones. In this case, the room impulse responses are of the form

$$\mathbf{a}_{ji}(n) = \left[\begin{array}{ccccccc} 0 & \ldots & 0 & a_{ji}(n) & 0 & \ldots & 0 \end{array} \right]^T \quad (15)$$

where the D_{ji} element is a_{ji} and D_{ji} is the propagation delay (in samples) from the ith source to the jth microphone.

In order to compensate for the propagation delays, we include delays d_{ij}, in the elements of the separation matrix so that we can time-align the undersired source from the mixtures and (hopefully) eliminate it. We therefore have

$$\mathbf{W}(n) = \left[\begin{array}{cc} w_{11}\delta(n-d_{11}) & w_{12}\delta(n-d_{12}) \\ w_{21}\delta(n-d_{21}) & w_{22}\delta(n-d_{22}) \end{array} \right] \quad (16)$$

Applying (16) to (7) we have as outputs

$$\begin{aligned} y_1(n) &= w_{11}(n)x_1(n-d_{11}) + w_{12}(n)x_2(n-d_{12}) \\ y_2(n) &= w_{21}(n)x_1(n-d_{21}) + w_{22}(n)x_2(n-d_{22}) \end{aligned} \quad (17)$$

Setting

$$\begin{aligned} d_{11} &= d_{22} = 0 \\ d_{12} &= D_{12} - D_{22} \\ d_{21} &= D_{21} - D_{11} \end{aligned} \quad (18)$$

and substituting into (17), we have

$$\begin{aligned} y_1(n) = {} & w_{11}(n)a_{11}(n)s_1(n-D_{11}) + \\ & w_{12}(n)a_{21}(n)s_1(n-D_{21}-D_{12}+D_{22}) + \\ & [w_{11}(n)a_{12}(n) + w_{12}(n)a_{22}(n)]s_2(n-D_{12}) \end{aligned} \quad (19)$$

and

$$\begin{aligned} y_2(n) = {} & w_{22}(n)a_{22}(n)s_2(n-D_{22}) + \\ & w_{21}(n)a_{12}(n)s_2(n-D_{12}-D_{21}+D_{11}) + \\ & [w_{22}(n)a_{21}(n) + w_{21}(n)a_{11}(n)]s_1(n-D_{21}). \end{aligned} \quad (20)$$

If we build (16) correctly, we can eliminate the undesired sources (third terms) in (19) and (20) thereby achieving separation. However, the output will be an "echoed" version of the desired source.

In order to estimate the relative delays, d_{12} and d_{21} in (16), we compute the cross-correlation between the mixture signals

$$p(m) = E[x_1(n)x_2(n+m)] \quad (21)$$

Estimate $d_{12}(n)$ and $d_{21}(n)$ from (21) and (22)
Compute $y_1(n)$,$y_2(n)$ from (17)
$z_1(n) = x_1(n)$, $z_2(n) = x_2(n-d_{12})$
$z_3(n) = x_1(n)$, $z_4(n) = x_2(n+d_{21})$
$\hat{\sigma}_i^2(n) = \lambda_2\hat{\sigma}_i^2(n-1) + (1-\lambda_2)z_i^2(n)$
$\hat{r}_{12}(n) = \lambda_2\hat{r}_{12}(n-1) + (1-\lambda_2)z_1(n)z_2(n)$
$\hat{r}_{34}(n) = \lambda_2\hat{r}_{34}(n-1) + (1-\lambda_2)z_3(n)z_4(n)$
$\alpha_i = 4y_i^3(n)$
$\beta_1 = -w_{11}(n)\hat{r}_{12}(n)z_1(n) - w_{12}(n)\hat{\sigma}_2^2(n)z_1(n) + w_{11}(n)\hat{\sigma}_1^2(n)z_2(n) + w_{12}(n)\hat{r}_{12}(n)z_2(n)$
$\beta_2 = -w_{21}(n)\hat{r}_{34}(n)z_3(n) - w_{22}(n)\hat{\sigma}_4^2(n)z_3(n) + w_{21}(n)\hat{\sigma}_3^2(n)z_4(n) + w_{22}(n)\hat{r}_{34}(n)z_4(n)$
$\gamma_1 = \left[w_{11}^2(n)\hat{\sigma}_1^2(n) + 2w_{11}(n)w_{12}(n)\hat{r}_{12}(n) + w_{12}^2(n)\hat{\sigma}_2^2(n)\right]^{-3}$
$\gamma_2 = \left[w_{21}^2(n)\hat{\sigma}_3^2(n) + 2w_{21}(n)w_{22}(n)\hat{r}_{34}(n) + w_{22}^2(n)\hat{\sigma}_4^2(n)\right]^{-3}$
$\mathbf{C}(n) = \left[\begin{array}{cc} -\alpha_1\beta_1\gamma_1 w_{12}(n) & \alpha_1\beta_1\gamma_1 w_{11}(n) \\ -\alpha_2\beta_2\gamma_2 w_{22}(n) & \alpha_2\beta_2\gamma_2 w_{21}(n) \end{array} \right]$
$\mathbf{W}(n+1) = \mathbf{W}(n) + \frac{\tilde{\mu}}{||\mathbf{C}(n)||_2^2}\mathbf{C}(n)$

Fig. 3. Extended KMA to include propagation delays.

where E is the expectation operator. Then indices associated with peaks in the cross correlation will determine the delay estimates in (18):

$$\begin{aligned} d_{12} &= \left|\arg_{-\infty<m<0} \max\left[p(m)\right]\right| \\ d_{21} &= \arg_{0<m<\infty} \max\left[p(m)\right]. \end{aligned} \quad (22)$$

In order to update $\mathbf{W}(n)$, we modify the normalized KMA (Fig. 2) to include appropriate delay estimates as described above. The complete separation algorithm for mixtures with unknown delays is given in Fig. 3.

4. RESULTS

In the following simulation results, we choose two source speech signals from the TIMIT speech corpus and digitally synthesize the mixtures according to (1). Algorithm parameters were selected as $\tilde{\mu} = 0.0005$ and $\lambda = 0.99995$.

As a reference, we first show results of normalized KMA when there are no propagation delays in the mixture. In this experiment, we randomly choose the linear mixing matrix,

$$\mathbf{A} = \left[\begin{array}{cc} 0.9501 & 0.6068 \\ 0.2311 & 0.4860 \end{array} \right] \quad (23)$$

and initialize the separation matrix as

$$\mathbf{W}(0) = \left[\begin{array}{cc} 1.0 & 0.0 \\ 0.0 & 1.0 \end{array} \right]. \quad (24)$$

Fig. 4 illustrates the learning curve for the speech separation under the linear mixture of (23). We note that after a few

seconds, SIRs on the order of 40dB are achieved indicating excellent separation.

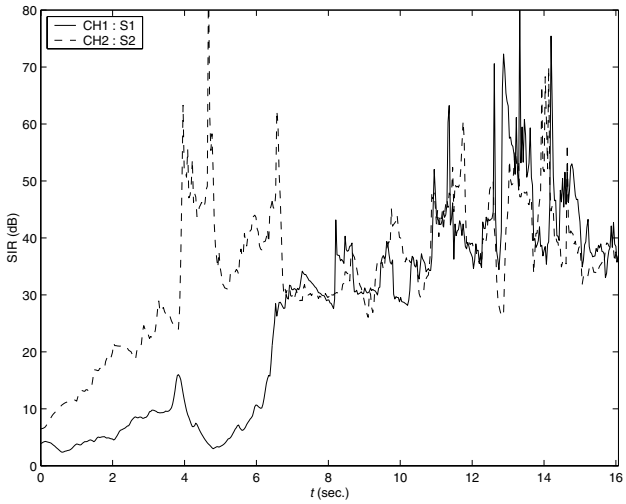

Fig. 4. Learning curve for speech separation (instantaneous mixture).

In the next experiment, we measure the performance of the algorithm under mixtures which with unknown delays. For this case we randomly choose the a_{ji} [same numbers as in (23)] and the propagation delays as

$$D_{11} = 10,\ D_{12} = 50,\ D_{21} = 70,\ D_{22} = 20. \qquad (25)$$

A plot of delay estimates using the cross-correlation method is illustrated in Fig. 5 and indicates accurate estimation of the relative delays. The resulting learning curve of the algorithm is given in Fig. 6. We note that after a ten seconds,

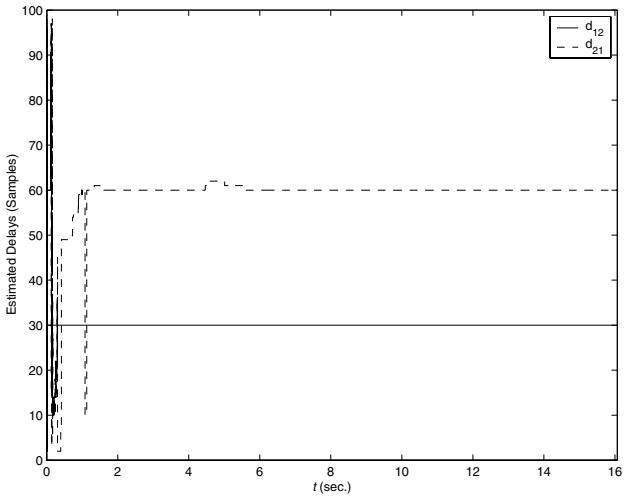

Fig. 5. Delay estimates using cross-correlation method.

SIRs on the order of 15-20dB are achieved indicating good separation.

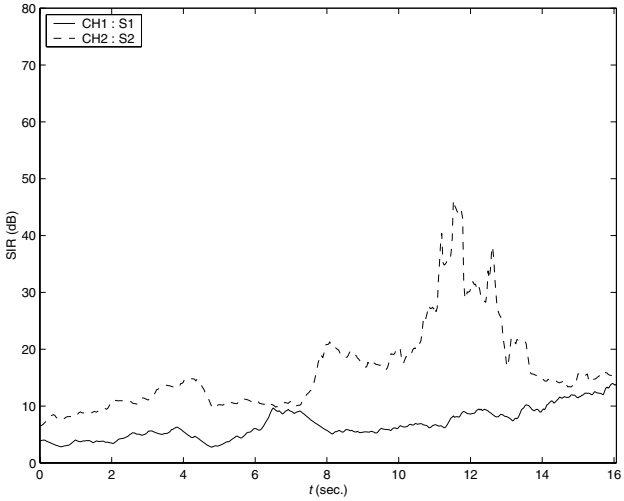

Fig. 6. Learning curve for speech separation (mixture contains scaled and delayed sources).

5. CONCLUSIONS

In this paper, we have extended a previously published algorithm for separation of two speech signals from two linear mixtures to now include arbitrary propagation delays of the sources in the mixtures. This extension assumes a more realistic mixing model. Results of the extension yield good separation quality with minimal computational overhead.

6. REFERENCES

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

[2] D. Wang and G. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 684–697, 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] K. Yen and Y. Zhao, "Adaptive co-channel speech separation and recognition," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 138–151, Mar. 1999.

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